Causal Pattern Mining in Highly Heterogeneous and Temporal EHRs Data

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Pranjul Yadav

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philiosophy

Dr. Vipin Kumar (Advisor) and Dr. Gyorgy Simon (Co-Advisor)

March, 2017

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school. First of all, I would like to thank my adviser, Professor Vipin Kumar, for his constant support, mentoring and guidance. The knowledge and wisdom he shared was very valuable during my thesis and would come quite handy, as I embark on my professional career.

I am also grateful to my Co-Adviser Dr. Simon for his knowledge and help towards my thesis. I am also thankful to my mentors Dr. Steinbach, Dr. Westra and Dr. Kuang for their valuable feedback and guidance throughout my dissertation.

I would also like to thank my lab mates Lisiane Prunelli, Sanjoy Dey, Gowtham Atluri, Sean Landman, Kevin Schiroo, Andrew Hangslebem, Anjali Srivastava, Alexander Hoff, Jia Li for their help towards my dissertation.

Dedication

I would like to dedicate this thesis to my parents and my sisters.

Abstract

The World Health Organization (WHO) estimates that the total healthcare spending in the U.S. is around 18% of its GDP for the year 2011. Even with such a high per-capita expenditure, the quality of healthcare in U.S. lags behind as compared to the healthcare in other industrialized countries. This inefficient state of the U.S. healthcare system is attributed to the current Fee-for-service (FFS) model. Under the FFS model, healthcare providers (doctors, hospitals) receive payments for every hospital visit or service rendered. The lack of coordination between the service providers and patient outcomes, leads to an increase in the costs associated with the healthcare management, as healthcare providers often recommend expensive treatments. Several legislations have been approved in the recent past to improve the overall U.S. healthcare management while simultaneously reducing the associated costs.

The HITECH Act, proposes to spend close to $30 billion dollars on creating a nationwide repository of electronic Health Records (EHRs). Such a repository would consist of patient attributes such as demographics, laboratories test results, vital information and diagnosis codes. It is hoped that this EHR repository will be a platform to improve care coordination between service providers and patients healthcare outcomes, reduce health disparities thereby improving the overall healthcare management system. Data collected and stored in the EHR (HITECH) and the need to improve care efficiency and outcome (ACT) would help to improve the current state of U.S. healthcare system. Data mining techniques in conjunction with EHRs can be used to develop novel clinical decision making tools, to analyze the prevalence and incidence of diseases and to evaluate the efficacy of existing clinical and surgical interventions.

In this thesis we focus on two key aspects of EHR data, i.e. temporality and causation. This becomes more important considering that the temporal nature of EHRs data has not been fully exploited. Further, increasing amounts of clinical evidence suggest that temporal nature is important for the development of clinical decision making tools and techniques. Secondly, several research articles hint at the the presence of antiquated clinical guidelines which are still in practice. In this dissertation, we first describe EHR along with the following terminologies : temporality, causation and heterogeneity.

Building on this, we then describe methodologies for extracting non-causal patterns in the absence of longitudinal data. Further, we describe methods to extract non-causal patterns in the presence of longitudinal data. We describe such methodologies in the context of Type-2 Diabetes Mellitus (T2DM). Furthermore, we describe techniques to extract simple and complex causal patterns from longitudinal data in the context of sepsis and T2DM. Finally, we conclude this dissertation, by providing a summary of our work along with future directions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The World Health Organization (WHO) estimates that the total healthcare spending in the U.S. is around 18% of its GDP for the year 2011. This represents a steady increase over the last few decades. Even with such a high per-capita expenditure, the quality of healthcare in U.S. lags behind as compared to the healthcare in other industrialized countries. This is corroborated by the fact that U.S. as compared to the other developed countries has relatively low life expectancy and high in-fact mortality rates.

The U.S. healthcare system is inefficient and wasteful. This can be attributed to the current Fee-for-service (FFS) model. Under the FFS model, healthcare providers(doctors, hospitals) receive payments for every hospital visit or service rendered. Hence, the focus is more on the services rendered to a patient and not on service outcomes. This lack of coordination between the service providers and patient outcomes, leads to an increase in the costs associated with the healthcare management, as healthcare providers often recommend expensive treatments.

Several legislations have been approved in the recent past to improve the overall U.S. healthcare management while simultaneously reducing the associated costs. These legislations aim to move from the FFS model towards the Accountable Care Organization (ACO) model. Under the ACO model, healthcare provider payments are closely tied with patient outcomes. In other words, healthcare providers are only paid if there is any improvement in patient health. Along with the ACO act, The Health Information Technology for Economic and Clinical Health (HITECH) Act is another legislation which aims to improve the overall healthcare management.

Under the HITECH Act, close to $30 billion dollars will be spent on creating a nationwide repository of electronic Health Records (EHRs). Such a repository [1] would consist of patient attributes such as demographics, laboratories test results, vital information and diagnosis codes. Further such information about patient health attributes will be collected over time. Storing patient healthcare records would enable better documentation of existing treatments and medical interventions. It is hoped that this EHR repository will be a platform to improve care coordination between service providers and patients healthcare outcomes, reduce health disparities thereby improving the overall healthcare management system. Further such a repository would lay the foundation for Evidence based Medicine.

Evidence based medicine is a mechanism to improve efficiency as required by the Accountable Care Act (ACA). Evidence based medicine is a medical practice which aims to bolster decision making by utilizing evidences from past conducted research for future patient healthcare recommendations and prescriptions. Such practices hypothesize that medical decision making, clinical guidelines should be based on best evidences as observed from research and not from an individual clinician's belief. EHR data could then be used to validate existing clinical guidelines as data related to medical prescriptions and recommendations along with the patient's health outcome over time is present in EHR. Further, EHRs could also be used for the development of novel clinical guideline and medical treatments.

The guidelines can be increasingly tailored to the patients to further reduce ineffective treatment and reduce waste, thereby providing the platform for the development of Precision Medicine. Precision medicine aims at the development and recommendation of customized medical treatments, clinical services and products. In such a model, patients are divided into subpopulations based on their genetic make-up or existing medical state. Different treatments are then prescribed for these subpopulations aiming to cure the same disease/outcome. An example could be a prescription of a drug which might lead to healthcare improvement in one population whereas an adverse side-effect in another sub-population. Hence, the drug would only be recommended for the subpopulation for which it has a beneficial effect.

Data collected and stored in the EHR (HITECH) and the need to improve care efficiency and outcome (ACT) would help to improve the current state of U.S. healthcare system. Evidence based medicine as the clinical framework and Clinical Decision Support (CDS) as the tool to manage and apply the accumulated knowledge, the time is ripe to apply data mining towards knowledgable discovery. Data mining techniques in conjunction with EHRs can be used to develop novel clinical decision making tools, to analyze the prevalence and incidence of diseases and to evaluate the efficacy of existing clinical and surgical interventions.

1.2 Scope

In this thesis, we aim to develop methods for extracting knowledge from EHR data. In particular, we would be using structured EHR datasets for analyzing our techniques and methodologies. Further, data obtained from parsing clinical notes using Natural Language Parsing (NLP) tools and techniques, human genetics data (bioinformatics) and medical imaging are beyond the scope of this thesis. Extracting knowledge from EHR data is challenging because of the following two reasons.

Firstly, Electronic Health Records (EHRs) is a rich source of longitudinal and time-series information consisting of patient demographic attributes, laboratories test results, vital information and diagnosis codes over time [2]. Traditional data mining techniques exist to analyze longitudinal datasets. However, there is still a need for the development of sophisticated techniques to analyze irregular time-series datasets [3, 4]. The need for such techniques becomes more relevant considering that temporal EHR data helps us to monitor the progression of patient's health from one medical state to a state of associated complications.

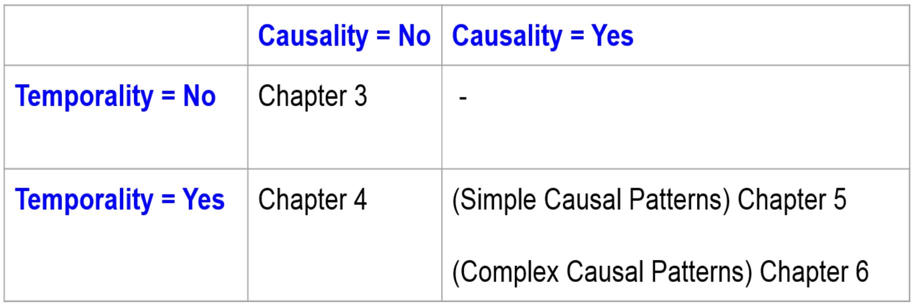
Secondly, major clinical research studies in the past have often focussed on association [5] thereby neglecting causation. In other words, studies have often analyzed the co-occurrence of symptoms, interventions and outcomes. However, the availability of longitudinal EHR data provides an opportunity to estimate the efficacy of clinical and surgical interventions for various outcomes of interest [6, 7, 8, 9].

This becomes more important considering our observations based on our recent survey literature. We observed that the temporal nature of EHRs data has not been fully exploited. Further, increasing amounts of clinical evidence suggest that temporal nature is important for the development of clinical decision making tools and techniques. Secondly, several research articles hint at the the presence of antiquated clinical guidelines which are still in practice. To overcome the aforementioned shortcomings associated with existing healthcare management,

In this thesis we focus on two key aspects of EHR data, i.e. temporality and causation. We would be describing these terminologies in greater detail in the next Section.

The overall structure of the thesis is as follows : In Section 2 we describe EHR along with the following terminologies : temporality, causation and heterogeneity. Building on this, in Section 3, we describe methodologies for extracting non-causal patterns in the absence of longitudinal data. In Section 4, we describe methods to extract non-causal patterns in the presence of longitudinal data. In Section 5, we describe techniques to extract simple causal patterns from longitudinal data. In Section 6, we describe methodologies to extract complex causal patterns from longitudinal data. Finally, we conclude the thesis in Section 7, by providing a summary of our work along with future directions.

Figure 1 2.2 provides a succinct representation of the layout of Chapters 3,4,5 and 6. Row labels indicate whether longitudinal data has been utilized in the development of the proposed techniques. Column labels indicate whether the methodologies are based on association or causation. All techniques proposed in this thesis are heterogeneity aware. Moreover, inferring causal inference in the absence of longitudinal data is not practical and hence would not be focussed upon.



	Causality = No	Causality = Yes
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	(Simple Causal Patterns) Chapter 5 (Complex Causal Patterns) Chapter 6

Figure 1.1: Thesis Succinct Representation

Chapter 2

Background

In this chapter, we will provide a brief description about EHRs data and its constituent data elements. Further we would also provide a brief discussion about the various terminologies we would be using in the following subsequent chapters i.e. causation, temporality and heterogeneity.

2.1 Electronic Health Records Data

Electronic Health Records consists of patient information such as demographics, laboratories test results, vitals information and diagnosis codes. Such information is usually collected when patients visits a healthcare provider. Data collected through such visits provide a mechanism to understand the disease incidence, prevalence and underlying disease mechanisms. Data elements such as demographics attributes, laboratories test results, diagnosis codes are usually stored in a databases. Information such as vitals signature are usually stored in flow-sheets which is a semi-structured data storage format. Clinical Notes are stored as paper records and hence require sophisticated NLP techniques to parse and extract information. Now we would be describing the data elements in greater detail :

- **Demographics Attributes:** Such data elements consist of patient information such as age, gender, ethnicity, marital status, medical insurance provider and others. Such information is usually collected and stored when patients visits the healthcare provider for the first time. This information is usually static in nature.

This attribute is highly heterogenous in nature. For example, gender, ethnicity are categorical attributes and age is a real-integer valued attribute.

- **Laboratories Test Results:** These elements are usually collected and stored based on a clinician's recommendation. Examples of such elements include readings for glucose, gFR, bilorubin etc. These data elements usually take continuous values.

- **Vital Signs:** They are usually collected whenever a patient visits a healthcare provider. Examples of such data elements include temperature, blood pressure, body weight, BMI etc. They also take continuous values.

- **Diagnosis Codes:** These data elements are usually stored when a patient is diagnosed with any disease. For example, a pre-defined code of 252.0 is usually assigned to a patient's health record if the patient is diagnosed with Type-2 Diabetes Mellitus.

- **Clinical Notes:**Such elements usually contain information typically recommended by a clinician . Examples include lifestyle recommendations, food recommendations and information about drug or alcohol consumption.

- Other data elements include information related to radiology reports, patient's genetic make-up and proteomics data of the patient. Such data elements are not collected and stored for common hospital visits.

2.1.1 Data Challenges

EHR datasets are prone to numerous challenges arising due to the nature of how such datasets are stored and collected. Examples of challenges associated with EHR datasets are censoring, missing data, irregular time-serious, class-imbalance problems, heterogeneity and biases. Now, we will discuss these challenges in greater detail.

- **Censoring:** This refers to the problem when information about a patient is only partially available. For example, consider the scenario where in a patient registers in a study , which started in 2012 and continued until 2016. Post some follow-ups

(2012-2014) the patient drops our of the study. Now there is no way to ascertain the medical state of the patient in 2015 and 2016. Further, there is no way to establish whether the study had a beneficial or detrimental effect on the patient. This kind of data is known as censored data. Censoring can take places in three ways i.e. right censored data, left censored data and interval censored data.

- **Missing data:** Missing data in EHRs arises due to numerous reasons. Firstly, missing data can arise due to fragmentation : a scenario when multiple healthcare providers only contain limited or partial information about a patients health status. Secondly, it can arise due to different and upcoming standards of storing EHR information as there might be a correct mapping algorithm from old to new standard. Thirdly, data can also be missing as a lot of information related to patient's health status is entered into clinical notes. Such information can only be partially extracted using NLP tools and techniques.
- **Irregular Timer-Series:** Comparing to other data sources such as those obtained in manufacturing settings or in climate sciences, EHR data often comprises of irregular time series elements. This arises as patients only visit healthcare providers when there an appropriate need arises. This leads to the occurrence of irregular spacing between consecutive observation readings. The challenge associated with such techniques further worsens as traditional data-mining techniques cannot handle irregular time-series datasets.
- **Class-Imbalance:** Considering the prevalence of diseases associated with the human body, some diseases are widely prevalent and some are not. This causes class imbalance issues when a particular disease of interest has not enough cases as compared to the controls in any chosen data cohort.
- **Heterogeneity:** EHRs are very heterogenous data sets considering that some attributes such as diagnosis codes are binary in nature, demographics attributes are usually categorical in nature and laboratories test results are continuous in nature. Another source of heterogeneity arises from the fact that a patient can progress to the same outcome (e.g. mortality) via varying disease progression paths. Further, the complexity arises as the probability of progression to the

outcomes is different along these paths.

- **Biases:** EHR datasets are often subjected to multiple biases and confounding effects. Such biases arise when data cohorts do not fully represent the general population characteristics. For example, an average age of 80 years in a data cohort does not represent the average age of the general population. Biases can also subjected to the way the cohort is chosen, the outcomes are designed and the geographical location where the study was conducted.

2.1.2 Data Opportunities

There are several opportunities associated with mining EHRs datasets. Examples of such opportunities include understanding disease progression, risk prediction, detecting adverse events, clinical guideline recommendations, phenotyping and estimating the effect of interventions. Now, we will describe the various opportunities in greater detail.

- **Understanding Disease Progression:** This refers to the problem of analyzing the progression of patient from one state of health to another state of health (associated complications). The aim of such analysis is to estimate the prevalence and incidence of disease across populations. Further, such analysis also provide a platform to understand how disease differ across geographical locations, ethnical and genetic make-up, across age groups and management of disease across healthcare providers.

- **Risk Prediction** This refers to the problem of estimating the probabilities of a patient's current risk or progression to any disease of interest. Clinical Decision Support models aim to assess such risk are usually developed using sophisticated data mining and machine learning techniques. Examples of such risk prediction models include Framingham score, Charlson score, etc.

- **Detecting Adverse Events** This refers to the problem of detecting adverse events associated with medical or surgical interventions. As EHRs consist of patient information collected across years, it is often feasible to estimate where any medical interventions lead to short term or long term adverse events.

- **Clinical Guideline Recommendations** These are the clinical protocols usually followed on a patient during a patient's inpatient or outpatient visits. For example, as per American Diabetes Association (ADA) guideline, it is often recommended that a patient visits a healthcare provider every 6 months to monitor his/her hemoglobin a1c readings. As EHRs collects information related to clinical guidelines, it serves as a platform to identify antiquated medical guidelines and development of novel ones.

- **Phenotyping Algorithms** Such algorithms are hand crafted or machine learned rules for identifying whether a patient is diagnosed with any disease. For example, two consecutive hemoglobin a1c readings (6 months apart) greater than 6.5 classifies a patient to be diabetic. EHRs can be used to validate existing phenotyping algorithms or development of novel ones.

- **Estimating the Effect of Interventions** This refers to the process of estimating the effect of medical or surgical interventions. As EHRs consists of patient longitudinal data spread across years, it is often possible to estimate the effect of interventions in short and long term.

2.2 Thesis Overview

Central to this thesis are the concepts of causation, temporal patterns and heterogeneity. Now we would be discussing them in greater detail.

2.2.1 Causal Patterns

Causal patterns are patterns that consist of two or more random variables as denoted by X and Y respectively, such that X causes Y. Our goal is then to estimate the effect of X on Y. In Fig 2.1, we demonstrate a simple causal pattern which only consist of two such random variables.

In Fig 2.2, we demonstrate a slightly complex causal pattern which consist of random variables X,Y and another random variable Z. Z is also referred to as a confounding random variable. Any estimation obtained, of the effect of random variable X on random variable Y, without incorporating the effect of random variable Z will be biased in nature.

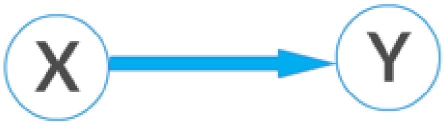


Figure 2.1: Causal Structure

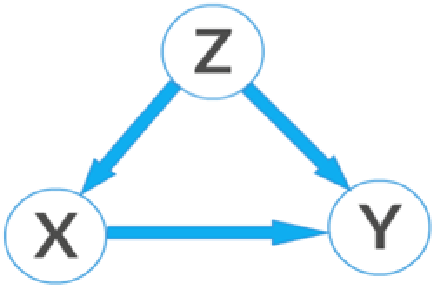


Figure 2.2: Causal Pattern with a Confounder

In Fig 2.3, we demonstrate a complex causal pattern which consist of random variables X,Y and other random variables denoted by U, V, Z and O respectively. In Chapter 5 and Chapter 6, we would be discussing ways to estimate the effect of X on Y while simultaneously incorporating the effect of other random variables in detail.

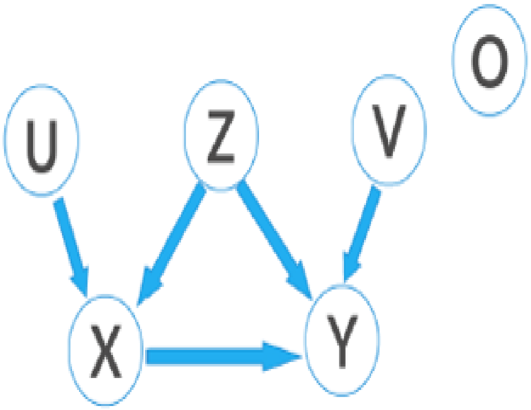


Figure 2.3: Causal Pattern with a Confounder and Other Variables

2.2.2 Temporal Information

EHRs are inherently temporal in nature as information about patients health state is often collected and stored over time. Collecting such information is very vital to accurately analyze the medical state of the patient because of numerous reasons. Firstly, collecting such information provides a mechanism to analyze the temporal progression (laboratories test results or vitals information or diseases) in patients health over time. Secondly, collecting such information and then analyzing them becomes more relevant as exposures over time matter (e.g. consistently low body temperature during surgery are usually associated with increased risk of postoperative complications). Thirdly, sequence in which events happen are usually predictive of certain outcomes as events close to the outcome are more important than events which occurred in the distant past.

2.2.3 Heterogeneity

EHRs consist of information collected of highly heterogenous disease mechanisms. These records provide a platform to explore diseases, which has multiple progression paths to various events of interest (e.g. mortality). For example, In Figure 2.4, we describe the progression of a patient from one state of health to another state of health (e.g. mortality). In our example, we consider three disease of interest i.e. Hypertension, Hyperlipidemia and T2DM and one outcome event i.e. mortality. As observed, even though the outcome is same (i.e. mortality), the risk estimated by the progression paths are vastly different.

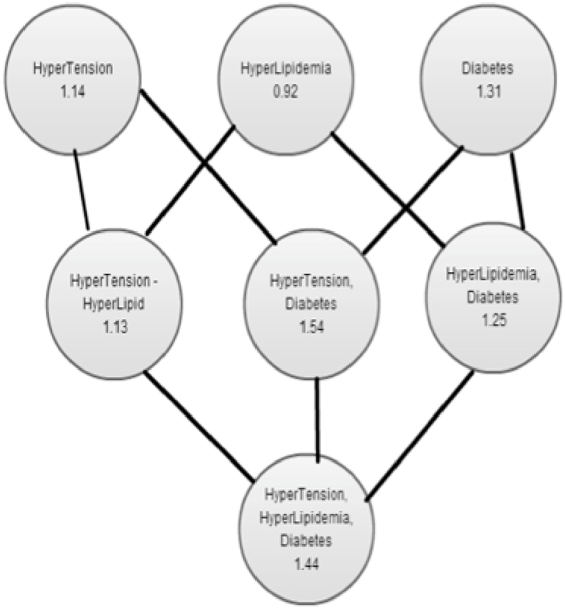


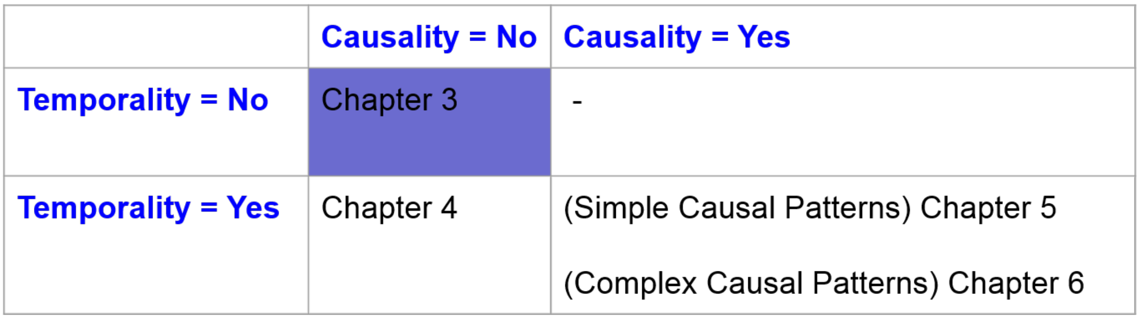
Figure 2.4: Progression and Risk Assessment of Co-morbid Conditions in Type 2 Diabetes Mellitus (T2DM)

Chapter 3

Non-Causal, Non-Temporal Pattern Mining

3.1 Introduction

In this chapter, we would discuss techniques, which aim to extract non-causal patterns in the absence of longitudinal data. illustrated in Figure 3.1. Such patterns are widely used to develop risk estimation indices and computing the probability of progression from one state of health to another state of health. In this chapter we will illustrate such patterns in the context of T2DM.



	Causality = No	Causality = Yes
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	(Simple Causal Patterns) Chapter 5 (Complex Causal Patterns) Chapter 6

Figure 3.1: Chapter 3 Description

3.2 Clinical Motivation

Diabetes mellitus (DM) affects 11.3% (25.6 million) of Americans age 20 or older and is the seventh leading cause of death in the United States [10]. There is considerable research on risk factors to predict and manage diabetic outcomes [10]. Without appropriate management of diabetes, patients are at risk for secondary diseases in almost every body system at later time points. Evidence based practice (EBP) guidelines for management and prevention of diabetic complications synthesize the latest scientific evidence. While EBP guidelines have been shown to improve care, they neither consider the patient's trajectory nor the sequence of events that lead up to the patient's current conditions. In this work, we show that such information is invaluable; a patient's risk of developing further complications depends on their trajectory thus far.

Simple disease models describing a single typical diabetes trajectory as a sequence of successively worsening conditions exist [11]. However, these models were aimed more at patient education than at a physiologically accurate description of the evolution of the underlying disease pathology. Such simple models obviously cannot form the basis of evidence based guidelines.

In heterogeneous diseases, analyzing the data on a per-subpopulation basis has been shown to elucidate more interesting patterns than analyzing the entire population [12]. In this chapter, we hypothesize, with abundant supporting evidence [13], that diabetes and the underlying metabolic syndrome follows multiple trajectories. We aim to develop a methodology that is capable of elucidating scientifically accurate diabetes trajectories retrospectively from the extensive clinical data repository of a large Midwestern health system. Specifically, we study a diabetic population and track changes to their health over time in terms of diabetes-related comorbidities as documented in the electronic health record (EHR).

Diabetes, its severity and the ensuing complications can be described most accurately through a large number of correlated EHR data elements, including associated diagnoses, laboratory results and vitals. The relationships among these data elements, known as multicollinearity, render efforts to track patients' conditions across time fraught with data overfitting issues. To contain the collinearity problem we summarize the patients' condition into a single dimension (a single score), which we term the

Diabetes Mellitus Complication Index (DMCI).

The development of severity indices from EHRs builds on a rich history. Even in the context of DM, several risk scores for diabetes from EHRs have been developed [14]. Most risk score models focus on predicting the risk of diabetes rather than the risk of the associated complications. Two risk scores have specifically focused on diabetes complications [15, 16] to predict outcomes; however their diabetes complication indices were limited to the use of complications based on International Classification of Diseases (ICD) codes alone [17] or asking patients if they were ever informed that they had DM complications [13]. In a diabetic population like ours, good predictors of the complications do not necessarily coincide with good predictors of diabetes given that the metabolic syndromes in our patients have already evolved past diabetes. The inclusion of additional variables, such as lab results and vital signs may provide useful information for early prediction of complications. This necessitates the development of a new diabetes complication index to be used in our effort to study patient trajectories.

In this chapter we make the following novel contributions. First, we develop DMCI which summarizes a patient's health in terms of post-diabetic complications into a single score. Second, through the use of this score, we track a patient's health and show that distinct trajectories in diabetes can be identified, demonstrating the need and laying the foundation for future clinical EBP guidelines that take trajectories into account.

3.3 Background

The novel DMCI was developed using Cox proportional hazards survival modeling techniques. Each of the 7 complications (CKD, CVD, CHF, PVD, IHD, Diabetic Foot, Ophthalmic) were modeled through a separate Cox regression model using patients who did not already present with the complication at baseline. Cox Proportional Hazard Models [18] are survival models which estimate the hazard $\lambda_j(t)$ for patient j at time t based on covariates Z_j and a baseline hazard $\lambda_o(t)$. The hazard function has the form as shown in equation 1.

$\lambda_j(t/Z_j) = \lambda_o(t)exp(Z_j\beta)$ .... (1)

The coefficient vector β is estimated through maximizing the partial likelihood. The partial likelihood can be maximized using the Newton-Raphson algorithm [19]. The partial likelihood [19] has the form as shown in equation 2.

L(β) =$\pi_{i:C_i=1}\theta_i\sum_{j:Y_j \geq Y_i\theta_j}$ .... (2)

θ_j has the form $exp(Z\beta)$ and C_i is an indication function. C_i is 1 if the event occurred and $C_i = 0$ if the event was censored. The baseline hazard is common to all patients. Besides the complications (except for the one we are modeling) age, gender, obesity, hypertension and hyperlipidemia diagnosis, laboratory test results and vitals, were included as covariates. Backwards elimination [20] was employed for variable selection.

Each of the 7 regression models (one for each complication) provided an estimate of the coefficients, which can be interpreted as the relative risk of developing the complication in question. For example, the first regression model estimates the risk of a patient developing CHF. Similarly the second regression model estimates the risk of patient developing IHD. In order to compute a patient's risk for developing diabetes induced complications, we compute a weighted average of patient's risk from the six regression models. Ophthalmic conditions are no longer considered as they have insufficient patient coverage (less than 100 patients). Patient's risk from individual regression model was computed using equation 3.

r$_{ij} = Z_i * \beta_j$ .... (3)

where r_{ij} denotes the i_{th} patient risk for the j_{th} complication, Z_i represents the covariates for the i_{th} patient and β_j are the coefficients estimated for the j_{th} complication.

Using this information, the i_{th} patient's risk (R_i) for any diabetes induced complication is then computed using equation 4.

$$\mathrm{R}_i = \sum_{j=1}^{6} w_j * r_{ij} \ldots\ldots\ldots\ldots\ldots (4)$$

This risk R_i is the risk of a patient developing diabetes related complications. We named this risk DMCI. The DMCI score is the weighted sum of the linear prediction from the six regression models. Concordance probability estimates [21] are used to determine the performance of the corresponding regression models. They are also used to weight the individual regression models. Table 1 represents weights assigned to each model, with the respective complication as the outcome.

Complication	Model Weight
CHF	0.787
IHD	0.569
CVD	0.694
PVD	0.688
CKD	0.758
FOOT	0.712

Table 3.1: Weights For Individual Regression Model

Therefore, the DMCI score can be thought of as approximately 6 times the relative risk a patient faces in developing a complication (any diabetic complication).

3.4 Methods

Using the DMCI score, the health status trajectory of every patient from 2009 onwards was calculated. Since individual patient trajectories might be susceptible to noise and outliers, we decided to group patients and their trajectories (time stamped sequence of DMCI scores) by complications. First, we considered a single complication at a time, creating seven categories: patients presenting with CKD, CVD, etc. at baseline. A

patient presenting with multiple complications falls into all applicable categories. Next, we considered pairs of complications: e.g. one possible category could consist of patients with IHD and diabetic foot problems.
For every category (sub-population of patients), the shape of the DMCI score trajectory was determined through segmented linear regression with 3 knots. One can think about these regression models as a straight line with one elbow (at $\hat{x}$). These trajectories can be expressed in the form below,

If $ifx<\hat{x}$ then y = a*x + b else If $ifx\geq\hat{x}$ then y = c*x + d

where in a,b,c,d ϵ R.

Residual sum of squares (RSS) was used as the objective function to obtain the coefficients of the segmented linear regression and the location of the elbow point. RSS has the form in equation 5

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \ldots\ldots\ldots\ldots\ldots (5)$$

In equation 5, y_i is the risk at i_{th} time stamp and $\hat{y_i}$ is the corresponding risk computed using segmented linear regression.

3.5 Data and Study Design

After Institutional Review Board (IRB) approval, a de-identified data set was obtained from a Midwest University's clinical data repository (CDR). The CDR contains over 2 million patients from a single Midwest health system that has 8 hospitals and 40 clinics. Data elements included various EHRs attributes, such as demographic information (age, gender), vital signs: systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, and body mass index (BMI); and laboratory test results: glomerular filtration rate (GFR), hemoglobin A1c, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), triglycerides and total cholesterol. Further ICD-9 codes related to both 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 Opthalmic complications were used in this study.

For our study, we used Jan. 1, 2009 as a baseline. The study cohort consists of patients with type 1 or type 2 DM at baseline, identified in billing transactions. Patients were included if they had at least two A1c results at least 6 months apart after baseline. Patients with no laboratory results or vitals before 2009 were excluded on the basis that they show no indication of receiving primary care at the health system. The final cohort consists of 13,360 patients. Patients' initial DMCI was determined at baseline, and their health (in terms of the DMCI score) was followed until last the follow-up. The mean time for follow-up was 1568 with a standard deviation of 263 days.

3.6 Results

Table 3.2 provides the count of patients in various cohorts. Table 3.3 provides the count for various populations with comorbidities.

Comorbidity	Count	Comorbidity	Count
IHD	4398	CHF	741
CVD	986	PVD	662
CKD	742	Foot	267

Table 3.2: Patient Counts for Single DM Comorbidity

Comorbidity	Count	Comorbidity	Count
IHD, CVD	457	IHD, PVD	379
IHD, CKD	361	IHD, Foot	662
IHD, CHF	478		

Table 3.3: Patient Counts for DM Comorbidities

Figure 3.2 presents the DMCI trajectory for varying subpopulations. The horizontal axis denotes time since baseline in days and the vertical axis corresponds to the DMCI score. Each curve in the graph represents a subpopulation defined by a single complication. For example, the bottommost curve corresponds to patients presenting with CHF at baseline. Their average risk of developing a complication (other than CHF, which they already have) is 4.4 at baseline, It increases steadily for approximately 550 days, at which point it reaches 4.7 and then it becomes flat (stops increasing materially going forward). As observed from the graph, the average risk associated with patients diagnosed with CKD is comparatively higher than that of patients diagnosed with CHF. Figure 3.2 shows that (i) subpopulations defined by various complications at baseline have a different average risk at baseline. This information is readily incorporated into existing indices and guidelines. The figure also shows that (ii) these patients have different patterns of risk moving forward. For example, the risk of developing a complication increases sharply for CHF patients for 550 days and then becomes flat. In contrast, the risk of IHD increases steadily (but at a lower rate) throughout the observation period; and CKD (topmost curve) increases at a much lower rate.

Complication	Min-Risk	Risk-25	Risk-50	Risk-75	Max-Risk
IHD	-6.02	1.86	3.92	5.98	22.68
CHF	-5.17	1.98	3.86	5.91	12.55
PVD	-5.23	2.07	3.90	6.20	14.37
CKD	-5.62	1.77	3.94	5.94	14.71
CVD	-6.72	2.16	4.00	6.04	14.37
Diabetic Foot	-4.64	2.08	3.95	6.08	14.37

Table 3.4: Distribution of Scores for Different Subgroups

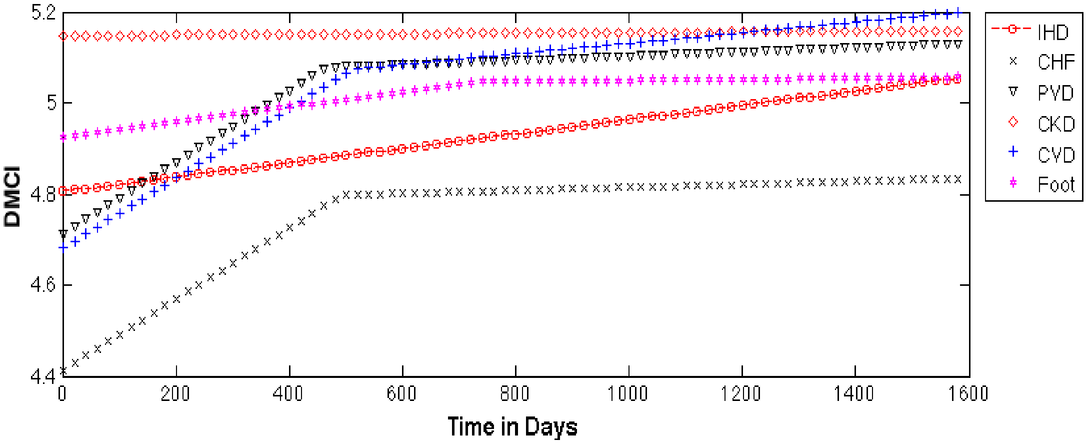


Figure 3.2: Health status trajectory for varying subpopulations

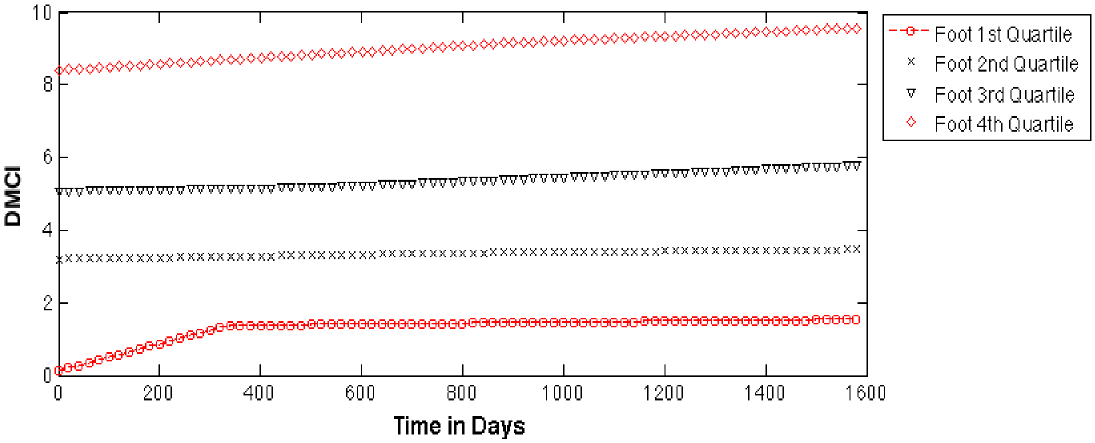


Figure 3.3: Shape of the individual quartiles for patients diagnosed with diabetic foot

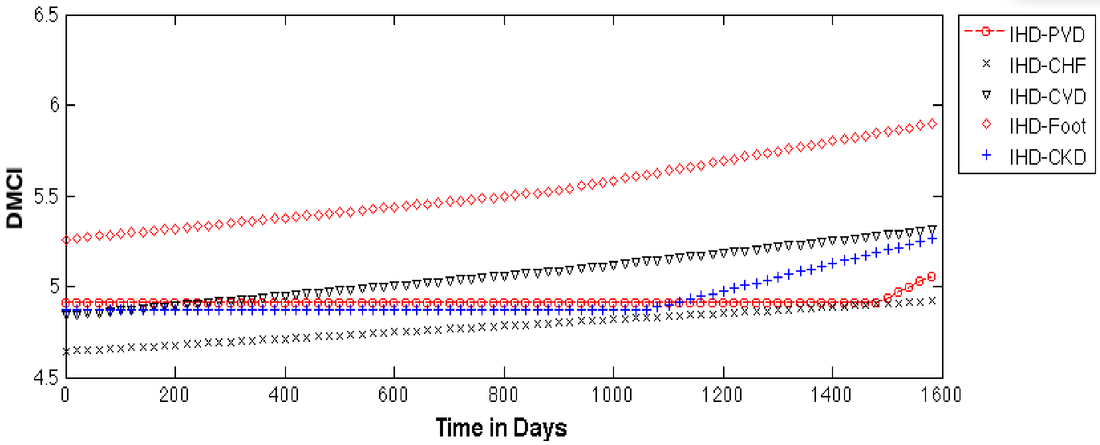


Figure 3.4: Shape of the individual quartiles for patients diagnosed with various complications

Figure 3.2 presents the average risk for each population. To illustrate the distribution of the risk, in Table 3.3, we provide the interquartile range of the DMCI score in each subpopulation. Using the information from table 1, the risk trajectories of patients belonging to the top 25 in their respective subgroups were analyzed. Figure 3.2 presents the average behavior for the highest-risk quartile.

In order to investigate whether the shape of the health-risk trajectory for each quartile within a subgroup is similar, the patterns for each quartile for multiple subpopulations were explored. In Figure 3.2, the shape of the individual quartiles for patients diagnosed with diabetic foot is depicted. The figure shows that having a different risk at baseline only tells a part of the story. These patients not only have different risks, but they also exhibit different progression patterns: their DMCI curves have different shapes.

Figure 3.3, depicts the trajectories of patients with IHD and an additional complication. The results suggest that even in a subpopulation defined by a single complication, significant heterogeneity exists, as evidenced by differing shapes of the trajectory curves.

3.7 Discussion

The purpose of this chapter was to model patients' progression towards diabetes complications through the use of a novel index, DMCI, derived from EHR data. The DMCI was used to stage the patients' health in terms of diabetic complications. Results clearly demonstrated the existence of multiple trajectories in diabetes thereby confirming the complex heterogeneity of the disease. Specifically, we divided patients into multiple (potentially overlapping) subpopulations based on their baseline complications and confirmed that patients with different baseline complications have different risks of developing additional complications. Second, we have also shown that these patient subpopulations differ not only in their risk but also in the temporal behavior of their risk: patients in certain subpopulations 'accrue 'risk at a higher rate initially and at a slower rate later, while the DMCI score in patients in other subpopulations increases at a steady rate throughout the follow-up period. Third, we have also demonstrated that the trajectories differ even within the same patient subpopulation. Patients presenting with additional complications (e.g. a second complication on top of IHD) have different risks and different trajectories. Finally, we have also shown that when we stratify patients within the same subpopulation by their baseline risk, they exhibit different trajectories. This can naturally be a consequence of these patients suffering from additional complications explaining their increased relative baseline risk.

These findings support a conclusion in a previous study that patient subgroups vary by level of severity. Dey et al. [12] used a national convenience sample of 581 Medicare-certified HHC agencies' EHRs for 270,634 patients to understand which patients are likely to improve in their mobility and found that mobility status at admission was the single strongest predictor of mobility improvement [12]. However, very different patterns were apparent when conducting the analysis within the level of severity for mobility at admission.

An interesting finding in our study is that patients with diabetic foot problems have the highest severity at base line, and more so when combined with IHD. This finding may be associated with the strict relationship between glycemic control and microvascular complications. Foot problems are associated both with nerve and vascular damage, creating a risk for infections. Uncontrolled glucose further exacerbates the potential for severe infections and potential amputations. Patients with diabetic foot complications are likely to continue having an increasing risk for additional problems, as foot problems are a leading cause of hospital admission, amputation, and mortality in diabetes patients [16].

Through our previous work [11] in investigating diabetic subpopulations and their risk of mortality, we have already gained an appreciation of the immense heterogeneity of diabetes and the metabolic syndrome. Studying trajectories expands this heterogeneity along a new dimension. While the preliminary work presented in this study merely offers a glimpse at the complexity of diabetes and its complications, it demonstrates the value of trajectories in understanding patient progression and possibly prognosis. Further research in this direction will undoubtedly lead to improvements in EBP guidelines by taking trajectories into account.

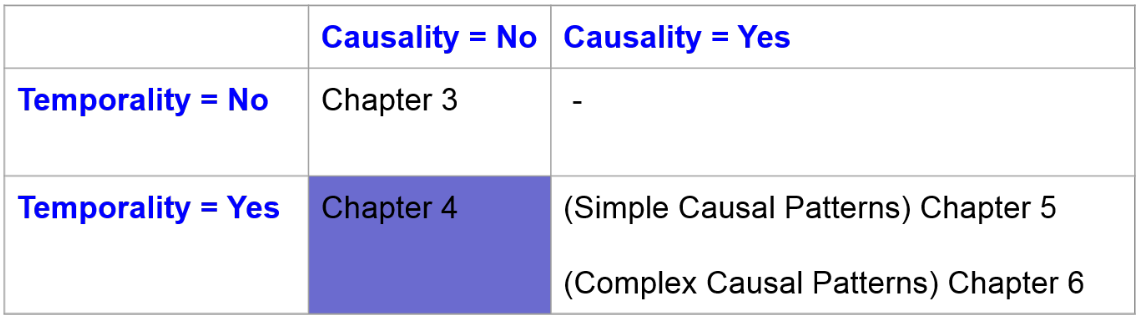
Limitations of this study include the secondary use of EHR data and its associated challenges. The data in this study represent care provided in a single health system; the study needs replication in additional health settings and under different clinical conditions. The DMCI score was developed from EHR data retrospectively and independent validation would be beneficial.

Chapter 4

Non-Causal Temporal Pattern Mining

4.1 Introduction

In this chapter, we would discussing techniques to extract and explore non-causal patterns in the presence of longitudinal data as illustrated in Figure 4.1. Such patterns are widely used to analyze the progression of patients from one state of health to another state of associated complications. Further, such analysis forms the basis for personalized and tailored health recommendations. In this chapter we will illustrate such patterns in the context of T2DM.



	Causality = No	Causality = Yes
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	(Simple Causal Patterns) Chapter 5 (Complex Causal Patterns) Chapter 6

Figure 4.1: Chapter 4 Description

4.2 Clinical Motivation

The use of large repositories of Electronic Health Records (EHR) data for assessing the risk of adverse outcomes, such as mortality or the development of new complications, is experiencing a rapid growth in popularity. The most common style of analysis for this purpose is based on longitudinal retrospective design, where patients are aligned on a particular point in time (e.g. enrollment into the study), called a *baseline*, their state of health at baseline is characterized by elements present in EHR data (*baseline characteristics*) and they are followed until *last follow-up*, at which point they suffer the adverse outcome in questions or are simply lost to follow-up (are *censored*).

Such studies have enjoyed great success. Strong epidemiological evidence has been discovered, which ultimately influenced health care policy. However, the acceptance and incorporation of these methods into clinical decision support systems is slow. The design underlying this methodology, where patients' risk is solely based on baseline characteristics, is incompatible with clinical practice. Providers constantly reevaluate patients' risks and adjust treatment accordingly. When the patient information shows no clear sign of improvement or deterioration, a common approach is to wait and see. As time progresses and the patient's condition further deteriorates, the outcome becomes more apparent and an appropriate intervention can be administered. When the patient's health has deteriorated to the final stages, the outcome can become obvious and also inevitable: there may be no time for a successful intervention. Knowing not the only the risk but also the expected timing of adverse events is important, allowing the care provider to have time to intervene. In this study, we look at a large diabetic population and aim to mimic the clinical process. We assess the patients' risk at every encounter, taking not only the prior conditions but also their sequence into account.

Our working hypothesis is that patients' health deteriorates following a (small or large) number of non-random mechanisms. These different mechanisms may affect organs or health indicators (blood sugar, lipid levels, blood pressure) differently, leading to different sequences of diagnoses. Therefore, the order in which the diagnoses appear in a patient's record can be suggestive of the underlying disease mechanism, allowing us to provide the patient a better prognosis.

Central to this idea is the concept of a trajectory. Formally, a *trajectory* is a sequence, a partially ordered set of conditions, through which patients commonly progress from a healthy state towards some outcome. We present a method for extracting trajectories, placing patients onto trajectories and assessing their risk of mortality based on the trajectories (potentially multiple trajectories) they follow and the extent to which they have progressed along each trajectory.

Our second key contribution is the introduction of the *forensic*-style analysis. In forensic investigation (of say an accident), investigators start from the outcome (accident) and trace events backwards in time. Factors that contributed to the accident tend to be most apparent closest to the time of the accident. In a similar vein, in our forensic-style analysis, we align patients on their outcome (death or censoring) and trace their conditions backwards over time.

The proposed forensics-style analysis offers several benefits. First, it directly answers our clinical question of time-to-death, as time in the model represents time-to-death, as opposed to time-since-enrollment in a typical study.

The second benefit concerns time-dependent covariates. As a patient's condition evolves (deteriorates), he develops new conditions, which were naturally not present at baseline. The predictors of the patient change. Traditional Cox models handle this by describing the patient with multiple records: every time the patient's state changes (a new condition is developed), a new record is added describing the new state along with the time when the change took place and the record became valid.

Suppose a patient develops disease A 2 years after entering the study, then condition B 5 years later (7 years into the study) and dies 4 years later (last follow-up is 11 years after enrollment). This patient would be described with two records: one having A as the sole covariate and valid time of 2yrs-7yrs; and a second record having both A and B as covariates and a valid time of 7yrs-11yrs, ending in death. If A is almost always followed by B when the patient dies, then A will appear to have no risk; all the risk is assigned to B. To better estimate the risk of A, we could assign the adverse outcome to both records (as opposed to correctly assigning it only to the second one), but then the patient appears to die twice *at two different times*, namely after 7 and 11 years, respectively. The resultant increase in the baseline hazard at 7 years is incorrect. If we measured time backwards from the last follow-up, both records would indicate the correct time of death, allowing us to correctly measure the risk of disease A.

We evaluated our method on the EHR data of a large health care system in the Midwestern United States in the context of the metabolic syndrome including type-II diabetes, its co-morbidities and complications.

In this chapter, we make the following contributions:

1. We propose modeling patients' risk of adverse outcome based on the trajectories they follow and the extent to which they have progressed along these trajectories; thus allowing us to take the sequence of events into account.
2. We introduce the *forensic-style* analysis, which aligns patients on last follow-up and measures time backwards. Measuring time backwards allows us to estimate the time-to-event more directly.
3. We modified the Cox proportional hazard model, using forensic-style analysis, to better model time-dependent covariates. Specifically, we modified how the outcome is designated, allowing the fitting algorithm to better estimate the risk of diseases in earlier stages of the trajectories.

The proposed method is a general methodology that can be applied to different time-to-event problems. Given the importance of diabetes and our expertise in diabetes, we describe and evaluate the method in the context of diabetes and the metabolic syndrome.

The chapter is organized as follows. Section II describes current state of the art techniques used to handle time-to-event data. Section III-A introduces terminology associated with trajectories. Section III-B discusses techniques for extracting frequent trajectories. In Section III-C we present our model and optimization framework. In Section IV, we discuss our results. Finally, Section V presents our conclusions.

Survival modeling techniques on time-to-event data have been explored widely in the past. Cox regression [22, 23] 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 [24], healthcare [25, 26, 27] and recommendation systems [28].

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 [29] and elastic-net regularized Cox models [30] have been developed, and have been further extended by regularizing them with convex combinations of L1 and L2 penalties [31]. We are not aware of regularization for time-dependent covariate Cox models [32], which would be a straightforward extension.

Chandan et. al [33] 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 [34]. 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 disease and intervention recurrences, and the hierarchical structure of medical knowledge and practices [35]. To handle the high-dimensionality of high-throughput genomic data, Kuang et al. [36] 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 [37] models have also been extended to handle censored data [38, 39, 40, 41]. 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. [42] 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 heart stroke.

Research has also been carried out on extending decision trees to handle censored data [43]. Ishwaran et al. [44] 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 [45, 46]. Techniques such as reverse survival [4] have also been explored in the past wherein they go further back in time.

4.3 Background

We consider seven **disease**s in the context of diabetes. These are hyperlipidemia (HL; high cholesterol), hypertension (HTN; high blood pressure), type-II diabetes mellitus (DM), chronic kidney disease (CKD), ischemic heart disease (IHD), cerebro-vascular disease (CVD), and congestive heart failure (CHF). These are chronic diseases; once the presence of the disease has been confirmed, they remain active. The patient may have the condition under control (i.e. a patient can have normal laboratory results), but the disease remains.

4.3.1 Trajectory Terminology

Not all mentions of these diseases in the patient's record indicate a new diagnosis. Often, these diagnosis are present for billing purposes, as they complicate treatment. To determine the precedence of the diseases in the trajectories, we need to focus on new (**incident**) diagnoses. The term 'incident' refers to the diagnosis creating a new incidence, as opposed to being a chronic condition in the background that complicates the treatment of a different disease. To identify incident diagnoses, we need to determine the **status** of diseases at any time point.

The disease is **confirmed** if we have evidence that the patient presents with the disease; it can be **ruled out** if we have evidence that the patient does not have the disease; or the status can be unknown otherwise (when we do not have evidence either way).

Definition 1 (Disease confirmed) *A disease is confirmed at time t and thereafter, if the patient's record has a diagnosis code, a prescribed medication or an abnormal lab result (if applicable) related to the disease.*

Definition 2 (Disease ruled out) *A disease is ruled out at time t and before, if no pertinent medication prescription or diagnosis code is present at or before t and a normal laboratory result is present at t.*

Definition 3 (Incident diagnosis) *A new disease diagnosis is incident at t if the disease is ruled out before t and confirmed after t.*

In plain language, a disease diagnosis is new (or an incident diagnosis) if we have evidence that it is new: it was absent before t and is present at t. For IHD, CVD and CHF, we do not have laboratory results to rule them out, so we assume that the first diagnosis of these diseases in our data is an incident diagnosis.

Definition 4 (Background disease) *Non-incident diagnosis of a confirmed disease.*

A background disease is a confirmed as a preexistent condition or a potentially preexisting condition that we cannot rule out. If a patient enters the study with (say) HTN, then HTN is a background disease (confirmed preexisting condition). If the first appearance of HL (high cholesterol) is a year after enrollment, but the patient does not have cholesterol measurements before the diagnosis, then HL is background (the patient may have had HL all along). If, however, we have a normal cholesterol measurement before the HL diagnosis, then the HL is incident, because we rule it out for (some part of) the first year.

Definition 5 (Precedence) *A disease A precedes disease B, $A \to B$, (or B follows A) in a patient, if the patient has an* incident *disease B at time t and A is a background or incident disease before t.*

Since B is an *incident* disease, we ruled it out before t, while A could not be ruled out before t, thus A occurred earlier than B.

Definition 6 (Trajectory) *A trajectory is a set of diseases, some incident, some background, with precedence information among them. In other words, a trajectory is a partially temporally ordered set of diseases.*

Example. $T = (A, B) \to C \to D$ is a trajectory with A and B being background diseases, whose ordering cannot be determined from our data and C and D are incident diseases, hence their ordering is known. The precedence information is transitive, so beside the depicted $A \to C$, $B \to C$ and $C \to D$ precedence relationships, $A \to D$ and $B \to D$ also hold. These diseases are chronic, hence at the time when the patient develops C, he also has A and B; and at the time he develops D, he also has A, B and C.

The central idea in our work is to place patients on trajectories, which requires that we define when a trajectory **applies** to a patient or **matches** a patient's trajectory.

Definition 7 (Sub-trajectory) *A trajectory S is a sub-trajectory of T if the diseases in S are a subset of the diseases in T and all precedence information in T that relates to the diseases in S holds true in S.*

Example. The trajectory $S = B \to C$ is a subtrajectory of T, as it contains a subset of the diseases B and C and all precedence relationships involving B or C in T, namely $B \to C$, also holds true in S.

Definition 8 (Prefix trajectory) *A trajectory P is a prefix trajectory of T if P is a subjectory of T and no disease in T precedes any of the diseases in P.*

Example. $S = (A, B) \to C$ is a prefix trajectory of T as none of the diseases in T precede the diseases in S. In contrast, $B \to C$ is not a prefix of T as there is a disease A in T that precedes C in T.

Definition 9 (Matching) *A trajectory T applies to a patient with trajectory X (or matches X) iff (i) there exists a prefix P of T that is a subtrajectory of X and (ii) there exists no disease d in X, such that d is not a part of P but is present in T.*

Example. Consider a patient trajectory $X = A \to B \to C \to D$. The trajectory $T = A \to B \to E$ matches X with prefix $P = A \to B$, because the only disease in T that is not in P (namely E) is not in X. The clinical motivation behind this definition is that the patient with trajectory X may be following T, just has not progressed to E yet. (He may also follow other trajectories that explain C and D).

4.3.2 Algorithm for Extracting the Frequent Trajectories

Our goal is enumerate all trajectories that patients frequently follow that end in mortality. Therefore, for trajectory extraction, we only consider patients who died and do not consider patients who remained alive after last follow-up.

We apply the venerable a-priori algorithm [47] to enumerate all sub-trajectories that occur in at least 4 patients. With 2814 total deaths, the support of 4 (support fraction of 4/2814) is the smallest support fraction that does not contain 0 in its 95% confidence interval. Trajectories that occurred in less than 3 patients can be random as 0 would be present in the confidence interval. The purpose of this 4-patient threshold was merely to establish a minimal reasonable standard for trajectories, it was not to extract an optimal set of trajectories. The discovered set of trajectories (library trajectories) will be passed to an optimization algorithm that will select the final trajectories.

In our application , we only have seven diseases thus the number of discovered trajectories is not a concern. With more diseases, the number of trajectories can grow exponentially, making the number of discovered trajectories a concern. Several remedies are available. The number of trajectories can be decreased through numerous heuristics, including the following:

- increasing the support threshold to correct for simultaneous hypothesis testing,
- increasing the support without any statistical justification,
- mining only maximal subsequences (sub-trajectories),
- mining approximately maximal subsequences (if the support of a trajectory differs only minimally from the support of one of its sub-trajectory, the sub-trajectory in question can be discarded), or
- frequent-set summarization techniques [48] can be easily adapted to sub-trajectories.

Heuristics to filter frequent item-sets and sequences have a rich literature and studying these heuristics is outside the scope for this work. We rely on the feature-selection facility of our modeling algorithm to select trajectories.

Output. The output of the algorithm is a library (set) of trajectories, which we call **library trajectories** that end in mortality and occur frequently in patients who suffered mortality. Some of these trajectories can be sub-trajectories of each other.

4.4 Methods

Given a potentially large and redundant set of library trajectories, discovered above, our goal in this section is to a develop a methodology for (i) selecting trajectories and (ii) to estimate the risk of mortality in patients, who may follow zero, one or more of the library trajectories.

.1. Data Format.

The trajectories are transformed into a binary *design matrix* X. The columns of X correspond to diseases along the trajectories: each disease along each trajectory is mapped to its own column. The rows of X correspond to patients during different time periods, active periods. Therefore, for the ith record, we have the associated trajectory information x_i (ith row in X), the beginning b_i and end e_i time of the active period, the patient id p_i and the outcome y_i. The indicator $A_i(t)$ signals whether record i is active at time t; it returns 1 for $b_i \geq t > e_i$. Note that in our forensic-style analysis, time is measured backwards, so $b_i > e_i$.

For each patient, the active time periods are defined by changes in the trajectory: whenever the patient develops a new disease which corresponds to progression along a trajectory, we add a new record with the appropriate timing information. Therefore each record represents a new state, where the patient has progressed further (has accumulated more diseases).

The outcome y_i is 1 if the patient p_i had an adverse outcome exactly b_i time after the beginning of the record. In contrast to Cox models with varying covariates, the outcome is 1 for all records of the patient. While it may appear that the patient had died multiple times, our definition of b_i (being measured from death) ensures that these "multiple" deaths coincide at the right time point. The baseline hazard can compensate for the multiplicity of deaths.

.2. Model.

The model is a variant of the Cox Proportional Hazards Regression model. Central to the model is the concept of **hazard**, which we define analogously to the Cox terminology, namely, as the instantaneous probability of death in exactly t time from an event.

$$\lambda_0(t)\exp(x_i\beta) \tag{4.1}$$

where $\lambda_0(t)$ is a time-dependent baseline hazard that is common across all patients and the trajectories x_i increase the hazard proportionally.

Given our design matrix X described earlier, the expression $x_i\beta$ expands into

$$x_i\beta = \sum_{L\in\mathcal{L}(b_i)} \sum_{d\in L(b_i)} \beta_{L,d} \tag{4.2}$$

where $\mathcal{L}(b_i)$ is the set of library trajectories that apply to the patient p_i at time b_i, the diseases d are the diseases confirmed for the patient at or before time b_i along the trajectory L, and $\beta_{L,d}$ are the coefficients. The sum $\sum_{d \in L} \beta_{L,d}$ is the (log) relative risk that having reached d along trajectory L confers on the patient. Notice that the (log) relative risk along a trajectory cumulates in a (log-)additive fashion, indicating that each events along the trajectory also confers a proportional hazard.

We can estimate the "probability" of death (technically, expected count of deaths) for patient p at time t as

$$\Lambda_p(t) = \sum_{\tau=0}^{t} \lambda_0(\tau) \exp(x_i \beta), \qquad \text{for } i: \; A_i(\tau) = 1, p_i = p \tag{4.3}$$

for all records i where the records is active at time τ and describes patient p.

.3. Likelihood.

The likelihood is the probability that for each patient p after developing each disease d, the outcome happens exactly time t after developing the disease.

$$\prod_i \left[\frac{\lambda_0(t) \exp(x_i \beta)}{\sum_j A_j(t) \lambda_0(t) \exp(x_j \beta)} \right]^{y_i} \qquad \text{for } i: \; t = b_i, \; j: \; b_j \geq t > e_j \tag{4.4}$$

Defining the vector of linear risk score u as $u = x\beta$, the log likelihood becomes

$$\ell(u) = \sum_i y_i \left[u_i - \log \sum_j A_j(b_i) \exp u_j \right] \tag{4.5}$$

.4. Optimization

Our goal with the optimization is (i) select a subset of the library trajectories for modeling and (ii) estimate their coefficients. We optimize β iteratively through a gradient boosting framework [49], adding a new trajectory in each iteration. Adding a library trajectory, say L, is equivalent to changing the corresponding set of coefficients in β, which we denote by β_L.

Performing boosting (gradient ascent in u-space), leads to the update

$$u^{(k+1)} = u^{(k)} + \gamma \frac{d\ell}{du^{(k)}}, \tag{4.6}$$

where γ is the learning rate, $u^{(k)}$ is the linear risk score u in the kth iteration and ℓ is the log likelihood function.

In iteration k, we need to find the trajectory that fits $d\ell/du^{(k)}$ the best. Let $\nabla\ell$ denote the gradient $d\ell/du^{(k)}$. We wish to find the trajectory L, with coefficient vector β_L, such that the quantity

$$\min \beta_L \quad (\nabla\ell - x_L\beta_L)'(\nabla\ell - x_L\beta_L)$$

is minimal across all trajectories. The prime sign (') denotes matrix (vector) transposition.

Once we find the optimal trajectory along with the optimal β_L, we can update the β vector.

The learning rate γ can be determined through line-search or can also be chosen as an arbitrary small number.

Stopping criterion. We stop adding trajectories, when the improvement of ℓ on either the training or a validation set is less then a pre-defined small positive number ε.

Initialization. We can either start with an empty set of trajectories, or we can provide a pre-selected set of trajectories resulting from a greedy coverage of the events in the patient trajectories. For our experiments, we started with an empty set.

Gradient. To derive $\nabla\ell$, we first separate out a particular component u_k from ℓ and then derive the partial derivative with respect to u_k.

$$\begin{aligned}\ell &= \sum_i \{y_i u_i - y_i \log\left[A_j(b_i)\exp u_j\right]\}\\ &= \left\{ y_k u_k - y_k \log\left[A_k(b_k)\exp u_k + \sum_{j\neq k} A_j(b_i)\exp u_j\right]\right\}\\ &+ \sum_{i\neq k}\left\{ y_i u_i - y_i \log\left[A_k(b_k)\exp u_k + \sum_{j} A_j(b_i)\exp u_j\right]\right\}\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell}{\partial u_k} &= y_k - y_k\frac{A_k(b_k)\exp u_k}{\sum_j A_j(b_k)\exp u_j} \qquad (4.7)\\ &\quad -\sum_{i\neq k} y_i \frac{A_k(b_i)\exp u_k}{\sum_j A_j(b_i)\exp u_j}\\ &= y_k - \sum_i y_i\frac{A_k(b_i)\exp u_k}{\sum_j A_j(b_i)\exp u_j} \qquad (4.8)\end{aligned}$$

The sum iterates over all records i that began during the active time of record k divided by the summed risk of records j that started when i was active. Thus the partial derivative can be restated in a more familiar form

$$\begin{aligned}\frac{\partial \ell}{\partial u_k} &= y_k - \sum_{\tau=b_k}^{e_k} \left(\sum_i \frac{y_i}{\sum_j A_j(\tau) \exp u_j} \right) \exp u_k, \qquad (4.9)\\ &\qquad\qquad\qquad\qquad \text{for } i: \; b_i = \tau.\\ &= y_k - \sum_{\tau=b_k}^{e_k} \lambda_0(\tau) \exp u_k \qquad (4.10)\end{aligned}$$

Unlike in the regular Cox models, the gradient is not the residual, only a part of the residual that the corresponding record is responsible for. For gradient boosting, it is not required that the gradient coincides with the residual. The form of the partial derivative, however, suggests a form for the cumulative hazard that parallels the Breslow estimate [50] in Cox models, which we presented in Eq. 4.3.

4.5 Data and Study Design

Data.

We use the clinical data repository of a large health care system situated in the Midwestern United States. Based on data availability, we selected 2005 to 2014 as the study period. We included all adult patients who developed type-II diabetes during this period. Mortality data from the state death registry was available for 8,000 of these patients. Our health care system has a large tertiary care arm, thus many of the patients may receive their primary care (and possibly diabetes care) outside this system leading to large gaps in the data. To exclude such patients, we required the study population to have at least 2 Hemoglobin A1c measurements at least 1 year apart. The final cohort consists of 2,814 *cases* (patients who died) and almost 2,000 *controls* (who were censored).

For these patients, we collected diagnoses, lab values, vitals and medication data. We use a combination of these data elements to determine whether a patient has each of the seven diseases at each point in time, as described earlier. Almost all patients had a history of obesity, so we dropped this variable.

In Table 4.1, we present our patient population statistics for both the cases and controls for the purpose of direct comparison. Cases and controls appear to have similar conditions at baseline, but cases deteriorate more rapidly. The table also shows that we have laboratory results and diagnoses for almost all patients.

4.6 Results

Our problem is not a traditional computer science problem hence methods to compare it against are very few, and mostly in biostatistics and epidemiology. We decided to evaluate our algorithm by showing that all innovations we claim improve the performance. We claim three innovations: (i) the use of trajectories, (ii) the forensic-style design: patients are aligned on the last follow-up and time is measured backwards from last-follow-up and (iii) designation of the outcome for the records belonging to the same patient. Accordingly, we will build four models, starting from the simplest model (the one that is typically used to solve this problem) and successively add our proposed features to it to isolate the effect of each of our contributions.

(1) Enrollment-Aligned Design. The typical approach to time-to-event problems is to conduct a retrospective study, where patients are aligned on their enrollment into the study and are followed until mortality or until they get lost to follow-up (until censoring). In other words, time in the model denotes time since enrollment measured in months. The baseline in our case translates into the first appearance of the patient in the EHR and last follow-up is the time stamp of the last piece of information in the EHR, or their date of death in the death registry.

The modeling method is Cox proportional hazards model with time-dependent covariates [32]. Patients are represented by multiple records, where each record describes a time-slice of the patients progression between two changes, i.e. the time points when new diagnoses appeared in the patient's record. As recommended for this design, only the last record of the cases (patients who died) is marked with death as an outcome. Naturally, none of the records of controls indicate a positive outcome. The predictors in this model are the seven diseases.

This is the simplest and most common model to solve our problem.

(2) Outcome-Aligned Design. The next simplest model aligns patients on outcome, which admittedly, is an unusual but reasonable design. Suppose patient i has follow-up T_i. We select a time T, which is larger than all T_i's and designate T as the last follow-up for all patients. Consequently, in this design, we align patients on their last follow-up (which happens at time T for all patients by design). Their enrollment time into the study will vary, it will be $T - T_i$ for patient i. This design assumes that the baseline hazard depends on time from death. This stands in sharp contrast with the assumption of the Enrollment-Aligned Design, which assumes that the baseline hazard depends on time from enrollment. Therefore, the Outcome-Aligned Design incorporates exactly one aspect of the proposed forensic-style analysis: alignment on outcome (i.e. alignment on the last follow-up).

The modeling algorithm is still Cox proportional hazard model with time-dependent covariates and the outcome for the patients is still designated in the usual way: only the last record has positive outcome (death) for the cases. Although we have aligned patients on last follow-up, naturally, not all of their records end at last follow-up, therefore designating all records of cases as positive would still appear as if the patients had died at multiple time points.

(3) Forensic-Style Design. This is the design proposed in this manuscript. Forensic-Style Design is similar to Outcome-Aligned Design in that patients are aligned on last follow-up, but it goes beyond by measuring time backwards. Measuring time backwards allows us to designate *all* records of cases (patients who died at last follow-up) as positive and still retain the correct time of death across all records.

Since the likelihood has a different meaning in this design, we use our own fitting algorithm from Section 4.5. we use seven "trajectories" each consisting of a single disease, thus our predictors are the diseases. We will refer to this model as 'fast w/o traj' (FAST without trajectories).

(4) Forensic-Style Analysis Via Survival Trajectories (FAST). The experiment is designed using the Forensic-Style Design, but instead of the seven diseases, we use trajectories as predictors. This is precisely the proposed methodology.

Evaluation Method

Given the time-to-event outcome, our evaluation metric is **survival concordance**. This is a widely used metric for time-to-event data. For any two patients, i and j, i

having a higher risk of death than j, survival concordance measures the probability that i dies earlier than j. Patient pairs, for which it is not possible to determine whether the higher risk patient dies earlier (e.g. he is still alive at last follow-up), are ignored. Ties (patient pairs with the same risk and same time-to-death) are also ignored. Note that i and j are different patients: two records of the same patient are not compared.

Survival concordance is similar to the C-statistic (or binary concordance, a.k.a area under the receiver operator curve) in that a random model, a model where the predicted risk is independent of the outcome, would give a concordance of .5, a perfect model would give 1. Survival concordance can also give a concordance value less than .5 if the estimated risk *decreases* with increasing actual risk.

To estimate the survival concordance in the presence of multiple records per-patient, we use 100-iterations of (grouped) bootstrap estimation. Our method is a computationally more efficient version of the robust estimator suggested in [51, Ch 8.2]. Although bootstrapping can increase the bias slightly as compared to jackknife [52, Ch. 11], but given the large number of patients (approx. 4,000), we are not overly concerned.

The sampling unit for the bootstrap resampling is a patient: *all* records of each patient are either included or excluded. In each bootstrap iteration, we use the out-of-bag (OOB) samples for testing and the resampled (bootstrapped) data set for training. 30% of the training set is left out for validation. Again, all records of each patient are either in the training set or in the validation set; we never split them between both. Boosting achieves regularization through early termination; hence we use the validation set to determine when to terminate the optimization process. The resultant model is then evaluated on the OOB sample.

In Figure 4.2, we present the survival concordance of the four models across the 100 bootstrap replications.

Effect of Aligning Patients on Outcome. The 'enrollment' and 'outcome' models use the same fitting algorithm (time-dependent Cox model), the same predictors (the seven diseases) and only differ in the study design: in the 'enrollment' model, patients are aligned on their enrollment into the study, while in the 'outcome' model, they are aligned on their last follow-up.

The benefit of aligning patients on last follow-up is clear. In the Enrollment-Aligned Design, time represents time-since-enrollment, which is not associated with death. On

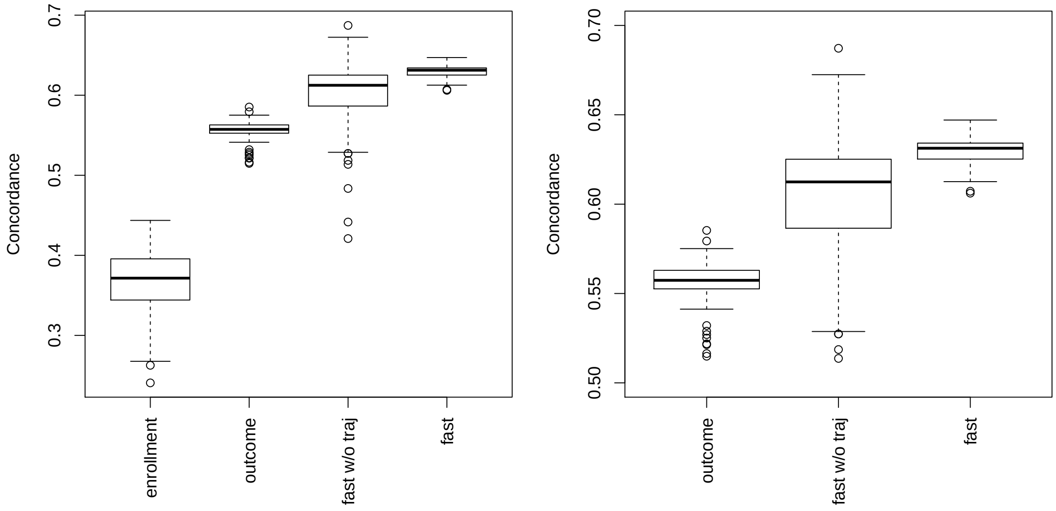


Figure 4.2: Concordance of the various designs estimated through bootstrapping.

the other hand, in case of the Outcome-Aligned Design, time is related to the time of death: time of death happens exactly at the same time point for all patients (by definition). Aligning patients on their time of death (or censoring) allows for more accurate description of the so-called *risk set*: the patients who were under observation at the time, having the potential for an event. (This is the denominator in the likelihood function.)

Effect of Outcome Designation. To assess the effect of the outcome designation, we can compare the 'outcome' model with the 'fast w/o traj'. Both of these models utilize a study design that aligns patients on the outcome; the difference between them lies in measuring time backwards and the outcome designation this change enables. The beneficial effect of this difference is very significant as observed from the paired t-test performed between survival concordance values of the two methods (p-value 1e-16).

Our choice of outcome designation was motivated by the following observation. Given a trajectory $a \rightarrow b \rightarrow c$ that ends in death, in the typical study design (Enrollment-Aligned or even Outcome-Aligned), when the patient only has a, or has a and b, his outcome is still designated as 'alive'. Since death rarely follows a or b without c, this designation leads the fitting algorithm to believe that a and b are protective. The result is negative coefficients for these diseases and a survival concordance less than .5 (see the 'enrollment' model).

What is surprising is that this observation holds true even without trajectories. HL and HTN occur in early stages of the metabolic syndrome and thus death rarely follows these conditions directly.

Effect of using trajectories. Finally, 'fast w/o trajectories' and 'fast' differ only in the use of trajectories. The use of trajectories is advantageous (p-value 7.5e-6).

When trajectories are not utilized, the model only has seven predictors. Estimating seven coefficients from 12k records contributed by 5k patients is trivial. On the other hand, it is suspected that these seven conditions affect the risk of mortality differently depending on the presence of other conditions.

When trajectories are utilized, the model is very flexible, allowing it to capture the clinical reality better. Unfortunately flexibility translates into increased model degrees of freedom, making the model susceptible to overfitting. The FAST algorithm is regularized to help it avoid or at least alleviate overfitting.

Comparison to Other Penalized Models. We have considered comparing FAST with other penalized regression models, however, there are two major obstacles. First, currently existing penalized regression implementations (most notably `glmnet`) do not support time-dependent covariates, rendering any comparison unfair.

Second, our trajectory-based boosting scheme resembles a grouped lasso [53] penalty (although they are not equivalent). Using penalized regression would allow us to draw upon the rich set of penalization techniques that have already been developed (e.g. structured lasso), but, these techniques have not been implemented for Cox models. Given that are focus is on the three innovations, rather than on studying regularization in this context, we decided that comparing regularization schemes is out of scope for this work, and did not implement these regularization schemes for Cox models with time-dependent covariates.

4.6.1 In-Depth Look at the FAST Results

Above, we have shown that the performance of FAST is substantially and (statistically) significantly better than any other model we have considered. In this section, we are going to show some of the resultant models.

Table 4.2 presents the coefficients of the three models that do not rely on trajectories. These are coefficients obtained from regular Cox models and thus their interpretation is as follows. For example, the relative risk of mortality that CHF (congestive heart failure) confers on a patient is $\exp(-.24) = .79$; a patient with CHF is 21% less likely to die than the average patient in our cohort. The coefficients of HL and HTN are 0 or NA because these diseases occur in nearly all patients. As a result, their risk is not reasonably estimable. ('fast w/o traj' did not select these variables, either.)

The model based on the Enrollment-Aligned Design indicates that CHF is protective from mortality; and all three of these models suggest that Chronic Kidney Disease (CKD) and Cerebro-Vascular Disease (CVD) are also protective. Based on clinical knowledge, these findings are patently wrong. The correct interpretation is that patients with CKD will most likely die from a different immediate cause and not from CKD itself. Ergo, these models do not tell us the risk of mortality conferred on the patient by CKD.

FAST Models.

In this section, we turn our attention to the FAST model. To assess the statistical significance of the coefficients, we ran 500 bootstrap replications, resulting in 500 models, each potentially using a different set of trajectories.

Most of the 500 models used only one (392 models) or two trajectories (34 models) and on the other extreme, there were models using 20, 22, and 28 trajectories (one model each). In Table 4.3, we present the trajectories that appeared in at least 10 models.

The table presents the trajectory, followed by the number of models that utilized this trajectory in parenthesis. We then present the average coefficients (across the models that utilized this trajectory) of the diseases along the trajectory and also the empirical p-value of the coefficient, which is the fraction of bootstrap iterations in which the sign of the coefficient in question was the opposite of the sign of the mean.

To illustrate the interpretation of these trajectories, consider for example, the last trajectory: $HTN \to HL \to DM$. (All trajectories end in death so we omit the outcome from the trajectory description.) Patients along this trajectory first develop hypertension (high blood pressure; HTN), then hyperlipidemia (high cholesterol; HL), followed by diabetes (DM) and they die without developing any diabetes complication. The relative risk of mortality conferred upon the patients by this trajectory (relative to the entire population) is exp(.09)=1.09 at the stage of HTN, exp(.09+.13)=1.24 by the time they progress to HL and exp(.09+.13+.20)=1.52 when they develop diabetes. The same patients could potentially follow other trajectories, as well, which could increase or decrease their relative risk.

Let us consider the first trajectory, $DM, HL, HTN \rightarrow CHF$, as a different example. Patients along this trajectory have pre-existent HL, HTN, and DM and develop CHF afterwards. In this trajectory, the sequence in which the initial HL, HTN and DM are developed is unknown, a patient who has developed them in any order matches this trajectory. Since patients can develop them sequentially (others can have them upon enrollment into the study), we can still estimate the effects of these conditions individually.

The sixth trajectory, $HTN \rightarrow HL \rightarrow DM \rightarrow CHF$, is a more specific version of the above $DM, HL, HTN \rightarrow CHF$ trajectory, where the ordering of HTN, HL and DM is fixed. If both trajectories are selected into a model, which is the case in 37 of the 41 models that selected the sixth trajectory, both trajectories apply to the patients, so the coefficients of the more specific trajectory can be viewed as modifiers to the general trajectory: in these patients the effects of HTN and HL are less severe but that of DM is more severe. If such a modification was not necessary, this trajectory would not be selected. This shows that the same conditions can have different effects depending on the order in which the diseases were developed.

The models that utilized trajectories were able to give more correct estimates for the diabetes complications (CVD, CHF, IHD, CKD). Recall that the models that did not utilize trajectories ended up estimating some of these severe conditions as protective (Table 4.2); this does not happen with trajectories: the coefficients of the diabetes complications are always positive, which is consistent with our clinical expectation.

4.7 Discussion

In this manuscript we presented Forensic-style Analysis based on Survival Trajectories (FAST). FAST makes two key contributions: it places patients onto disease trajectories to assess their risk of progression to an adverse outcome (mortality in our study) and it performs a forensic-style analysis, where patients are aligned on their last follow-up and time is measured backwards. Measuring time backwards allows a third ancillary contribution: we can designate the outcome as positive for all records of cases (patients who ultimately died), potentially leading to better estimates of the effects of diseases that occur early in the progression.

The typical method for solving this problem comes from the domain of epidemiology as a combination of study design (longitudinal retrospective: patients are aligned on enrollment and followed longitudinally) and a modeling algorithm Cox proportional hazards model with time dependent covariates.

When the patients' state does not evolve over time, namely they entered the study with a set of diseases and had the same set of diseases at last follow-up, our method simplifies to this above typical method and measuring time backwards or forwards, aligning patients on enrollment or outcome will not make a difference. This is situation for the vast majority of the studies in existence.

When the patients' state evolves over time, our proposed method starts to differ. To isolate the effect of our innovations, we successively enhanced this baseline method (which we referred to as Enrollment-Aligned Design) by adding our contributions one at a time. We have thus demonstrated the benefit of aligning patients on outcome when we believe that time-to-death is important; we demonstrated the benefit of our outcome designation and we have also isolated the beneficial effect of using trajectories.

We then took a deeper look at our proposed methodology to show that it conforms with medical knowledge. We have shown that unlike the models that did not utilize trajectories, the trajectory based model identified diabetes complications as harmful. The Outcome-Aligned model identified the heart diseases (IHD and CHF) as harmful, and indeed these diseases often lead to mortality, but only the trajectory based method identified CKD and CVD as factor that increase the risk of mortality.

Another hypothesis was the models with the traditional outcome designation may incorrectly designate early risk factors as protective. Many patients entered the study with HTN and HL causing the design matrix of our study to become near-singular, thus the coefficients of HTN and HL could not be estimated at all. Some trajectories use one or the other, thus through the trajectories it because possible to estimate their effect (but only within a trajectory). Given the high disease burden of our study population, we cannot expect positive coefficients for all diseases along all trajectories: some patients with some diseases have below-average risk of death. However, we found some trajectories, e.g. the last two in Table 4.3, which have early diseases and yet have all positive coefficients.

	Dead	Censored
Total Number of Patients	**2814**	**1976**
Average Age At First Encounter	63.83	61.83
At Last Encounter	72.41	66.22
Mean Number of Follow-up Years	8.58	4.35
Patient diagnosis history		
DM at first encounter	1032	713
by last encounter	2496	1753
HL at first encounter	1630	832
by last encounter	2631	1636
HTN at first encounter	1572	1084
by last encounter	2803	1885
CHF at first encounter	96	78
by last encounter	875	159
IHD at first encounter	206	243
by last encounter	1039	447
CVD at first encounter	80	82
by last encounter	604	152
CKD at first encounter	67	150
by last encounter	1093	343
Male	1383	1062
Female	1431	914
Number of patients with lab results		
Blood pressure	2771	1887
a1c	2814	1976
Lipid panel	2666	1553
GFR	2812	1128
Number of patients with abnormal lab results		
Blood pressure	2811	1511
a1c	2741	1556
Lipid panel	2535	1243
GFR	450	306

Table 4.1: Demographics Statistics of Patient population

covariate	enrollment	outcome	fast w/o traj
HL	–	–	0.00
HTN	–	–	0.00
DM	14.56	16.24	8.21
CKD	-0.45	-0.21	-0.07
IHD	0.42	0.05	0.00
CVD	-0.28	-0.37	-0.01
CHF	-0.24	0.10	0.00

Table 4.2: Coefficients of the non-trajectory based models

	HL	HTN	DM	CKD	IHD	CVD	CHF
$DM, HL, HTN \rightarrow CHF$ (263)							
coefs	-0.23	-0.31	0.13	–	–	–	1.21
p-val	0.04	0.00	0.06	–	–	–	0.00
$DM, HL, HTN \rightarrow CKD$ (192)							
coefs	-0.25	-0.30	0.12	1.09	–	–	–
p-val	0.03	0.01	0.04	0.00	–	–	–
$DM, HL \rightarrow HTN \rightarrow CKD$ (41)							
coefs	-0.10	-0.11	-0.04	0.76	–	–	–
p-val	0.02	0.02	0.27	0.00	–	–	–
$HL, HTN \rightarrow DM \rightarrow CKD$ (30)							
coefs	-0.13	-0.05	0.11	0.88	–	–	–
p-val	0.13	0.43	0.47	0.00	–	–	–
$HTN \rightarrow DM, HL \rightarrow CKD$ (22)							
coefs	-0.07	-0.07	0.01	0.83	–	–	–
p-val	0.18	0.14	0.50	0.00	–	–	–
$HTN \rightarrow HL \rightarrow DM \rightarrow CHF$ (16)							
coefs	-0.05	-0.01	0.15	–	–	–	0.86
p-val	0.25	0.50	0.25	–	–	–	0.00
$HL, HTN \rightarrow IHD$ (16)							
coefs	-0.03	0.07	–	–	0.56	–	–
p-val	0.38	0.12	–	–	0.00	–	–
$DM, HL, HTN \rightarrow IHD$ (14)							
coefs	-0.10	-0.15	0.03	–	0.69	–	–
p-val	0.36	0.29	0.43	–	0.14	–	–
$HL \rightarrow HTN \rightarrow CVD \rightarrow DM$ (14)							
coefs	0.02	0.03	0.13	–	–	0.70	–
p-val	0.36	0.18	0.00	–	–	0.00	–
$HTN \rightarrow HL \rightarrow DM$ (10)							
coefs	0.09	0.13	0.20	–	–	–	–
p-val	0.00	0.00	0.00	–	–	–	–

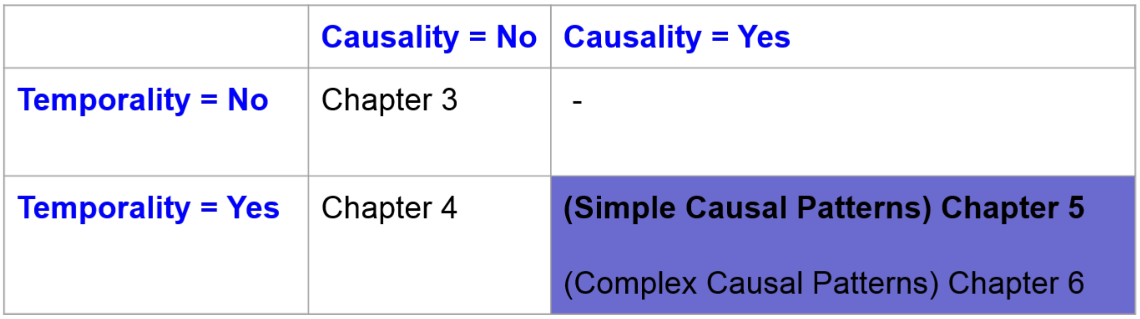
Table 4.3: Trajectories and their coefficients that were utilized in at least 10 models

Chapter 5

Simple Causal Pattern Mining

5.1 Introduction

In this chapter, we would discussing techniques to extract simple causal patterns in the presence of longitudinal data as illustrated in Figure 5.1.



	Causality = No	Causality = Yes
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	**(Simple Causal Patterns) Chapter 5** (Complex Causal Patterns) Chapter 6

Figure 5.1: Chapter 5 Description

By simple causal patterns, we imply those patterns which consists of three groups of random variables. Group X consists of random variable, whose intervention we wish to computer, Group Y consists of random variables (i.e. outcome variables). Group Z consists of random variables which are also known as confounding variables. Figure 5.2 illustrates such patterns.

Such patterns are widely used to estimate the effect of medical or surgical interventions on outcomes of interest. Further, such analysis forms the basis for identification of

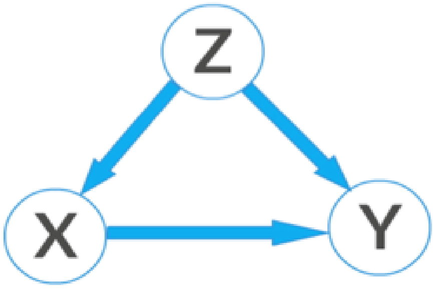


Figure 5.2: Simple Causal Pattern

antiquated guidelines, development of new guidelines and discovery of adverse events. In this chapter we will illustrate such patterns in the context of T2DM.

5.2 Clinical Motivation

According to the Center for Disease Control and Prevention, the incidence of sepsis or septicemia has doubled from 2000 through 2008, and hospitalizations have increased by 70% for these diagnoses[54]. In addition, severe sepsis and shock have higher mortality rates than other sepsis diagnoses, accounting for an estimated mortality between 18% and 40% [55, 56]. During the first 30 days of hospitalization, mortality can increase to a range of 10% to 50% [57] depending on the patients risk factors. Patients with severe sepsis or septic shock are sicker, have longer hospital stays, and are more frequently discharged to other short-term hospital or long-term care institutions than patients with other conditions.

The use of evidence-based practice (EBP) guidelines, such as the Surviving Sepsis Campaign (SSC) [58], could lead to an earlier diagnosis, and consequently, earlier treatment. However, these guidelines have not been widely incorporated into clinical practice. The SSC is a compilation of international recommendations for the management of severe sepsis and shock [58]. Many of these recommendations are interventions to prevent further system deterioration during and after diagnosis. Even when the presence of sepsis or progression to sepsis is suspected early in the course of treatment, timely implementation of adequate treatment management and guideline compliance are still a challenge [59]. Therefore, the effectiveness of the guideline in preventing clinical complications for this population is still unclear to clinicians and researchers alike.

The majority of studies have focused on early detection and prevention of sepsis and little is known about the compliance rate to SSC and the impact of compliance on the prevention of sepsis-related complications. Further, the measurement of adherence to individual SSC recommendations rather than the entire SSC is, to our knowledge, limited [60] . The majority of studies have used traditional randomized control trials with analytic techniques such as regression modeling to adjust for risk factors known from previous research [61]. Data-driven methodologies, such as data mining techniques and machine learning, have the potential to identify new insights from electronic health records (EHRs) that can strengthen existing EBP guidelines.

The national mandate for all health professionals to implement interoperable EHRs by 2015 provides an opportunity for the reuse of potentially large amounts of EHR data to address new research questions that explore patterns of patient characteristics, evidence-based guideline interventions, and improvement in health [62, 63, 64]. Furthermore, expanding the range of variables documented in EHRs to include team-based assessment and intervention data can increase our understanding of the compliance with EBP guidelines and the influence of these guidelines on patient outcomes. In the absence of such data elements, adherence to guidelines can only be inferred; it cannot be directly observed.

In this chapter, we present a methodology for using EHR data to estimate the compliance with the SSC guideline recommendations and also estimate the effect of the individual recommendations in the guideline on the prevention of in-hospital mortality and sepsis-related complications in patients with severe sepsis and septic shock.

5.3 Methods

Missing Data: Any observation that took place before the estimated onset of sepsis (TimeZero) was considered baseline observation. Simple mean imputation was the method of choice for imputing missing values. Imputation was necessary for lactate (7.7%), temperature (3%), and WBC (3%). There was no missing data for the other variables and for the outcomes of interest. Central venous pressure was not included as a baseline characteristic due to the high number of missing values (54%).

Propensity Score Matching : Patients who received an intervention may be in worse health than patient who did not receive an intervention. For example, patients whose lactate was measured may have more apparent (and possibly advanced) sepsis than patients whose lactate was not measured. To compensate for such disparities, propensity score matching (PSM) was employed. The goal of PSM is to balance the data set in terms of the covariates between the exposed and unexposed groups. This is achieved by matching exposed patients with unexposed patients on their propensity (probability) of receiving the intervention. This ensures that at TimeZero, pairs of patients, one exposed and one unexposed, is at the same state of health and only differs in their exposure to the recommendation. PSM is a popular technique for estimating treatment effects [65, 66].

To compute the propensity of patients to receive treatment, a logistic regression model was used, where the dependent variable is exposure to the recommendation and the independent variables are the covariates. The linear prediction (propensity score) of this model was computed for every patient. A new (matched) population was created from pairs of exposed and unexposed patients with matching propensity scores. Two scores match if they differ by no more than a certain caliper (.1 in our study) [67]. The effect of the recommendation was estimated by comparing the incident fraction among the exposed and unexposed patients in the matched population.

PSM nested inside Bootstrapping Simulation: In order to incorporate the effect of additional sources of variability arising due to estimation in the propensity score model and variability in the propensity score matched sample, 500 bootstrap samples were drawn from the original sample [68]. In each of these bootstrap iterations, the propensity score model was estimated using the above caliper matching techniques and the effect of the recommendation was computed with respect to all outcomes. In recent years, bootstrap simulation has been widely employed in conjunction with PSM to better handle bias and confounding variables17. For each recommendation and outcome, the 500 bootstrap iterations result in 500 estimates of the effect (of the recommendation on the outcome), approximating the sampling distribution of the effect.

5.4 Data and Cohort

Data from the EHR of a health system in the Midwest was transferred to a clinical data repository (CDR) at the University of Minnesota which is funded through a Clinical Translational Science Award. After IRB approval, de-identified data for all adult patients hospitalized between 1/1/09 to 12/31/11 with a severe sepsis or shock diagnosis was obtained for this study.

The sample included 186 adult patients age 18 years or older with an ICD-9 diagnosis code of severe sepsis or shock (995.92 and 785.5*) identified from billing data. Since 785.* codes corresponding to shock can capture patients without sepsis, patients without severe sepsis or septic shock, and patients who did not receive antibiotics were excluded. These exclusions aimed to capture only those patients who had severe sepsis and septic shock, and were treated for that clinical condition. The final sample consisted of 177 patients. Variables of Interest

The fifteen predictor variables (baseline characteristics) were collected. These include socio-demographics and health disparities data: age, gender, race, ethnicity, and payer (Medicaid represents low income); laboratory results: lactate and white blood cells count (WBC); vital signs: heart rate (HR), respiratory rate (RR), temperature (Temp), mean arterial blood pressure (MAP); and diagnoses for respiratory, cardiovascular, cerebrovascular, and kidney-related co-morbid conditions. ICD-9 codes for comorbid conditions were selected according to evidence in the literature. Co-morbidities were aggregated from the patientś prior problem list to detect preexisting (upon admission) respiratory, cardiovascular, cerebrovascular, and kidney problems. Each category was treated as yes/no if any of the ICD-9 codes in that category were present.

The outcomes of interest were in hospital mortality and development of new complications (respiratory, cardiovascular, cerebrovascular, and kidney) during the hospital encounter. New complications were determined as the presence of ICD-9 codes on the patientś billing data that did not exist at the time of the admission.

This chapter aims to analyze compliance with the SSC guideline recommendations in patients with severe sepsis or septic shock. Therefore, the baseline (”TimeZero”) was defined as the onset of sepsis and the patients were under observation until discharged. Unfortunately, the timestamp for the diagnoses is dated back to the time of admission;

hence the onset of sepsis needs to be estimated. The onset time for sepsis was defined as the earliest time during a hospital encounter when the patient meets at least two of the following six criteria: $MAP < 65, HR > 100, RR > 20, temperature < 95or > 100.94, WBC < 4or > 12, andlactate > 2.0$. The earliest time when two or more of these aforementioned conditions were met, a TimeZero flag was added to the time of first occurrence of that abnormality, and the timing of the SSC compliance commenced.

Guideline Compliance SSC guideline recommendations were translated into a readily computable set of rules. These rules have conditions related to an observation (e.g. MAP ¡ 65 Hgmm) and an intervention to administer (e.g. give vasopressors) if the patient meets the condition of the rule. The SSC guideline was transformed into 15 rules, one for each recommendation in the SSC guideline, and each rule was evaluated for each patient.

We call the treatment of a patient compliant with a specific recommendation, if the patient meets the condition of the corresponding rule any time after TimeZero and the required intervention was administered; the treatment is non-compliant if the patient meets the condition of the corresponding rule after TimeZero, but the intervention was not administered (any time after TimeZero); and the recommendation is not applicable to a treatment if the patient does not meet the condition of the corresponding rule. In estimating compliance (as a metric) with a specific recommendation, we simply measure the number of compliant encounters to which the recommendation is applicable. We also estimate the effect of the recommendation on the outcomes. We call a patient exposed to a recommendation, if the recommendation is applicable to the patient and the corresponding intervention was administered to the patient. We call a patient unexposed to a recommendation if the recommendation is applicable but was not applied (the treatment was non-compliant). The incidence fraction in exposed patients with respect to an outcome is the fraction of patients with the outcome among the exposed patients. The incidence fraction of the unexposed patients can be defined analogously. We define the effect of the recommendation on an outcome as the difference in the incidence fractions between the unexposed and exposed patients. The recommendation is beneficial (protective against an outcome) if the effect is positive, namely, the incidence faction in the unexposed is higher than the incidence fraction in the unexposed patients.

5.5 Results

Table 1 shows the baseline characteristics of the study population. Results are reported as total count for categorical variables, and mean with inter-quartile (25%-75%) range for continuous variables. As shown in Table 1, the majority of patients were male, Caucasian, and had Medicaid as the payer. Before the onset of sepsis, Cardiovascular co-morbidities (56.4%) were common, the mean HR (101.3) was slightly above the normal, as well as lactate (2.8), and WBC (15.8). The mean length of stay for the sample was 15 days, ranging from less than 24 hours to 6 months. TimeZero was within the first 24 hours of admission, and patients at that time were primarily (86.4%) in the emergency department.

Feature	Mean
Total Number of Patients	177
Average Age	61
Gender(Male)	102
Race(Caucasian)	97
Ethnicity(Latino)	11
Payer(Medicaid)	102
White Blood cell	15.8
Lactate	28
Mean blood Pressure	73.9
Temperature	98.4
Heart Rate	101.3
Respiratory Rate	20.6
Cardiovascular	100
Cerebrovascular	66
Respiratory	69
Kidney	62

Table 5.1: Demographics statistics of patient population

Fifteen rules from the SSC recommendations were identified. Figure 5.3 presents a description of these rules along with the number of patients whose treatment was compliant with the recommendation in question. Y means the treatment of the patient was compliant, N indicates non-compliant and N/A means the recommendation (rule) was not applicable or could not be calculated. Using this information, rules LactateFluid, GlucoseInsulin, MAP, MAPFluids, CVPFluids, Albumin, and Diuretic were removed as patient coverage, either in the exposed or unexposed group, was insufficient. Rules BCulture, Antibiotic, Lactate, BGlucose, Vasopressor, CVP, RespDistress, and Ventilator were included.

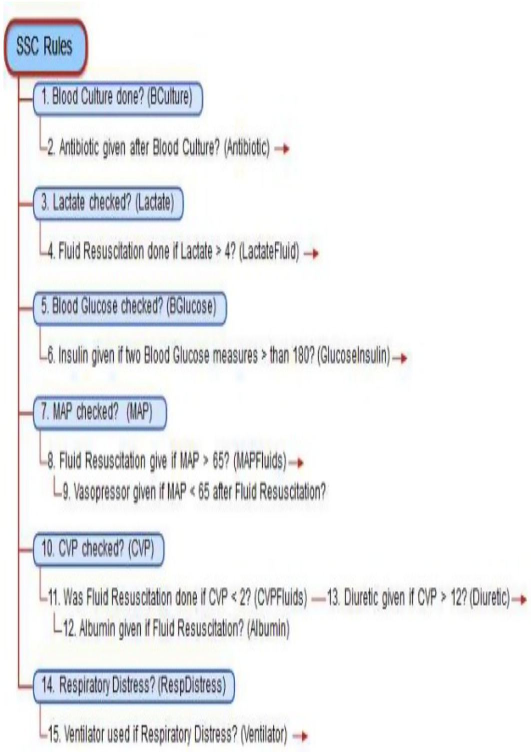

SSC Rules
1. Blood Culture done? (BCulture)
2. Antibiotic given after Blood Culture? (Antibiotic)
3. Lactate checked? (Lactate)
4. Fluid Resuscitation done if Lactate > 4? (LactateFluid)
5. Blood Glucose checked? (BGlucose)
6. Insulin given if two Blood Glucose measures > than 180? (GlucoseInsulin)
7. MAP checked? (MAP)
8. Fluid Resuscitation give if MAP > 65? (MAPFluids)
9. Vasopressor given if MAP < 65 after Fluid Resuscitation?
10. CVP checked? (CVP)
11. Was Fluid Resuscitation done if CVP < 2? (CVPFluids)
13. Diuretic given if CVP > 12? (Diuretic)
12. Albumin given if Fluid Resuscitation? (Albumin)
14. Respiratory Distress? (RespDistress)
15. Ventilator used if Respiratory Distress? (Ventilator)

Figure 5.3: SSC rules for measuring guideline compliances

In Figure 5.4, the difference in the mean rate of progression to complications between the exposed and unexposed groups is depicted. Since we used bootstrap simulation, for each rule-complication pair, 500 replications were performed resulting in 500 estimates for the effect. These estimates are presented as boxplots. The panes (groups of boxplots) correspond to the complications and the boxes within each pane correspond to the recommendation (rule). For example, the effect of the Ventilator rule (Recommendation 15: patients in respiratory distress should be put on ventilator) on Death is shown in the rightmost box (Ventilator) in the bottom-most pane (Death). Since all effects in the boxplot are above 0, namely the number of observed complications in the unexposed group is higher than in the exposed; compliance with the Ventilator rule reduces the number of deaths. Therefore, the corresponding recommendation is beneficial (on mortality). The use of Ventilator was also beneficial in reducing the Respiratory and Cardiovascular complications. Similarly, checking glucose was found beneficial in reducing Respiratory and Cardiovascular complications.

To further ensure the validity of the results, we examine the propensity score distribution in the exposed and unexposed group. As an illustration, Figure 5.5 depicts the propensity score distribution for a randomly selected bootstrap iteration to measure the effect of Ventilator on Death. The horizontal axis represents the propensity score, which is the probability of receiving the interventions, and the vertical axis represents the density distribution, namely the proportion of patients in each group with a particular propensity for being put on Ventilator. Figure 2 shows substantial overlap between the propensity scores in the exposed and unexposed group.

5.6 Conclusion

The overall purpose of this study was to use EHR data to determine compliance to the Surviving Sepsis Campaign (SSC) guideline and measure its impact on inpatient mortality and sepsis complications in patients with severe sepsis and septic shock. Results showed that compliance with many of the recommendations was outstanding: MAP was measured in all patients, and the recommendations related to checking blood culture, lactate, blood glucose, and respiratory distress were followed in the overwhelming majority of the patients.

On the other hand, the treatment of a large number of patients was not compliant with the CVP Fluids recommendation (to perform fluid resuscitation in patients with CVP below 2). This may be due to a study design artifact, where the rule only considered interventions initiated after TimeZero (estimated onset of sepsis), while the fluid resuscitation may have taken place earlier. Alternatively, the apparently poor compliance could also be explained with issues related to the coding of fluids: during data validation, we found that the majority of fluids were not coded in the system.

Our study also demonstrates that retrospective EHR data can be used to evaluate the effect of compliance with guideline recommendations on outcomes. We found a number of SSC recommendations that were statistically significantly protective against more than one complication: Ventilator and BGlucose were protective against Death (not BGlucose), Respiratory and Cardiovascular.

Other recommendations, BCulture, Antibiotic, Vasopressor, Lactate, CVP, and RespDistress, showed results less consistent with our expectation. For instance, Vasopressor used to treat low MAP, appears to increase cerebrovascular complications. While this finding is not strictly statistically significant, it may be congruent with the fact that small brain vessels are very sensitive to changes in blood pressure. Low MAP can cause oxygen deprivation, and consequently brain damage.

Ventilator, Vasopressor, and BGlucose showed protective effects against Respiratory complications. The SSC guideline recommends the implementation of ventilator therapy as soon as any change in respiratory status is noticed. This intervention aims to protect the patient against further system stress, restore hypoxia, help with perfusion across the main respiratory-cardio vessels, and decrease release of toxins due to respiratory efforts.

Our study is a proof-of-concept study demonstrating that EHR data can be used to estimate the effect of guideline recommendations. However, for several combinations of recommendations and outcomes, the effect was not significant. We believe that the reason is that guidelines represent workflows and the effect of the workflow goes beyond the effects of the individual guideline recommendations. For example, by considering the recommendations outside the context of the workflow, we may ignore whether the intervention addressed the condition that triggered its administration. If low MAP triggered the administration of vasopressors, without considering the workflow, we do not know whether MAP returned to the normal levels thereafter. Thus we cannot equate an adverse outcome with the failure of the guideline, it may be the result of the insufficiency of the intervention. Moving forward, we are going to model the workflows behind the guidelines and apply the same principles that we developed in this work to estimate the effect of the entire workflow.

Another limitation of the study concerns timing. For this analysis, guideline compliance was considered only after TimeZero (the estimated onset), since compliance with SSC is only necessary in the presence of suspected or confirmed sepsis. There is no reason to suspect sepsis before TimeZero. However, some interventions may have started earlier, without respect to sepsis. For example, 100% of the patients in this sample had antibiotics (potentially preventive antibiotics), but only 99 (55%) patients received it after TimeZero.

A third limitation is that EHR does not provide date and time for certain ICD-9 diagnoses. During a hospital stay, all new diagnoses are recorded at the day of admission. We know whether it was present on admission or not, thus we know whether it is a pre-existing or new condition, but do not know precisely when the patient developed this condition during the hospitalization. For this reason, we are unable to detect whether the SSC guideline was applied before or after a complication occurred, thus we may underestimate the beneficial effect of some of the recommendations. For example, high levels of lactate is highly related to hypoxia and pulmonary damage9. If these patients were checked for lactate after pulmonary distress, we would consider the treatment compliant with the Lactate recommendation, but we would not know that the respiratory distress was already present at the time of the lactate measurement and we would incorrectly count it as a complication that the guideline failed to prevent.

This study demonstrated that retrospective EHR data could be used to estimate compliance with individual guideline recommendations in the SSC guideline. Further, EHR data can be used to estimate the effect of guideline adherence on sepsis-related complications in patients with severe sepsis and septic shock. We found that most treatment courses we observed were compliant with many guideline recommendations and were able to demonstrate these recommendations have significant beneficial (protective) effect on some outcomes. Since guidelines encapsulate a workflow, which goes beyond a mere collection of recommendations, further study is needed to prove the beneficial effect of the entire SSC workflow.

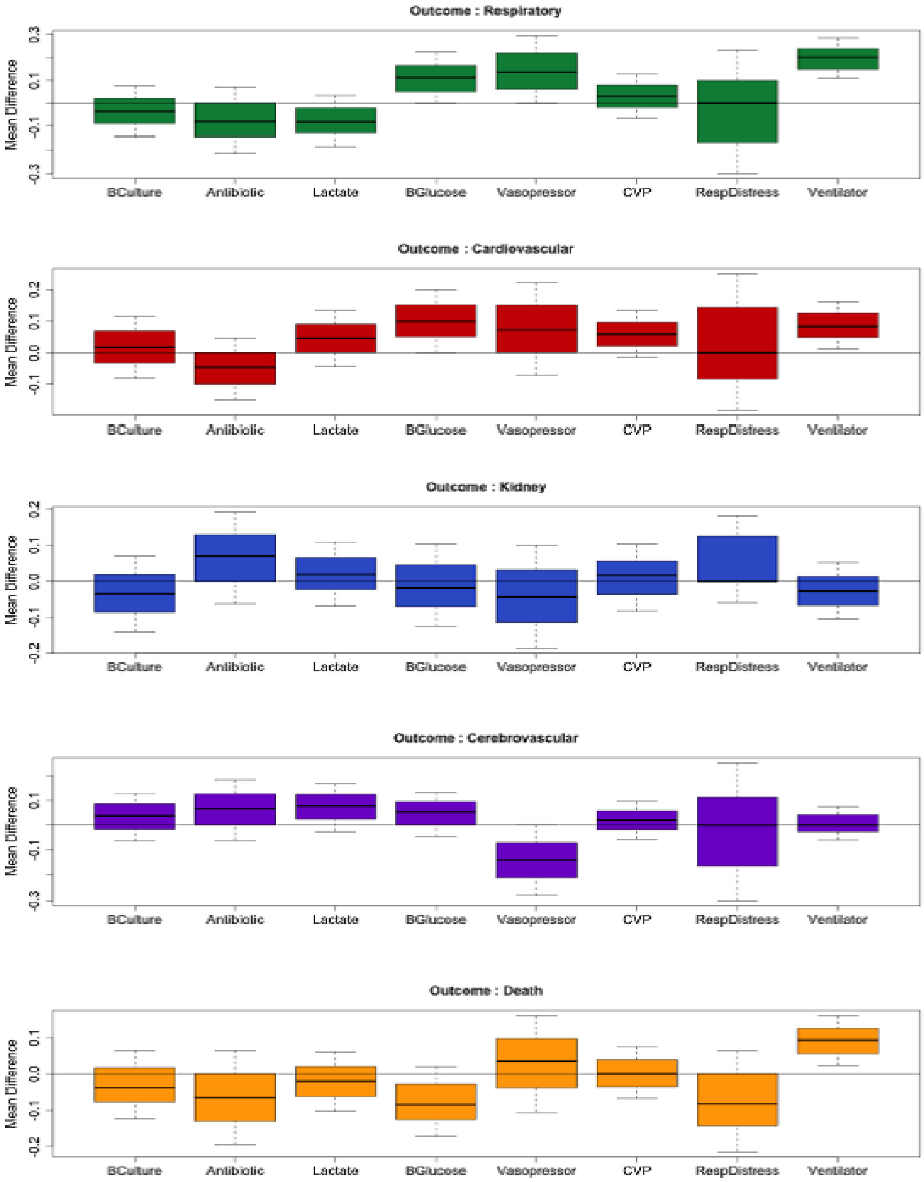


Figure 5.4: Causal effect estimation for different outcomes across subpopulations

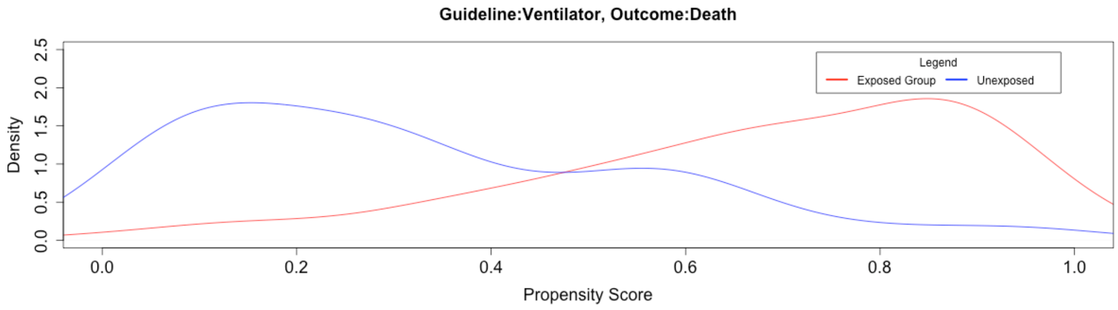


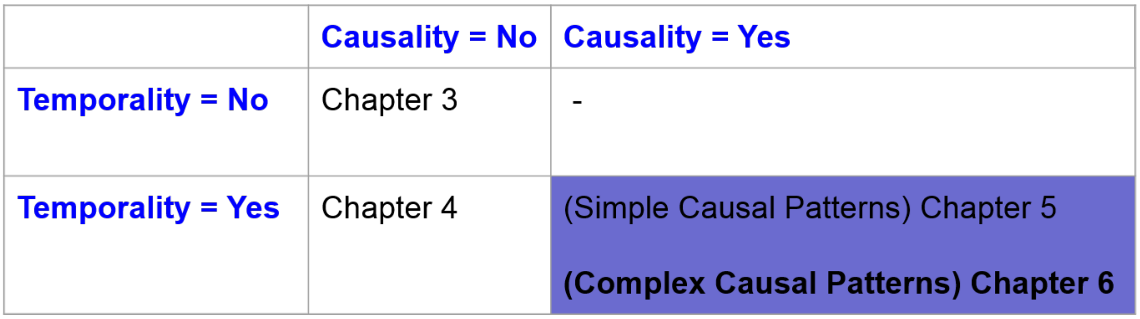
Figure 5.5: Propensity Score Overlap

Chapter 6

Complex Causal Pattern Mining

6.1 Introduction

In this chapter, we would discussing techniques to extract simple complex patterns in the presence of longitudinal data as illustrated in Figure 6.1.



	Causality = No	**Causality = Yes**
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	(Simple Causal Patterns) Chapter 5 **(Complex Causal Patterns) Chapter 6**

Figure 6.1: Chapter 6 Discussion

By complex causal patterns, we imply those patterns which consists of random variables as denoted in 6.2. These random variables have been discussed in detail in the later part of this chapter. Such patterns are widely used to estimate the efficacy of medical and clinical interventions for outcomes of interest. In Chapter 5, we discussed simple causal patterns wherein only the effect of confounding variables was incorporated to estimate the effect of random variable X on random variable Y. In this Chapter 6, we will also incorporate the effect of random variables U,V and O while incorporating the effect of X on Y. In this chapter we will illustrate such patterns in the context of T2DM.

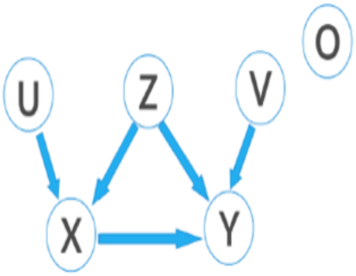


Figure 6.2: Complex Causal Pattern

6.2 Clinical Motivation

Effective management of human health remains a major societal challenge as evidenced by the rapid growth in the number of patients with multiple chronic conditions. Type-II Diabetes Mellitus (T2DM), one of those conditions, affects 25.6 million (11.3%) Americans of age 20 or older and is the seventh leading cause of death in the United States [10]. Effective treatment of T2DM is frequently complicated by diseases comorbid to T2DM, such as high blood pressure, high cholesterol, and abdominal obesity. Currently, these diseases are treated in isolation, which leads to wasteful duplicate treatments and suboptimal outcomes.

Finding optimal treatment for patients who suffer from multiple associated diseases, each of which can have multiple available treatments is a complex problem. We could simply use techniques based on association, but a reasonable algorithm would likely find that the use of a drug is associated with some unfavorable outcome. This does not mean that the drug is harmful; in fact in many cases, it simply means that patients who take the drug are sicker than those who do not and thus they have a higher chance of the unfavorable outcome. What we really wish to know is whether a treatment *causes* an unfavorable outcome, as opposed to being merely associated with it.

The difficulty in quantifying the effect of interventions on outcomes stems from subtle biases. Suppose we wish to quantify the effect of a cholesterol-lowering agent, statin, on diabetes. We could simply compare the proportion of diabetic patients in the subpopulation that takes statin and the subpopulation that does not and estimate the effect of statin as the difference between the two proportions. This method would give the correct answer only if the statin-taking and non-statin-taking patients are identical in all respects that influence the diabetes outcome. We refer to this situation as treated and untreated patients being *comparable*. Unfortunately, statin taking patients are not comparable to non-statin-taking patients, because they take statin to treat high cholesterol, which by and in itself increases the risk of diabetes. High cholesterol *confounds* the effect of statin. Many different sources of bias exist, confounding is just one of the many. In this manuscript, we are going to address several different sources of bias, including confounding.

Two key challenges arise when trying to estimate the effect of multiple interventions First, is the issue of *comparability*: to estimate the effect of intervention, we need two groups of patients who are identical in all relevant aspects except that one group receives the intervention and the other group does not. For a single intervention, the first group is typically the sicker patients who still do not get treated and the second group consists of the healthier patient who get treatment. They are reasonably in the same state of health. However, when we go from a single intervention to multiple intervention and try to estimate their *joint* effect, comparability no longer exists. A patient requiring multiple simultaneous interventions is so fundamentally different from a patient who does not need any intervention that they are not comparable.

Analogously, to care providers adjusting drugs one or two at a time rather than introducing then to the treatment regiment in large sets, we will estimate the effect of large intervention sets sequentially. Not only does this approach provide more reliable estimates it directly helps provide chose the appropriate treatment under many conditions.

The other key challenge in finding optimal intervention sets for patients with combinatorial sets of diseases is the combinatorial search space. Even if we could trivially extend the methods for quantifying the effect of a single intervention to a set of concurrent interventions, we would have to systematically explore a combinatorially large search space. The association rule mining framework [69] provides an efficient solution for exploring combinatorial search spaces, however, it only detects associative relationships. Our interest is in causal relationships.

In this manuscript, we propose causal rule mining, a framework for transitioning from association rule mining towards causal inference in subpopulations. Specifically, given a set of interventions and a set of items to define subpopulations, we wish to find all subpopulations in which effective intervention combinations exist and in each such subpopulation, we wish to find all intervention combinations such that dropping any intervention from this combination will reduce the efficacy of the treatment. We call these *closed intervention sets*, which are not to be confused with closed item sets. As a concrete example, interventions can be drugs, subpopulations can be defined in terms of their diseases and for each subpopulation (set of diseases), our algorithm would return effective drug cocktails of an increasing number of constituent drugs. Leaving out any drug from the cocktail will reduce the efficacy of the treatment.

To address the exploration of the combinatorial search space, we propose a novel frequency-based anti monotonic pruning strategy enabled by the closed intervention set concept. The essence of anti-monotonic property is that if a set I of interventions does not satisfy a criterion, none of its supersets will. The proposed pruning strategy based on closed intervention sets allows for additional pruning beyond the support based pruning strategy used by the Apriori algorithm [69].

Underneath our combinatorial exploration algorithm, we utilize the Rubin-Neyman model of causation [70]. This model sets two conditions for causation: a set X of interventions causes a change in Y iff X happens before Y and Y would be different had X not occurred. The unobservable outcome of what would happen had a treated patient not received treatment is a *potential outcome* and needs to be estimated. We present and compare five methods for estimating these potential outcomes and describe the biases these methods can correct.

Typically the ground truth for the effect of drugs is not known. In order to assess the quality of the estimates, we conduct a simulation study utilizing five different synthetic data sets that introduce a new source of bias. We will evaluate the effect of the bias on the five proposed methods underscoring the statements with rigorous proofs when possible.

We also evaluate our work on a real clinical data set from the Mayo Clinic. We have data for over 52,000 patients with 11 years of follow-up time. Our outcome of interest is 5-year incidence of T2DM and we wish to extract patterns of interventions for patients suffering from combinations of common co-morbidities of T2DM. First, we evaluate our methodology in terms of the computational cost, demonstrating the effectiveness of the pruning methodologies. Next, we evaluate the patterns qualitatively, using patterns involving statins. We show that our methodology extracted patterns that allow us to explain the controversial patterns surrounding statin [71].

Contributions. (1) We propose a novel framework for extracting causal rules consisting of multiple interventions with multiple outcomes in subpopulations of interest. (2) We introduce the concept of closed intervention sets to extend the concept of quantifying the effect of a single intervention to a set of concurrent interventions thus sidestepping the patient comparability problem. Closed intervention sets also allow for a pruning strategy that is strictly more efficient than the traditional pruning strategy used by the Apriori algorithm [69]. (3) We compare five methods of estimating causal effect from observational data that are applicable to our problem and rigorously evaluate them on synthetic data and mathematically prove (when possible) why they work.

6.3 Background

Consider a set $\mathcal{X}$ of **items**, which are single-term predicates evaluating to ‘true’ or ‘false’. For example, $\{age > 55\}$ can be an item. A k-**itemset** is a set of k items, evaluated as the conjunction (logical ’and’) of its constituent items. Consider a dataset D = $\{ d_1, d_2 .....d_n \}$, which consists of n **observations**. Each observation, denoted by D_j is a set of items. An itemset X=$\{x_1, x_2, \ldots, x_k\}$ ($X \subset \mathcal{I}$) **supports** an observation D_j if all items in X evaluate to ‘true’ in the observation. The **support** of X is the fraction of the observations in D that support X. An itemset is **frequent** if its support exceeds a pre-defined minimum support threshold.

A association rule is a logical implication of form $X \Rightarrow Y$, where X and Y are disjoint itemsets. The support of a rule is (XY) and the **confidence** of the rule is $(Y|X)$.

Given an **intervention** itemset X and an **outcome** item Y, such that X and Y are disjoint, a causal rule is an implication of form XY, suggesting that X *causes* a change in Y. Let the itemset S define a **subpopulation**, consisting of all observations that support S. This subpopulation consists of all observations for which all items in S evaluate to 'true'. The **causal rule** $XY|_S$ implies that the intervention X has causal effect on Y in the subpopulation defined by S. The quantity of interest is the **causal effect**, which is the change in Y in the subpopulation S caused by X. We will formally define the metric used to quantify the causal effect shortly.

Rubin-Neyman Causal Model. X has a causal effect on Y if (i) X happens earlier than Y and (ii) if X had not happened, Y would be different [70].

Our study design ensures that the intervention X precedes the outcome Y, but fulfilling the second conditions requires that we estimate the outcome for the same patient both under intervention and without intervention.

Potential Outcomes. Every patient in the dataset has two potential outcomes: Y_0 denotes their outcome had they not had the intervention X; and Y_1 denotes the outcome had they had the intervention. Typically, only one of the two potential outcomes can be observed. The observable outcome is the **actual** outcome (denoted by Y) and the unobservable potential outcome is called the **counterfactual** outcome.

Using the definition of counterfactual outcome, we can now define the metric for estimating the change in Y caused by X. **Average Treatment response on the Treated** (ATT) [72] is a widely known metric in the causal literature and is computed as follows: (X Y—$_S$) $= [Y_1 - Y_0]_{X=1} = [Y_1]_{X=1} - [Y_0]_{X=1}$, where denotes the expectation and the $X = 1$ in the subscript signals that we only evaluate the expectation in the treated patients ($X = 1$).

As we mentioned before, computing ATT for a large set X of interventions may be difficult or even impossible because of the difficulty of finding *comparable* patients. Thus we introduce the concept of **differential causal effect**, denoted by ΔATT, which quantifies the excess ATT that a new intervention x exerts on top of the set X' of interventions the patient already receives:

$$(xY|_S) = [Y_1 - Y_0]_{S,X'x=1}.$$

[η - closed intervention set] An intervention set X is η-closed iff

$$\forall x \in X, \quad |ATT(xY|_{S,X \setminus x})| > \eta.$$

Intuitively, an intervention set X is η-closed if removing any intervention x reduces its effect by at least η. In practical terms we would only add drug x to a cocktail X' if the resultant cocktail $X'x$ is more effective than all of its subsets by at least η.

Biases. Beside X, numerous other variables can also exert influence over Y, leading to biases in the estimates. The quintessential tool for eliminating or reducing these biases is the causal graph, depicted in Figure 6.3. The nodes of this graph are sets of variables that play a causal role and edges are causal effects. This is not a correlation graph (or dependence graph), because for example, U and Z are dependent given X, yet there is no edge between them.

Variables (items in $\mathcal{I}$) can exert influence on the effect of X on Y in three ways: they may only influence X, they may only influence Y or them may influence both X and Y. Accordingly, variables can be put into four categories:

V are variables that directly influence Y and thus have *direct effect* on Y

U are variables that only influence Y through X and thus have *indirect effect* on Y;

Z are variables that influence both X and Y and are called *confounders*; and finally

O are variables that do not influence either X or Y and hence can be safely ignored.

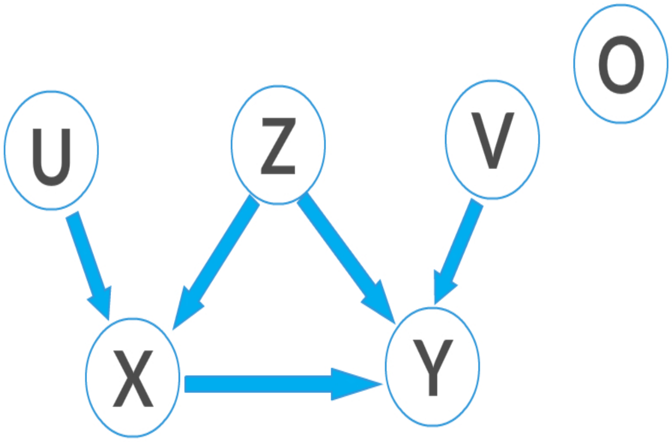


Figure 6.3: Rubin-Neyman Causal Model

Much of the causal inference literature assumes that the causal graph is known and true. In other words, we know apriori which variables fall into each of the categories, U, Z, V and O. In our case, only X and Y are specified and we have to infer which category each other variable (item) belongs to. Since this inference relies on association (dependence) rather than causation, the discovered graph may have errors resulting in misclassifications of variables into the wrong category. For example, because of the marginal dependence between U and Y, variables in U can easily get misclassified as Z. Such misclassifications do not necessarily lead to biases, but they can cause loss of efficiency.

Problem Formulation. Given a data set D, a set $\mathcal{S}$ of **subpopulation-defining** items, a set $\mathcal{X}$ of **intervention** items, a minimal support threshold θ and a minimum differential causal effect threshold η, we wish to find all subpopulations S ($S \subset \mathcal{S}$) and all intervetions X ($X \subset \mathcal{X}$), X and S are disjoint, such that the causal rule $XY|_S$ is frequent, namely it has support $((XYS) > \theta)$ and its intervention set X is η-closed.

Causation has received substantial research interest in many areas. In computer science, Pearl [73] and Rosenbaum[74] laid the foundation for causal inference. However, several fields such as cognitive science, econometrics, epidemiology, philosophy and statistics have built their respective methodologies [75, 76, 77].

At the center of causation is a causal model. Arguably, one of the earliest and popular models is the Rubin-Neyman causal model [70]. Under this model X causes Y, if X occusr before Y; and without X, Y would be different. Beside the Rubin-Neyman model (counterfactual analysis), there are several other causal models, such as Structural Equation Modeling [76], graphical models (causal graphical models [78]), and the symbiosis between counterfactual and graphical models. In our work, we use the potential outcomes framework from the Rubin-Neyman model and we use causal graphical models to identify and correct for biases.

Causal graphical models are tools to visualize causal relationships among variables. Nodes of the causal graph are variables and edges are causal relationships. Most methods assume that the causal graph structure is a priori given, however, methods have been proposed for discovering the structure of the causal graph [79, 80]. In our work, the structure is partially given: we know the relationships among groups of variables, but we have to assign each variable to the correct group based on data.

Knowing the correct graph structure is important, because substructures in the graph are suggestive of sources of bias. To correct for biases, we are looking for specific substructures. For example, causal chains can be sources of overcorrection bias and "V"-shaped structures can be indicative of confounding or endogenous selection bias [77]. Many other interesting substructures have been studied [81, 82, 83]. In our work, we consider three fundamental such structures: direct causal effect, indirect causal effect and confounding. Of these, confounding is the most severe and has received the most research interest.

Numerous methods exist to handle confounding, which includes propensity score matching (PSM) [67], structural marginal models [77] and g-estimation [76]. The latter two being specifically used for time-varying interventions [77].

Propensity score matching is used to estimate the effect of an intervention on an outcome. The propensity score is the propensity (probability) of a patient receiving the intervention given his baseline characteristics and the propensity score is used to create a new population that is free of confounding. Many PSM techniques exist and they typically differ in how they use the propensity score to create this new population [84, 85].

Applications of causal modeling are not exclusive to social and life sciences. In data mining, Lambert et al. [86] investigated the causal effect of new features on click through rates and Chan et al. [87] used doubly robust estimation techniques to determine the efficacy of display advertisements. Causal Modeling techniques have also been widely utilized in EHRs [88] . In particular, Prunelli et al.[8] used PSM based techniques to estimate the effect of 3-hour bundle sepsis guidelines on patient mortality and associated complications.

Even extending association rules mining to causal rule mining has been attempted before [89, 90, 91]. Li et al. [89] used the odds ratio to identify causal patterns and later extended their technique [91] to handle large data set. Their technique, however, is not rooted in a causal model and hence offers no protection against computing systematically biased estimates. In their proposed causal decision trees [92], they used the potential outcomes framework, but still have not addressed correction for various biases, including confounding.

6.4 Methods

6.4.1 Pruning Metrics

We can now present our algorithm for causal pattern mining. At a very high level, the algorithm comprises of two nested frequent pattern enumeration [93] loops. The outer loop enumerates subpopulation-defining itemsets S using items in $\mathcal{S}$, while the inner loop enumerates intervention combinations using items in $\mathcal{X} \setminus \mathcal{S}$. More generally, $\mathcal{X}$ and $\mathcal{S}$ can overlap but we do not consider that in this chapter. Effective algorithms to this end exists [94, 95], we simply use Apriori [69].

Once the patterns are discovered, the Δ ATT of the interventions are computed, using one of the methods from Section 6.4.2 and the frequent, effective patterns are returned.

On the surface, this approach appears very expensive, however several novel, extremely effective pruning strategies are possible and we describe them below.

Potential Outcomes Support Pruning. Let X be an intervention k-itemset, S be a subpopulation-defining itemset, and let X and S be disjoint. Further, X_{-i} be an itemset that evaluates to 'true' iff all items except the ith item are 'true' but the ith item is 'false'. Using association rule mining terminology, all items in X except the ith item are present in the transaction.

[Potential Outcomes Support Pruning] We only need to consider itemsets X such that

$$\begin{aligned} \min\{(S, X), \quad (\{S, X_{-1}), \ldots, \\ (S, X_{-k})\} \quad > \theta. \end{aligned}$$

In order to be able to estimate the effect of $x \in X$ in the subpopulation S, we need to have observations with x 'true' and also with x 'false' in S.

Potential Outcome Support Pruning is anti-monotonic.

Proof: Consider a causal rule $XY|_S$. If the causal rule $XY|_S$ is infrequent, then

$$(XS) < \theta \quad \vee \quad \exists i, (X_{-i}S) < \theta.$$

If $(X_{-i}S)$ had insufficient support, then any extension of it with an intervention item x will continue to have insufficient support, thus the $XxY|_S$ rule will have insufficient support. Likewise, if (XS) had insufficient support, then any extension of it with an intervention item x will also have insufficient support.

Pruning based on Differential Causal Effect. Pruning based on differential causal effect eliminates an intervention set X such that $\forall x \in X, |ATT(xY|_{S,X\setminus x})| \leq \eta$. In other words, we eliminate the drug x from the cocktail that does not increase its causal effect by at least η. While this pruning is not anti-monotonic, it is effective and produces the most clinically relevant patterns. Consider interventions a,b and an intervention set X, it is possible that $\Delta ATT(aY|_X) < \eta$ and $\Delta ATT(bY|_{Xa}) > \eta$ in which case we incorrectly prune Xab. However, b is likely contributing much more to the effect of Xab than a, thus the drug Xb likely captures most of the beneficial effect of Xab.

6.4.2 Causal Estimation Methods

ΔATT, our metric of interest, with respect to a single intervention x in a subpopulation S is defined as

$$(xY|_S) = [Y_1 - Y_0]_{S,Xx=1},$$

which is the expected difference between the potential outcome under treatment Y_1 and the potential outcome without treatment Y_0 in patients with S who actually received treatment. Since we consider treated patients, the potential outcome Y_1 can be observed, the potential outcome Y_0 cannot. Thus at least one of the two must be estimated. The methods we present below differ in which potential outcome they estimate and how they estimate it.

For the discussion below, we consider the variables X, Z, U and V from the causal graph in Figure 6.3. X is a single intervention, U, V and Z can be sets of items. For regression models, we will denote the matrices defined by U, V and Z in the subpopulation S as U, V and Z (same letter as the variable sets).

Counterfactual Confidence (CC). This is the simplest method. We simply assume that the patients who receive intervention $X = 1$ and those who do not $X = 0$, do not differ in any important respect that would influence Y. Under this assumption, Y_1 in the treated is simply the actual outcome in the treated and the potential outcome Y_0 is simply the actual outcome in the non-treated ($X = 0$). Thus

$$\begin{aligned} &= ((X=1)Y|_S) - ((X=0)Y|_S), \\ &= (Y|S, X=1) - (Y|S, X=0) \end{aligned}$$

In the followings, to improve readability, we drop the S subscript. All evaluations take place in the S subpopulations.

Direct Adjustment (DA). We cannot estimate Y_0 in the treated ($X = 1$) as the actual outcome Y in the untreated, because the treated and untreated populations can significantly differ in variables such as Z and V that influence Y. In Direct Adjustment, we attempt to directly remove the effect of V and Z by including them in a regression model. Since a regression model relates the means of the predictors with the mean of the outcome, we can remove the effect of V and Z by making their means 0.

Let R be a generalized linear regression model, predicting Y via a link function g

$$g(Y|V, Z, X) = \beta_0 + \beta_V V + \beta_Z Z + \beta_X X.$$

Then the (link-transformed) potential outcome under treatment is $g(Y_1) = \beta_0 + \beta_V V + \beta_Z Z + \beta_X$ and the potential outcome without treatment is $g(Y_0) = \beta_0 + \beta_V V + \beta_Z Z$. The ATT is then

$$\begin{aligned} = \; & \left[g^{-1}(Y_1|V, Z, X=1)\right]_{X=1} - \\ & \quad \left[g^{-1}(Y_0|V, Z, X=0)\right]_{X=1}. \end{aligned}$$

where $g^{-1}(Y_1|V, Z, X = 1)$ is prediction for an observation with the observed V and Z but with X set to 1. The $(\cdot)_{X=1}$ notation signifies that these expectation of the predictions are taken only over patients who actually received the treatment.

The advantage of DA (over CC) is manyfold. First, it can adjust for Z and V as long the model specification is correct, namely the interaction terms that may exist among Z and V are specified correctly. Second, we get correct estimates even if we ignore U, because U is conditionally independent of Y given X. This unfortunately only is a theoretical advantage, because we have to infer from the data whether a variable is a predictor of Y and U is marginally dependent on Y, so we will likely adjust for U, even if we don't need to.

Counterfactual Model (CM). In this technique, we build an explicit model for the potential outcome without treatment Y_0 using patients with $X = 0$. Specifically, we build a model g(Y—V,Z,X=0)=$\beta_0 + \beta_V V + \beta_Z Z$. and estimate the potential outcome as g(Y$_0|V,Z) = g(Y|V,Z,X=0)$. The differential causal effect is then = (Y—X=1) - $\left[g^{-1}(Y_0|V,Z)\right]_{X=1}$.

Similarly to Direct Adjustment, the Counterfactual Model does not depend on U. However, in case of the Counterfactual Model, we are only considering the population with $X = 0$. In this population, U and Y are independent, thus we will not include U variables into the model.

Propensity Score Matching (PSM). The central idea of Propensity Score Matching is to create a new population, such that patients in this new population are comparable in all relevant respects and thus the expectation of the potential outcome in the untreated equals the expectation of the actual outcome in the untreated.

Patients are matched based on their propensity of receiving treatment. This propensity is computed as a logistic regression model with treatment as the dependent variable

$$\log \text{odd}(X) = \beta_0 + \beta_V V + \beta_Z Z.$$

Patient pairs are formed, such that in each pair, one patient received treatment and the other did not and their propensities for treatment differ by no more than a user-defined caliper difference ρ.

The matched population has an equal number of treated and untreated patients, is balanced on V and Z, and thus the patients are comparable in terms of their baseline risk of Y. Hopefully, the only factor causing a difference in outcome is the treatment.

For estimating , the potential outcome without treatment is estimated from the actual outcomes of the patients in the matched population who did not receive treatment:

$$\begin{aligned} &= [Y_1 - Y_0] \\ &= (Y|X=1,M) - (Y|X=0,M), \end{aligned}$$

where M denotes the matched population.

Among the methods we consider, propensity score matching most strictly enforces the patient comparability criterion, however, it is susceptible to misspecification of the propensity regression model, which can erode the quality of the matching.

Stratified Non-Parametric (SN). In the stratified estimation, we directly compute the expectation via stratification. The assumption is that the patients in each stratum are comparable in all relevant respects and only differ in the presence or absence of intervention. In each stratum, we can estimate the potential outcome Y_0 in the treated as the actual outcome Y in the untreated.

$$\begin{aligned} &= [Y_1 - Y_0]_{X=1} \\ &= \sum_l P(l|X=1)\left[P(Y_1|l, X=1) - P(Y_0|l, X=1)\right] \\ &= \sum_l P(l|X=1)\left[P(Y|l, X=1) - P(Y|l, X=0)\right], \end{aligned}$$

where l iterates over the combined levels of V and Z. If we can identify the items that fall into U, then we can ignore them, otherwise, we should include them as well into the stratification.

The stratified method makes very few assumptions and should arrive at the correct estimate as long as each of the strata are sufficiently large. The key disadvantage of the stratified method lies in stratification itself: when the number of items across which we need to stratify is too large, we may end up dividing the population into excessively many small subpopulations (strata) and become unable to estimate the causal effect in many of them thus introducing bias into the estimate.

6.5 Data and Study Design

In this study we utilized a large cohort of Mayo Clinic patients with data between 1999 and 2010. We included all adult patients (69,747) with research consent. The baseline of our study was set at Jan. 1, 2005. We collected lab results, medications, vital signs and status, and medication orders during a 6-year *retrospective period* between 1999 and the baseline to ascertain the patient's baseline comorbidities. From this cohort, we excluded all patients with a diagnosis of diabetes before the baseline (478 patients), missing fasting plasma glucose measurements (14,559 patients), patients whose lipid health could not be determined (1,023 patients) and patients with unknown hypertension status (498 patients). Our final study cohort consists of 52,139 patients who were followed until the summer of 2010.

Patients were phenotyped during the retrospective period. Comorbidities of interest include Impaired Fasting Glucose (IFG), abdominal obesity, Hypertension (HTN; high blood pressure) and hyperlipidemia (HLP; high cholesterol). For each comorbidity, the phenotyping algorithm classified patients into three broad levels of severity: normal, mild and severe. Normal patients show no sign of disease; mild patients are either untreated and out of control or are controlled using first-line therapy; severe patients require more aggressive therapy. IFG is categorized into normal and pre-diabetic, the latter indicating impaired fasting plasma glucose levels but not meeting the diabetes criteria yet. For this study, progression to T2DM within 5 years from baseline (i.e. Jan 1, 2005) was chosen as our outcome of interest. Out of 52,139 patients 3627 patients progressed to T2DM , 41028 patients did not progress to T2DM and the remaining patients (7484) dropped out of the study.

6.6 Results

In this section, we present three evaluations of the proposed methodology. The first evaluation demonstrates the computational efficiency of our pruning methodologies, isolating the effect of each pruning methods: (i) Apriori support-based pruning, (ii) Potential Outcome Support Pruning, and (iii) Potential Outcome Support Pruning in conjunction with Effective Causal Rule Pruning. In the second section, we provide a qualitative evaluation, looking at patterns involving statin. We attempt to use the extracted patterns to explain the controversial findings that exist in the literature regarding the effect of statin on diabetes. Finally, in order to compare the treatment effect estimates to a ground truth, which does not exits for real drugs, we simulate a data set using proportions we derived from the Mayo Clinic data set.

6.6.1 Pruning Efficiency

In our work, we proposed two new pruning methods. First, we have the Potential Outcome Support Pruning, which aims to eliminate patterns for which the ATT is not estimable. Second, we have the Effective Causal Rule Pruning, where we eliminate patterns that do not improve treatment effectiveness relative to the sub-itemsets.

In Figure 6.4 we present the number of patterns discovered using (i) the traditional Apriori support based pruning, (ii) our proposed Potential Outcome Support Pruning (POSP), and (iii) POSP in conjunction with Effective Causal Rule Pruning (ECRP). In Figure 6.4, the x-axis represents the minimum support threshold parameter(θ) and the y-axis represents the number of patterns generated.

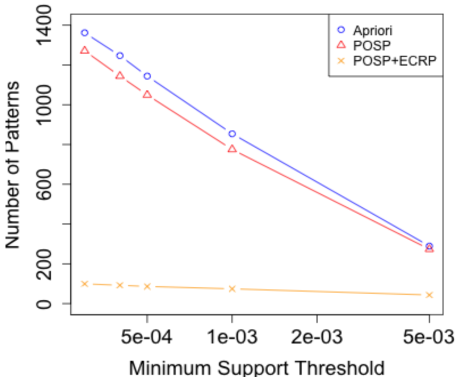


Figure 6.4: Comparison of Pruning Techniques

Apiori support based pruning only eliminates patterns based on support threshold. POSP eliminates patterns for which the differential causal effect is not estimable due to low sample size. POSP in conjunction with ECRP eliminates patterns for which the differential causal effect is not estimable or the estimated is below the chosen threshold. In other words, it eliminates patterns where the treatment is simply ineffective and thus has low clinical relevance.

6.6.2 Statin

In this section, we demonstrate that the proposed causal rule mining methodology can be used to discover non-trivial patterns from the above diabetes data set.

In recent years, the use of statins, a class of cholesterol-lowering agents, have been prescribed increasingly. High cholesterol (hyperlipidemia) is linked to cardio-vascular mortality and the efficacy of statins in reducing cardio-vascular mortality is well documented. However, as evidenced by a 2013 BMJ editorial [71] devoted to this topic, statins are surrounded in controversy. In patients with normal blood sugar levels (labeled as NormalFG), statins have a detrimental effect, they increase the risk of diabetes;

yet in pre-diabetic patients (PreDM), it appears to have no effect. What we demonstrate below is that this phenomenon is simply disease heterogeneity.

First, we describe how this problem maps to the causal rule mining problem. Our set of interventions ($\mathcal{X}$) consists of statin and our subpopulation defining variables consist of the various levels of HTN, HLP and IFG ($\mathcal{S}$). Our interest is the effect of statin (x) on T2DM (Y) in all possible subpopulations S, $S \subset \mathcal{S}$.

In this setup, HLD, which is associated with statin use (statins treat HLD) and T2DM, is a confounder (Z). A cholesterol drug, other than statin, (say) fibrates, are in the U category: they are predictive of statin (patients on monotherapy who take fibrates do not take statins), but have no effect on Y, because its effect is already incorporated into the hyperlipidemia severity variables that defined the subpopulation. Variables that only influence diabetes but not statin use (say HTN) would fall into the V category. All subpopulations have variables that fall into Z and U and some subpopulation may also have V.

The HLP variable uses statin as part of its definition, thus we constructed two new variables. The first one is HLP1, a variable at the borderline between HLP-Normal and HLP-Mild, consisting of untreated patients with mildly abnormal lab results (these fall into HLP-Normal) and patients who are diagnosed and receive a first-line treatment (they fall into HLP-Mild). Comparability is the central concept of estimating causal effects and these patients are comparable at baseline. Similarly, we also created another variable, HLP2, which is at the border of HLP-Mild and HLP-Severe, again consisting of patients who are comparable in relevant aspects of their health at baseline.

S	CC	DA	CM	PSM	SN
PreDM	0.145	0.022	0.010	0.022	0.017
NormFG	0.060	0.023	0.034	0.017	0.029
HLP1	0.078	0.019	0.014	0.010	0.010
HLP2	0.021	-0.013	-0.010	-0.021	-0.015
PreDM,HLP1	0.067	0.018	0.021	0.004	0.002
PreDM,HLP2	0.001	-0.038	-0.031	-0.048	-0.043
NormFG,HLP1	0.043	0.020	0.015	0.014	0.013
NormFG,HLP2	0.017	-0.002	-0.002	-0.005	-0.004

Table 6.1: ATT due to statin in various subpopulations S as estimated by the 5 proposed methods.

Table 6.1 presents the ATT estimates obtained by the various methods proposed in Section 3.4 for some of the most relevant subpopulations. In this case, the intervention set X consists of a single intervention, statin, thus reduces to the usual definition of ATT. Negative ATT indicates beneficial effect and positive ATT indicates detrimental effect.

Counterfactual confidence (CC) estimates statin to be detrimental in all subpopulations. While statins are known to have detrimental effect in patients with normal glucose levels [71], it is unlikely that statins are universally detrimental, even in patients with severe hyperlipidemia, the very disease it is supposed to treat.

The results between DA, CM, PSM and SN are similar, with PSM and SN having larger effect sizes in general. The picture that emerges from these results is that patients with severe hyperlipidemia appear to benefit from statin treatment even in terms of their diabetes outcomes, while statin treatment is moderately detrimental for patients with mild hyperlipidemia.

Bootstrap estimation was used to compute the statistical significance of these results. For brevity, we report the results only for PSM. The estimates are significant in the following subpopulations: NormFG, PreDM+HLP2 (p-values are ¡.001) and NormFG+HLP1 (p-value .05).

The true ATT in these subpopulations is not know. To investigate the accuracy that the various methods achieve, we use simulated data set that is largely based on this example [71, 96].

6.6.3 Synthetic Data

In this section, we describe four experiments utilizing synthetic data sets, each of which introduces a new potential source of bias. Our objective is to illustrate the ability of the five methods from Section 6 for adjusting for these biases. We compare their ATT estimates to the true ATT we used to generate the data set and discuss reasons for their success or failure.

The rows of Table 6.2 correspond to the synthetic data sets in increasing order of the biases we introduced and the columns corresponds to the methods: Conf (confidence), CC (Counterfactual Confidence), DA (Direct Adjustment), CM (Counterfactual Model), PSM (Propensity Score Matching) and SNP (Stratified Non-Parametric).

Some of these methods, DA, CM, PSM and SNP take the causal graph structure into account while estimating ATT. Specifically, they require the information whether a variable is a confounder (Z), has a direct effect (V), an indirect effect (V), or no effect (O).

In all of the data sets, we use a notation consistent with Figure 1: Z is the central disease with outcome Y; X is the intervention of interest that treats Z; V is another disease with direct causal effect on Y, but V is independent of X; and U is a third disease, which can be treated with X, but has no impact on Y. All data sets contain 5000 observations. When X is a single intervention, we use ATT and interchangeably.

I. Direct Causal Effect from V. We assume that every patient in the cohort has disease Z at the same severity. They are all comparable w.r.t. Z. 30% of the patients are subject to the intervention X aimed at treating Z, while others are not. Untreated patients face a 25% chance of having Y, while treated patients only have 10% chance. Some patients, 20% of the population, also have disease V, which directly affects Y: it increases the probability of Y by 5%.

In this example the true ATT is -.15, as X reduces the chance of Y by 15%. Our causal graph dictates that X and V be marginally independent, hence this this effect is homogeneous across the levels of V. (Otherwise V would become predictive of X and it would become a confounder. Confounding is discussed in experiments III-V.) All methods estimated the ATT correctly, because ATT does not depend on V. We can demonstrate this by stratifying on V and using the marginal independence of X and V.

$$\begin{aligned}
&= [(Y|X=1) - (Y|X=0)] \\
&= \sum_{v \in V} (V=v)\,[(Y|V=v, X=1) - (Y|V=v, X=0)] \\
&= \sum_{v \in V} [(Y, V=v|X=1) - (Y, V=v|X=0)] \\
&= (Y|X=1) - (Y|X=0)
\end{aligned}$$

where v denotes the levels of V.

II. Indirect Causal Effect. The setup for this experiment is the same as for the 'Direct Causal Effect' experiment, except we have disease U instead of V. Just like Z, disease U is also treated by X, but U has no direct effect on Y; its effect is indirect through X. U is thus independent of Y given X. The true ATT continues to be -.15. Again, the ATT does not depend on U, hence all methods estimated it correctly. To demonstrate that ATT does not depend on U, we use stratification and the conditional independence of Y and U.

$$\begin{aligned}
&= [(Y|X=1) - (Y|X=0)] \\
&= \sum_{u \in U} [(Y|U=u, X=1)(U=u|X=1) \\
&\qquad -(Y|U=u, X=0)(U=u|X=0)]
\end{aligned}$$

$$\begin{aligned}
&= \sum_{u \in U} [(Y|X=1)(U=u|X=1) \\
&\qquad -(Y|X=0)(U=u|X=0)] \\
&= (Y|X=1)\sum_{u}(U=u|X=1) - \\
&\qquad (Y|X=0)\sum_{u}(U=u|X=0) \\
&= (Y|X=1) - (Y|X=0)
\end{aligned}$$

III. Confounding. In this experiment, we consider the simplest case of confounding, involving a single disease Z, a single treatment X and outcome Y. 20% of the patients have disease Z and 95% of the diseased patients are treated with X, while 5% are not. All treated patients have Z. 25% of the untreated patients ($Z = 1$ and $X = 0$) have outcome Y; 10% of the treated patients ($Z = 1$ and $X = 1$) have the outcome; and only 5% of the healthy patients ($Z = 0$) have it. The true ATT is -.15.
In the presence of confounding, the counterfactual confidence and ATT do not coincide. With z denoting the levels of Z and (z) being a shorthand for $(Z = z)$,

$$\begin{aligned}
&= [(Y|X=1) - (Y|X=0)] \\
&= \sum_{z}(z)\left[(Y|X=1,z) - (Y|X=0,z)\right],
\end{aligned}$$

while the counterfactual confidence (CC) is

$$CC \quad = \quad (Y|X=1) - (Y|X=0)$$

$$\begin{aligned} = & \sum_z [(Y|X=1,z)(z|X=1) \\ & \quad -(Y|X=0,z)(z|X=0)] \,. \end{aligned}$$

When $(z|X) \neq (z)$, these quantities do not coincide. However, any method that can estimate $(Y|X,Z)$ for all levels of Z and X will arrive at the correct ATT estimate. We used logistic regression in our implementation of the Direct Adjustment method, which can estimate $(Y|X,Z)$ when X and Z have no interactions. Note that the causal graph admits interaction between X and Z, thus model misspecification can cause biases in the estimate.

IV. Confounding with Indirect Effect. In addition to the Confounding experiment, we also have an indirect causal effect from U. We now have two diseases, Z and U, each of which can be treated with X. 20% of the population has Z and independently, 20% has U. 25% of the patients who have Z and have no treatment ($X = 0$) have Y, while only 10% of the treated ($X = 1$) patients have it, regardless of whether the patient has U. (If the probability of Y was affected by U, it would be another confounder, rather than have an indirect effect.)
X has a beneficial ATT of -.15 in patients with $Z == 1$ (and $X == 1$) and has no effect in patients with $Z = 0$ (who get X because of U). Thus the true ATT=-.0833.
In this experiment, the counterfactual model was the best-performing model. The counterfactual model estimates the ATT through the definition

$$= [(Y_1|X=1) - (Y_0|X=1)] \,,$$

where Y_0 is the potential outcome the patient would have without treatment $X = 0$ and $(Y_0|X = 1)$ is the counterfactual probability of Y (the probability of Y had they not received X) in the population who actually got $X = 1$. Note that the potential outcome $Y_1|X = 1$ in the patients who actually got $X = 1$ is the observed outcome $Y|X = 1$. With u and z denoting the levels of U and Z, respectively and (u) being a shorthand for $(U = u)$,

$$\begin{aligned} = & \ [(Y|X=1) - (Y_0|X=1)] \\ = & \sum_u \sum_z (u,z) [(Y|X=1,u,z) - (Y_0|X=1,u,z)] \\ = & \sum_z (z) \sum [(Y|X=1,z) - (Y_0|X=1,z)] \end{aligned}$$

$$= \sum_z (z) \sum [(Y|X=1,z) - (Y|X=0,z)] \, ,$$

which coincides with the data generation mechanism, hence the estimate is correct.

In the derivation, step 2 holds because U and Z are independent given X and step 3 uses the fact that the counterfactual model estimates $P_0(Y|X = 1, z, u)$ from the untreated patients, thus

$$(Y_0|X=1,z,u) = (Y|X=0,z,u) = (Y|X=0,z).$$

V. Confounding with Direct and Indirect Effects. In this experiment, we have three diseases: our index disease Z, which is a confounder; U having an indirect effect on Y via X; and V having a direct effect on Y. 20% of the population has each of Z, V and U independently. 95% of patients with Z or U get the intervention X. 25% of the untreated patients with Z get Y, while only 10% of the treated patients do, regardless of whether they have U. Patients with V face a 5% in their chance of experiencing outcome Y.
X has a beneficial ATT of -.15 in patients with $Z = 1$ and have no effect in patients with $Z = 0$ (who get X because of U). Whether a patient has V does not influence the effect of X. The true ATT is thus -.0833.
None of the methods estimated the effect correctly, but Propensity Score Matching came closest. Analytic derivation of why it performed well is outside the scope of this thesis, but in essence, its success is driven by its ability to maximally exploit the independence relationships encoded in the causal graph. It can ignore V when it constructs the propensity score model, because X and V are independent (when Y not given); and it can ignore U and V when it computes the ATT in the propensity matched population. On the other hand, the causal graph admits interaction among U, Z and X, thus a logistic regression model as the propensity score model can be subject to model misspecification.
The Stratified Non-Parametric method, which is essentially just a direct implementation of the definition of ATT, underestimated the ATT by almost 25%. The reason lies in the excessive stratification across all combinations of the levels of U, V, and Z. Even with just three variables, most strata did not have sufficiently many patients (either treated

or untreated) to estimate $(Y|X, u, v, z)$. In the discussion, we will describe remedies to overcome this problem.

	Conf	CC	DA	CM	PSM	SN
I.	+.110	-.150	-.150	-.150	NA	-.150
II.	+.099	-.150	-.150	-.150	-.151	-.149
III.	+.099	+.047	-.136	-.136	-.136	-.136
IV.	+.077	+.024	-.019	-.083	-.068	-.064
V.	+.072	+.038	-.037	-.105	-.074	-.067

Table 6.2: The ATT estimates by the 6 methods in the five experiments.

6.7 Discussion

In this section, we proposed the causal rule mining framework, which transitions pattern mining from finding patterns that are associated with an outcome towards patterns that cause changes in the outcome. Finding causal relationships instead of associations is absolutely critical in health care, but also has appeal beyond health care.

The numerous biases that arise in establishing causation make quantifying causal effects difficult. We use the Neyman-Rubin causal model to define causation and use the potential outcome framework to estimate the causal effects. We correct for three kinds of potential biases: those stemming from direct causal effect, indirect causal effect and confounding. We compared five different methods for estimating the causal effect, evaluated them on real and synthetic data and found that three of these methods gave very similar results.

We have demonstrated on real clinical data that our proposed method can effectively enumerate causal patterns in a large combinatorial search space due to the two new pruning methods we developed for this work. We also demonstrated that the patterns discovered from the data were very rich and we managed to illustrate how the effect of statin is different in various subpopulations. The results we found are consistent with the literature but go beyond what is already known about statin's effect on the risk of diabetes.

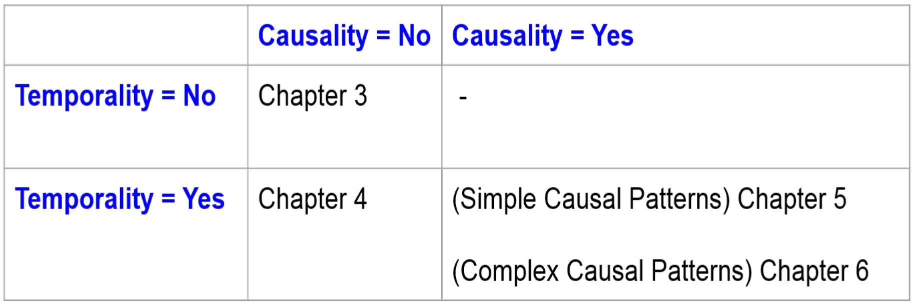
The discussions and experimental results provided in this chapter provide some general guidance on when to use the different methods we described. We recommend counterfactual confidence if no confounding is suspected as counterfactual confidence is computationally efficient and can arrive at the correct solution even when direct effects and indirect effects are present. In the presence of confounding, propensity score matching gave the most accurate results, but due to the need to create a matched population, it has built-in randomness, increasing its variance. Moreover, the counterfactual model as well as the propensity score model are susceptible to model misspecification. If unknown interactions among variables are suspected, we recommend the stratified non-parametric method. With this technique, model misspecification is virtually impossible, however, its sample size requirement is high. The stratified model is suboptimal if we need to stratify across many variables. Stratifying across many variables can fragment the population into many strata too small to afford us with the ability to estimate the effects correctly. If the estimates use some strata but not others, they may be biased.

Chapter 7

Conclusion and Future Work

In this thesis we aim to extract insightful information from Electronic Health Records (EHRs). In particular, we proposed techniques to model irregular time-series datasets from EHR. Further, we also proposed techniques which can be used to estimate the effect of medical and surgical interventions. To summarize, In Chapter 1 we present a brief overview of the existing state of healthcare in U.S. and how the current state is lacking behind other developed nations w.r.t several key statistics. We also provided a discussion on how the HITECH and ACA acts aim to improve the current state of U.S. healthcare. In chapter 2, we introduced EHRs and provided a brief discussion about the kind of data elements which are stored in an EHR. We then discussed the various challenges and opportunities associated with modeling EHRs.

In chapter 3, we introduced techniques which can be used to extract associative patterns from EHRs in the absence of longitudinal data. We discussed our methodologies within the context of T2DM. In chapter 4, we presented novel methodologies to extract associative patterns from EHRs in the presence of longitudinal data. Further, we also discussed how events which occurred close to the outcome are more predictive as compared to events which happened much earlier. In chapter 5, we presented techniques which can be used to extract simple causal patterns. Patterns in which we aim to estimate the effect of interventions on outcomes while only incorporating the effect of confounders. In chapter 6, we presented techniques which can be used to estimate complex causal patterns. Such patterns were extracted in the context of T2DM. In 7.1 we provide a brief summarization of the layout of our thesis.



	Causality = No	Causality = Yes
Temporality = No	Chapter 3	-
Temporality = Yes	Chapter 4	(Simple Causal Patterns) Chapter 5 (Complex Causal Patterns) Chapter 6

Figure 7.1: Thesis Discussion

7.1 Future Work

There are several directions in which this thesis can be extended in the future. In this thesis we used domain knowledge to identify variables whether they are confounding variables or exogenous variables. In future, more sophisticated techniques can be developed to automatically identify whether the variables are confounding, exogenous or independent. Such techniques can be developed using partial association tests and other similar statistical framework. Such automatic classification of variables will help us in more accurate estimation of the causal effect of medical and surgical interventions on various outcomes of interest.

Another interesting area for future research lies in the field of online causal estimation. Currently, we perform offline causal estimation i.e. any estimation of any medical intervention on outcomes of interest is performed using the entire datasets. This is not only time consuming but also a repetitive process as the whole analysis has to be computed when a new data set arises. By online causal estimation we imply techniques where in by using the new data to slightly tailor the existing model (computed using old data). Such techniques can then have various associated opportunities. Such techniques have the potential to create alarms when the new intervention estimates deviate from the old estimates.

We also hope that the techniques developed in this thesis for diseases such as T2DM and sepsis can also be applicable for other diseases. Further the techniques and results obtained on our data cohorts should be verified using data cohorts across geographies.

This would ensure the validity of our techniques as well as the validity of our models. Moreover, sophisticated techniques still need to be developed to handle challenges associated with modeling EHRs such as censoring, irregular time-series and missing data.

References

[1] Pranjul Yadav and Vinith Samala. Benchmarking over a semantic repository. In *Advanced Computing (ICoAC), 2010 Second International Conference on*, pages 51–59. IEEE, 2010.

[2] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records: A survey. *arXiv preprint arXiv:1702.03222*, 2017.

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

[4] Pranjul Yadav, Michael Steinbach, Lisiane Pruinelli, Bonnie Westra, Connie Delaney, Vipin Kumar, and Gyorgy Simon. Forensic style analysis with survival trajectories. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 1069–1074. IEEE, 2015.

[5] Pranjul Yadav, Kevin Viken, Vipin Kumar, Michael S Steinbach, and Maneesh Bhargava. Interrogation of bronchoalveolar lavage fluid in acute respiratory distress syndrome using multiplexed proseek immunoassay. In *C49. Respiratory Failure: Clinical And Translational Aspects Of Vili And Lung Protectice Mv*, pages A5246–A5246. Am Thoracic Soc, 2016.

[6] Bonnie L Westra, Sean Landman, Pranjul Yadav, Michael Steinbach, et al. Secondary analysis of an electronic surveillance system combined with multi-focal interventions for early detection of sepsis. *Applied Clinical Informatics*, 8(1):47–66, 2017.

[7] Pranjul Yadav, Lisiane Prunelli, Alexander Hoff, Michael Steinbach, Bonnie Westra, Vipin Kumar, and Gyorgy Simon. Causal inference in observational data. *arXiv preprint arXiv:1611.04660*, 2016.

[8] Lisiane Pruinelli, Pranjul Yadav, Andrew Hangsleben, Jakob Johnson, Sanjoy Dey, Maribet McCarty, Vipin Kumar, Connie W Delaney, Michael Steinbach, Bonnie L Westra, et al. A data mining approach to determine sepsis guideline impact on inpatient mortality and complications. *AMIA Summits on Translational Science Proceedings*, 2016:194, 2016.

[9] Pranjul Yadav, Michael Steinbach, M Regina Castro, Pedro J Caraballo, Vipin Kumar, and Gyorgy Simon. Frequent causal pattern mining: A computationally efficient framework for estimating bias-corrected effects.

[10] Centers for Disease Control, Prevention, et al. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, 201(1), 2011.

[11] Barbara A Ramlo-Halsted and Steven V Edelman. The natural history of type 2 diabetes: practical points to consider in developing prevention and treatment strategies. *Clinical Diabetes*, 18(2):80, 2000.

[12] Sanjoy Dey, Jeremy Weed, Joanna Fakhoury, Jacob Cooner, György J Simon, Michael Steinbach, Bonnie L Westra, and Vipin Kumar. Data mining to predict mobility outcomes for older adults receiving home health care. In *AMIA*, 2013.

[13] Tiinamaija Tuomi, Nicola Santoro, Sonia Caprio, Mengyin Cai, Jianping Weng, and Leif Groop. The many faces of diabetes: a disease with increasing heterogeneity. *The Lancet*, 383(9922):1084–1094, 2014.

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

[15] B Graeme Fincke, Jack A Clark, Mark Linzer, Avron Spiro III, Donald R Miller, Austin Lee, and Lewis E Kazis. Assessment of long-term complications due to type

2 diabetes using patient self-report: The diabetes complications index. *The Journal of ambulatory care management*, 28(3):262–273, 2005.

[16] Bessie Ann Young, Elizabeth Lin, Michael Von Korff, Greg Simon, Paul Ciechanowski, Evette J Ludman, Siobhan Everson-Stewart, Leslie Kinder, Malia Oliver, Edward J Boyko, et al. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *The American journal of managed care*, 14(1):15, 2008.

[17] Nicholas A Avitabile, Ajaz Banka, and Vivian A Fonseca. Glucose control and cardiovascular outcomes in individuals with diabetes mellitus: lessons learned from the megatrials. *Heart failure clinics*, 8(4):513–522, 2012.

[18] Per Kragh Andersen and Richard David Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

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

[20] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[21] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

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

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

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

[25] Kenji Ikeda, Hiromitsu Kumada, Satoshi Saitoh, Yasuji Arase, and Kazuaki Chayama. Effect of repeated transcatheter arterial embolization on the survival

time in patients with hepatocellular carcinoma. an analysis by the cox proportional hazard model. *Cancer*, 68(10):2150–2154, 1991.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[47] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[48] G Simon, P Caraballo, T Therneau, S Cha, M Castro, and P Li. Extending association rule summarization techniques to assess risk of diabetes mellitus. 2015.

[49] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[50] JA Anderson and A Senthilselvan. Smooth estimates for the hazard function. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 322–327, 1980.

[51] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model.* Springer Science & Business Media, 2000.

[52] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[53] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[54] Margaret Jean Hall, Sonja N Williams, Carol J DeFrances, and Aleksandr Golosinskiy. Inpatient care for septicemia or sepsis: a challenge for patients and hospitals. 2011.

[55] Celeste M Torio and Roxanne M Andrews. National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160, 2006.

[56] Keith A Wichterman, Arthur E Baue, and Irshad H Chaudry. Sepsis and septic shocka review of laboratory models and a proposal. *Journal of Surgical Research*, 29(2):189–201, 1980.

[57] Merete Storgaard, Jesper Hallas, Bente Gahrn-Hansen, Svend S Pedersen, Court Pedersen, and Annmarie T Lassen. Short-and long-term mortality in patients with community-acquired severe sepsis and septic shock. *Scandinavian journal of infectious diseases*, 45(8):577–583, 2013.

[58] R Phillip Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive care medicine*, 39(2):165–228, 2013.

[59] Roberta Capp, Cheryl Lynn Horton, Sukhjit S Takhar, Adit A Ginde, David A Peak, Richard Zane, and Keith A Marill. Predictors of patients who present to the emergency department with sepsis and progress to septic shock between 4 and 48 hours of emergency department arrival. *Critical care medicine*, 43(5):983–988, 2015.

[60] HM Nguyen, A Schiavoni, KD Scott, and MA Tanios. Implementation of sepsis management guideline in a community-based teaching hospital–can education be potentially beneficial for septic patients? *International journal of clinical practice*, 66(7):705–710, 2012.

[61] Matthew R Dettmer, Nicholas M Mohr, and Brian M Fuller. Sepsis-associated pulmonary complications in emergency department patients monitored with serial lactate: an observational cohort study. *Journal of critical care*, 30(6):1163–1168, 2015.

[62] Lemuel R Waitman, Judith J Warren, EL Manos, and Daniel W Connolly. Expressing observations from electronic medical record flowsheets in an i2b2 based

clinical data repository to support research and quality improvement. In *AMIA Annu Symp Proc*, volume 2011, pages 1454–1463, 2011.

[63] HHS Centers for Medicare & Medicaid Services (CMS) et al. Medicare and medicaid programs; electronic health record incentive program. final rule. *Federal register*, 75(144):44313, 2010.

[64] Earl Steinberg, Sheldon Greenfield, Dianne Miller Wolman, Michelle Mancher, Robin Graham, et al. *Clinical practice guidelines we can trust*. National Academies Press, 2011.

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

[66] Kei Miyata, Hirofumi Ohnishi, Kunihiko Maekawa, Takeshi Mikami, Yukinori Akiyama, Satoshi Iihoshi, Masahiko Wanibuchi, Nobuhiro Mikuni, Shuji Uemura, Katsutoshi Tanno, et al. Therapeutic temperature modulation in severe or moderate traumatic brain injury: a propensity score analysis of data from the nationwide japan neurotrauma data bank. *Journal of neurosurgery*, 124(2):527–537, 2016.

[67] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[68] Peter C Austin and Dylan S Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319, 2014.

[69] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[70] J. Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, pages 271–299, 2008.

[71] R. Huupponen and J. Viikari. Statins and the risk of developing diabetes. *BMJ*, 346, 2013.

[72] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[73] V. Didelez and I. Pigeot. Judea pearl: Causality: Models, reasoning, and inference. *Politische Vierteljahresschrift*, 42(2):313–315, 2001.

[74] P. Rosenbaum. *Observational studies.* Springer, 2002.

[75] D. Freedman. From association to causation via regression. *Advances in applied mathematics*, 18(1):59–110, 1997.

[76] W. Bielby and R. Hauser. Structural equation models. *Annual review of sociology*, pages 137–161, 1977.

[77] J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

[78] F. Elwert. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013.

[79] D. Heckerman. A bayesian approach to learning causal networks. In *UAI*, pages 285–295. Morgan Kaufmann Publishers Inc., 1995.

[80] D. Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[81] G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

[82] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2-3):163–192, 2000.

[83] S. Mani, P. Spirtes, and G. Cooper. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*, 2012.

[84] J. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[85] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[86] D. Lambert and D. Pregibon. More bang for their bucks: Assessing new features for online advertisers. In *Workshop on Data mining and audience intelligence for advertising*, pages 7–15. ACM, 2007.

[87] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *ACM-SIGKDD*, pages 7–16. ACM, 2010.

[88] Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6):1102–1112, 2011.

[89] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, and B. Sun. Mining causal association rules. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 114–123. IEEE, 2013.

[90] P. Holland and D. Thayer. Differential item performance and the mantel-haenszel procedure. *Test validity*, pages 129–145, 1988.

[91] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):14, 2015.

[92] J. Li, S. Ma, T. Le, L. Liu, and J. Liu. Causal decision trees. *arXiv preprint arXiv:1508.03812*, 2015.

[93] B. Goethals. Survey on frequent pattern mining. *Univ. of Helsinki*, 2003.

[94] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. Freespan: frequent pattern-projected sequential pattern mining. In *ACM-SIGKDD*, pages 355–359. ACM, 2000.

[95] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[96] M Regina Castro, Gyorgy Simon, Stephen S Cha, Barbara P Yawn, L Joseph Melton III, and Pedro J Caraballo. Statin use, diabetes incidence and overall mortality in normoglycemic and impaired fasting glucose patients. *Journal of General Internal Medicine*, pages 1–7.