Mining Electronic Health Records (EHR): A Survey
Do Not Circulate

Pranjul Yadav, University of Minnesota - Twin Cities
Michael Steinbach, University of Minnesota - Twin Cities
Vipin Kumar, University of Minnesota - Twin Cities
Gyorgy Simon, University of Minnesota - Twin Cities

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).

Additional Key Words and Phrases: Electronic Health Records,Data Mining

ACM Reference Format:
Pranjul Yadav,Michael Steinbach,Vipin Kumar,Gyorgy Simin,2016. Mining Electronic Health Record (EHRs) : A Survey *ACM Trans. Embedd. Comput. Syst.* 1, 1, Article 1 (April 2016), 70 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Numerous recent studies have found the health care system in the United States to be the most expensive in the world, yet trailing behind most advanced economies in quality [Schuster et al. 1998; Wang et al. 2012]. The cost of health care is steadily increasing both in absolute cost and as a percentage of the GDP, soon reaching unsustainable levels [Levit et al. 2003; Reinhardt et al. 2004]. To curb this trend, simultaneous improvements in quality and decrease in cost are necessary. Many believe that advanced analytics holds the key to achieve these opposing goals.

The foundation for analytics is data. The US government has mandated health care providers to implement electronic health record (EHR) systems with a primary goal of documenting patients' care. This effort has already been a success. Multiple studies have demonstrated that EHRs have reduced clinical errors [Singh et al. 2008; Singh and Graber 2010; Singh et al. 2012; Abramson et al. 2011; Agrawal and Wu 2009], improved chronic illness care [Dorr et al. 2007; Meigs et al. 2003; Shortell et al. 2009; Rollman et al. 2002; Crosson et al. 2007] 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 [Manolio 2009]. Also, since EHR systems store detailed care information about potentially very large patient populations with long follow-up times, they are exceptionally well positioned to serve as a research platform enabling advanced analytics.

The motivation for advanced analytics comes from another government mandate that transforms the current fee-for-service payment model to a new model based on population health management. Under the new model, primary care providers are responsible for managing entire patient populations with their payments tied to care quality. Since care providers are no longer paid for services rendered, but rather for the outcomes of the service, they are incentivized to increase their efficiency through implementing better care practices, thus creating an opportunity for analytics.

The cornerstone for modern medicine is evidence-based practice [Sackett 2000]. At the heart of evidence-based practice lies a large knowledge base of best practice recommendations that have been put forth by committees of well-established care providers and (ideally) validated through randomized clinical trials. These best practice recommendations, if followed, enable providers to increase their efficiency and reduce waste. Generating potential evidence, which of course needs validation through the rigorous clinical trials framework, is one outlet for advanced analytics to impact practice. It is not the only outlet, but it is certainly one that is tested and proven.

The traditional vehicle for evidence creation is the Randomized Clinical Trial (RCT) [Matthews 2006]. RCTs are controlled studies that select subjects based on well-defined criteria and procedures (known as the protocol) and then randomly assign these subjects to treatment and control groups. RCTs are considered the gold standard for evidence creation, but they suffer from some well-known limitations, namely, high cost, small sample size [Schulz and Grimes ; Chan et al. 2008; Passamani et al. 1985; Van Spall et al. 2007; Olschewski et al. 1992; Moore et al. 1996; Gurwitz et al. 1992], a limited ability to account for confounding factors [Rothwell ; Sacco et al. 1995] or comorbidities [Fortin et al.], and short duration [Kakkar et al. 2008]. High cost relegates RCTs to a *confirmatory* role, where previously hypothesized recommendation can be confirmed. Thus, less expensive methodologies are often employed to generate new recommendations.

One of these methodologies is observational studies, which are analyses carried out on previously existing data, such as EHR data. While observational studies are less expensive than RCTs, their results are also less reliable. Subjects for RCTs are selected specifically to minimize any difference between the treatment and control groups, yielding a trial population with potentially very little resemblance to the original subject population. On the other hand, data in EHR system was not even collected with analytics as the primary purpose. Thus carrying out observational studies on EHR data requires techniques that address shortcomings of EHR data.

Currently this need is mostly filled by a number of health-related analytical disciplines. Most notably, health sciences research, epidemiology and biostatistics have embraced observational studies and have successfully developed techniques to extract reliable knowledge from this abundant but noisy preexisting data. In contrast to the above analytic disciplines, data mining is known for its flexible modeling approach and has a proven track record of extracting actionable knowledge from large

collections of data in a wide range of collections such as e-commerce, social-media, recommender-systems, etc. The emerging discipline of clinical data mining aims leverage the strengths of data mining and health-related analytical disciplines to develop methods that better meet the challenge of extracting knowledge out of large collections of health-related data, primarily EHR data. Clinical data mining stands at the intersection of epidemiology, biostatistics, data mining and machine learning, borrowing methodologies from epidemiology and biostatistics for study design and to overcome the EHR data challenges.

Advanced analytics, through flexible modeling, has great potential in health care. Mining EHR data can 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. EHR data can also provide superior public health surveillance information on chronic conditions such as asthma [Fiks et al. 2009; Davis et al. 2010] and Type-2 Diabetes Mellitus (T2DM) [Weber et al. 2008; OConnor et al. 2011]. It can also help in comparing risks due to community factors such as economic disparity [Bilheimer and Klein 2010; Roblin et al. 2009].

With all three ingredients of success for data mining—data, motivation and outlets for results—having come together, we believe that an explosive growth in the adoption of clinical data mining is imminent. Given the unique challenges that EHR data poses, a wave of innovation in data mining will ensue. In this survey, beside reviewing existing work on clinical data mining involving EHR data, we also summarize background knowledge on study design, the characteristics and challenges of EHR data, and techniques developed by other disciplined to address these challenges. (Although needed for the survey, these sections present concepts and techniques that are not well known in data mining, but have potential uses beyond EHR data.) It is our hope that this survey will contribute to the acceleration of innovation that will allow data mining to find the high level of success in health care that it has already achieved in other areas.

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 and machine learning community only to a very limited extent. Section 5 introduces metrics that are used to measure the outcome of a treatment or study, while Section 6 describes study designs, which are the cornerstone of EHR data mining. In Section 7, we introduce techniques developed in other fields, most notably epidemiology and biostatistics, that address many of the challenges we discussed in Section 4. 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. In Section 9, we provide an overview of clinical research carried out in various applications areas, which were introduced in Section 2. Finally, in Section 10, we analyze the current state-of-the-art and analyze our findings. In Section 11, we discuss the future of EHRs in relation to data mining and in Section 12, we present a short conclusion. In the Appendix, we present 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

An appropriate place to start our discussion of clinical data mining is to describe some clinical questions that data mining can help answer. The fundamental question of medicine is to decide on the treatment that is most suitable and effective for a particular patient and data mining has the potential to help address this question in numerous ways. In this section, we present a sequence of increasingly complex clinical applications, starting with the simplest epidemiological questions, continuing with common clinical research tasks such as automatic identification of study cohorts, risk prediction and risk factor discovery, all the way to complex applications such as discovering knowledge related to best clinical practices from data.

2.1. Understanding the Natural History of Disease

The most basic epidemiological inquires are not data mining problems per se, they are concerned with question like: 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*, or in other words, what are the *precursors and consequences* of this condition?

Studying the sequences of diseases clearly enters the realm of data mining and it has been applied to studying the progression of a patient's medical state over time, which is also 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), T2DM and eventually towards diabetes associated 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, cardio-vascular complications or peripheral complications. Even though the outcome (complications in this case) is the same, disease progression paths leading to the outcome might be 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, discovery of biomarkers or the development of novel risk estimation indices.

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.

2.2. Cohort Identification

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 others who most likely do not (*controls*). This can be achieved through phenotyping algorithms, either hand-crafted or machine learned. Phenotyping algorithms characterize the disease in terms of patient characteristics observable from the EHR data and classify patients as likely having the disease, likely not having the disease, or disease status is uncertain. Phenotyping algorithms are at the heart of modern cohort discovery; the problem of

identifying patients with certain diseases and certain other clinical characteristics.

Traditionally, cohort identification was carried out through chart reviews, where nurse abstractors have painstakingly reviewed patients' medical records to determine whether the patient meets the criteria for inclusion into the cohort. However the scale enabled by EHRs renders manual chart review impractical. Instead, electronic phenotyping algorithms are applied, with manual chart review relegated to 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.

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.

2.3. Risk Prediction/Biomarker Discovery

With a cohort in hand, we can build predictive models, opening up a wide range of opportunities for data mining. These models can predict the risk of disease, e.g. estimating the probability of developing a condition of interest in 5 years (*risk prediction*) helping health care provider focus their limited resources or investigate which predictors are relevant (*biomarker discovery*, *risk factor discovery*) in developing the outcome in question. Accurate knowledge of risk factors can help guide preventive efforts or focus interventions.

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 [Ng et al. 2014; Tran et al. 2014b; Rekatsinas et al.]. 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.

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.

2.4. Predicting the next complication: What and When

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 disease 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 or for disease such as sepsis, where it takes couple of hours for a patient to progress from healthy state to a critical one. This becomes more relevant as patients follow multiple health trajectories, often leading to life threatening conditions including mortality [Hripcsak et al. 2015; Park and Ghosh 2014a]. In such a scenario, predicting the next complication is a challenging problem. For example, predicting the next hospitalization for a patient diagnosed with T2DM. Clinical decision support systems informed by such models will lead to improved patient care, thereby improving overall health care.

2.5. Quantifying the effect of Intervention

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*.

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 demographics attributes such as gender, age, ethnic make-up, socio-economic status, etc.

2.6. Constructing Evidence Based Guidelines

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 the 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, however 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.

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 [Field et al. 1992]. Evidence based guidelines (EBG) try to guide decision making by identifying best clinical practices,that are meant to improve the quality of patient care [Barretto et al. 2003]. 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.

2.7. Adverse Event Detection

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 have tremendous potential for both detecting and to predicting such adverse events.

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. 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.

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.

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 a priori.

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 [Farri et al. 2012], creation of clinical sense inventory of clinical abbreviations and acronyms [Moon et al. 2014] and development of tailored NLP methods to extract information from operative notes [Wang et al. 2012].

3.3. Flowsheets

In between the two extremes (structured tables and unstructured clinical notes) lies the (semi-structured) flow sheet format [Weed 1968]. 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. Many of those challenges are also frequently encountered in other areas: noise, high dimensionality, sparseness, non-linear relationships, complicated dependencies between variables, and extraction of higher level features from low level information. Less frequent, but still not uncommon elsewhere, are issues related to data integration across multiple sites (medical providers) and / or multiple types of data sets (e.g. clinical vs. claim data). As in other domains, it is also important to incorporate domain knowledge, including knowledge about the relationships. For example, the blood pressure of a patient on medication for hypertension needs to be interpreted in that context.

However, the single most important challenge, the one that arguably impacts data mining methodologies at the most, is missing data, and this will be the focus of our discussion in the rest of this section. 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 increased risk of outcome 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 un-fragmented 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 [Frydman 1995], dementia [Joly et al. 2002], nepropathy [Andersen 1988] and mortality [Andersen 1988].

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 [Kaplan and Meier 1958], bayesian non-parametric methods [Kaplan and Meier 1958] which involves prior belief about the shape of the survival function, semi parametric proportional hazards regression with fixed covariates or time dependent covariates [Wei et al. 1989], additive hazards regression model and parametric regression models using weibull distribution or log logistic distribution [Cox 1972].

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 accompanying 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. Biases and Confounding Effects

Studies performed using EHRs often have biases and confounding effects [Moher et al. 1998; Schulz et al. 1995]. 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 [Gruber 1986; Burkman 1981; Nordin 1977; Seltzer et al. 1974]. 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 [Wacholder et al. 1992]. 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 [Vernick et al. 1984; JACQUELINE et al. 1989]. 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 [Wacholder et al. 1992].

Confounding is another issue which might undermine the internal validity of any study [Abramson and Abramson 2001; Walline 2001]. 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 [Ory 1977; Schwingl et al. 1999; Jain 1976]. 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 [Walline 2001]. For example, if there are few subjects who consume tobacco, then it would help to remove these subjects from the study. Similarly, pairwise matching [Walline 2001] and stratification [Mantel and Haenszel 1959] 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 [Lang and Secic 2006]. 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 readings, smoking history and various other risk factors.

5. 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 [Hennekens et al. 1987; Last et al. 2001; Sackett et al. 1996; Sterne and Smith 2001]. Incidence rate [Last et al. 2001] 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 [Last et al. 2001] 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 [Sackett et al. 1996] 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.

6. STUDY DESIGN

EHR data is mere collection of database tables that need to be transformed into an analysis matrix that is amenable to data mining in a manner that allows us to answer the question we set out to study. Suppose we wish to construct a risk prediction model that can predict the 5-year risk of a particular disease D for any patient. To construct such a model, we would take a cross section of the patient population in year Y, allowing us to have a representative sample from which this "any patient" may come from. Y is ideally at least 5 years before the last date in the EHR, so that we have sufficient (5-year) follow-up for many patients. Next, we reconstruct all patients' state of health in year Y, collecting their medical history before Y; and then follow them five years forward (until year $Y + 5$) and establish their outcomes, namely whether they developed D. The analysis matrix would have patients as its rows and patient characteristics in year Y as columns; and would have an outcome for D, as well. Traditional predictive modeling techniques are directly applicable to such a matrix for carrying out various research related tasks.

The way we transformed that EHR data follows a particular *study design* that allows us to answer our question. Each study question can require a different study design. In this section, we review some of the most commonly used study designs and the questions they allows us to answer. For additional details, the interested reader is referred to [Grimes and Schulz 2002].

Study designs form a hierarchy which is depicted in Figure 1. Accordingly, study designs can be primarily classified into two major groups i.e. experimental and observational.

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 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** [Funai et al. 2001], 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 [Kelsey 1996; Rothman and Greenland 1986]. Observational studies can be further categorized as analytical (if there is a comparison group (i.e. case and control)) or descriptive (no comparison group).

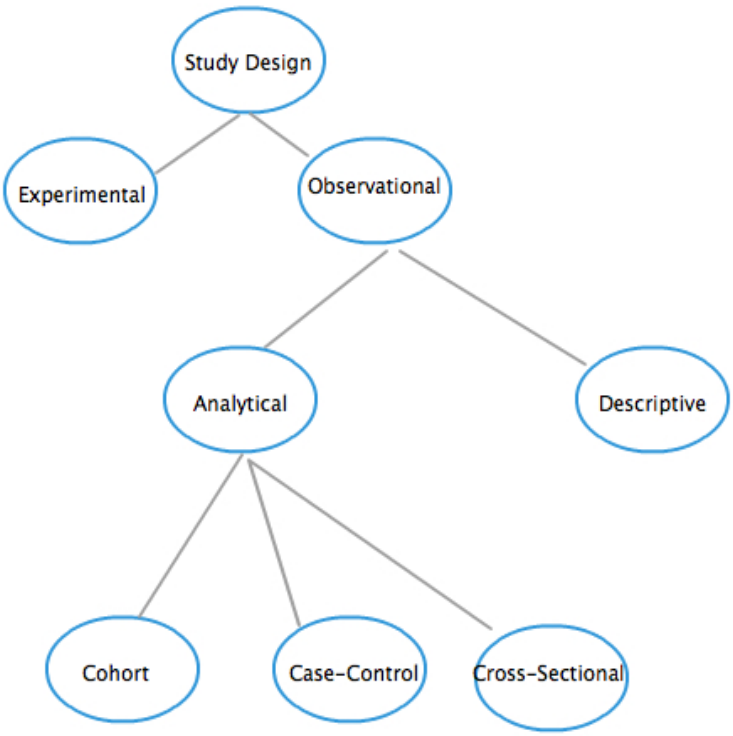

Fig. 1. Study Design Classification

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 [Last et al. 2001] 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** [Doll et al. 2000; Hannaford and Kay 1998; Kim et al. 2000; Huang et al. 1999]. 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** [Last et al. 2001]. 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.

6.1. 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 [Torrey et al. 2000; XU et al. 2011; Osaki and Minowa 2001; Avidan et al. 2001].

These study designs are very useful in the investigation of diseases that have a long latency period, such as cancer and T2DM 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.

6.2. Cohort Studies

Cohort studies are also known as incidence, longitudinal,forward-looking, follow-up, concurrent or prospective studies [Lilienfeld and Stolley 1994]. 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 [Beral et al. 1999; Seman et al. 1999; Colditz et al. 1996]. 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 [Walline 2001]. 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 [Sackett 1979]. 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 Peripheral Vascular Disease (PVD) and Ischemic Heart Disease (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.

6.3. 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 [Lee 1994]. 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 [Mann 2003].

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 [Johnson and Hall 1988; Andersen et al. 2006; Coffey et al. 1992]. Due to this, metrics such as prevalence are widely used in these study designs [Lee 1994]. 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 [Levin 2006]. 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).

6.4. Descriptive Studies

Descriptive studies are designed to describe the existing distribution of variables, without regard to causal or other hypotheses [Walline 2001; Hennekens et al. 1987]. In such studies, apart from age and gender, other characteristics such as race, occupation and recreational activities are often described[Gabert and Stueland 1993; Krane et al. 1988; Jaremin et al. 1996; Marshall et al. 1997; Giordano et al. 2002; Weiss et al. 2005; Anderson et al. 1991]. 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 [Tough et al. 2000; Bider et al. 1999; Dunn and Macfarlane 1996; Steegers-Theunissen et al. 1998]. 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 prediabetic 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.

7. APPROACHES

Having defined some metrics of interest in Section 5, and having selected a study design from Section 6, to transform our EHR data into an analytics matrix that data mining methods can operate on, we appear ready to answer the clinical question we set out to solve. While the data format may suggest that we can directly apply our existing data mining techniques, the data itself creates some challenges, which we described in Section 4 in detail. Addressing these challenges require analytical approaches that have not been handled by traditional data mining challenges. Most notably, we describe approaches for addressing censoring, the irregular temporal nature of the data and biases and confounding.

Approaches are concepts and ideas that provide high-level solutions to these challenges. Within these high-level solutions, concrete data mining techniques can be developed. Our focus in this section is to describe these high-level ideas; later, partly in this section as illustrative examples, but mostly in the subsequent sections, we will discuss some concrete analytics techniques that use these approaches.

Chung et al. [Chung et al. 1991] provided a survey of statistical methods to analyze the length of time until an event of interest occurs. Such modes have been widely used to analyze the survival times (i.e. time until death) of medical patients, analyze the time until recidivism and many other applications. In their paper, they summarized the statistical literature on survival analysis. Kleing et al. [Klein and Moeschberger 2005] provides a comprehensive overview of various techniques used to handle survival and censored data. In their book, they discuss the various sources where survival/censored data is generated and provides multiple examples in medical settings. They then discuss various quantities and models of interest such as the survival function and the hazard function, sampling schemes for censored data. They conclude the book, by discussing several advanced topics such as estimating the hazard function, excess mortality, bayesian non-parametric techniques, gamma-frailty models and other techniques for univariate and multivariate methods. Dahejia et al. [Dehejia and Wahba 2002] discussed causal inference and sample selection bias in non experimental settings. In their paper, they discussed the use of propensity score-matching methods

for non-causal studies and discussed several methods by implementing them using data from the National Supported Work experiment.

7.1. 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? To answer, the aforementioned clinical questions, techniques are required which can handle censoring, which is frequently present in EHRs. Techniques to handle censored data, can be divided into three major categories: non-parametric, semi-parametric or 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 [Kaplan and Meier 1958] and Nelson- Aalen estimators [Cox 1992]. Rihal et al. [Rihal et al. 2002] used Kaplan-Meir estimators for incidence and prognostic implications of acute renal failure in patients undergoing percutaneous coronary intervention (PCI). Dormandy et al. [Dormandy et al. 2005] 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. [Rossing et al. 1996] used Nelson-Aalen estimators for analyzing the predictors of mortality in insulin dependent diabetes. Ekinci et al. [Ekinci et al. 2011] used such non-parametric techniques for exploring salt intake consumption and mortality in patients with diagnosed with T2DM.

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) [Keiding et al. 1997] 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. [Babuin et al. 2008] 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 [Wilson et al. 2008] used AFT models to predict cardiovascular risk by using predictors such as age, gender, 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. [Yadav et al. 2015] 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 [Vinzamuri et al. 2014] 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.

7.2. 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. These models handle unevenly spaced (irregular) EHRs by assuming a correlation structure among multiple clinical observations of a patient recorded at different time points.

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. [Hernán et al. 2000] 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. [Yu et al. 2010] used marginal models to estimate the effect of medication adherence on health outcomes among patients diagnosed with T2DM. Nandi et al. [Nandi et al. 2012] used such models to estimate the direct effect of adverse childhood social

conditions on onset of heart disease, T2DM and stroke. King et al. [King 2012] have discussed the use of marginal models for T2DM related research.

Conditional models [Laird and Ware 1982] 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 our 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. [Yamaoka and Tango 2005] used conditional models to evaluate the efficacy of lifestyle education for preventing T2DM in individuals at high risk. They observed that lifestyle education intervention reduced glucose levels by 0.84 mmol/l as compared to the control group. Mezuk et al. [Mezuk et al. 2008] examined the bi-directional prospective relationships between depression and T2DM using a random effects model. Nouwen [Nouwen et al. 2010] 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 patients diagnosed with T2DM have a 24% increased risk of developing depression.

7.3. 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) [Peikes et al. 2008] 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. [Tao et al. 2010] used PSM to determine 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. [Austin 2007] provided a systematic review of the use of propensity score matching in the cardiovascular surgery literature. Polkinghorne et al. [Polkinghorne et al. 2004] used PSM to analyze the inception and intervention rate of native arteriovenous fistula (AVF). Yasunaga et al. [Yasunaga et al. 2013] investigated postoperative outcomes after laparoscopic or open distal gastrectomy in Japan. Exact, 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 groups of interest.

Short et al. [Short et al. 2011] 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. [Kuss et al. 2010] 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 [Hogan and Lancaster 2004] 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, the subjects are usually weighted. 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 [Robinson et al. 2011] used inverse probability weighting for examining whether lower serum levels of serum 25-hydroxyvitamin are associated with increased risk of developing type 2 diabetes.

8. CLINICAL DATA MINING METHODOLOGIES

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 from biostatistics and epidemiology 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 outcome of interest 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

[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.

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).

	Supervised	Unsupervised
Time Aware	Cohort and Retrospective Studies (8.3.2)	Temporal Descriptive Studies (8.1.2)
Time Agnostic	Cross-Sectional Studies (8.2),Cohort and Retrospective Studies (8.3.1)	Atemporal Descriptive Studies (8.1.1)

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. In Table-1, we present the structure of the following subsections.

Tan et al. [Tan et al. 2006] provides a comprehensive broad, yet in-depth overview of the field of data mining. Their book discusses various techniques related to clustering, classification and pattern identification. Clustering can be broadly divided as a division of data into groups of similar objects. Berkhin et al. [Berkhin 2006] survey concentrates on clustering algorithms from a data mining perspective. They believed that clustering is a data modeling techniques that provides concise summarization of the data. In their survey articles, they demonstrate how clustering techniques can be used in a broad range of applications. Hipp et al. [Hipp et al. 2000] presented a survey on efficient algorithms widely used for association rule mining related tasks. In their survey, they first explained the fundamentals of association rule mining and demonstrated the similarities and differences among various association rule mining related techniques. They also observed that that the run-time behaviors of the various association rule mining techniques are much more similar to the contrary. Han et al. [Han et al. 2007] provided a brief overview of frequent pattern mining related activities across frontiers such as sequential pattern mining, structure pattern mining, correlation mining and associative classification. They believed that frequent pattern mining research has had a deep impact on data mining methodologies.

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 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. EHRs are inherently temporal in nature, but such data is usually converted into atemporal by taking a snapshot of a patient's data at a point in time.

In this section, we would be discussing how Unsupervised techniques 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). Next, we would discuss how these techniques make use of the approaches we discussed in Section 5.

Clustering: Gotz et al. [Gotz et al. 2011a] 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. [Roque et al. 2011] stratified patients using hierarchical clustering, where the distance between patient records was computed using the cosine similarity of diagnosis codes. Bauer-Mehren et al. [Bauer-Mehren et al. 2013] 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. [Doshi-Velez et al. 2014] 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. [Kalankesh et al. 2013] noted that representing the medical state of a patient with diagnosis codes can lead to sparse clusters since EHRs contains large number of diagnosis codes often running into thousands. To overcome this problem, they used Principal Component Analysis (PCA) [Dunteman 1989] to reduce the dimensionality, thereby making the structure more amenable for visualization and clustering. Marlin et al. [Marlin et al. 2012] developed a probabilistic clustering method to mitigate the effects of unevenly spaced data, which is inherent in EHRs. They used unsupervised learning techniques for automatically uncovering insightful patterns from physiologic time-series data.

Association Analysis: Association rule mining techniques [Zhang and Zhang 2002] such as Apriori have also been used on EHR data to identify associations among clinical concepts (medications, laboratory results and problem diagnoses). These techniques have the ability to discover associations and interpretable patterns from EHRs data. However, the performance of such techniques often deteriorates when there are large number of clinical variables present in EHRs. Wright et al. [Wright et al. 2010a] 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. [Cao et al. 2005] used co-occurrence statistics to identify direct and indirect associations among medical concepts. Holmes et al. [Holmes et al. 2011] 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. [Shin et al. 2010] 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. [Hanauer et al. 2009] used statistical tests to observe common pathways for diseases such as granuloma annulare and osteoarthritis.

Overcoming Challenges: 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 (using HMM to simplify away time so that standard clustering is applicable).

Temporal Abstraction Framework: The temporal abstraction framework has been frequently used to prepare 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. [Shahar 1997] 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. [Sacchi et al. 2007] 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. [Jin et al. 2008] 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. [Batal et al. 1994] 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. [Ghassempour et al. 2014] 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.

Sequential Rule Mining: Researchers have explored sequential association rule mining techniques for identifying causal relationships between diagnosis codes. Hanauer and Ramakrishnan [Hanauer and Ramakrishnan] 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 [Liao and Chen 2013] proposed a sequential pattern mining approach to mine sequences with a gap constraints. Such gaps represent the delay between two concepts. Hripsack et al. [Hripcsak and Albers 2013] 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.

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 [Caban and Gotz 2015], cohort analysis and exploration [Zhang et al. 2014; Gotz et al. 2012a], exploring comorbidities [Gotz et al. 2012b; Sun et al. 2010], exploring concepts [Cao et al. 2011b], clinical decision support [Gotz et al. 2011b], cohort identification [Cao et al. 2011a], disease network visualization [Perer and Sun 2012] and temporal frequent event sequences [Perer and Wang 2014].

Overcoming Challenges: 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. Although 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. [White et al. 2013] 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. [Iyer et al.] 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. [Haerian et al. 2012] hypothesized that adverse events might be caused by the patient's underlying medical condition. Along similar lines, Vilar et al. [Vilar et al. 2012] used disproportionality based techniques to analyze adverse drug events related to pancreatitis, Li et al.[Li et al.] used penalized logistic regression to analyze associations between ADRs and Epstein et al. [Epstein et al.] 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 [Pathak et al. 2013a; Carroll et al. 2011; Xu et al. 2011] have been widely explored. Supervised pattern mining approaches using the temporal abstraction framework have been used for predicting Heparin Induced Thrombocytopenia (HIT) [Batal et al. 1994]. Batal and Hauskrecht [Batal and Hauskrecht 2010] 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 [Batal et al. 2012]. 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 [Lucas et al. 2004]. Zhao et al. [Zhao and Weng 2011] 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. [Sverchkov et al. 2012] 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.

Overcoming Challenges: 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 inretrospective 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.

8.3.1. 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 [Mani et al.], potentially preventable events [Sarkar and Srivastava 2013], 30 day hospital readmissions [Cholleti et al. 2012; Park and Ghosh 2014b], post-hospitalization VTE risk [Kawaler et al. 2012] [Zhai et al. 2014], T2DM risk forecasting [Mani et al. 2012], atrial fibrillation [Karnik et al. 2012], 5 year long life expectancy risk calculation [Mathias et al. 2013], risk of depression using diagnosis codes [Huang et al. 2014], survival of heart-lung transplant patients [Oztekin et al. 2009], breast cancer survivability [Sarvestani et al. 2010], 30 day mortality in patients suffering with cardio-vascular diseases, risk of retinopathy in patients suffering from type 1 diabetes mellitus (T1DM) [Skevofilakas et al. 2010], mortality in patients suffering from acute kidney injury [Matheny et al.], mortality prediction in ICU [Herasevich et al. 2013] and risk of dementia [Maroco et al. 2011]. For these analyses, almost all flavors of common predictive modeling techniques (decision trees [Mani et al. 2012; Sarvestani et al. 2010] ,[Austin et al. 2012], ensemble techniques (e.g. bagging,boosting,random forests) [Cholleti et al. 2012; Kawaler et al. 2012; Mani et al.],[Karnik et al. 2012],nave Bayes [Kawaler et al. 2012; Karnik et al. 2012; Sarvestani et al. 2010], linear regression, support vector machines[Mani et al. 2012] and logistic regression [Zhai et al. 2014; Mani et al. 2012; Cholleti et al. 2012; Huang et al. 2014; Chang et al. 2011] have been used. These techniques have also been used for identification of regional differences in breast cancer survival rates despite guidelines [Ito et al. 2009], comparison of cancer survival rates across continents [Coleman et al. 2008],comparison of cancer and survival patients over time, exploring relationships between hospital surgical volumes and 5 year relationship of stomach cancers [Nomura et al. 2003], comparing dosage volumes of warfarin in European-American and African-American [Ramirez et al. 2012], postpartum depression rates in Asian-American subgroups (Indian, Chinese, Filipino, Japanese, Korean, Vietnamese) [Goyal et al.], 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. [Ghalwash and Obradovic] proposed predictive modeling technique to find a suitable duration of the hemoadsorption (HA) therapy control and observed that their method led to substantial monetary savings. Sun et al. [Sun et al. 2014] 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. [Wang et al. 2015] 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. [Panahiazar et al. 2015] 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. [Wang et al. 2014] 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.

Overcoming Challenges: 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.

8.3.2. 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. Other techniques which incorporate the effect of time include dynamic bayesian networks, sequential pattern mining, etc.

Survival Modeling: Wells et al. [Wells et al. 2008] 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 [Vinzamuri and Reddy 2013] 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. [Vinzamuri et al. 2014] 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 [Cox 1992; Vinzamuri and Reddy 2013] 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 [Wooldridge 1992], healthcare [Ikeda et al. 1991; Liang et al. 1990; Lumley et al. 2002] and recommendation systems [Kapoor et al. 2014].

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 [Tibshirani et al. 1997] and elastic-net regularized Cox models [Simon et al. 2011a] have been developed, and have been further extended by regularizing them with convex combinations of L1 and L2 penalties [Zhang and Lu 2007]. We are not aware of regularization for time-dependent covariate Cox models [Therneau and Crowson 2014], which would be a straightforward extension.

Reddy et al. [Vinzamuri et al. 2014] 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 [Vinzamuri and Reddy 2013]. 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 [Gopakumar et al. 2014]. To handle the high-dimensionality of high-throughput genomic data, Kuang et al. [Zhang et al. 2013] 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 [Hearst et al. 1998] models have also been extended to handle censored data [Khan and Zubek 2008; Evers and Messow 2008; Shivaswamy et al. 2007; Van Belle et al. 2007; Shiao and Cherkassky 2014]. 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. [Khosla et al. 2010] 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 [Gordon and Olshen 1985]. Ishwaran et al. [Ishwaran et al. 2008] 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 [Kattan et al. 1998; Snow et al. 1994]. Techniques such as reverse survival [rev] 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 models, other methods that can incorporate temporal information also exist. Dynamic Bayesian networks (DBN) have been used to model temporal relationships among EHR variables [Rana et al. 2015]. Nachimuthu et al. [Nachimuthu et al. 2010] 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. [Sandri et al. 2014] 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. [Rose et al. 2005] used DBN's to assist physicians in monitoring the weight of patients suffering from chronic renal failure, Gatti et al. [Gatti et al. 2011] used it to model heart failure and Peelen et al. [Peelen et al. 2010] used hierarchical DBN's for modeling organ failure. Expectation-Maximization was used to learn conditional probabilities in these DBN's.

Sequential Pattern Mining: 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 T2DM [Batal et al. 2012]. Sengupta et al. [Sengupta and Naik 2013] used similar techniques for detecting sequential rules associated with the early identification of brain tumors. Simon et. al. [Simon et al. 2013a] 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.

Overcoming Challenges: 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.

9. CLINICAL DATA MINING APPLICATIONS

In section 2, we discuss the various applications areas where EHRs have been used to extract meaningful information. In section 8, we introduced and discussed various data-mining techniques which have been widely used in the recent past for exploring and extracting information associated with EHRs. In this section, we will analyze the research carried out in various application areas in conjunction with the data-mining techniques. Such an analysis, will provide examples of how research has been

traditionally carried out in various application areas using existing data mining techniques. This analysis can then be used to carry out research in novel application areas.

Jensen et al. [Jensen et al. 2012] 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. Reddy and Agarwal [Reddy and Aggarwal 2015] provided a comprehensive overview of different aspects of healthcare data analytics. In their book, they explored EHRs in the context of 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.

9.1. Understanding the Natural History of Disease

Research in past have focussed on exploring disease progression patterns and comorbidity analysis. Jensen et al. [Jensen et al. 2014] 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. [Teno et al. 2001] 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. [Murtagh et al. 2008] 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. Wang et al. [Wang et al. 2014] 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.

In the past, researchers have used comorbidity analysis for observing how alcohol usage is associated with depression, anxiety and personality disorders [Roque et al. 2011]. Doshi-Velez et al. used patient stratification techniques to observe comorbidities in patients suffering from autism spectrum disorders [Doshi-Velez et al. 2014]. Wright et al. [Wright et al. 2010b] employed the Apriori framework to detect associations between clinical concepts (laboratories test results, medications) and problem lists. Cao et al. [Cao et al. 2005] used a statistical framework to detect an association between diseases such as 'myasthenia gravis 'and 'cushingoid facies'. Similarly Holmes et al. [Holmes et al. 2011] studied the comorbidities for rare diseases such as Kaposi sarcoma, toxoplasmosis and Kawasaki disease. Using the association

rule mining framework, Shin et al. [Shin et al. 2010] explored the comorbidities associated with hypertension such as non-insulin dependent T2DM, cerebral infection and chronic renal failure. Dasgupta et al. [Dasgupta and Chawla] 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.

9.2. Cohort Identification

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 [Friedlin et al. 2010], rheumatoid arthritis [Liao et al. 2010], pneumonia [Liao et al. 2010], critical care [Halpern et al. 2014] and asthma [Meystre et al. 2009]. Kandula et al. [Kandula et al. 2011] 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. [Rasmussen et al. 2014] 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. Castelli et al. [Castelli et al. 1977] 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. [Newton et al. 2013] 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. [Overby et al. 2013] 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. [Pathak et al. 2013b] identified various challenges 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. [Schram et al. 2014] 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. [Boland et al. 2013] 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. [Gotz et al. 2014] 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. [Wang et al. 2013] 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. [Peissig et al. 2012] 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. [Pathak et al. 2012] 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. [Ho et al. 2014c; Ho et al. 2014b; 2014a] proposed tensor factorization methods to derive phenotypes. Schulam et al. [Schulam et al. 2015] 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. [Hu et al. 2012] 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.

9.3. Risk Prediction/Biomarker Discovery

Such applications usually employ supervised techniques for risk prediction and biomarker discovery. Greenland et al. [Greenland et al. 2004] 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. [Knaus et al. 1991] 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. [Sarkar et al. 2012] 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. [Letham et al. 2013] used Bayesian model and Markov chain Monte Carlo sampling to develop interpretable predictive models using EHRs data. Ebadollahi et al [Ebadollahi et al. 2010] 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 [Feldman and Chawla 2014] 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. [Ngufor et al. 2015] 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. [Byrd et al. 2014] constructed a system to automatically identify heart failure diagnostic criteria. Kamkar et al. [Kamkar et al. 2015] 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. [Tran et al. 2014a] 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. [Lakshmanan et al. 2013] presented an approach for mining clinical care pathways correlated with patient outcomes that involves a combination of clustering, process mining and frequent pattern mining. Vijayakrishnan et al. [Vijayakrishnan et al. 2014] 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. Schrom et al. used association rule mining along with propensity score matching to investigate how statin can lead to overt diabetes in certain subpopulations [Schrom et al. 2013a]. Supervised techniques, such as survival association rule mining [Simon et al. 2013a], 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. [Harpaz et al. 2013] used statistical techniques to identify a chemical biomarker rasburicase, that results in adverse events related to pancreatitis. Vellanki et al. [Vellanki et al. 2014] 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, one can discover learning patterns (biomarkers) more efficiently as compared to the parametric methods.

9.4. Predicting the next complication: What and When

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. [Yadav et al. 2015] 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. [Zhao and Weng 2011] 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 [Algar et al. 2003], cardiac arrest [Detsky et al. 1986], bariatric

surgery [Finks et al. 2011], carotid endarterectomy [McCrory et al. 1993], acute cough [Hay et al. 2004], breast reconstruction [McCarthy et al. 2008], pulmonary resection [Epstein et al. 1993; Ferguson and Durkin 2003; 2002; Ferguson et al. 1988], knee replacement [SooHoo et al. 2006], lumbar decompression [Benz et al. 2001], orthopedic surgery [Ozkalkanli et al. 2009], hysteroscopic surgery [Propst et al. 2000] and febrile neutropenic cancer [Klastersky et al. 2006].Lui and Hauskrecht [Liu and Hauskrecht 2014; 2013a; Liu et al. 2013; Liu and Hauskrecht 2013b] 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. [Panahiazar et al. 2014] 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.

9.5. Quantifying the effect of Intervention

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. [Schrom et al. 2013a] 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 T2DM, and identified novel risk factors. Campbell [Campbell 1996] 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. [Prochaska et al. 2008] analyzed the effects of risky behaviors such as smoking, alcohol abuse, physical inactivity, and poor diet on human health. Ronsmans and Campbell [Ronsmans and Campbell 2011] presented the evidence of the effect of health interventions on mortality reduction from hypertensive diseases in pregnancy. Law et al. [Law et al. 2003] 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.

9.6. Constructing Evidence Based Guidelines

Data mining techniques such as association rule mining, sequential rule mining and regression approaches can be used to develop and test existing guidelines[Liu and Rubin 2012],[Eibling et al. 2014]. EBG's have been developed for diseases treated in the emergency department [Agrawal and Kosowsky 2009], medication therapy for upper respiratory tract infection [Zeng et al. 2014], ear, nose, diabetes mellitus type 2 (T2DM) [Heselmans et al. 2013], prosthodontics [Bidra 2014], 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 [Kinnunen-Amoroso 2013] and Singapore[Goh et al. 2014]). Pivovarov et al. [Pivovarov et al. 2014] 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.

9.7. Adverse Event Detection

Extensive research on ADRs has been performed in the context of cardiovascular complications, pancreatic complications [Vilar et al. 2012; Harpaz et al. 2013] and allergies [Epstein et al.]. Supervised techniques, such as disproportionality analysis, logistic regression, Bayesian inference and NLP techniques have frequently been used to discover ADRs [Iyer et al. ; Haerian et al. 2012]. Besides EHR data, some research has also used weblogs to identify ADRs [White et al. 2013]. Similarly, research has been conducted using statistical analysis to identify certain medications which lead to adverse effects [Haerian et al. 2012; White et al. 2013; Coloma et al. 2013]. Bobo et al. [Bobo et al. 2014] 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. [Pathak et al. 2013c; Shen et al. 2014] 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. Nachimuthu et al. [Nachimuthu et al. 2010] used them to identify fluctuations in glucose levels while Rose et al. [Rose et al. 2005] used them for patients admitted for hemodialysis. Such techniques can lead to a reduction in overall health care costs,access barriers [Davis et al. 2014; van der Velde et al. 2012], unnecessary hospital admissions, frequency of primary care visits and improvement in illness prevention and care co-ordination [Khan et al. 2012].

10. DISCUSSION

Applications	Descriptive		Cross Sectional	Retrospective or Case-Control	
	Atemporal	Temporal		Time Agnostic	Time Aware
Understanding the Natural History of Disease	[Dunteman 1989; Zhang and Zhang 2002; Cao et al. 2005; Holmes et al. 2011; Hanauer et al. 2009]	[Hanauer and Ramakrishnan ; Liao and Chen 2013; Hripcsak and Albers 2013; Sacchi et al. 2007; Jin et al. 2008; Batal et al. 1994]	[Dasgupta and Chawla]		[Nachimuthu et al. 2010; Verduijn et al. 2007; Batal et al. 2012]

Cohort Identification	[Roque et al. 2011; Bauer-Mehren et al. 2013]	[Gotz et al. 2014; Schulam et al. 2015]	[Wang et al. 2013; Peissig et al. 2012; Pathak et al. 2012; Nadkarni et al. 2014]	[Carroll et al. 2011; Xu et al. 2011; Sarvestani et al. 2010]	[Albers et al. 2014]
Risk Prediction/Biomarker Discovery	[Gotz et al. 2011a; Vellanki et al. 2014]		[Sarkar et al. 2012; Letham et al. 2013; Ebadollahi et al. 2010; Feldman and Chawla 2014; Ngufor et al. 2015; Byrd et al. 2014]	[Collins et al. 2011a; Simon et al. 2011b; Mani et al. ; Matheny et al. ; Pakhomov et al. 2011; Fung et al. 2008; Wells et al. 2008; Vinzamuri et al. 2014; van der Heijden et al. 2014; Paxton et al. 2013; Zhao and Weng 2011; Maroco et al. 2011; Breault et al. ; Wang et al. 2015; Sun et al. 2014]	[Gatti et al. 2011]
Predicting the next complication: What and When				[Westra et al. 2011; Skevofilakas et al. 2010; Oztekin et al. 2009; Ghalwash and Obradovic]	[Sandri et al. 2014; Vinzamuri and Reddy 2013; Peelen et al. 2010]
Quantifying the effect of Intervention				[Schrom et al. 2013a]	
Patient Medical Trajectories	[Wang et al. 2014]	[Ghassempour et al. 2014]			[Yadav et al. 2015]

Constructing evidence based guidelines		[Pivovarov et al. 2014]			
Adverse Event Detection			[Sathyanarayana et al. 2014; White et al. 2013; Iyer et al. ; Haerian et al. 2012; Pathak et al. 2013a; Carroll et al. 2011; Xu et al. 2011; Bobo et al. 2014]		

Table 1 provides a succinct representation of the major research carried out using data mining techniques in conjunction with EHRs. The columns in Table-1 represent the methodologies categorized into groups,which we presented in section 8. The rows in Table-1 correspond to the major application areas,which we broadly discussed in section 9. 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 will analyze why certain application areas of EHRs are widely popular and why other application areas are virtually unexplored. We also discuss why certain data mining techniques are widely utilized by clinical researchers and the underlying reasons for such usage. We aim to discuss these aforementioned observations using our expertise across domains such as computer science, biostatistics and epidemiology. Lastly, we conclude this section by presenting our views on what we believe the future holds for data mining in conjunction with EHRs.

A quick glance at Table 1 reveals that substantially more work has been carried out in retrospective or case-control settings as compared to the descriptive 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 carried out in descriptive settings often leads to the discovery of redundant, simple statistical observations and widely known facts. This can be attributed to the fact that high dimensionality and associated heterogeneity of EHR data often lead to increased complexity thereby requiring large amounts of EHRs data to discover meaningful relationships through descriptive techniques. With EHRs still in its infancy, existing data repositories are moderate in size.

Another significant observation from Table 1 is that risk prediction 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. T2DM) 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. Another popular application area is understanding the natural history of the disease. This is attributed to the fact that such areas substantially overlaps with existing disciplines such as epidemiology. We also observe that there are certain application areas that are sparsely filled. Examples of such areas as observed in Table-1 are quantifying the effect of interventions and constructing evidence based guidelines. This is attributed to the fact that performing research in these areas is complicated as it requires large interdisciplinary teams and sophisticated techniques.

We also observe that utilizing data mining techniques in certain application areas are infeasible. Examples include the usage of descriptive data mining techniques for predicting complications, quantifying the effect of interventions and analyzing adverse event detections. This can be attributed to the fact that in such research activities patients are already grouped into cases and controls and therefore such research activities would be classified as case-control, retrospective or cross-sectional studies. We also observed that current research activities does not utilize the temporality associated with EHR data. For example, descriptive atemporal studies are more frequently conducted as compared to descriptive temporal techniques. Similar is the case with retrospective or case-control studies. We identified couple of reasons for this phenomenon. 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.

11. FUTURE OF EHR MINING

The current landscape of health care and the drivers that shape it virtually ensure that mining EHR data will play an increasingly important role in the future. Major examples of these drivers include the transition from the current reimbursement-based health care model to the Accountable Care Organization model; personalizing care to make it safer, more efficient and to reduce waste; and shared decision making, patients' desire to become more involved in their own care. All of these drivers require strong analytics based on large populations to be successful.

The cornerstone of modern medicine is clinical evidence and generating hypotheses for clinical evidence is one role that data mining is likely to play. This role is not a replacement of clinical trials but a synergistic role,where data mining can create high-quality hypotheses that can be validated through clinical trials. Clinical trials are expensive, thus the number of patients participating in a study is kept to the minimum required to validate the hypothesis. Secondary use of such small data sets for hypothesis generation is impractical, creating an opportunity for mining EHR data.

Even before the broad availability of EHR-based clinical data, large-scale observational studies over entire populations have been carried out. However, these studies have typically utilized claims data, which paints a less comprehensive picture of the patient than EHR (clinical) data. Claims data are recordings of charges from providers to the insurance companies, including prescriptions, diagnoses and laboratory test orders; and they do not contain vital signs, patient-provided health information (e.g. smoking status), the results of the lab tests and even the diagnoses are limited to a small set of the most costly conditions. Despite the well-documented shortcomings of

claims data, studies utilizing them exhibit many of the key characteristics of EHR data, including the intermittent generation of longitudinal data, censoring, confounding and the need for a robust study design. Methods developed for observational studies are adopted to EHR mining and they form the epidemiological foundation of EHR mining.

The evolution of EHR data mining will involve the integration of these epidemiological concepts with traditional data mining methodologies, developing them further in many directions.

One of these directions is data representation. EHR data is rich, it consists of highly heterogeneous data, collected from a wide range of sources (structured and unstructured clinical data, images, omics data, mobile health) potentially longitudinally at varying frequencies and resolution. The emergence of such data type heterogeneity necessitates that we revisit even fundamental questions like how data is best represented for modeling purposes.

Other directions include analyzing temporal and sequence data, handling missing data, and causal inference. Both data mining and biostatistics/epidemiology have methods to address these issues but they need to be further developed to suit EHR data better.

The characteristics of EHR data that drive the development of new data mining techniques are not unique to EHR data. Heterogeneity is present and poses challenges in many areas, including earth science and climate; time-to-event data finds its origins in failure analysis; and censoring along with the intermittent interactions with customers also happens in customer relationship management and recommendation systems. Mirroring how techniques for censored data and causal inference are being adopted from biostatistics and epidemiology into clinical data mining, new developments in EHR mining will likely find applications in many other data mining and analytics domains.

12. CONCLUSION

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.

13. 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 [Anderson et al. 2003]. 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 [Rydén et al. 2007] 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 [Lévesque et al. 2010], 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 T2DM, 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 [Elixhauser et al. 2008], 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 [Li et al. 2013].

Another important goal for phenotyping is to harmonize disease definitions across time. The clinical criterion for diabetes has changed [Wareham and ORahilly 1998] 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 [Conway et al. 2011; Li et al. 2013; Chen et al. 2013].

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.

13.1. 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.

13.2. 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 subpopulation. 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.

13.2.1. Framingham Score. Let us start our discussion of cohort studies towards risk estimation with the venerable Framingham Diabetes Score [Wilson et al. 2007] 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 non-diabetic. 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 [Collins et al. 2011b], but the Framingham score is the one with widest acceptance in clinical practice.

13.2.2. 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 [Wilson et al. 2007]) 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 13.2, 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 [Simon et al. 2013b] 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.

13.2.3. 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 [Macedo et al. 2014] Most studies have found statin to have no effect on progression to overt diabetes,some found it to be beneficial [Taylor et al. 2013] and some found it to be detrimental [Rajpathak et al. 2009].

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 [Schrom et al. 2013b] 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.

13.2.4. 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.

ACKNOWLEDGMENTS

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

(????). http://www.un.org/esa/population/techcoop/DemEst/manual10/chapter8.pdf

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

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

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

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

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

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

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

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

Frederick A Anderson, H Brownell Wheeler, Robert J Goldberg, David W Hosmer, Nilima A Patwardhan, Borko Jovanovic, Ann Forcier, and James E Dalen. 1991. 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 (1991), 933–938.

Robert N Anderson, National Center for Health Statistics (US), and others. 2003. Deaths: leading causes for 2001. (2003).

Peter C Austin. 2007. 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 (2007), 1128–1135.

Peter C Austin, Douglas S Lee, Ewout W Steyerberg, and Jack V Tu. 2012. 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 (Sept. 2012), 657–673. DOI:http://dx.doi.org/10.1002/bimj.201100251

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

Luciano Babuin, Vlad C Vasile, Jose A Rio Perez, Jorge R Alegria, High-Seng Chai, Bekele Afessa, and Allan S Jaffe. 2008. 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 (2008), 759–765.

Sistine A Barretto, Jim Warren, Andrew Goodchild, Linda Bird, Sam Heard, and Markus Stumptner. 2003. Linking guidelines to Electronic Health Record design for improved chronic disease management. In *AMIA Annual Symposium Proceedings*, Vol. 2003. American Medical Informatics Association, 66.

Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. 2012. 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* (2012), 280. DOI:http://dx.doi.org/10.1145/2339530.2339578

Iyad Batal and Milos Hauskrecht. 2010. Mining Clinical Data using Minimal Predictive Rules. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010 (Jan. 2010), 31–35. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041386\&tool=pmcentrez\&rendertype=abstract

Iyad Batal, Lucia Sacchi, and Riccardo Bellazzi. 1994. Multivariate Time Series Classification with Temporal Abstractions. (1994).

Anna Bauer-Mehren, Paea Lependu, Srinivasan V Iyer, Rave Harpaz, Nicholas J Leeper, and Nigam H Shah. 2013. Network analysis of unstructured EHR data for clinical research. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science* 2013 (Jan. 2013), 14–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3845760\&tool=pmcentrez\&rendertype=abstract

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

Valerie Beral, Carol Hermon, Clifford Kay, Philip Hannaford, Sarah Darby, and Gillian Reeves. 1999. 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 (1999), 96–100.

Pavel Berkhin. 2006. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer, 25–71.

D Bider, A Livshitz, I Tur Kaspa, A Shulman, J Levron, and J Dor. 1999. 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 (1999), 221–226.

Avinash S Bidra. 2014. Evidence-based prosthodontics: fundamental considerations, limitations, and guidelines. *Dental clinics of North America* 58, 1 (Jan. 2014), 1–17. DOI:http://dx.doi.org/10.1016/j.cden.2013.09.001

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

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. 2014. An electronic health record driven algorithm to identify incident antidepressant medication users. *Journal of the American Medical Informatics Association* 21, 5 (2014), 785–791.

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

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. http://www.ncbi.nlm.nih.gov/pubmed/12234716

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

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

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

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

Hui Cao, Marianthi Markatou, Genevieve B Melton, Michael F Chiang, and George Hripcsak. 2005. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* (Jan. 2005), 106–110. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560759\&tool=pmcentrez\&rendertype=abstract

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

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

Robert J Carroll, Anne E Eyler, and Joshua C Denny. 2011. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2011 (Jan. 2011), 189–96. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243261\&tool=pmcentrez\&rendertype=abstract

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

An-Wen Chan, Asbjø rn Hróbjartsson, Karsten J Jø rgensen, Peter C Gø tzsche, and Douglas G Altman. 2008. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ (Clinical research ed.)* 337 (Jan. 2008), a2299. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2600604\&tool=pmcentrez\&rendertype=abstract

Ying-Jui Chang, Min-Li Yeh, Yu-Chuan Li, Chien-Yeh Hsu, Chao-Cheng Lin, Meng-Shiuan Hsu, and Wen-Ta Chiu. 2011. Predicting hospital-acquired infections by scoring system with simple parameters. *PloS one* 6, 8 (Jan. 2011), e23137. DOI:http://dx.doi.org/10.1371/journal.pone.0023137

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

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

2012 (Jan. 2012), 103–11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540449\&tool=pmcentrez\&rendertype=abstract

Ching-Fan Chung, Peter Schmidt, and Ana D Witte. 1991. Survival analysis: A survey. *Journal of Quantitative Criminology* 7, 1 (1991), 59–98.

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

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

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. 2008. Cancer survival in five continents: a worldwide population-based study (CONCORD). *The lancet oncology* 9, 8 (Aug. 2008), 730–56. DOI:http://dx.doi.org/10.1016/S1470-2045(08)70179-7

Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. 2011a. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine* 9, 1 (Jan. 2011), 103. DOI:http://dx.doi.org/10.1186/1741-7015-9-103

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

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. 2013. 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 (Jan. 2013), 13–23. DOI:http://dx.doi.org/10.1007/s40264-012-0002-x

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, and others. 2011. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. In *AMIA annual symposium proceedings*, Vol. 2011. American Medical Informatics Association, 274.

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

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

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

Dipanwita Dasgupta and Nitesh V Chawla. DISEASE AND MEDICATION NETWORKS: AN INSIGHT INTO DISEASE-DRUG INTERACTIONS. (????).

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

Melinda M Davis, Jillian M Currey, Sonya Howk, Molly R DeSordi, Linda Boise, Lyle J Fagnan, and Nancy Vuckovic. 2014. 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 (Jan. 2014), 69–78. DOI:http://dx.doi.org/10.1111/jrh.12027

Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002), 151–161.

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. 1986. Predicting cardiac complications in patients undergoing non-cardiac surgery. *Journal of general internal medicine* 1, 4 (1986), 211–219.

Richard Doll, Richard Peto, Jillian Boreham, Isabelle Sutherland, and others. 2000. Smoking and dementia in male British doctors: prospective study. *Bmj* 320, 7242 (2000), 1097–1102.

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, and others. 2005. 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 (2005), 1279–1289.

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

Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. 2014. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 133, 1 (Jan. 2014), e54–63. DOI:http://dx.doi.org/10.1542/peds.2013-0819

Alistair Dunn and Alison Macfarlane. 1996. 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 (1996), F10–F19.

George H Dunteman. 1989. *Principal components analysis*. Number 69. Sage.

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

David Eibling, Marvin Fried, Andrew Blitzer, and Gregory Postma. 2014. Commentary on the role of expert opinion in developing evidence-based guidelines. *The Laryngoscope* 124, 2 (Feb. 2014), 355–7. DOI:http://dx.doi.org/10.1002/lary.24175

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

A Elixhauser, C Steiner, and L Palmer. 2008. Clinical classifications software (CCS). *Book Clinical Classifications Software (CCS)(Editor edˆ eds)* (2008).

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. DOI:http://dx.doi.org/10.1136/amiajnl-2013-001756

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

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

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

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

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

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

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

Marilyn J Field, Kathleen N Lohr, and others. 1992. *Guidelines for Clinical Practice:: From Development to Use*. National Academies Press.

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. 2009. Impact of electronic health record-based alerts on influenza vaccination for children with asthma. *Pediatrics* 124, 1 (2009), 159–169.

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, and others. 2011. Predicting risk for serious complications with bariatric surgery: results from the Michigan Bariatric Surgery Collaborative. *Annals of surgery* 254, 4 (2011), 633–640.

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. DOI:http://dx.doi.org/10.1370/afm.516

Jeff Friedlin, Marc Overhage, Mohammed A Al-Haddad, Joshua A Waters, J Juan R Aguilar-Saavedra, Joe Kesterson, and Max Schmidt. 2010. Comparing methods for identifying pancreatic cancer patients

using electronic data sources. In *AMIA Annual Symposium Proceedings*, Vol. 2010. American Medical Informatics Association, 237.

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

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

Glenn Fung, Shipeng Yu, Cary Dehing-Oberije, Dirk De Ruysscher, Philippe Lambin, Sriram Krishnan, and R Rao Bharat. 2008. Privacy-Preserving Predictive Models for Lung Cancer Survival Analysis. *Practical Privacy-Preserving Data Mining* (2008), 40.

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

E. Gatti, D. Luciani, and F. Stella. 2011. A continuous time Bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal* 24, 4 (Dec. 2011), 496–515. DOI:http://dx.doi.org/10.1007/s10696-011-9131-2

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)*.

Shima Ghassempour, Federico Girosi, and Anthony Maeder. 2014. Clustering Multivariate Time Series Using Hidden Markov Models. *International journal of environmental research and public health* 11, 3 (2014), 2741–2763.

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

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. 2014. Ministry of health clinical practice guidelines: diabetes mellitus. *Singapore medical journal* 55, 6 (June 2014), 334–47. http://www.ncbi.nlm.nih.gov/pubmed/25017409

Shivapratap Gopakumar, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2014. Stabilizing Sparse Cox Model using Clinical Structures in Electronic Medical Records. *arXiv preprint arXiv:1407.6094* (2014).

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

David Gotz, Harry Stavropoulos, Jimeng Sun, and Fei Wang. 2012a. ICDA: a platform for intelligent care delivery analytics. In *AMIA annual symposium proceedings*, Vol. 2012. American Medical Informatics Association, 264.

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

David Gotz, Jimeng Sun, Nan Cao, and Shahram Ebadollahi. 2011a. Visual cluster analysis in support of clinical decision intelligence. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2011 (Jan. 2011), 481–90. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243281\&tool=pmcentrez\&rendertype=abstract

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

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

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. DOI:http://dx.doi.org/10.1111/j.1552-6909.2012.01352.x

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

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

Stephen B Gruber. 1986. Clinical Epidemiology: The Architecture of Clinical Research. *The Yale journal of biology and medicine* 59, 1 (1986), 77.

J H Gurwitz, N F Col, and J Avorn. 1992. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA : the journal of the American Medical Association* 268, 11 (Sept. 1992), 1417–22. http://www.ncbi.nlm.nih.gov/pubmed/1512909

K Haerian, D Varn, S Vaidya, L Ena, H S Chase, and C Friedman. 2012. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical pharmacology and therapeutics* 92, 2 (Aug. 2012), 228–34. DOI:http://dx.doi.org/10.1038/clpt.2012.54

Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. 2014. Using Anchors to Estimate Clinical State without Labeled Data. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 606.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery* 15, 1 (2007), 55–86.

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. DOI:http://dx.doi.org/10.1136/amiajnl-2012-001117

David A Hanauer, Daniel R Rhodes, and Arul M Chinnaiyan. 2009. Exploring clinical associations using '-omics' based enrichment analyses. *PloS one* 4, 4 (Jan. 2009), e5203. DOI:http://dx.doi.org/10.1371/journal.pone.0005203

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

Rave Harpaz, Santiago Vilar, William Dumouchel, Hojjat Salmasian, Krystl Haerian, Nigam H Shah, Herbert S Chase, and Carol Friedman. 2013. 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 (May 2013), 413–9. DOI:http://dx.doi.org/10.1136/amiajnl-2012-000930

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

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

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

Vitaly Herasevich, Daryl J Kor, Arun Subramanian, and Brian W Pickering. 2013. Connecting the dots: rule-based decision support systems in the modern EMR era. *Journal of clinical monitoring and computing* 27, 4 (Aug. 2013), 443–8. DOI:http://dx.doi.org/10.1007/s10877-013-9445-6

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

Annemie Heselmans, Stijn Van de Velde, Dirk Ramaekers, Robert Vander Stichele, and Bert Aertgeerts. 2013. 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 (Jan. 2013), 83. DOI:http://dx.doi.org/10.1186/1748-5908-8-83

Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. 2000. Algorithms for association rule mininga general survey and comparison. *ACM sigkdd explorations newsletter* 2, 1 (2000), 58–64.

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

Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014a. Extracting Phenotypes from Patient Claim Records Using Nonnegative Tensor Factorization. In *Brain Informatics and Health*. Springer, 142–151.

Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014b. 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*. ACM, 115–124.

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

Antony B Holmes, Alexander Hawson, Feng Liu, Carol Friedman, Hossein Khiabanian, and Raul Rabadan. 2011. Discovering disease associations by integrating electronic clinical data and medical literature. *PloS one* 6, 6 (Jan. 2011), e21132. DOI:http://dx.doi.org/10.1371/journal.pone.0021132

George Hripcsak and David J Albers. 2013. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association : JAMIA* 20, e2 (Dec. 2013), e311–8. DOI:http://dx.doi.org/10.1136/amiajnl-2013-001922

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

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

Sandy H Huang, Paea LePendu, Srinivasan V Iyer, Ming Tai-Seale, David Carrell, and Nigam H Shah. 2014. Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association : JAMIA* (July 2014). DOI:http://dx.doi.org/10.1136/amiajnl-2014-002733

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

Kenji Ikeda, Hiromitsu Kumada, Satoshi Saitoh, Yasuji Arase, and Kazuaki Chayama. 1991. 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 (1991), 2150–2154.

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

Yuri Ito, Akiko Ioka, Hideaki Tsukuma, Wakiko Ajiki, Tomoyuki Sugimoto, Bernard Rachet, and Michel P Coleman. 2009. Regional differences in population-based cancer survival between six prefectures in Japan: application of relative survival models with funnel plots. *Cancer science* 100, 7 (July 2009), 1306–11. DOI:http://dx.doi.org/10.1111/j.1349-7006.2009.01170.x

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. DOI:http://dx.doi.org/10.1136/amiajnl-2013-001612

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

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

B Jaremin, E Kotulak, M Starnawska, W Mroziński, and E Wojciechowski. 1996. 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 (1996), 405–416.

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. 2014. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications* 5 (2014).

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

Huidong Warren Jin, Jie Chen, Hongxing He, Graham J Williams, Chris Kelman, and Christine M O'Keefe. 2008. 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 (July 2008), 488–500. DOI:http://dx.doi.org/10.1109/TITB.2007.900808

Jeffrey V Johnson and Ellen M Hall. 1988. 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 (1988), 1336–1342.

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

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. 2008. 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 (July 2008), 31–9. DOI:http://dx.doi.org/10.1016/S0140-6736(08)60880-6

Leila Kalankesh, James Weatherall, Thamer Ba-Dhfari, Iain Buchan, and Andy Brass. 2013. Taming EHR data: using semantic similarity to reduce dimensionality. *Studies in health technology and informatics* 192 (Jan. 2013), 52–6. http://www.ncbi.nlm.nih.gov/pubmed/23920514

Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. 2015. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *Journal of biomedical informatics* 53 (2015), 277–290.

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

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

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

Shreyas Karnik, Sin Lam Tan, Bess Berg, Ingrid Glurich, Jinfeng Zhang, Humberto J Vidaillet, C David Page, and Rajesh Chowdhary. 2012. 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 (Jan. 2012), 5562–5. DOI:http://dx.doi.org/10.1109/EMBC.2012.6347254

Michael W Kattan, Kenneth R Hess, and J Robert Beck. 1998. 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 (1998), 363–373.

Emily Kawaler, Alexander Cobian, Peggy Peissig, Deanna Cross, Steve Yale, and Mark Craven. 2012. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2012 (Jan. 2012), 436–45. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540493\&tool=pmcentrez\&rendertype=abstract

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

Jennifer L Kelsey. 1996. *Methods in observational epidemiology*. Vol. 26. Oxford University Press, USA.

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

Wajahat Ali Khan, Maqbool Hussain, Muhammad Afzal, Muhammad Bilal Amin, and Sungyoung Lee. 2012. 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 (Jan. 2012), 1274–7. DOI:http://dx.doi.org/10.1109/EMBC.2012.6346170

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

Kyung S Kim, Willis L Owen, Deborah Williams, and Lucile L Adams-Campbell. 2000. A comparison between BMI and Conicity index on predicting coronary heart disease: the Framingham Heart Study. *Annals of epidemiology* 10, 7 (2000), 424–431.

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

Maritta Kinnunen-Amoroso. 2013. How occupational health care professionals experience evidence-based guidelines in Finland: a qualitative study. *Journal of evaluation in clinical practice* 19, 4 (Aug. 2013), 612–6. DOI:http://dx.doi.org/10.1111/j.1365-2753.2012.01850.x

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

John P Klein and Melvin L Moeschberger. 2005. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

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. 1991. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest Journal* 100, 6 (1991), 1619–1636.

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

Oliver Kuss, Benita von Salviati, and Jochen Börgermann. 2010. 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 (2010), 829–835.

Nan M Laird and James H Ware. 1982. Random-effects models for longitudinal data. *Biometrics* (1982), 963–974.

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

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

John M Last, International Epidemiological Association, and others. 2001. *A dictionary of epidemiology*. Vol. 141. Oxford Univ Press.

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

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

Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2013. An interpretable stroke prediction model using rules and Bayesian analysis. (2013).

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

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

Katharine Levit, Cynthia Smith, Cathy Cowan, Helen Lazenby, Art Sensenig, and Aaron Catlin. 2003. Trends in US health care spending, 2001. *Health Affairs* 22, 1 (2003), 154–164.

Dingcheng Li, Gyorgy Simon, Christopher G Chute, and Jyotishman Pathak. 2013. Using Association Rule Mining for Phenotype Extraction from Electronic Health Records. *AMIA Summits on Translational Science Proceedings* 2013 (2013), 142.

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. DOI:http://dx.doi.org/10.1136/amiajnl-2013-001718

Kung-Yee Liang, Steven G Self, and Xinhua Liu. 1990. The Cox proportional hazards model with change point: An epidemiologic application. *Biometrics* (1990), 783–793.

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

Vance Liao and Ming-Syan Chen. 2013. Efficient mining gapped sequential patterns for motifs in biological sequences. *BMC systems biology* 7 Suppl 4 (Jan. 2013), S7. DOI:http://dx.doi.org/10.1186/1752-0509-7-S4-S7

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

Yueyi I Liu and Daniel L Rubin. 2012. The role of informatics in health care reform. *Academic radiology* 19, 9 (Sept. 2012), 1094–9. DOI:http://dx.doi.org/10.1016/j.acra.2012.05.006

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

Zitao Liu and Milos Hauskrecht. 2013b. Sparse linear dynamical system with its application in multivariate clinical time series. *arXiv preprint arXiv:1311.7071* (2013).

Zitao Liu and Milos Hauskrecht. 2014. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial intelligence in medicine* (2014).

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

Peter J F Lucas, Linda C van der Gaag, and Ameen Abu-Hanna. 2004. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine* 30, 3 (March 2004), 201–14. DOI:http://dx.doi.org/10.1016/j.artmed.2003.11.001

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

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

Subramani Mani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny. 2012. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2012 (Jan. 2012), 606–15. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540444\&tool=pmcentrez\&rendertype=abstract

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. DOI:http://dx.doi.org/10.1136/amiajnl-2013-001854

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

Teri A Manolio. 2009. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. (2009).

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

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

João Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendonça. 2011. 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 (Jan. 2011), 299. DOI:http://dx.doi.org/10.1186/1756-0500-4-299

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

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. DOI:http://dx.doi.org/10.1177/0272989X10364246

Jason Scott Mathias, Ankit Agrawal, Joe Feinglass, Andrew J Cooper, David William Baker, and Alok Choudhary. 2013. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *Journal of the American Medical Informatics Association : JAMIA* 20, e1 (June 2013), e118–24. DOI:http://dx.doi.org/10.1136/amiajnl-2012-001360

John NS Matthews. 2006. *Introduction to randomized controlled clinical trials*. CRC Press.

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

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

James B Meigs, Enrico Cagliero, Anil Dubey, Patricia Murphy-Sheehy, Catharyn Gildesgame, Henry Chueh, Michael J Barry, Daniel E Singer, and David M Nathan. 2003. A Controlled Trial of Web-Based Diabetes Disease Management The MGH Diabetes Primary Care Improvement Project. *Diabetes Care* 26, 3 (2003), 750–757.

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

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

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

Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. 2014. 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 (2014), 299–307.

W S Moore, B Young, W H Baker, J T Robertson, J F Toole, C L Vescera, and V J Howard. 1996. Surgical results: a justification of the surgeon selection process for the ACAS trial. The ACAS Investigators. *Journal of vascular surgery* 23, 2 (Feb. 1996), 323–8. http://www.ncbi.nlm.nih.gov/pubmed/8637110

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

Senthil K Nachimuthu, Anthony Wong, and Peter J Haug. 2010. Modeling Glucose Homeostasis and Insulin Dosing in an Intensive Care Unit using Dynamic Bayesian Networks. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010 (Jan. 2010), 532–6. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041323\&tool=pmcentrez\&rendertype=abstract

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

Arijit Nandi, M Maria Glymour, Ichiro Kawachi, and Tyler J VanderWeele. 2012. 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 (2012), 223.

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, and others. 2013. 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 (2013), e147–e154.

Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. 2014. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics* 48 (2014), 160–170.

Che Ngufor, Sudhindra Upadhyaya, Dennis Murphree, Nageswar Madde, Daryl Kor, and Jyotishman Pathak. 2015. A Heterogeneous Multi-Task Learning for Predicting RBC Transfusion and Perioperative Outcomes. In *Artificial Intelligence in Medicine*. Springer, 287–297.

Etsuko Nomura, Hideaki Tsukuma, Wakiko Ajiki, and Akira Oshima. 2003. Population-based study of relationship between hospital surgical volume and 5-year survival of stomach cancer patients in Osaka, Japan. *Cancer science* 94, 11 (Nov. 2003), 998–1002. http://www.ncbi.nlm.nih.gov/pubmed/14611678

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

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

M Olschewski, M Schumacher, and K B Davis. 1992. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled clinical trials* 13, 3 (June 1992), 226–39. http://www.ncbi.nlm.nih.gov/pubmed/1320559

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

Yoneatsu Osaki and Masumi Minowa. 2001. Factors associated with earthquake deaths in the great Hanshin-Awaji earthquake, 1995. *American journal of epidemiology* 153, 2 (2001), 153–156.

Casey Lynnette Overby, Jyotishman Pathak, Omri Gottesman, Krystl Haerian, Adler Perotte, Sean Murphy, Kevin Bruce, Stephanie Johnson, Jayant Talwalkar, Yufeng Shen, and others. 2013. 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 (2013), e243–e252.

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

Asil Oztekin, Dursun Delen, and Zhenyu James Kong. 2009. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *International journal of medical informatics* 78, 12 (Dec. 2009), e84–96. DOI:http://dx.doi.org/10.1016/j.ijmedinf.2009.04.007

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. 2011. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. *The Annals of Family Medicine* 9, 1 (2011), 12–21.

Serguei V S Pakhomov, Nilay D Shah, Holly K Van Houten, Penny L Hanson, and Steven A Smith. 2011. 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 (Jan. 2011), 1080–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243229\&tool=pmcentrez\&rendertype=abstract

M Panahiazar, V Taslimitehrani, N Pereira, and J Pathak. 2015. 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*, Vol. 216. IOS Press, 40.

Maryam Panahiazar, Vahid Taslimitehrani, Naveen L Pereira, and Jyotishman Pathak. 2014. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Studies in health technology and informatics* 210 (2014), 369–373.

Yubin Park and Joydeb Ghosh. 2014a. Ensembles of $(\{alpha\})$-Trees for Imbalanced Classification Problems. *Knowledge and Data Engineering, IEEE Transactions on* 26, 1 (2014), 131–143.

Yubin Park and Joydeep Ghosh. 2014b. A Hierarchical Ensemble of α-Trees for Predicting Expensive Hospital Visits. In *Brain Informatics and Health*. Springer, 178–187.

E Passamani, K B Davis, M J Gillespie, and T Killip. 1985. A randomized trial of coronary artery bypass surgery. Survival of patients with a low ejection fraction. *The New England journal of medicine* 312, 26 (June 1985), 1665–71. DOI:http://dx.doi.org/10.1056/NEJM198506273122603

Jyotishman Pathak, Abel N Kho, and Joshua C Denny. 2013a. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association : JAMIA* 20, e2 (Dec. 2013), e206–11. DOI:http://dx.doi.org/10.1136/amiajnl-2013-002428

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

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

Jyotishman Pathak, Richard C Kiefer, and Christopher G Chute. 2013c. Mining Anti-coagulant Drug-Drug Interactions from Electronic Health Records Using Linked Data. In *Data Integration in the Life Sciences*. Springer, 128–140.

Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. 2013. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2013 (Jan. 2013), 1109–15. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900132\&tool=pmcentrez\&rendertype=abstract

Linda Peelen, Nicolette F de Keizer, Evert de Jonge, Robert-Jan Bosman, Ameen Abu-Hanna, and Niels Peek. 2010. Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit. *Journal of biomedical informatics* 43, 2 (April 2010), 273–86. DOI:http://dx.doi.org/10.1016/j.jbi.2009.10.002

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

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, and others. 2012. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association* 19, 2 (2012), 225–234.

Adam Perer and Jimeng Sun. 2012. MatrixFlow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, Vol. 2012. American Medical Informatics Association, 716.

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

Rimma Pivovarov, David J Albers, George Hripcsak, Jorge L Sepulveda, and Noémie Elhadad. 2014. Temporal trends of hemoglobin A1c testing. *Journal of the American Medical Informatics Association* 21, 6 (2014), 1038–1044.

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

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

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

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

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. 2012. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics* 13, 4 (March 2012), 407–18. DOI:http://dx.doi.org/10.2217/pgs.11.164

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

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. 2014. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *Journal of biomedical informatics* 51 (2014), 280–286.

Chandan K Reddy and Charu C Aggarwal. 2015. *HealtHcare Data analytics*. Vol. 36. CRC Press.

Uwe E Reinhardt, Peter S Hussey, and Gerard F Anderson. 2004. US health care spending in an international context. *Health Affairs* 23, 3 (2004), 10–25.

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.

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, and others. 2002. Incidence and prognostic importance of acute renal failure after percutaneous coronary intervention. *Circulation* 105, 19 (2002), 2259–2264.

Jennifer G Robinson, JoAnn E Manson, Joseph Larson, Simin Liu, Yiqing Song, Barbara V Howard, Lawrence Phillips, James M Shikany, Matthew Allison, J David Curb, and others. 2011. Lack of association between 25 (OH) D levels and incident type 2 diabetes in older women. *Diabetes Care* 34, 3 (2011), 628–634.

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

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

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

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. 2011. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology* 7, 8 (Aug. 2011), e1002141. DOI:http://dx.doi.org/10.1371/journal.pcbi.1002141

C. Rose, C. Smaili, and F. Charpillet. 2005. 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)*. IEEE, 5 pp.–598. DOI:http://dx.doi.org/10.1109/ICTAI.2005.7

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

KJ Rothman and S Greenland. 1986. Modern epidemiology.Boston: Little Brown and Company. *Study design for comparison of orthoses* 273 (1986), 101–106.

Peter M Rothwell. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 365, 9453 (????), 82–93. DOI:http://dx.doi.org/10.1016/S0140-6736(04)17670-8

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

Lucia Sacchi, Cristiana Larizza, Carlo Combi, and Riccardo Bellazzi. 2007. Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery* 15, 2 (June 2007), 217–247. DOI:http://dx.doi.org/10.1007/s10618-007-0077-7

R L Sacco, D E Kargman, Q Gu, and M C Zamanillo. 1995. Race-ethnicity and determinants of intracranial atherosclerotic cerebral infarction. The Northern Manhattan Stroke Study. *Stroke; a journal of cerebral circulation* 26, 1 (Jan. 1995), 14–20. http://www.ncbi.nlm.nih.gov/pubmed/7839388

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

David L Sackett. 2000. *Evidence-based medicine*. Wiley Online Library.

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

Micol Sandri, Paola Berchialla, Ileana Baldi, Dario Gregori, and Roberto Alberto De Blasi. 2014. Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. *Journal of biomedical informatics* 48 (April 2014), 106–13. DOI:http://dx.doi.org/10.1016/j.jbi.2013.12.008

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

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

A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh, and M. Salehi. 2010. Predicting breast cancer survivability using data mining techniques. In *2010 2nd International Conference on Software Technology and Engineering*, Vol. 2. IEEE, V2–227–V2–231. DOI:http://dx.doi.org/10.1109/ICSTE.2010.5608818

Aarti Sathyanarayana, Jyotishman Pathak, Rozalina McCoy, Santiago Romero-Brufau, Maryam Panaziahar, and Jaideep Srivastava. 2014. Clinical Decision Making: A Framework for Predicting Rx Response. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 1185–1188.

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. 2014. The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology* 29, 6 (2014), 439–451.

John R Schrom, Pedro J Caraballo, M Regina Castro, and György J Simon. 2013a. Quantifying the effect of statin use in pre-diabetic phenotypes discovered through association rule mining. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2013 (Jan. 2013), 1249–1257. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900142\&tool=pmcentrez\&rendertype=abstract

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

Peter Schulam, Fredrick Wigley, and Suchi Saria. 2015. Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

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

Kenneth F Schulz and David A Grimes. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 365, 9467 (????), 1348–53. DOI:http://dx.doi.org/10.1016/S0140-6736(05)61034-3

Mark A Schuster, Elizabeth A McGlynn, and Robert H Brook. 1998. How good is the quality of health care in the United States? *Milbank quarterly* 76, 4 (1998), 517–563.

Pamela J Schwingl, Howard W Ory, and Cynthia M Visness. 1999. 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 (1999), 241–249.

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

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. 1999. Lipoprotein (a)-cholesterol and coronary heart disease in the Framingham Heart Study. *Clinical chemistry* 45, 7 (1999), 1039–1046.

Dipankar Sengupta and Pradeep K Naik. 2013. SN algorithm: analysis of temporal clinical data for mining periodic patterns and impending augury. *Journal of clinical bioinformatics* 3, 1 (Jan. 2013), 24. DOI:http://dx.doi.org/10.1186/2043-9113-3-24

Y Shahar. 1997. A framework for knowledge-based temporal abstraction. *Artificial Intelligence* 90, 1-2 (Feb. 1997), 79–133. DOI:http://dx.doi.org/10.1016/S0004-3702(96)00025-2

Feichen Shen, Dingcheng Li, Hongfang Liu, Yugyung Lee, Jyotishman Pathak, Christopher G Chute, and Cui Tao. 2014. Using semantic web technologies for quality measure phenotyping algorithm represen-

tation and automatic execution on EHR data. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*.

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

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. 2010. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthcare informatics research* 16, 2 (June 2010), 77–81. DOI:http://dx.doi.org/10.4258/hir.2010.16.2.77

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

Philip M Short, Samuel IW Lipworth, Douglas HJ Elder, Stuart Schembri, Brian J Lipworth, and others. 2011. Effect of β blockers in treatment of chronic obstructive pulmonary disease: a retrospective cohort study. *Bmj* 342 (2011), d2549.

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

Gyorgy J. Simon, Vipin Kumar, and Peter W. Li. 2011b. 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*. ACM Press, New York, New York, USA, 823. DOI:http://dx.doi.org/10.1145/2020408.2020550

Gyorgy J Simon, John Schrom, M Regina Castro, Peter W Li, and Pedro J Caraballo. 2013a. Survival association rule mining towards type 2 diabetes risk assessment. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2013 (Jan. 2013), 1293–302. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900145\&tool=pmcentrez\&rendertype=abstract

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

Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, and others. 2011a. Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of statistical software* 39, 5 (2011), 1–13.

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

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

Hardeep Singh, Aanand Dinkar Naik, Raghuram Rao, and Laura Ann Petersen. 2008. Reducing diagnostic errors through effective communication: harnessing the power of information technology. *Journal of General Internal Medicine* 23, 4 (2008), 489–494.

Marios Skevofilakas, Konstantia Zarkogianni, Basil G Karamanos, and Konstantina S Nikita. 2010. 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 (Jan. 2010), 6713–6. DOI:http://dx.doi.org/10.1109/IEMBS.2010.5626245

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

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

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

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

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

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. 2014. Predicting changes in hy-

pertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association* 21, 2 (2014), 337–344.

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

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, and others. 2006. *Introduction to data mining*. Vol. 1. Pearson Addison Wesley Boston.

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

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

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

Terry Therneau and Cindy Crowson. 2014. Using time dependent covariates and time dependent coefficients in the cox model. *Red* 2 (2014), 1.

Robert Tibshirani and others. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine* 16, 4 (1997), 385–395.

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

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

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

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

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

Maarten van der Heijden, Marina Velikova, and Peter J F Lucas. 2014. Learning Bayesian networks for clinical time series analysis. *Journal of biomedical informatics* 48 (April 2014), 94–105. DOI:http://dx.doi.org/10.1016/j.jbi.2013.12.007

E T van der Velde, H Foeken, T A Witteman, L van Erven, and M J Schalij. 2012. 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 (Feb. 2012), 66–70. DOI:http://dx.doi.org/10.1007/s12471-011-0234-x

Harriette G C Van Spall, Andrew Toren, Alex Kiss, and Robert A Fowler. 2007. 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 (March 2007), 1233–40. DOI:http://dx.doi.org/10.1001/jama.297.11.1233

Pratibha Vellanki, Thi Duong, Svetha Venkatesh, and Dinh Phung. 2014. Nonparametric Discovery of Learning Patterns and Autism Subgroups from Therapeutic Data. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 1828–1833.

Marion Verduijn, Lucia Sacchi, Niels Peek, Riccardo Bellazzi, Evert de Jonge, and Bas A J M de Mol. 2007. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. *Artificial intelligence in medicine* 41, 1 (Sept. 2007), 1–12. DOI:http://dx.doi.org/10.1016/j.artmed.2007.06.003

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

Rajakrishnan Vijayakrishnan, Steven R Steinhubl, Kenney Ng, Jimeng Sun, Roy J Byrd, Zahra Daar, Brent A Williams, Shahram Ebadollahi, Walter F Stewart, and others. 2014. 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 (2014), 459–464.

Santiago Vilar, Rave Harpaz, Lourdes Santana, Eugenio Uriarte, and Carol Friedman. 2012. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. *PloS one* 7, 7 (Jan. 2012), e41471. DOI:http://dx.doi.org/10.1371/journal.pone.0041471

Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. 2014. Active Learning based Survival Regression for Censored Data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 241–250.

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

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

Jeffrey J Walline. 2001. Designing Clinical Research: an Epidemiologic Approach. (2001).

Fei Wang, Jimeng Sun, and Shahram Ebadollahi. 2012. Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5, 1 (2012), 54–69.

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

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

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

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

Zheng Wang, Prithwish Chakraborty, Sumiko R Mekaru, John S Brownstein, Jieping Ye, and Naren Ramakrishnan. 2015. 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*. ACM, 1285–1294.

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

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

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

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

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

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

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

Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz. 2013. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association : JAMIA* 20, 3 (May 2013), 404–8. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001482

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

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

Jeffrey M Wooldridge. 1992. Some alternatives to the Box-Cox regression model. *International Economic Review* (1992), 935–955.

Adam Wright, Elizabeth S Chen, and Francine L Maloney. 2010a. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics* 43, 6 (Dec. 2010), 891–901. DOI:http://dx.doi.org/10.1016/j.jbi.2010.09.009

Adam Wright, Elizabeth S Chen, and Francine L Maloney. 2010b. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics* 43, 6 (Dec. 2010), 891–901. DOI:http://dx.doi.org/10.1016/j.jbi.2010.09.009

Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B Peterson, Qingxia Chen, Subramani Mani, Mia A Levy, Qi Dai, and Josh C Denny. 2011. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2011 (Jan. 2011), 1564–72. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243156\&tool=pmcentrez\&rendertype=abstract

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

Pranjul Yadav, Lisiane, Andrew, Sanjoy, Bonnie Katherine, Vipin Connie, and Gyorgy Simon Michael. 2015. Modelling Trajectories for Diabetes Complications. *SDM* (2015).

Kazue Yamaoka and Toshiro Tango. 2005. Efficacy of Lifestyle Education to Prevent Type 2 Diabetes A meta-analysis of randomized controlled trials. *Diabetes care* 28, 11 (2005), 2780–2786.

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

Andrew P Yu, Yanni F Yu, and Michael B Nichol. 2010. 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 (2010), 1038–1045.

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

Haijun Zhai, Patrick Brady, Qi Li, Todd Lingren, Yizhao Ni, Derek S Wheeler, and Imre Solti. 2014. 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 (Aug. 2014), 1065–71. DOI:http://dx.doi.org/10.1016/j.resuscitation.2014.04.009

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

Hao Helen Zhang and Wenbin Lu. 2007. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94, 3 (2007), 691–703.

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

Zhiyuan Zhang, David Gotz, and Adam Perer. 2014. Iterative cohort analysis and exploration. *Information Visualization* (2014), 1473871614526077.

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