Discovery of Type 2 Diabetes Trajectories from Electronic Health Records

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

Wonsuk Oh

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Gyorgy J Simon

September, 2020

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school.

I would like to express my sincere gratitude to my advisor, Dr. Gyorgy J. Simon, for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Besides my advisors, I would like to express my deepest appreciation to my committee members, Drs. Vipin Kumar, Erich Kummerfeld, and Sisi Ma, for agreeing to serve on my committee and sharing their valuable time reviewing my dissertation.

I am highly indebted to Pedro J. Caraballo, MD and M. Regina Castro, MD for their valuable time reviewing every manuscript carefully, giving constructive feedback, enthusiastically providing necessary clinical information, and their great support in completing this endeavor.

I would like to extend my sincere thanks to lab mates for their continuous support and friendship: Era K. Oh, Pranjul Yadav, Zhen Hu, Jia Li, Mengdie Wang, Xinpeng Shen, and Haoyu Yang.

I would like to acknowledge Mayo Clinic and Fairview Health Services for the providing the invaluable and necessary data to finish my dissertation.

I am indebted to all my roommates, friends I made during in Twin Cities. Special thanks to Kiwook Ha, Young-hwan Lee, Dongwoo Kim, Sangok Yoo, Chang Hyuk Kim, Minhae Cho, Hunkwan Park, Hyungjin Choi, Kwanhee Lee, Kwangsung Oh, Akinori Kitsuki, Jisu Lee, Sangyoon Lee, Woongsun Jeon, and Jihyeon Lee

Most importantly, I would like to express my sincere appreciation to my parents,

my wife, my son, my sister, parents-in-law, and brother-in-law. This dissertation would not have been accomplished without their support and love.

Dedication

To my mother, Misun Kim.

Abstract

Type 2 diabetes (T2D) is one of the fastest growing public health concerns in the United States. There were 30.3 million patients (9.4% of the US populations) suffering from diabetes in 2015. Diabetes, which is the seventh leading cause of death in the United States, is known to be a non-reversible (incurable) chronic disease, leading to severe complications, including chronic kidney disease, amputation, blindness, and various cardiac and vascular diseases. Early identification of patients at high risk is regarded as the most effective clinical tool to prevent or delay the development of diabetes, allowing patients to change their life style or to receive medication earlier. In turn, these interventions can help decrease the risk of diabetes by 30-60%.

Many studies have been conducted aiming at the early identification of patients at high risk in the clinical settings. These studies typically only consider the patient's current state at the time of the assessment and do not fully utilize all available information such as patient's medical history. Past history is important. It has been shown that laboratory results and vital signs can differ between diabetic and non-diabetic patients as many as 15-20 years before the onset of diabetes. We have also shown in our study that the order in which patients develop diabetes-related comorbidities is predictive of their diabetes risk even after adjusting for the severity of the comorbidities.

In this thesis, we develop multiple novel methods to discover T2D trajectories from Electronic Health Records (EHR). We define trajectory as an order of in which diseases developed. We aim to discover typical and atypical trajectories where typical trajectories represent predominant patterns of progressions and atypical trajectories refer to the rest of the trajectories. Revealing trajectories can allow us to divide patients into subpopulations that can uncover the underlying etiology of diabetes. More importantly, by assessing the risk correctly and by a better understanding of the heterogeneity of diabetes, we can provide better care.

Since data collected from EHR poses several challenges to directly identify trajectories from EHR data, we devise four specific studies to address the challenges: First, we propose a new knowledge-driven representation for clinical data mining, second, we

demonstrate a method for estimating the onset time of slow-onset diseases from intermittently observable laboratory results in the specific context of T2D, third, we present a method to infer trajectories, the sequence of comorbidities potentially leading up to a particular disease of interest, and finally, we propose a novel method to discover multiple trajectories from EHR data. The patterns we discovered from above four studies address a clinical issue, are clinically verifiable and are amenable to deployment in practice to improve the quality of individual patient care towards promoting public health in the United States.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Type 2 diabetes (T2D) is one of the fastest growing public health concerns in the United States[1]. There were 30.3 million patients (9.4% of the US populations) suffering from diabetes in 2015[2]. Diabetes, which is the seventh leading cause of death in the United States, is known to be a non-reversible (incurable) chronic disease[3, 4], leading to severe complications[5, 1], including chronic kidney disease, amputation, blindness, and various cardiac and vascular diseases. Early identification of patients at high risk is regarded as the most effective clinical tool to prevent or delay the development of diabetes through life style change or early pharmaceutical intervention. In turn, these interventions can help decrease the risk of diabetes by 30-60%[6, 7].

Many studies[8, 9] have been conducted aiming at the early identification of patients at high risk in the clinical settings. These studies typically only consider the patient's current state at the time of the assessment and do not fully utilize all available information such as patient's medical history. Past history is important. It has been shown that laboratory results and vital signs can differ between diabetic and non-diabetic patients as many as 15-20 years before the onset of diabetes[10]. We have also shown in our study that the order in which patients develop diabetes-related comorbidities is predictive of their diabetes risk even after adjusting for the severity of the comorbidities[11].

Understanding patients' disease progression over time has the potential to enable new levels of personalized intervention strategies, but it requires rich clinical data and novel machine learning methods.

The recent adaptation of electronic health records (EHR)[12, 13] allows us to access

large volumes of rich, longitudinal clinical data inexpensively. These data consist of information about patients' health, including diagnoses, medications, vital signs, and laboratory results, thus combining them into a model would result in a detailed model that can improve diagnosis, prognosis, and if interpretable, even our understanding of diseases and patient population[14, 15, 16, 11, 17]. However, the EHR is not designed and collected for research purposes, and, thus, the use of data, as it exists in the EHR, poses several challenges including unreliable diagnosis code[18], missing not at random[19, 20, 21], and inaccurate time stamps[11, 22]. Before we can develop methods for extracting disease progression patterns from EHR data, we have to address these challenges. In this thesis, we developed a methods for estimating and validating the onset time of diseases, and have developed novel data representations, particularly for clinical data mining purposes.

The focus of this thesis is on discovering T2D trajectories from EHR data. We define a trajectory as the natural order in which diseases develop. We aim to discover typical and atypical trajectories where typical trajectories represent predominant patterns of progressions and atypical trajectories refer to the rest of the trajectories. Knowledge about trajectories can help us divide patients into subpopulations, reveal the underlying etiology of diabetes, and more importantly, it can help us assess the risk of diabetes and its complication more correctly. In aggregate, this knowledge can lead to a better understanding of the heterogeneity of diabetes, and subsequently, to better and more individualized care.

Since the transferability of models and the reproducibility of findings[23, 24, 25] are the key concern of the machine learning-oriented studies, we evaluate and validate the proposed methods on cohorts from two large healthcare systems in the Upper Midwest United States.

Learning trajectories can be approached from two different perspectives. First, trajectories can be viewed as simply a sequences of events, and existing sequence learning methods[26] can be applied. Second, the diseases along a trajectory can also be viewed as causally linked events, where the development of a diseases causes progression to the next disease. This second perspective allows us to use causal structure discovery[27, 28] methods for trajectory mining. However, despite the recent success of these models in

medical researches[29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39], these methods are not designed to solve the problem of extracting sequences themselves from partially observable EHR data and are therefore unable to extract full sequences and may extract sequences with incorrect ordering caused by inaccurate time stamps. In this thesis, we develop a novel methodology that successfully addresses these issues.

1.1 Dissertation organization

The remainder of this thesis is organized into six chapters: In Chapter 2 we review recent literature regarding the key challenges in using data from EHRs for clinical research purpose, data representations and trajectories; in Chapter 3 we propose a new knowledge-driven representation for clinical data mining and study which characteristics make representations most suitable for particular clinical analytics tasks including trajectory mining; in Chapter 4 we demonstrate a method for estimating the onset time of slow-onset diseases from intermittently observable laboratory results in the specific context of T2D; in Chapter 5 we propose a new method for inferring T2D trajectories, defined as a sequence of comorbidities (i.e., hyperlipidemia, hypertension, and impaired fasting glucose), using EHR; in Chapter 6 we propose a new computational method for learning disease trajectories from EHR data that can help revealing some of the underlying mechanisms and their associated risk of developing diabetes; and finally in Chapter 8 we conclude the thesis by providing a summary of our work with contributions to Science section discusses the benefits of this particular research.

Chapter 2

Background

2.1 Electronic Health Records

The widespread adoption of Electronic Health Records (EHR)[12, 13] in recent years offers us the opportunity to unlock the potential of personalized medicine. The EHR allows us to inexpensively obtain rich clinical data consisting of laboratory results, diagnoses, medications, vital signs and social descriptors for large patient populations. These data elements all contribute information about patients' health, thus combining them into a model would result in a detailed model[14, 15, 16, 11, 17] that can take into account individual differences, and consequently, can improve diagnosis and prognosis compared to population-based models. However, since the EHR is not designed for the purpose of research, there are several challenges in discovering trajectories.

2.1.1 Challenges of using EHR data for the Discovery of Type 2 Diabetes Trajectories

The secondary use of EHR data poses numerous challenges. So many, in fact, that reviewing them comprehensively is outside the scope of this work. In this section, we focus on the challenges that are anticipated in our work of discovering trajectories.

First, missing data is prevalent in EHR, but most of them are informative missing data[19, 20, 21] (also called missing not at random). Informative missing data refers to the missingness whereby the probability of a missing value is dependent on the value

that is missing. As an example, a majority of missing laboratory test results imply the absence of clinical signs or symptoms of the target disease since laboratory tests won't be ordered without clinical justification.

Second, some essential information is stored in an unstructured format (if at all) hidden within narrative clinical notes[40, 41, 42]. For example, lifestyle interventions are the first line of defense in both the prevention and the management of diabetes, yet related information is unavailable in structured format in the EHR.

Third, we cannot determine the onset time of disease precisely[11, 42]. The onset time of disease is the earliest time when the patient meets the diagnostic criteria for the disease. Many analytical methods rely on the onset time of disease, but we can only access the recorded time of disease, the time when the disease is discovered and recorded into EHR. The failure of extracting accurate onset times and the use of the recorded times as an alternative is of key importance especially for trajectory mining because there can be a large gap between the onset time and the recorded time of disease.

Fourth, we can directly observe only part of the disease progressions since patients' medical history in EHR can be insufficient to cover the entire disease progressions and can be fragmented across multiple providers[43, 44]. As an example, the development of slow-onset conditions, such as T2D and its comorbidities can take decades. Even with 13 years of follow-up, we can only observe partial trajectories, that is, the development of only a few new conditions. Therefore, if we tried to observe full trajectories under this constraint, we would focus on patients with the fastest progression, possibly biasing the results. We will discuss how these challenges affect trajectory mining later in Section 2.4.

2.2 Type 2 Diabetes

2.2.1 Pathophysiology of diabetes

Insulin secretion and the regulation of glucose and lipid metabolism are the key concepts for understanding the pathogenesis of diabetes[45, 46]. Human brain, liver and skeletal muscle are the organs that can store glucose[45]. The brain is known to be insulin-independent glucose uptake organ i.e. the amount of uptake is not governed by the amount of inulin in the blood, while the liver and skeletal muscle are known to be

insulin-dependent glucose uptake organ. When the plasma glucose is concentrated on the blood, insulin is released from the beta cell of the pancreases to accelerate uptake of glucose and fatty acids into insulin sensitive tissue. This uptake will give negative feedback to the beta cell and will decrease secretion of the insulin. This feedback loop leads to the homeostasis of glucose concentration on the blood.

Diabetes[47, 48, 49] is characterized by loss of homeostasis of glycemic control by reduced insulin secretion or insulin action or both. Diabetes is generally categorized into two categories; Type 1 and 2. In Type 1 Diabetes, the malfunction of autoimmune system results in a lack of beta cells of the islets of Langerhans causing a loss of ability to secret insulin. In contrast, in Type 2 Diabetes, insulin is present in sufficient amounts, but insulin sensitive tissues such as skeletal muscle cells, adipose tissues and liver fail to react to insulin. Although when we refer to 'diabetes', we commonly refer to these two types, in fact, diabetes has various subtypes including maturity-onset diabetes in the young (MODY)[50] and latent autoimmune diabetes of adults (LADA)[51].

2.2.2 Prevalence and significance of Type 2 Diabetes

In the United States, Type 2 Diabetes affects 29.1 million people or 9.3% of the US population with an additional 8.1 million people undiagnosed in 2014[52] and it is growing compared with 2011 (25.8 million)[53]. The rate of growth is higher than expected. As of 2001, the prevalence of diabetes was expected to grow to 7.2% by 2050[54], but has already reached 9.2% in 2014. Newer studies expect the prevalence of diabetes to be 21% in 2050[55].

Diabetes can lead to various complications, including blindness, chronic kidney disease, kidney failure and various cardiac and vascular diseases[5, 1] and it is the seventh leading cause of death in the United States[1]. The total cost for US diabetic patients in 2012[56] is estimated $245 billion. Average estimated medical cost for each patient is estimated $13,700 per year, and diabetes takes into account $7,900 out of $13,700 medical expense.

2.3 Data representation

EHR systems[12] store information about entire populations and offer long follow-up times. Since EHR is not designed for supporting clinical analytics, data directly collected from EHR is not ready to be used in trajectory mining methods. A data representation[57, 58] is a transformation of data into a format amenable to the particular analytic technique. In this section, we will briefly review methods for data representation, and discuss their applicability to trajectory mining.

Representations can be categorized as outcome-specific or outcome-independent. Outcome-specific data representations are specific to a particular study end point (outcome) and are not applicable to different end points, while outcome-independent representations do not consider an outcome. For example, a diabetes risk score is an outcome-specific data representation which would not be used to estimate the risk of other diseases, cardiovascular disease as an example; but a comorbidity index is an outcome-independent data representation and it can be used for numerous outcomes (readmission, re-hospitalization, short-term mortality, etc).

2.3.1 Outcome-specific Representation—Severity Score

Disease severity score[59] (also disease severity scale[60] or disease severity index[61]) quantifies disease burden with respect to some outcome of interest. For example, the Framingham diabetes score[46] associates disease burden, defined by a handful of risk factors, with the risk of developing diabetes (an outcome). Traditionally, regression analysis is extensively used: the Framingham score is derived from a logistic regression model with 8-year diabetes status as the dependent variable and various diabetes risk factors as the independent variables. Severity score is a dimensionality-reducing representation, as it summarizes numerous original risk factors into a single number, which is proportional to the burden conferred by those risk factors on some outcome. Although the construction of severity scores through regression models requires an outcome, they can still be outcome-independent by using mortality as a generic outcome[62].

2.3.2 Outcome-independent Representations

Outcome-independent representations transform the original data into a new set of features, typically with a different dimensionality. Many currently existing representations, such as principal component analysis (PCA)[63] and nonnegative matrix factorization (NMF)[64, 65] have the specific aim of reducing the problem dimensionality. PCA is a statistical procedure that transforms a set of features into a new set of orthogonal features (called principal components), and NMF factorizes the original matrix into two matrices having only non-negative values, in a way that each subsequent component captures maximal amount of the residual information. Dimensionality reduction is achieved by using only the first few components.

Deep neural networks (DNNs)[66] are computational models that are inspired by neural networks in animal brains and have recently achieved considerable success. Much of this success is attributed to the data representation of these techniques, which is known as de-noising autoencoders (DAE)[67, 68]. DAEs consist of successive layers of transformations, where the outcome of each layer is the input to the next. Each layer is thought to extract higher-level features than the previous. The criterion for goodness of the transformation is the reconstruction error, which is a measure of how well an autoencoder can reconstruct the original data from its output. Autoencoders can perform dimensionality reduction or expansion. Requirements of Data representations for the Discovery of Type 2 Diabetes Trajectories

2.3.3 Requirements of Data representations for the Discovery of Type 2 Diabetes Trajectories

An ideal data representation for disease trajectory mining would have the following characteristics. First, data representation needs to show high performance on analytics based on binary or event data. This is because our goal is to discover trajectories where a trajectory is a sequence of "events" towards a certain outcome. Second, data representation needs to show high interpretability. Our long-term goal is to reveal T2D trajectories that can be practically applied in a clinical setting where its success depends on interpretability of data representation.

None of the above data representation meets both requirements. Outcome independent representations, such as PCA and NMF, tend to show lower performance and lower interpretability because the aim of the outcome independent representations is to minimize reconstruction error in general. These representations show lower performance than outcome-dependent representations with the same number of features. On the other hand, outcome-dependent representations, severity scores in particular, show high performance and good interpretability because the aim of the representations is to quantify the severity of target disease. This allows us to have high performance by minimizing outcome specific information loss and guaranteed interpretability since scores tell the severity of patient's condition directly. However, the outcome-dependent representations are not suitable to trajectory mining. Severity score, as an example of outcome-dependent representations, is a projection of high-dimensional space onto a linear subspace in accordance with the risk of an impending adverse event or mortality. Accordingly, different conditions from separate physiological mechanisms can be projected onto the same point. For this reason, the use of the outcome-dependent representations shows a limited ability to discover disease trajectory.

2.4 Trajectory mining

In this section, we focus on the methods for learning trajectories from data. The common methods are *sequential pattern mining* and *causal structure learning*. We will review these methods, and discuss how the unique characteristics of data from EHR can affect the applicability of these methods for the discovery of disease trajectories.

2.4.1 Sequential pattern mining

The goal of sequential pattern mining[26] is to discover the complete set of subsequences that appeared frequently in a database of sequences where a sequence is a set of ordered elements and a subsequence is an ordered subset of the elements in the sequence. As an example of disease progressions, elements are diseases and subsequences are partial trajectories. Thus, with sequential pattern mining, we can extract partial trajectories of disease progression. However, our interest is not to discover partial trajectories but to infer full trajectories from partial trajectories.

Since there can be an excessive number of sequential patterns, much effort is directed at reducing the number of patterns. First, some studies proposed methods for discovering so called closed sequential patterns. A sequential pattern s is closed if there is no supersequence with the same support as s. The closed sequential patterns can have the same information as non-closed sequential patterns with fewer patterns. Therefore, disease trajectories based on closed sequential patterns guarantee to have minimal subsets which can lead to improving interpretability as well. Second, some studies proposed methods for discovering sequential patterns with user-specified constraints. Minimum frequency of each pattern for sequential patterns and sliding window approach are the common approaches to define user-specified constraints. Domain knowledge is also used for deriving user-specific constraints frequently, such as 'diabetes is a non-reversible chronic disease so that patient's conditions are getting worse over time". These user-specified constraints allow us to reduce the search space and to discover only sequential patterns that are of the interest. For this reason, the user-specified constraints allow us to discover disease trajectories where each progression pair has particular clinical meaning.

Sequential pattern mining has been applied to reveal short sequences of events in clinical settings including readmission[38, 39] and adverse events[32, 35, 36], however, its application to trajectory mining is limited for the following reasons. First, we can only access the recorded time of the disease rather than the onset time of the disease. Sequential pattern mining relies on accurate timestamps to determine the order of events. However, some diseases can be discovered through the diagnosis of their complications, and the order of events can be invalid. Second, patient visits are intermittent so that we could observe multiple events occurring at the same time from EHR data. In reality, they could have developed at different time points but were merely discovered and recorded at the same time. This lack of ordering can also lead to the discovery of invalid trajectories or to the failure to discover some valid trajectories. Third, we cannot access patient's entire medical history in EHR. Even with 13 years follow-up history in our data, we cannot observe full trajectories. Therefore, we can only extract multiple partial and truncated trajectories from sequential pattern mining. Finally, sequential pattern mining does not have the concept of independent trajectories, and can hence capture multiple overlapping trajectories by chance.

Sequential pattern mining, as it exists today, requires overcoming the challenges in using EHR data for trajectory mining. Our method is based on sequential pattern mining but with appropriate adaptations. Structure learning

2.4.2 Structure learning

Another class of methods, that we could use for trajectory mining, is structure learning[27, 28]. The goal of structure learning is a little different from finding trajectories; it is a framework for learning the structure of graphs where graphs consist of a finite set of nodes and a set of directed edges. Most structure learning methods focus on revealing structure for causal inference. Our interest, however, is not to discover the causal structure but to discover precedence of associated events.

There are two major approaches to structure learning: constraint-based and score-based. Constraint-based structure learning is a set of processes that conducts conditionally independent tests to identify a set of edges based on the evaluation of faithfulness assumption, and to find the best directed acyclic graph (DAG) that satisfies the constraints. Score-based structure learning is a set of iterative processes that find a graph with maximal score. Some studies have shown the effectiveness of structure learning to solve various clinical tasks including monitoring[69], screening, diagnosis and prognosis[29, 30, 31, 33, 34, 37]. However, only a few studies so far have attempted to discover disease trajectories. In principle, we could discover causal chains and treat them as trajectories, however, this is overly restrictive as we are primarily interested in precedence: it is highly unlikely that comorbidities in diabetes cause each other; it is more likely that the comorbidities are clinical manifestations of a common underlying metabolic degradation. Also, the uncertainty in the timestamps and the observed sequence of events makes reconstruction of a causal graph challenging; our methodology, if successful, could provide the much needed precedence information that these structure learning methods require.

Our trajectories are not intended to be causal. We are primarily interested in precedence. In a way, our method, which will be explained in the later section, is complementary to the causal discovery methods: the (hopefully) robust precedence information extracted by our method makes structure learning possible (or easier).

Chapter 3

A new knowledge-driven representation for clinical data mining

3.1 Introduction

The widespread adoption of electronic health records (EHR)[12] enables new kinds of analytics such as explicitly modeling population heterogeneity or identifying benefit groups for an intervention[14, 15, 16, 70, 11, 71]. It is well understood that different analytics tasks and techniques operate optimally on different types of data[57, 58]. For example, association pattern mining requires binary or categorical data[72] and most regression models assume that the predictor variables have an additive effect[73]. Data, as it exists in the EHR, is not ideal for many analytics tasks.

A data representation is a transformation of data into a format amenable to a particular analytic technique. Data transformations are not new, e.g., log or rank transformations of non-normally distributed variables[74, 75] have been a mainstay for decades. The recent success of deep learning in some applications[17, 76, 77, 78] has put data representation into the spotlight and is, at least in part, attributed to the underlying data representation[67, 68]. There are other common data representations, such as dimensionality reduction[64, 65] through principal component analysis[63], and

other techniques, such as phenotyping[79, 80], which are not even commonly thought of as a data representations. In this work, we propose a data representation, which is specific to the clinical domain and represents data at a high level and enriches it with clinical knowledge.

Specifically, SEV augments the original data with a set of ordered or partially ordered binary variables, combining information about patients' state from multiple perspectives: therapies, diagnoses, and whether or not the laboratory results or vital signs are normal and/or achieve a typical therapeutic target. These variables are (at least partially) ordered: the variable 'patient is under control with first-line oral therapy', represents a lower severity than the variable 'patient is not under control despite last-line therapy'. These variables are highly interpretable, as they follow clinical reasoning and incorporate clinical knowledge.

To make the discussion concrete, we carry out our study in the context of type 2 diabetes (T2D). Diabetes is a common disease with severe complications[81], affecting 29.1 million Americans[52]. T2D can be prevented or delayed through lifestyle modifications and/or pharmacological treatment[6, 7], hence identifying patients at high risk is of high importance. From a technical perspective, T2D is an ideal evaluation platform, as it exhibits common challenges: T2D is heterogeneous; risk factors are correlated and not necessarily additive; and the time frame between the risk factors and the onset of diabetes can be as long as 20 years, which makes missing data inevitable[82, 44, 5, 11, 10].

We encode diabetes risk factors, hyperlipidemia, hypertension, and obesity as SEVs (a set of SEVs for each disease) and perform two clinical tasks related to type 2 diabetes. The first task is to predict the onset of diabetes using a Cox model and the second one is to model population heterogeneity in terms of the risk of T2D incidence using association pattern mining. We will compare SEV to five other data presentations, including the original data. The main objective is to study the characteristics of the data representations.

3.2 Severity Encoding Variables

Severity Encoding Variables (SEV) is our proposed outcome-independent representation. The purpose of SEV is to summarize the numerous facets of a disease into a single

hierarchical variable. Nodes at the same level in the hierarchy are fully or partially ordered.

The construction of the hierarchy replicates the clinical reasoning steps of determining the severity of a certain disease. Reasoning involves a sequence of questions: (i) are lab results and vital signs present and normal, (ii) has an intervention been initiated, and if it has, how aggressive is it (first-line treatment, combination therapy, etc.), and (iii) has a diagnosis been recorded. Accordingly, the first split (at the root) produces three nodes: patient with missing, normal, and abnormal lab results. Next, we reason about medications. Each of the three nodes can be split indicating whether treatment has been initiated and how aggressive those treatments are. The final question splits the nodes based on the presence of diagnoses.

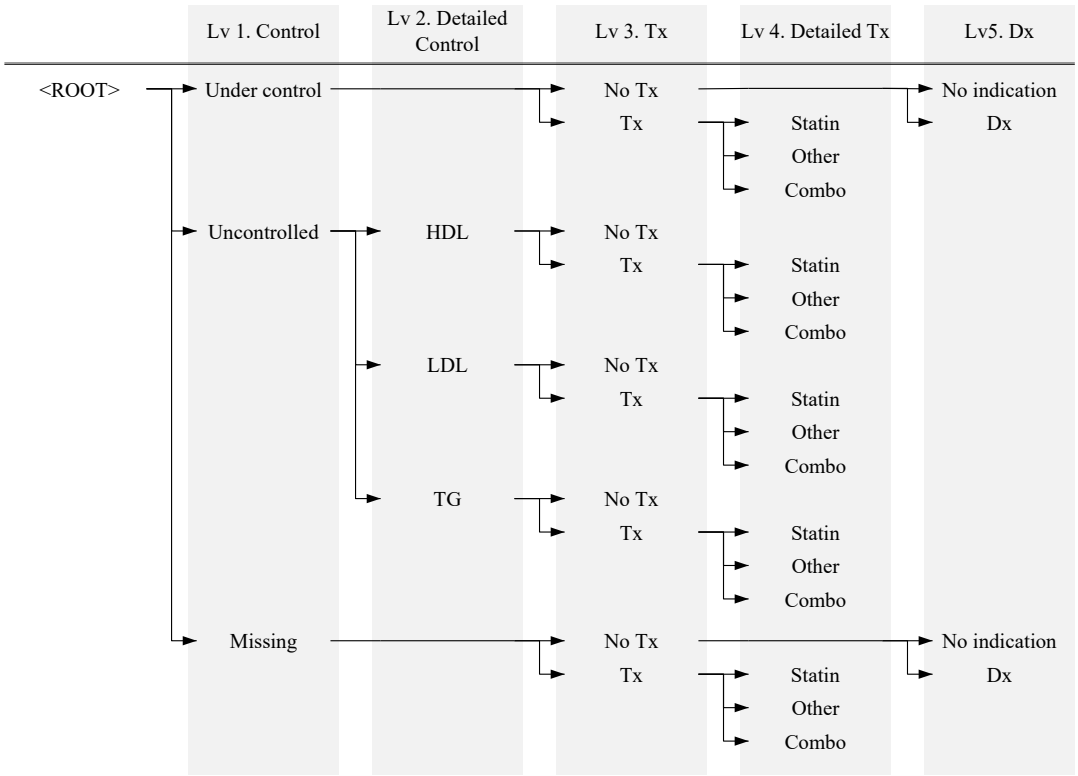


Figure 3.1: Sample Severity Encoding Variable hierarchy for hyper-lipidemia. Abbreviations used: Treatment (Tx), Diagnosis (Dx), High-density lipoprotein (HDL), Low-density lipopro-tein (LDL), Triglycerides (TG).

Figure 3.1 illustrates the SEV for hyperlipidemia. At the root of the hierarchy, we ask whether lab results (LDL, HDL and TG) are normal (if they are not missing) and which (if any) are abnormal. At the next level, we reason using medications. For example, does a patient under control use medications? If medications are used, are they first-line medications (statins in case of HL), other drugs, or combinations of drugs? On the last level, we reason using diagnoses. Naturally, diagnoses are most helpful if no other indication of disease exists.

For analysis, the hierarchy can be cut at any level and the nodes at that level are taken as binary variables. For example, cutting the hierarchy at the top-most level results in a set of three binary variables: 'patient is under control', 'patient is not under control', and 'laboratory results are missing'. These variables are partially ordered: being under control could (but does not have to) indicate lower severity than not being under control, but 'lab results are missing' is not comparable to the other two in terms of severity. Cutting the hierarchy at the (say) third level yields 10 leaves and incorporates information about medication use. One of these leaves would be 'patient has abnormal LDL despite medication'. By changing the level at which the hierarchy is cut, we can increase the number of leaves (and information content).

SEV is a framework for representing diseases as hierarchies induced by a sequence of clinical decisions; it is not a set algorithm for modeling all diseases. Recall that SEV is outcome independent; once a SEV is constructed, it can be used for multiple study end-points. The diseases that we build SEVs for are predictors of the outcome and the construction of the SEV can (and possibly should) depend on the disease that we build the SEV for. Depending on the disease in question, a different ordering of the same clinical questions could yield a more clinically meaningful hierarchy, and other diseases may incorporate altogether different questions (for example, stage and grade of cancer). We have not observed substantial changes in predictive performance in terms of the ordering of the questions.

Table 3.1 shows how SEV relates to other existing data representations. The table presents a categorization of existing and the proposed data representations along three axes: whether they are outcome-specific or outcome-independent; whether they are dimensionality reducing or expanding; and whether they are data-driven or knowledge-driven. All methods except SEV are data driven.

Table 3.1: Categorization of the data representations.

	Outcome-Specific	Outcome Independent	
		Data-Driven	Knowledge-Driven
Dimensionality-Reducing	SS	PCA, DAE-9	SEV
Dimensionality-Expanding		DAE-34	SEV

3.3 Materials and Methods

3.3.1 Data, Cohort Construction and Study Design

Mayo Clinic, located in Rochester, MN, provides primary care to a large population. Resources available at Mayo Clinic are described elsewhere[83]. After IRB approval, a cohort of 75,317 patients aged 18 or older on 01/01/2005 with research consent was constructed. The cohort was followed from the baseline of 01/01/2005 until the end of 2015. To determine patients' baseline status, we retrospectively collected diagnoses of obesity, hyperlipidemia, hypertension, and prediabetes; laboratory test results for lipid panels and fasting plasma glucose (FPG); vital signs (blood pressure, and body mass index [BMI]); demographic information (age, gender); and medications for hypertension and hyperlipidemia. From the cohort, we excluded patients with preexisting diabetes at or before baseline (11,897 patients) and suspicion of diabetes (3 patients with fasting plasma glucose $>$ 125 ml/dL and 2 patients taking anti-diabetic drugs), resulting in a final cohort of 63,415 patients. A description of the cohort is provided in Table 3.2.

3.3.2 Comparative representation

Severity Scores (SS): A severity score is computed for each diabetes risk factor (obesity, hypertension, hyperlipidemia, pre-diabetes) quantifying the risk factor's contribution to diabetes. While all features could be combined into a single severity score (analogously to the Framingham score), we compute a severity score for each risk factor, combining only the features that are related to the specific risk factor. Modeling the risk factors separately allows us to retain the relationships among them.

For each risk factor, the corresponding SS is the linear prediction from a Cox model,

Table 3.2: Study variables for demographics, prediabetic, obe-sity, hyperlipidemia and hypertension.

Risk Factor	No. (%) of sample
Demographic	
Age	41.8 ± 15.5
Sex: male	42.3 %
Prediabetic	
Impaired Fasting Glucose (IFG)	2.9 %
Fasting Plasma Glucose (FPG) level, mg/dL	94.8 ± 8.5
Missing FPG	34.5 %
Obesity	
Obesity	2.4 %
Body mass index (BMI)	28.0 ± 7.5
Missing BMI	34.7 %
Hyperlipidemia (HLD)	
HLD	24.5 %
High-Density Lipoprotein (HDL) level, mg/dL	54.9 ± 12.3
Low-Density Lipoprotein (LDL) level, mg/dL	112.0 ± 23.9
Triglycerides (TG) level, mg/dL	130.1 ± 54.0
Missing HLD related lab	39.6 %
Statin	39.6 %
Fibrate	0.7 %
Bile-acid Resins	0.1 %
Other HLD drugs	0.4 %
Hypertension (HTN)	
HTN	21.6 %
Systolic Blood Pressure (SBP), mm Hg	122.1 ± 15.8
Diastolic Blood Pressure (DBP), mm Hg	73.5 ± 9.7
Missing BP	15.1 %
ACE inhibitors	3.6 %
ARBs	1.3 %
Beta blockers	6.1 %
Calcium channel blocker	1.9 %
Diuretic	1.1 %
Antihypertensive drugs in peripheral vascular disease	0.0 %
Other HTN drugs	0.2 %

whose independent variables are the data elements that describe the risk factor in question and the dependent variable is diabetes outcome. Missing blood pressure measurements were imputed using mean imputation and a bias-correcting indicator variable

signaling whether imputation was performed for each patient was added.

Principal Component Analysis (PCA) In this study, logistic principal component analysis (PCA)[84] is applied to the risk factors, resulting in a single set of principal components. We kept the first 9 principal components because additional components are unable to explain significant amounts of variation. PCA is thus a dimensionality-reducing, outcome-independent representation.

Deep autoencoder (DAE): For this study, we used two configurations, tuned via cross-validation. Both used the hyperbolic tangent activation function, had two hidden layers with 20 nodes on the first layer and had 9 and 34 nodes on the second layer, respectively. The first configuration (DAE-9) has the lowest reconstruction error among configurations that reduce the dimensionality of the problem, while the 34-node configuration (DAE-34) has the lowest reconstruction error among all configurations. DAE-9 is a dimensionality-reducing representation, while DAE-34 is a dimensionality-expanding representation.

Severity Encoding Variables (SEV): A severity encoding was constructed for each of the four risk factors of diabetes independently. The hierarchy was cut at the leaf level, making it dimensionality-expanding (there are more nodes in the hierarchy than original features).

3.3.3 The two tasks

Regression Analysis: The objective is to measure the impact of the data representations on the predictive performance of estimating patients' 8-year risk of T2D. Risk factors (lab results, vital signs, diagnoses (ICD-9 billing code rolled up into categories), and prescriptions rolled up into NDF-RT pharmaceutical subclasses) are determined at baseline and are transformed into the five new representations. The sixth representation is RAW, the original (untransformed) data. Six Cox proportional hazard models are constructed using age, gender and each of the six data representations as independent variables. Backwards elimination is applied.

Association Pattern Analysis The central concept in association pattern mining is an *item*, which is a binary variable such as 'presence of hyperlipidemia diagnosis' or 'LDL $\geq$ 130 mg/dL'. Items are combined into conjunctive sets, called *itemsets* (e.g. 'LDL $\geq$ 130 mg/dL AND diagnosis of hyperlipidemia'). The association of an itemset

with the outcome is measured through *confidence*, which is the fraction of patients presenting with the outcome among patients who present with all conditions in the itemset (fraction of patients who developed diabetes among those with LDL $\geq$ 130 mg/dL and diagnosis of hyperlipidemia in our example). Association pattern mining systematically enumerates all itemsets and computes their confidence. In the Classification Based on the Association (CBA) framework[85], the risk of diabetes for a patient is the confidence of the highest-confidence rule that applies to that patient.

Continuous variables (age, severity scores, scores from PCA and DAE) are categorized into deciles (with backwards elimination discarding superfluous categories) and laboratory results and vital signs are dichotomized using the American Diabetes Association[86] and World Health Organization[87] cutoffs. Of interest are the number of patterns and their predictive performance. A data representation that can achieve higher predictive performance with a lower number of rules is preferable.

3.3.4 Evaluation Methodology

We used bootstrap estimation with 1,000 replications and paired t-tests were used to compare the models. The evaluation metric is concordance, which is the probability that, for a random pair of patients where one remained free of diabetes longer than the other, the estimated risk for the patient who remained diabetes-free longer is lower. Since all tasks were carried out using the same algorithms, the model's ability to predict diabetes directly relates to the representation's ability to retain diabetes-related information.

3.4 Results

3.4.1 Regression analysis

Figure 3.2 shows the concordance of the various data representations as box plots. The top, middle, and bottom line in each box correspond to the upper quartile, median, and the lower quartiles of the concordances estimated from the 1,000 bootstrap replications, respectively. The representations are ordered left to right by the number of features they produce.

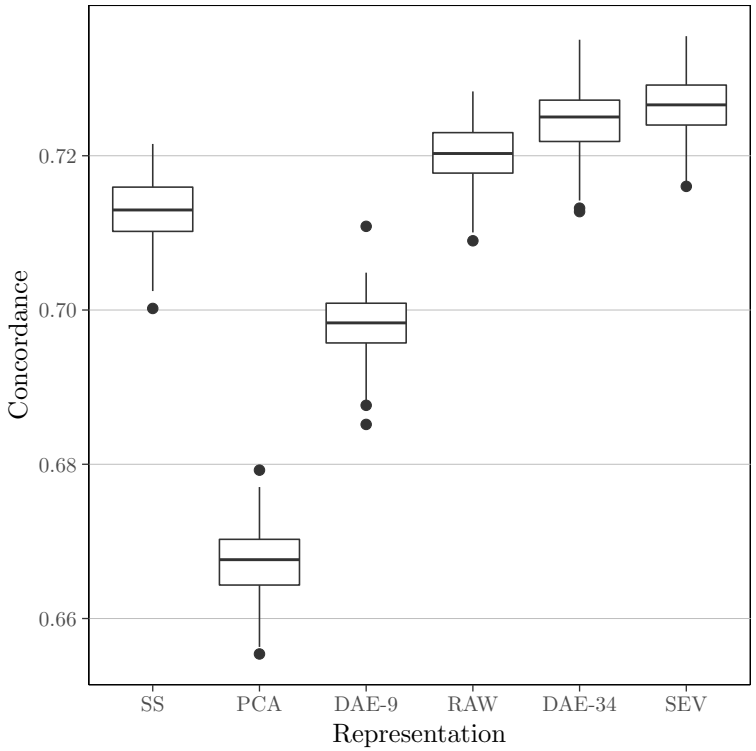


Figure 3.2: Performance comparison of data representations for the regression task.

While all performance differences are statistically significant, some are not substantial. Our population consists of relatively healthy patients, hence all methods achieved high discrimination. A more clinically meaningful question is to accurately estimate diabetes in risk patients who are at relatively high risk and may actually benefit from an intervention. To this end, we consider patients with Framingham score of at least 20 and in Figure 3.3, we present the predictive performance of the Cox model on the 6 data representations on these 2,493 patients.

3.4.2 Association analysis

Association rule mining can discover an exponentially large number of patterns, many of which can be coincidental. The parameter that controls the number of patterns is *Minimum Support in Cases (minsupC)*, the number of cases (patients who developed diabetes) to whom the pattern applies. Figures 3.4 and 3.5 display the concordance and

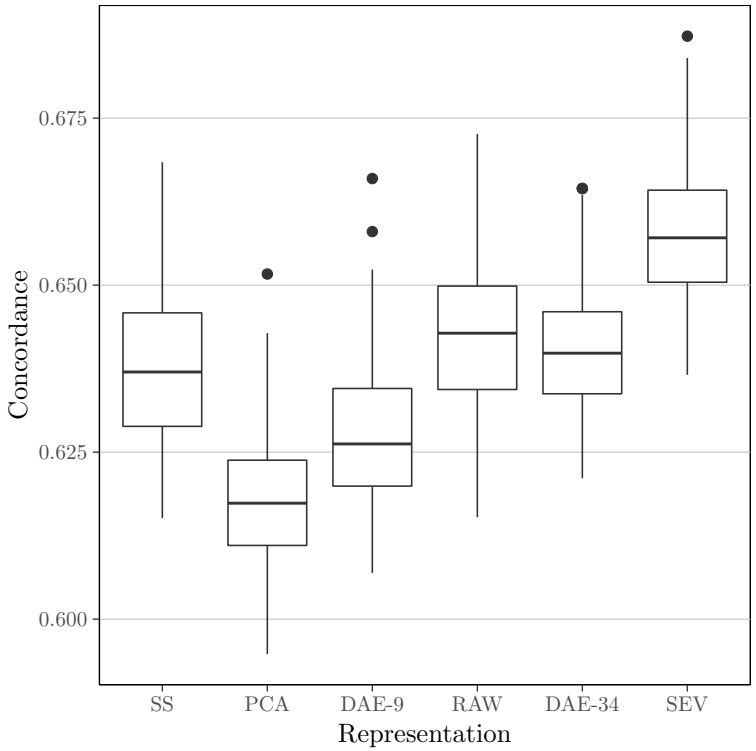


Figure 3.3: Comparison of concordance on subpopulation with Framingham score ≥ 20.

number of patterns discovered as a function of minsupC.

3.5 Discussion

As the paradigm for clinical studies continues to shift toward precision medicine, the range of tasks that clinical data analysis is used for will broaden. Since these newer tasks may operate optimally with different data representations, understanding existing and developing new data representations will become increasingly important. In this manuscript, we proposed a new data representation, Severity Encoding Variables, which represents diseases at a high level and is enriched with clinical knowledge. We compared SEV to five other existing data representations using two clinical tasks.

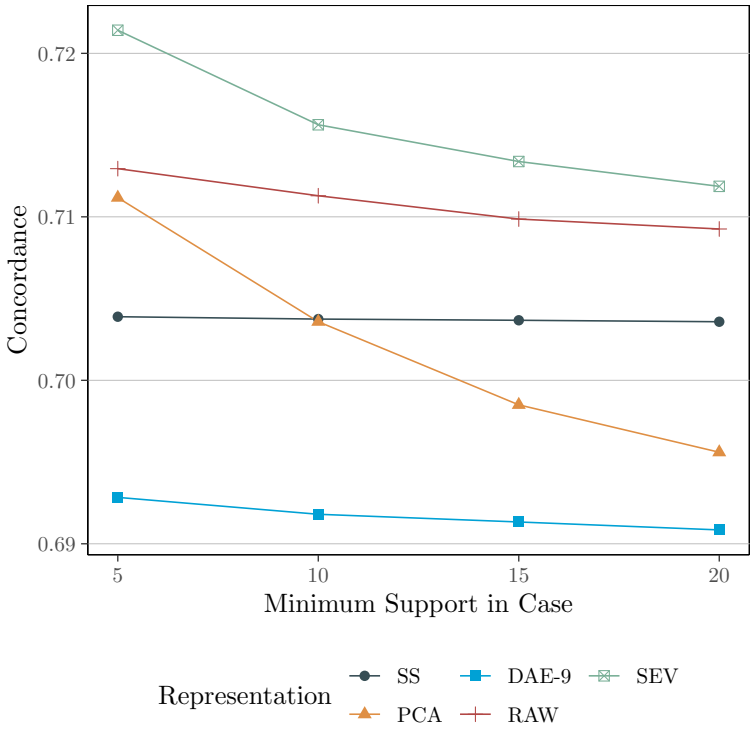


Figure 3.4: Comparison of the predictive performance of the association patterns discovered using the various data representations as a function of the minimum support in cases (minsupC).

3.5.1 Assessing the Risk of Incident Diabetes through Regression

The key concern in regression is information loss. The two dimensionality expanding methods, SEV and DAE-34, achieved the highest performance, as they can extract more information (e.g. SEV encodes some interactions and Deep Autoencoders can encode non-linearities). While the performance difference between these two methods in the entire population was minimal (although statistically significant), when we focused on the subpopulation with very high Framingham score (20 or higher), the performance gap widened substantially and SEV outperformed DAE by 20%. Given their high risk of developing diabetes, this is precisely the group of patients for which we need to estimate the risk accurately so that we can effectively target preventive measures to the patients most in need.

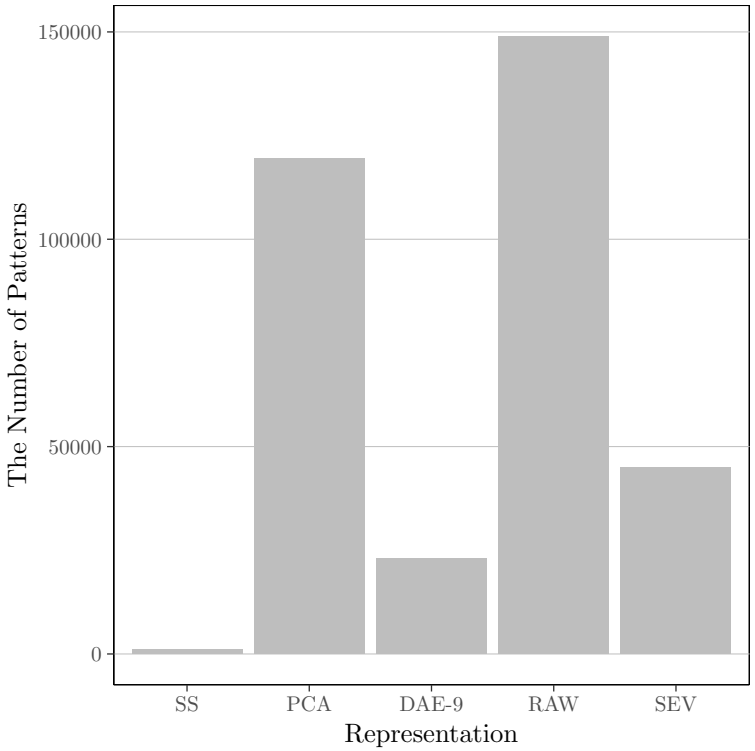


Figure 3.5: The number of association patterns discovered using the various data representations. (minsupC=5).

Mechanistically, SEV's performance advantage stems primarily from interactions. It can distinguish between patients who have similar lab results at baseline but are in very different states of severity: e.g. patients who are not yet pharmaceutically treated are very different from those who are already undergoing combination therapy at baseline. Despite having similar (abnormal) lab results, the latter patients are at a disproportionately higher risk and interaction among the various facets of the disease are required to model this correctly. Second, SEV can handle missing data without imputation, identifying that the presence of the diagnosis code is more important in patients who have no available lab results than in patients where the lab results already suggest the presence of the disease.

While interactions among various facets of a disease partly explain how SEV achieves high performance, selecting the right interactions is important. Some classification

methods, such as decision trees or association rules, are capable of automatically discovering interactions, however, as our experiment with association rules demonstrates, finding the right combination of interactions is non-trivial.

Dimensionality-reducing data representations did not perform well. Dimensionality reduction can reduce noise and can also lead to information loss. Given that our problem is "tall", the number of patients far exceeds the number of variables, dimensionality reduction led to information loss. Among the dimensionality-reducing methods, SS takes the diabetes outcome into account, and hence managed to preserve most of the outcome-related information, achieving a reasonable performance with the smallest number of features. PCA and DAE-9 are outcome-independent, and have suffered greater outcome-related information loss than SS despite having more features.

3.5.2 Modeling Patient Population Heterogeneity through Association Pattern Mining

On this task, SEV performed substantially (and statistically significantly) better than others. The association mining algorithm itself performs dimensionality expansion by forming combinations of the features the data representation provides. To find high-risk patients, we typically focus on patterns that occur in small patient groups, which can yield less reliable risk estimates and higher predisposition to overfitting (finding patterns that happen to randomly coincide with diabetes). Different data representations offer different mechanisms to reduce overfitting. The severity scores reduce the number of items an itemset can have. For example, for SS, there are only 5 axes (demographics, obesity, hypertension, hyperlipidemia and prediabetes), each of which is categorized into multiple bins. Since a patient cannot fall into two different bins along the same axis, the maximal number of conditions in a pattern is 5, which seriously limits the number of patterns. Some patterns have as many as 11 conditions in the RAW representation.

SEV, the data representation that achieved the highest performance on association pattern mining, applies a different mechanism. SEV uses the same dichotomization as RAW, but SEV combined these dichotomized variables into predefined "sub-patterns". For instance, the SEV item 'lipids under control' is a combination of three RAW items: LDL is normal AND HDL is normal AND TG is normal. These higher-level items constrain the space of possible patterns (based on clinical knowledge) and thus reduce

the tendency for overfitting.

3.5.3 Generalizability

We tested the data representations with a regression model and association pattern mining to highlight certain characteristics of the SEV representation. We believe that these results generalize to other classification methods, as well. First, the SEV representation offers a high-level clinical description of the diseases enhancing clinical interpretability of the models. Second, SEV can improve predictive performance by automatically handling missing lab results and by incorporating clinically meaningful high-order interactions. Third, as we have mentioned earlier, some methods have the ability to discover interactions, and discovering high-order interaction is non-trivial. Currently, there are no classification methods that can do all three well.

3.5.4 Limitations

Unlike the data-driven representations, the construction of the SEV requires clinical expertise. Most of the effort is spent on classifying diagnoses into categories and determining pharmaceutical subclasses for drugs. This effort is not specific to SEVs; even the RAW representation had access to these higher-level categorizations. The effort that is specific to SEV is determining whether lab results and vital signs are normal and whether a drug is first-line or last-line medication. This information is often readily available from practice guidelines, such as the American Diabetes Association guidelines for diabetes. The effort to include this information is small, but non-negligible. However, SEV is outcome-independent, thus once a hierarchy for a risk factor or disease is defined, it can be used for numerous outcomes without the need to change it.

3.6 Conclusions

For both regression and association pattern mining, SEV provides the highest performance, substantially higher than the other data representations in a high-risk subpopulation, where accurate risk assessment is particularly important to appropriately target preventive measures. Besides having the highest performance, SEV produces clinically interpretable models and can also handle missing values.

Chapter 4

Estimation of onset time for diseases

4.1 Introduction

The recent adoption of the electronic health record (EHR)[12, 79] provides us with an opportunity to use it for advanced analytics. Many of these analyses rely on the onset time of diseases. For example, accurate onset time is required for time-to-event outcome analyses, and also for analyses that are concerned with sequences in which diseases develop. Especially for slow-onset diseases like hyperlipidemia (HLD), hypertension (HTN), and type 2 diabetes mellitus (T2D), onset time is not directly observable from the EHR[11]. EHR store the time when a problem was discovered or recorded, which can be significantly (years) different from the onset time, when the disease actually started. In this study, we construct a model that can reliably estimate the onset time for slow onset diseases from intermittently observable EHR data elements, most notably from laboratory results. We demonstrate this method through estimating the onset time of T2D from HbA1c and as a concrete application of this methodology, we use the estimated onset time of T2D to optimize the time when a patient needs to come back for diabetes screening.

T2D is a progressive metabolic disease, defined by chronically elevated blood sugar levels. The American Diabetes Association (ADA)[88] defines diabetes as a condition in which glycated hemoglobin A1c (HbA1c) levels exceed 6.5%. T2D is a fast growing

public health concern in the United States[81]. Approximately, 29.1 million Americans (9.3 % of the total population) are suffering from diabetes in 2014[52]. Out of 1,000 cases 7.8 are newly diagnosed this year, and this number is estimated to nearly double to 15 out of 1,000 cases by 2050[55]. Diabetes has severe consequences, with a significant impact on the quality of life[89, 90]. It is known to be the leading cause of kidney failure and blindness and the seventh-leading cause of death in the United States[81]. Since diabetes is a non-reversible progressive chronic disease[3, 4, 91], prevention and timely diagnosis are of key importance. Early identification of patients at high risk allows for intervention through lifestyle change or early drug therapy, which has shown to be very effective, reducing new incidents of diabetes by 30 to 60%[6, 7].

Several diabetes guidelines[88, 92] are available for use in clinical practice. These guidelines cover comprehensive procedures for diabetes diagnosis and management including risk assessment[46, 93, 14, 94], intervention strategies, and timing of follow-up visits. Despite the paramount importance of prevention, surprisingly, these guidelines only have loose recommendations for follow-up times for patients who are not yet diabetic. Although there are studies that examined the follow-up times for cost-effectiveness analysis[95] or for progression to diabetes[96], none of them addresses glucose level changes over time. In our work, we aim to provide better recommendations for follow-up times for non-diabetic patients (who naturally do not receive diabetes drugs). A suitable follow-up time reduces waste by not requiring healthy patients to have diabetes testing too early but it still affords the opportunity to intervene should the patient progress to near-diabetes. As a first step, in this paper, we aim to estimate the onset time of diabetes, the earliest time when a non-diabetic patient has HbA1c in excess of 6.5%.

Modeling the trajectory of HbA1c from EHR data poses several key challenges. First, patient visits are intermittent, so the actual progression of HbA1c is unobservable: we may not be able to directly observe the HbA1c level of a patient at a particular time t. Even though we may not be able to observe the HbA1c level directly, the patient can still contribute partial information to the model: if the patient was out of control (HbA1c $\geq 6.5\%$) at an earlier time T ($T < t$), then we expect the HbA1c to be out of control at t; or if the patient was under control (HbA1c $< 6.5\%$) at a later time T ($T > t$), then the patient should be under control at t. Not all models can make use of

this partial information. Second, since HbA1c tends to increase over time, patients will either become diabetic or receive preventive diabetes drugs and hence get censored. This virtually guarantees that censoring in our study is not random. Third, some essential data elements, most notably patient education and change in lifestyle interventions, are missing from the structured EHR. Lifestyle interventions are the first line of defense in both the prevention and the management of diabetes, yet related information is unavailable in structured format in the EHR.

The natural choice for estimating onset time is by modeling HbA1c trajectory using a linear regression model or by modeling time-to-event using a proportional hazards model. Unfortunately, both models constructed from EHR data lack the capabilities to address many of the above challenges. In fact, if we built a linear model to predict HbA1c, we would erroneously find that HbA1c actually decreases over time. HbA1c would not typically decrease without intervention; it appears to decrease because of non-random censoring and patients undertaking lifestyle interventions. A proportional hazards model we constructed to directly estimate the onset time of diabetes also provides poor estimates for the same reasons: the non-informative censoring assumption is violated and the intervention is not observable and hence is unaccounted for.

In this manuscript, we propose a novel approach that uses a Bayes network[97, 98, 28] to model HbA1c level progression in non-diabetic patients and to subsequently estimate the optimal amount of time before a patient's HbA1c level gets out of control. Our model addresses all four of the above challenges. It models the (unobservable) actual and observed HbA1c level separately. While we can only observe a patient's HbA1c intermittently, we can estimate the (hidden) actual HbA1c at any time. Our model also takes partial information into account. If we observe a patient to have HbA1c in excess of 6.5% at time T, then his latent HbA1c level is expected to be above 6.5% any time afterwards until the patient receives intervention; and similarly, if we observe a patient to have HbA1c level less than 6.5% at time T, we expect his latent HbA1c to be less than 6.5% at any time before T. This separation between the latent and observed HbA1c is the key to adjusting for non-random censoring. In addition, we include a latent variable for the lifestyle intervention, allowing us to further reduce bias.

We evaluated the HbA1c progression model on a cohort from a large healthcare system in the Upper Midwest United States. We demonstrated that the resultant model

Table 4.1: Study population

Description	Count
Inclusion:	
Primary care patients with age 18 years or older	+ 157,945 = 157,945
Exclusion:	
No diabetes-related observation(s) during retrospective period.	− 87,871 = 70,074
No diabetes-related observation(s) during follow-up.	− 12,826 = 57,248
Patient already presents with T2D at baseline.	− 37,075 = 20,173
Not having at least two HbA1c measurements during follow-up.	− 14,299 = 5,874

reflects the actual changes in HbA1c level well, and we also showed that the model has the ability to accurately estimate the time to the onset of diabetes.

4.2 Materials and Methods

4.2.1 Data

A retrospective observational study design is used to construct predictive models for glycated hemoglobin (HbA1c) level using patient's baseline characteristics. We collected clinical data from a large healthcare system in the Upper Midwest region of the United States. Vital signs, diagnoses (ICD-9-CM) and laboratory results are available after 2006, while medications are available after 2010. The baseline for each patient is established on a patient's first observation date on or after January 1, 2011. We use a retrospective period (2006-baseline) to establish the patient's baseline characteristics and we track HbA1c measurement during the follow-up period (baseline-December 31, 2013) until loss to follow-up or until the patient shows indication of T2D including T2D medication prescription, out-of-control HbA1c level (≥ 6.5) or a related diagnosis code.

The construction of the study cohort is described in Table 4.1. We include all primary care patients aged 18 years or older at baseline (157,945 patients). We exclude patients who do not have either HbA1c measurement or antidiabetic medications during the retrospective (87,871 patients) or during the follow-up period (12,826 patients). The lack of diabetes-related observations prevents us from determining the patient's baseline

diabetes status. We also exclude patients who are diabetic or who receive diabetes drugs at baseline (37,075 patients), and patients not having at least two HbA1c measurements after 2011 (14,299 patients). Type 1 diabetes patients are not included after the cohort selection. Our final cohort has 5,874 patients.

Table 4.2: Demographics, comorbidity, and HbA1c level data at baseline

Risk Factor	No. (%) of sample
Demographic	
Age, mean ± SD	57.0 ± 14
Sex: male	47.2 %
Comorbidity	
Hyperlipidemia (HLD)	82.3 %
Hypertension (HTN)	70.8 %
Obese (BMI $\geq$ 30)	55.5 %
Glycated hemoglobin (HbA1c)	
Normal ($<$ 5.7 percent)	40.8 %
Prediabetes (5.7-6.4 percent)	59.2 %

For the study cohort, we collected demographic information, diagnoses codes (ICD-9-CM) of comorbidities, laboratory results, vital signs, and medications. The ICD-9 codes are used for identifying type-2 diabetes mellitus [ICD-9-CM 250.x0 and 250.x2], hypertension [ICD-9-CM from 401.xx to 405.xx] and hyperlipidemia [ICD-9-CM 272.0x, 272.1x, 272.2x and 272.4x]. The American Diabetes Association (ADA)[88] guideline is followed to determine whether an HbA1c level is normal. Normal HbA1c values are $<$ 5.7%. Obesity is defined as a body mass index (BMI) equal or greater than 30. All missing laboratory results are treated as normal (the physician saw no need to order them). Since ICD-9-CM codes by themselves are insufficient to capture all types of diseases[99, 100, 101, 102], we use phenotypes (combinations of diagnoses codes, abnormal vital signs/laboratory results and medication usage) to identify comorbidities. Table 4.2 presents the baseline characteristics of our cohort.

4.2.2 Bayesian network

In this manuscript, we construct a progression model for glycated hemoglobin A1c (HbA1c) level using a Bayesian network. Bayes networks explicitly describe dependence

relationships among variables, allowing us to incorporate our prior beliefs. Our central prior belief is that HbA1c does not decrease without lifestyle change, therapy, or some other kind of intervention and that prediabetic patients receive lifestyle intervention, which is not recorded in their EHR.

HbA1c progression model

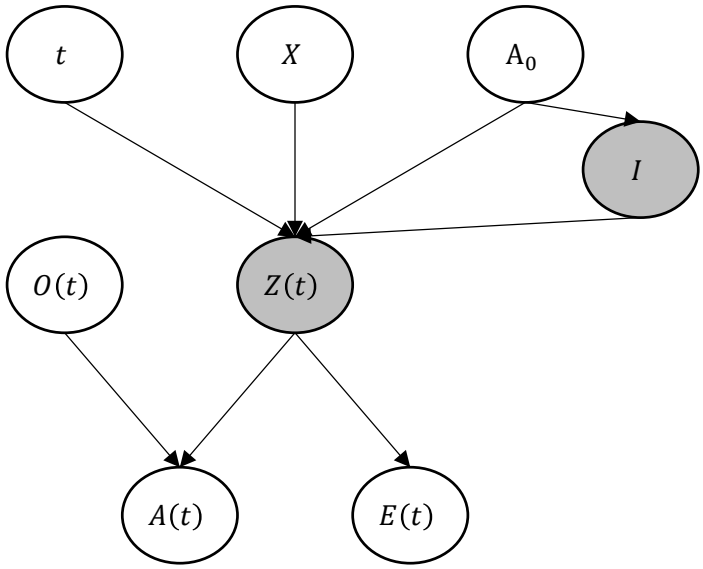


Figure 4.1: Glycated hemoglobin (HbA1c) level progression model.

We present the graph for our HbA1c progression model in Figure 4.1 as a template graph. Template graphs offer a concise representation of temporal graphs, where nodes in the large rectangle (template) repeat at every time point, while nodes outside the template are time-independent. Clear circle nodes are observable and shaded circle nodes are latent (not directly observable). The formulation in Figure 4.1 assumes that $Z(t)$ is independent of $Z(\tau)$, $t \neq \tau$, given time t and the observed variables. This is a reasonable assumption, because $Z(t)$ directly incorporates t as a predictor. Time is always measured relative to baseline in years.

Nodes X and A_0 describe the patient's baseline characteristics: A_{0j} denotes the observed baseline HbA1c level and X_j denotes the baseline comorbidities for patient j.

Since baseline characteristics are measured at time 0, they are time-independent.

We model the possible intervention a patient may have received as a latent variable I_j, which depends on the baseline HbA1c level and the baseline comorbidities. Patients who become prediabetic (HbA1c ≥ 5.7) or suffer from multiple comorbidities are likely to receive advice to change their lifestyle (exercise, eating habits) in order to prevent or delay progression to overt diabetes. The fact that the patient received such advice is not recorded in our data and whether the patient complies with such advice is also unobservable. Specifically, we model the latent intervention variable I as

$$
\begin{aligned}
& p(I_j = i \mid A_{0,j}) \\
& = \begin{cases} 1, & \text{if } i \neq i_{5.7} \text{ and } i \neq i_{6.0} \text{ and } A_{0,j} < 5.7 \\ 1, & \text{if } i = i_{5.7} \text{ and } 5.7 \leq A_{0,j} \text{ and } A_{0,j} < 6.0 \\ 1, & \text{if } i = i_{6.0} \text{ and } 6.0 \leq A_{0,j} \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{4.1}
$$

Note that $i_{5.7}$ is the latent intervention administrated when $A_0 \geq 5.7$ and $i_{6.0}$ is the latent intervention administrated when $A_0 \geq 6.0$. We adopt two-level lifestyle intervention since those who have HbA1c over 5.7[88] need to consider lifestyle intervention and those with HbA1c over 6.0[87] receive more aggressive lifestyle intervention with cautious follow-up schedule. Also, with increased baseline HbA1c, the effect of intervention could diminish. Intervention as defined above is deterministic; it is formulated as a probabilistic model to fit into the Bayes network framework.

In order to address the problem of non-random censoring, the cornerstone of our methodology is the separation between the unobservable actual HbA1c level at time t, $Z_j(t)$, and the observed HbA1c level at time t, $A_j(t)$. We can compute the hidden HbA1c level at any time t, while the observed HbA1c $A_j(t)$ level is only available at one time point T_j for each patient; it is unknown at any other time point.

We assume $Z_j(t)$ is identical to a linear combination of time t, the baseline HbA1c

level $A_{0,j}$, the patient's baseline comorbidities X_j, and the latent intervention I_j.

$$\begin{aligned} & p(Z(t) = z \mid A_0, X, I = i) \\ & = \begin{cases} 1, & \text{if } z = A_0 + X\xi + t\beta + i_{5.7}\delta_1 + i_{6.0}\delta_2 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4.2}$$

where $Z(t)$ is forced to take the value predicted by the linear model; ξ, β, and δ are coefficients of X, t, and the interventions, respectively.

The observed HbA1C level $A_j(t)$ of patient j at time t is identical to $Z_j(t)$ with Gaussian noise when it is observed (i.e. $O_j(t) = 1$ and is unknown otherwise.

$$\begin{aligned} & p(A_j(T) = a \mid O_j(t) = o, Z_j(t) = z) \\ & = \begin{cases} \Phi(z - a, \sigma^2), & \text{if } a = z \text{ and } o = 1 \\ 1, & \text{if } a \text{ is unknown and } o = 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4.3}$$

Notice that ϕ and Φ denote the probability density function (PDF) and the cumulative distribution function (CDF) of the standard normal distribution. $A_j(t)$ is unknown at all time points $t \neq T_j$ and needs to coincide with $Z_j(t)$ at $t = T_j$. We would like to point out that this does not require $A_j(t)$ to exactly coincide with the model prediction from Eq (4.2); there is a Gaussian error term in Eq (4.3) that allows for differences between the model prediction and $Z_j(t)$ and thus $A_j(t)$.

Finally, $E_j(t) = 1$ denotes that patient j has had an event at time t or before, namely there exists a time $T_j \leq t$, such that $Z_j(T_j) \geq 6.5$.

$$\begin{aligned} & p(E_j(t) = e \mid Z_j(t) = z) \\ & = \begin{cases} 1, & \text{if } e = 1 \text{ and } 6.5 \leq z \\ 1, & \text{if } e = 0 \text{ and } z < 6.5 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4.4}$$

Similarly to the intervention, $E_j(t)$ is also deterministic; we simply use a probabilistic notation to cast it into the Bayes network formalism.

The separation of the actual and observed HbA1c level is the key concept of our Bayesian network model. If the patient j already had an event, then $Z_j(t)$ must be $\geq$ 6.5; otherwise $p(E_j(t))$ becomes 0. Similarly, $Z_j(t)$ must be < 6.5 for patients under observation who will not suffer an event at T_j, i.e. $A_j(T_j) < 6.5$. The likelihood, which we will discuss in the following section, becomes 0 if this assertion does not hold.

Parameters learning

In the previous section, we described the structure of our network model; this section is concerned with estimating the values of the parameters ξ, β, δ. The maximum likelihood estimation (MLE) is widely used for estimating the parameters of a network model. Specifically, the data likelihood is

$$\begin{aligned} &\mathcal{L}(\xi, \beta, \delta \mid A_0, X, O(t), A(t), E(t)) \\ &= \prod_t \prod_j p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta) \end{aligned} \tag{4.5}$$

where the probability density function (PDF) of the data, $p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta)$, is the marginal distribution over the latent variables $Z(t)_j$ and I_j as following:

$$\begin{aligned} &p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta) \\ &= \sum_z \sum_i p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t), Z_j(t) = z, I_j = i; \xi, \beta, \delta) \end{aligned} \tag{4.6}$$

With the probabilities of A_0, X, $O(t)$ being constant in the sense that they do not depend on the parameters and with the probabilities of $A(t)$, $E(t)$, and $Z(t)$ given in

equations (4.2), (4.3) and (4.4), respectively, we can rewrite the PDF of the data as

$$
\begin{aligned}
&p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta) \\
&= \begin{cases}
\Phi(z-a, \sigma^2), & \text{if } [I_j \text{ is consistant with } A_{0,j}] \text{ AND} \\
& \quad [z = A_0 + X\xi + t\beta + i_{5.7}\delta_1 + i_{6.0}\delta_2] \text{ AND} \\
& \quad [a = z \text{ and } o = 1] \text{ AND} \\
& \quad [(e = 1 \text{ and } 6.5 \leq z) \text{ or } (e = 0 \text{ and } z < 6.5)] \\
1, & \text{if } [I_j \text{ is consistant with } A_{0,j}] \text{ AND} \\
& \quad [z = A_0 + X\xi + t\beta + i_{5.7}\delta_1 + i_{6.0}\delta_2] \text{ AND} \\
& \quad [a \text{ is unknown and } o = 0] \text{ AND} \\
& \quad [(e = 1 \text{ and } 6.5 \leq z) \text{ or } (e = 0 \text{ and } z < 6.5)] \\
0, & \text{otherwise}
\end{cases}
\end{aligned} \tag{4.7}
$$

We seek the parameter values ξ, β, θ that maximize the likelihood in Eq. (4.5) using the PDF (4.7). Any off-the-self algorithm can be used to compute the parameters[103, 104].

Prediction

Our goal is to estimate predicted HbA1c level at time t. Since our model has hidden nodes that we cannot observe directly, we need to find $Z(t) = z$ that maximizes $p(Z(t) \mid A_0, X, O(t), E(t))$ for given time t. Recall that the latent intervention variables are required for estimating $Z(t)$, but we cannot directly observe them, thus we seek

$$
\begin{aligned}
&\arg\max_z p(Z_t = z \mid A_0, X, O(t), E(t)) \\
&= \sum_i [p(Z_(t) = z \mid A_0, X, O(t), E(t), I = i) \times p(I = i \mid A_0, X)]
\end{aligned} \tag{4.8}
$$

4.3 Results

The goal of this work is to estimate the onset time of diseases, diabetes in our concrete application, through modeling patients' HbA1c trajectory via Bayesian network. Onset time is the earliest time when a patient's HbA1c exceeds 6.5. Unfortunately, due to the

intermittent nature of patient visits, the actual onset time is observable only for relatively few patients (492 out of 5,874). Thus we are unable to evaluate the performance of the proposed model on the entire cohort. Instead, we evaluate our model in two parts. In the first part, we evaluate the performance of the proposed model on the entire population on two related tasks, for which the outcomes are available for all patients. Our method was not designed for these tasks, the purpose of this evaluation is merely to demonstrate that our method can achieve reasonable performance on the entire cohort even when we compare it to standard techniques optimized for that task. Specifically, we evaluate the proposed model's ability to predict HbA1c levels and compare it to a multivariate regression model, which is the natural choice for modeling a continuous outcome. The second task is predicting the (but not the onset time) of diabetes and compare our method with the Cox proportional hazard model. In the second part, we evaluate our model for the intended purpose of the algorithm, namely for predicting the onset time of diabetes. Naturally, we can only use the 492 patients, for whom the onset time is known. We compare our model with the Cox proportional hazards model, using median survival time as the estimate for onset time. The 10-folds cross-validation is used to evaluate model accuracy. All implementation and analysis is conducted with the use of R version 3.3.1.

4.3.1 Evaluations on the Entire Population

Besides the proposed model, we built two additional models on the same cohort we used to construct our proposed model. The first model is a multivariate linear regression model with the observed HbA1c as the dependent variable and baseline HbA1c, baseline comorbidities, follow-up time and the latent interventions. Analogously to the proposed model, we considered obesity (Obese), high cholesterol (hyperlipidemia; HLD) and high blood pressure (hypertension; HTN) as comorbidities. The second model is a Cox proportional hazards model predicting the onset of diabetes at the end of follow-up using the same set of independent variables as the multivariate linear regression model (except time). Time is incorporated into the baseline hazard.

First, we interpret the models. Table 4.3, 4.4 and 4.5 shows the predictors, the coefficients, and the empirical p-values estimated through bootstrap estimation with 1000 replications, for the three models. We interpret the coefficients of the multivariate

Table 4.3: Predictors and coefficient estimates from the multivariate linear regression model.

	Coefficient	p-value
(Intercept)	1.556	$< .001$
A_0	0.729	$< .001$
time t	-0.008	0.346
$i_{5.7}$	0.001	0.967
$i_{6.0}$	0.040	0.027
Obese	0.037	$< .001$
HLD	-0.025	0.109
HTN	0.013	0.265

Table 4.4: Predictors and coefficient estimates from the Cox proportional hazards model.

	Coefficient	p-value
A_0	3.622	$< .001$
$i_{5.7}$	0.109	.732
$i_{6.0}$	0.225	.323
Obese	0.444	.002
HLD	-0.814	$< .001$
HTN	0.121	.382

Table 4.5: Predictors and coefficient estimates from the Bayesian network model.

	Coefficient	p-value
time t	0.320	$< .001$
$i_{5.7}$	-0.089	.008
$i_{6.0}$	0.298	$< .001$
Obese	-0.048	$< .001$
HLD	-0.098	$< .001$
HTN	-0.074	$< .001$

linear regression model as follows. A unit increase in baseline HbA1c (A_0) increases the HbA1c at time t by .729; being obese increases the HbA1c at time t by .037 and receiving the latent intervention does not change the HbA1c significantly. The other variables can be interpreted similarly. Of particular interest is the coefficient of time: during each additional year of follow-up time, the HbA1c level increases by -0.008, on average (i.e. it decreases by .008). This negative coefficient is problematic for two reasons. First,

our understanding is that HbA1c increases unless the patient receives interventions. Second, this model cannot be used to estimate the time to onset of diabetes, because the estimated HbA1c level is predicted to decrease over time, suggesting that the patient will never develop diabetes. This is a clear example demonstrating the challenges of modeling the HbA1c trajectory from EHR data. Additionally, the model obviously contradicts clinical knowledge as well.

Table 4.4 shows the Cox proportional hazards model. The effects of coefficients can be interpreted in the standard way: a unit increased in A_0, which is a baseline HbA1c level, is associated with 3.622 increase in the log hazard rate.

Table 4.5 shows that the coefficients for Eq (4.2) in the Bayesian network model. The interpretation of the coefficient in the Bayesian network model is analogous to that of the multivariate linear model: a year increase in follow-up time t increases the HbA1c level by 0.320, on average. Assuming this rate of increase in HbA1c, a patient would progress from almost prediabetic levels (say, HbA1c of 5.5%) to overt diabetes (HbA1c $\geq$6.5%) in 3 years, which is reasonably consistent with screening interval in current guidelines[95, 88].

Estimating HbA1c level

Figure 4.2 depicts a scatter plot of observed HbA1c levels against estimated HbA1c levels generated by the methods: the proposed Bayes network model (denoted by circles) and by the linear regression model (denoted by crosses). The x-axis is the observed HbA1c levels, and the y-axis is the estimated HbA1c levels. Points on the diagonal line have no error between the observed and the estimated, and being far away from the line indicates a larger error.

Both models offer reasonable performance; although the linear regression model has the slightly higher correlation (.569) with the actual HbA1c values than the proposed model (.501). We wish to emphasize that the proposed model was not designed to provide as accurate HbA1c prediction as possible; but rather our goal was to construct a model that is physiologically feasible and offers good prediction of diabetes onset time. We argue that despite the small deficit in predictive performance, the proposed model is actually better. First, as we discussed earlier, the linear model cannot predict HbA1c levels in excess of 6.5; thus it is totally inadequate for predicting the onset time of

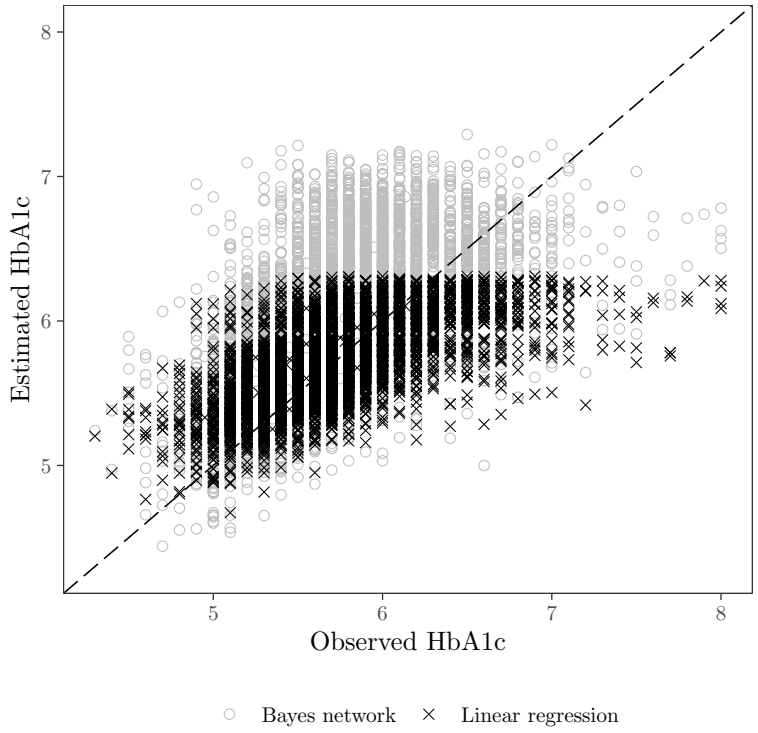


Figure 4.2: Scatter plot of observed HbA1c levels against estimated HbA1c levels.

diabetes: its prediction is that no patient will ever develop diabetes. Second, the Bayes Network model overestimates the HbA1c levels for some patients, most prominently in patients, for whom HbA1c levels improved over time. (These patients are in the top left triangle of the figure.) For example, among the 259 patients for whom the Bayes Network model overestimated the HbA1c level by more than 1 unit, the HbA1c level actually improved, it decreased on average from 6.1 to 5.5 over 1.9 years. Again, poor predictive performance for these patients is hardly surprising as the Bayesian network model was constructed under the explicit assumption that HbA1c deteriorates over time[105, 106]. For these patients, linear regression model performed better. On the other hand, HbA1c is expected to increase in the absence of interventions[6, 7] by the nature of the progressive disease. When patients' observed HbA1c levels increase, the Bayesian Network model outperforms the linear regression model. For instance, the Bayesian network estimates HbA1c level correctly (within .2 unit difference) for 130 out

of 414 patients who have estimated HbA1c level over 6.5.

Since the multivariate linear regression model has no ability to predict diabetic (> 6.5) HbA1c levels, we cannot use it to estimate diabetes onset time, and thus we exclude it from further evaluation.

Estimating the Onset of Disease

In this section, we evaluate the ability of the proposed model to predict the onset of diabetes (whether a patient develops diabetes by the end of follow-up). As we mentioned before, we can determine whether a patient developed diabetes by the end of his follow-up for all 5,874 patients, but we cannot determine the precise onset time of diabetes. We consider a patient diabetic if he has a diabetes diagnosis code or observed HbA1c $> 6.5\%$ at last follow-up. We compared our model to the Cox proportional hazards model using the concordance index (c-index)[107] as the evaluation metric. The c-index (also known as AUC) in our case is the probability that for a randomly chosen pair of patients, where one has developed diabetes by the end of his follow-up, while the other has not, the patient with the disease has a higher estimated risk. For the Bayes network the estimated risk is the estimated HbA1c; for the Cox model it is the cumulative hazard—both computed at last follow-up. The c-statistic ranges between 0.5 and 1.0; 0.5 denotes a random model and 1.0 corresponds to perfectly discriminating predictions. The c-index value for the Bayesian network and the Cox proportional hazards models were 0.776 (95% confidence interval: 0.754 and 0.802) and 0.736 (95% confidence interval: 0.705 and 0.766), respectively. The confidence interval was computed using bootstrap estimation with 100 replications. This result demonstrates that the Bayesian network model has better ability to predict the onset of diabetes than the Cox model.

4.3.2 Estimating T2D Onset Time

We now return to our original problem of estimating the onset time of diabetes, the time when the HbA1c level exceeds 6.5. Unfortunately, for the vast majority of patients, the onset time is not observable. We identified 492 patients who have an observed HbA1c level between 6.3 and 6.7 and we simply call the corresponding observation time as the observed onset time, although the actual onset time can be slightly different. We

estimate onset time as the earliest time when a patient's estimated HbA1c exceeds 6.5. We can evaluate our method by comparing the estimated and the observed onset time.

We cannot expect the estimated onset time to match the observed onset time exactly. Changes in HbA1c are not instantaneous; researchers tend to think of HbA1c as a 3-month running average of the blood glucose level. Accordingly, our observed onset time could differ from the actual onset time by almost 2 months (the time it takes for the HbA1c to progress from 6.3 to 6.5 according to our model). Therefore, we believe that the best possible estimate is within 2 months of the actual onset time. Thus these observed onset times are not exact, but are appropriate as a "silver standard".

Evaluating the proposed method against the "silver standard"

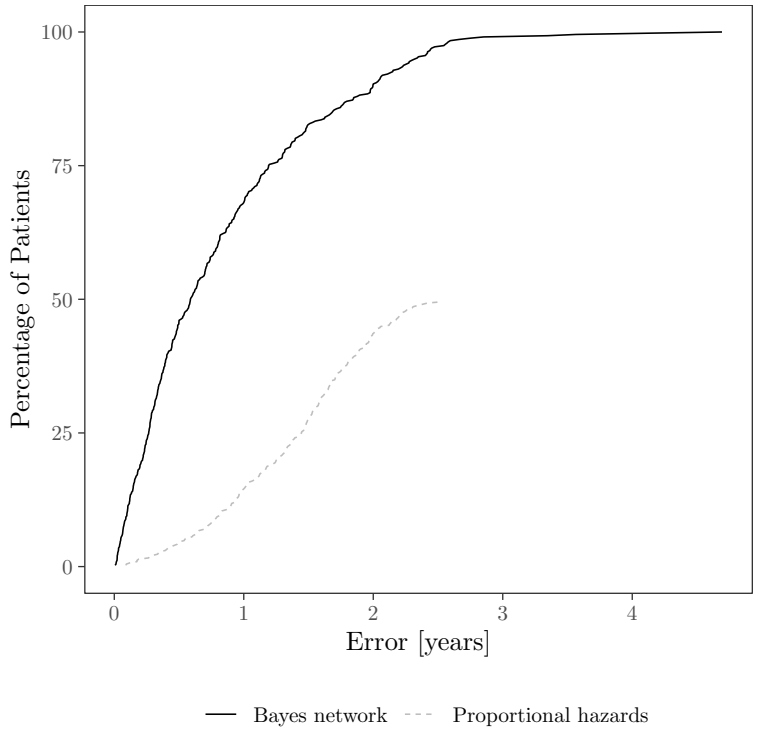


Figure 4.3: The cumulative density of the prediction error in years.

Figure 4.3 shows the cumulative density distribution of the prediction error. We computed the prediction error as the absolute difference between the predicted onset

time and the onset time observed from the EHR data. The x-axis represents the prediction error (measured in years), and the y-axis represents the cumulative probability of that prediction error denotes the percentage of patients who have a prediction error smaller than the corresponding value on the x-axis. For example, 43.5% of patients (y-axis) have a prediction error less than half year (0.5 on the x-axis). The solid black line represents the cumulative density curve of the prediction error from the Bayesian network. The figure shows that a quarter of the patients (22.0 %) have a prediction error less than three months, which is our theoretical lowest error; and almost half of the patients (43.5 %) have an error less than half year.

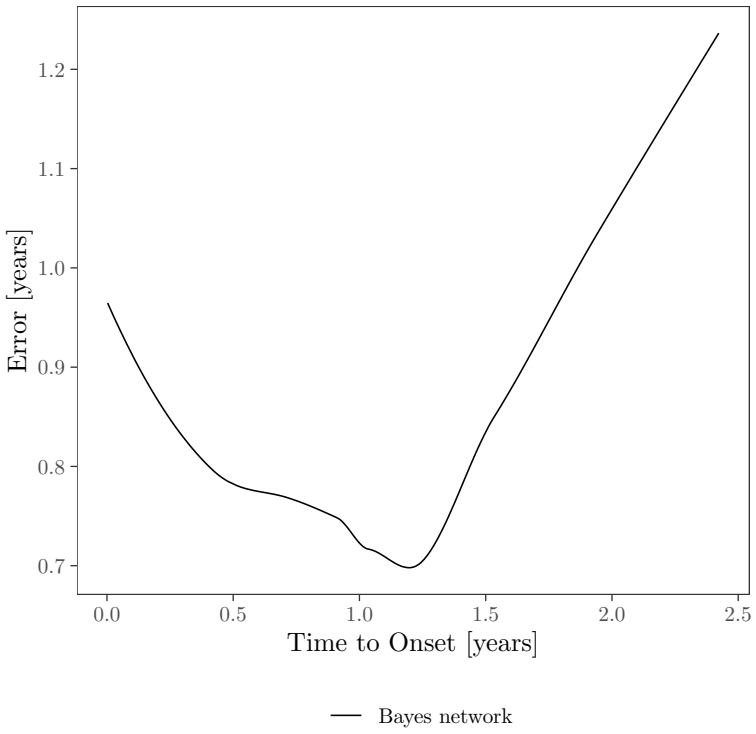


Figure 4.4: The cumulative density of the prediction error in years.

Figure 4.4 shows the prediction error from the Bayes network model as a function of the observed time to onset. The x-axis represents the observed time to onset (in years), and the y-axis represents the average prediction error for patients with that particularly observed time to onset. The plot reveals that the Bayesian network method

is most reliable (has the lowest error on average) when the onset time is a bit over a year. This is reassuring because achieving half a year prediction error in patients whose observed onset time is 3 months is trivial. Instead, most of the prediction errors are from patients with observed onset time less than three months and longer than two years. Patients who fall out of control within three months from the baseline could be diabetic or almost diabetic at baseline since the HbA1c level generally does not change much in such a short amount of time. Naturally, the uncertainty in prediction grows as the follow-up time increases beyond 2 years.

Comparative evaluation of the proposed method

The gray dashed line in Figure 4.3 represents the cumulative density of the prediction error from the Cox model. We use median survival time, which is the earliest time when the survival probability is equal or less than 0.5, as the estimated onset time.

Not only does the Cox model show significantly lower prediction ability as compared with the Bayesian network model, but it failed to make a prediction for 261 out of the 492 patients, because these patients had a survival probability higher than 0.5 at the end of follow-up. The cumulative density of the prediction error from the Cox model reveals that only 1.2% of the patients have a prediction error equal or less than three months and 4.5% of the patients have a prediction error equal or less than half year. Since the 3-year to 5-year interval is regarded as cost-effective screening interval[95], the Cox model with 53.0% of the patients having a prediction error of more than 3-years, has limited utility in this application.

4.4 Summary and Discussion

Electronic health records (EHR) contain rich, longitudinal, and large volumes of EHR data, which offers us a new way of conducting clinical research. Many research questions rely on the accurate estimation of onset time for diseases. Because of the inherent limitations of EHR data, we cannot always directly extract this information from EHR as they contain the recorded time for diseases rather than the onset time. For slow-onset diseases, the onset time and the recorded time can be substantially different; they can differ by years. In this paper, we propose a novel model to estimate the onset time of

chronic diseases primarily based on the defining laboratory results. Additionally, we demonstrated our model through a concrete application of estimating the onset time of Type 2 diabetes (T2D) to optimize the time when a patient needs to come back for diabetes screening.

While T2D management guidelines exist and address most aspects of diabetes diagnosis and care, including recommendations for follow-up time, they provide little guidance for follow-up times for patients who do not have diabetes yet. Asking these patients to visit providers earlier than necessary is wasteful; while delaying their visit until they have developed diabetes narrows the window of opportunity for prevention. In this paper, we take a first step towards developing recommendations for non-diabetic patients by constructing an EHR-based model that can accurately predict the onset time of diabetes.

First, in order to accurately predict the onset time of diabetes based on estimated HbA1c levels, the estimated HbA1c levels need to be accurate. The model offers good performance for patients with increasing HbA1c levels, showing less than .2 unit prediction error in 130 out of 414 patients who have estimated HbA1c level over 6.5. These are the patients who are most likely to develop diabetes in the near future and thus accurate prediction is most critical for these patients. Conversely, the model did not deliver good performance for patients with decreasing HbA1c levels, given that the model was deliberately penalized for predicting decreasing HbA1c levels. As we mentioned earlier, in patients prone to develop diabetes, who are exhibiting signs of insulin resistance, decreasing HbA1c levels are most likely a result of unobserved interventions, rather than the natural (intervention-free) course of the disease. These patients are not at as high risk of developing diabetes as others, thus our small prediction error is acceptable.

The performance of our model is particularly reassuring when compared to multivariate linear regression, the standard method for modeling continuous measures, such as HbA1c. Linear regression, with no ability to compensate for informative censoring (patients who are removed from the study for taking antidiabetic drugs are very likely to develop diabetes) constructed a model that predicted decreasing HbA1c levels. Not only is this model not supported by our knowledge of the pathophysiology of metabolic degeneration, but it also cannot be used for predicting onset times, because the decreasing HbA1c levels suggest that the patient will never develop diabetes.

We further evaluated our method in terms of its ability to predict the onset time for diabetes on a smaller cohort in which diabetes onset times are known. We found that we could estimate the onset time within 6 months error for almost half of the patients; and within 3 months accuracy for almost quarter of the patients. Our prediction error was lowest for patients who progressed to diabetes in a bit over a year. Our prediction error increased sharply for patients with an onset time 2+ years into the future. This result suggests that the prediction model is most appropriate for prediabetic patients, precisely the patients where correct follow-up time is most crucial.

An alternative to the proposed model could be a Cox proportional hazards model using one of the standard methods for estimating time to onset of diabetes. Using median survival as an estimate for time to onset of diabetes, the Cox model offered poor performance with a prediction error in excess of 3 years for almost half of the population. Periods between 3 to 5 years are considered cost-effective follow-up intervals[95], thus an error of 3 years for half of the patients renders this alternative impractical.

In conclusion, we successfully developed a predictive model to estimate onset time of chronic diseases primarily based on the defining laboratory results. We demonstrated this method through the concrete application of estimating the optimal follow-up time for effective T2D screening. To the best of our knowledge[108, 109], this is the first study of modeling the trajectory of HbA1c level and estimating the onset time of diabetes. The methodology is not specific to T2D; we believe that it can be generalized to other slow onset diseases in a straightforward manner.

Chapter 5

Discovery of disease trajectories towards a certain outcome

5.1 Introduction

The advent of electronic health record (EHR) systems has paved the way to perform large-scale data analytics to uncover new medical knowledge that was previously inaccessible. EHR systems store information about entire populations and offer long follow-up times. In this study, we work with data from a premier healthcare provider in the Midwestern United States, which pioneered the adoption of EHR systems in the region, allowing us access to nearly 13 years of follow-up time for a relatively large number of patients. Such long follow-up, in turn, allows us to study disease trajectories that lead to type-2 diabetes.

Type 2 diabetes (T2D) is one of the fastest growing public health concerns in the United States[81]. There are 29.1 million patients (9.3% of the US populations) suffering from diabetes in 2014[52]. Diabetes, which is the seventh leading cause of death in the United States, is known to be a non-reversible (incurable) chronic disease[3, 4], leading to severe complications[81, 5], including chronic kidney disease, amputation, blindness, and various cardiac and vascular diseases. Early identification of patients at high risk is regarded as the most effective clinical tool to prevent or delay the development of T2D, allowing patients to change their life style or to receive medication earlier. In turn, these interventions can help decrease the risk of diabetes by 30-60%[6, 7]. Many risk

models[110, 111, 46] aiming at early identification of patients at high risk are widely used in the clinical settings. There models typically only consider the patient's current state at the time of the assessment and ignore the trajectory, the sequence of events that led up to the state.

The motivating hypothesis for our work is to study whether the trajectory influences the risk of diabetes. Diabetes is a heterogeneous disorder involving complex biological mechanisms. In our study, we discovered multiple trajectories to diabetes that can help some of the underlying mechanisms and their associated risk of developing diabetes.

Access to 13 years of follow-up allows us to make inferences about trajectories leading up to T2D. Since many diseases are progressive (worsen over time), Electronic Health Record (EHR) data with its large sample size and long follow-up time allows us the opportunity to study the progression of these diseases. However, due to the nature of EHR data, unlocking this potential is challenging. The challenge stems from two compounding factors. First, EHR data was not designed to be a research platform; thus some critical data elements are not directly observable and need to be inferred. Second, chronic conditions have slow onset and as a result, the onset time is not only unobservable, but is difficult to estimate accurately. The purpose of this manuscript is two-fold: (1) we first describe the challenges we faced in using EHR data and the methods we developed to overcome those challenges; and (2) we then describe the interesting findings we uncovered.

Specifically, we define trajectories as sequences in which patients develop comorbidities as they progress towards T2D. Besides T2D, we consider three important comorbidities: hyperlipidemia (HLD, high cholesterol or unbalance of the various lipids), hypertension (HTN, high blood pressure) and impaired fasting glucose (IFG, elevated fasting plasma glucose). We infer the typical (most frequent) trajectory and enumerate the atypical trajectories that our data support. We build predictive models to determine whether following an atypical trajectory is associated with different risk of diabetes.

We perform our analysis on a large community-based cohort derived from EHR system in the Rochester Epidemiology Project[112] consisting of patients who received their primary care at Mayo Clinic. The data has nearly 13-years of unfragmented follow-up, making it the largest and cleanest EHR-derived data set of its kind. In this manuscript, we show that a single typical trajectory exists and it is consistent with the

trajectory that is commonly used for diabetes patient education. We enumerate several atypical trajectories that cover approximately 27% of the diabetes cases observed in our data set and assess the excess risk (if any) they confer on patients following them.

5.2 Data and challenges

5.2.1 Data

The study cohort consists of Mayo Clinic primary care patients residing in Olmsted County, MN. During the study period from 1999-2013, when complete EHR data was available, we have 70k patients with research consent. Informed consent was obtained from patients during each visit and consent information was stored in the EHR. Demographic information, diagnosis codes encoded as ICD-9-CT, laboratory results, vital signs and medication data were collected for this period.

5.2.2 Challenges

To establish trajectories, sequences in which the disease develop, we should only consider new (incident) diagnoses (as opposed to preexisting conditions) along with their onset dates. Surprisingly, this information is difficult to infer from the EHR system for the following reasons.

Secondary Use of EHR data EHR systems were originally developed for documenting patients' state for reimbursement purposes. The presence of diagnosis codes in the EHR are driven by billing rules. They may be present because the corresponding condition was tested, possibly newly discovered, or was complicating the treatment of other conditions. There is no designation in the EHR whether a diagnosis is incident or preexisting. Moreover, diagnoses may be missing (no reimbursement was requested for the condition) and can be false positive (the patient was merely tested for a condition).

Slow-onset conditions The second issue concerns the onset date. The development of T2D as well as the comorbidities that commonly precede it can take decades. The signs for these conditions are subtle and can remain undetected for years. Establishing the onset time for these conditions is challenging. Instead of trying to estimate the onset

Table 5.1: Study population.

Description	Count
Inclusion:	
Patients age $\geq$ 18 at 2005/01/01	+ 69,747 = 69,747
Exclusion:	
Diabetic patients	− 389 = 69,358
Patients with unknown glucose	− 14,559 = 54,717
Patients with unknown lipid	− 1,023 = 53,862
Patients with unknown BP	− 498 = 53,598
Non-diabetic patients who did not survive 5 years	− 10,089 = 43,509

date, we only assume that it happened before the earliest recording date. Another issue regarding the slow progression is that even with 13 years of follow-up, we can only observe partial trajectories, that is, the development of only a few new conditions. Therefore, if we tried to observe, rather than infer, the sequences, we would focus on patients with the fastest progression, possibly biasing the results.

5.3 Study design

A retrospective observational study design is adapted. We use January 1, 2005 as the baseline for our study. The period prior to the baseline, i.e. 1999-2004, is called a pre-baseline periods. We use the pre-baseline period to determine patients' baseline diabetes status and comorbidities by retrospectively examining their medical history through laboratory measurements, vitals, and diagnoses. Of particular interest is the presence of T2D related comorbidities HLD, HTN and IFG at the baseline. We set a follow-up period of 2005-2013 to follow the patients and record whether they developed diabetes. The incidence of T2D and its date during the follow-up period were determined via chart review.

The construction of the study cohort is described in Table 5.1. We included all adult patients with research consent and no diabetes diagnosis code at baseline. There are 69,747 such patients. From this cohort, we excluded patients with a high suspicion of diabetes (389 patients with fasting plasma glucose $>$ 125 mg/dL or those taking diabetes medications), unknown glucose value (14,559 patients), undetermined lipid status (1,023

patients) and unknown blood pressure (498 patients). Our final study cohort consists of 43,509 patients, and 4,795 of the 43,509 patients (11%) developed diabetes during the follow-up period.

Table 5.2: Study variables for impaired fasting glucose, hypertension and hyperlipidemia accounting for severity.

Risk Factor	Description	Count
Impaired fasting glucose (IFG)		
ifg.no	Fasting plasma glucose (FPG) ≤ 100	35,110
ifg.pre1	$100 <$ FPG ≤ 110	6,797
ifg.pre2	$110 <$ FPG ≤ 125	1,602
Hypertension (HTN)		
htn.no	No indication of hypertension	29,603
htn.untx	No drug is needed and only one blood pressure result is elevated	5,355
htn.tx	Treatment needed	8,551
Hyperlipidemia (HLD)		
hld.no	No indication of hyperlipidemia	12,092
hld.untx	No therapeutic need, but some indication of hyperlipidemia exists (lab or diagnosis)	25,439
hld.tx	Treatment needed	5,978
Obesity		
obese.no	BMI < 25	13,061
obese.overweight	$25 \leq$ BMI < 30	10,642
obese.obese	Diagnosis or BMI ≥ 30	12,188

To determine whether a patient has a particular comorbidity at the baseline, we use phenotyping algorithms. Phenotyping algorithms[14, 113] are simple classifiers that infer the presence of a disease based on diagnoses, lab results, vitals and medications. Specifically, in this study, we constructed three ordinal variables for IFG, HTN, and HLD as combinations of diagnosis, abnormal laboratory results (or vitals) and medication. The ADA guidelines were followed to determine whether a laboratory result (or vital sign) is normal. Table 5.2 lists these variables and the number of patients. Except for HLD, the majority of patients do not have a comorbidity at baseline in our cohort. Table 5.3 presents the baseline characteristics of the remaining variables.

Table 5.3: Baseline characteristics for variables not in Table 5.2.

Risk Factor	Description	Count
Demographic		
age	Age (mean±SD)	46 ± 16
gender	Gender (% male)	42.16 %
tobacco	Smoking status (past or current smoker %)	14.92 %
Diagnoses		
renal	Renal disease (prevalence %)	1.40 %
ihd	Ischemic heart disease (prevalence %)	6.31 %
cvd	Cardiovascular disease (prevalence %)	2.02 %
pvd	Peripheral vascular disease (prevalence %)	1.10 %
chf	Congestive heart failure (prevalence %)	0.92 %
carotid	Carotid artery disease (prevalence %)	0.86 %

5.4 Methods

5.4.1 Extracting the typical and atypical trajectories

We define a diabetes trajectory as a sequence of comorbidities (i.e., HDL, HTN, and IFG) potentially leading up to diabetes. The ordering of these comorbidities is denoted by an arrow ($\rightarrow$). For example, suppose we have three comorbidities A, B and C, and the trajectory $A \rightarrow B \rightarrow C$ indicates that A is followed by B and B is followed by C. These conditions are generally assumed to follow many different sequences (trajectories). We call the trajectory followed by most patients, *typical*, and label all other trajectories, *atypical*.

We only know that at baseline a patient has already developed a set of comorbidities (say) A, B and C, but we could not directly observe in which order these comorbidities were developed. We can, however, estimate it. Suppose B follows A, $A \rightarrow B$. If B indeed follows A, every time we encounter B, we should also encounter A. Therefore the probability $p(A \mid B)$ should be high. Accordingly, we define the probability of $A \rightarrow B$ as

$$p(A \rightarrow B) = p(A \mid B) \tag{5.1}$$

Let us extend this to calculate the probability of an entire trajectory. We define the

probability of a trajectory as the likelihood of observing the data set under the assumption that it was generated by the trajectory in question. Suppose there are four comorbidities A, B, C and D, and the trajectory is $A \to B \to C \to D$. Patients following this trajectory may have progressed to different stages: some patients may have progressed all the way to D, others to C, some to only A or B, and yet others may not present with any symptoms yet, but will follow the trajectory once they start progressing. In patients, who have already progressed to D along this trajectory, we should see A, B and C with very high probability, i.e. $p(A, B, C \mid D)$ should be high. In other patients following the same trajectory, who have only progressed to C, we should see A and B with high probability, i.e. $p(A, B \mid C)$ should be high. We define the probabilities for patients who have only progressed to A or B analogously, giving us the probability of the trajectory as

$$p(A \to B \to C \to D) = p(A, B, C \mid D) \times p(A, B \mid C) \times p(A \mid B) \tag{5.2}$$

Note that the same patient can be counted multiple times. For example, a patient presenting with A and B at baseline, is counted not only for the sequence $p(A \to B \to C \to D)$, but also for $p(A \to B \to D \to C)$, as well as for $p(B \to A \to C \to D)$ among others. Therefore, the likelihood does not coincide with the percentage of patients following this trajectory.

5.4.2 Type 2 diabetes risk modeling with trajectories

To address the association between the different trajectories and the risk of developing diabetes, we constructed a multivariate logistic regression model for diabetes outcome using demographics (Table 5.3), glucose level, staged comorbidities (Table 5.2), as well as three trajectories (Table 5.5). Data analysis was conducted in R version 3.2.3.

5.5 Results

In this section, we show the typical trajectory extracted from our data and subsequently enumerate the atypical trajectories. We then investigate whether the atypical trajectories are associated with increased risk of developing T2D.

5.5.1 The typical trajectory

Table 5.4: The five most likely trajectories.

No.	Trajectory	Likelihood
1	HLD → HTN → IFG → T2D	0.100
2	HLD → HTN → T2D → IFG	0.067
3	HLD → IFG → HTN → T2D	0.058
4	HTN → HLD → IFG → T2D	0.044
5	HLD → IFG → T2D → HTN	0.040

Table 5.5: Typical and atypical trajectories.

No.	Trajectory	HLD	HTN	IFG	Count	T2D	p(T2D)
1	Typical	N	N	N	8,795	235	0.037
2	Typical	Y	N	N	16,307	948	0.058
3	Typical	Y	Y	N	8,297	950	0.115
4	Typical	Y	Y	Y	3,485	1,362	0.391
5	Atypical with skipping HLD	N	Y	N	1,711	92	0.054
6	Atypical with skipping HLD	N	Y	Y	413	84	0.203
7	Atypical with skipping HTN	Y	N	Y	3,328	992	0.298
8	Atypical with skipping HLD and HTN	N	N	Y	1,173	132	0.113

In Table 5.4, we present the five most likely trajectories, selected based on the likelihood. The likelihoods are generally low, because the probability of progression to T2D itself is low. The most likely trajectory in our study cohort is p(HLD → HTN → IFG → T2D), which coincides with the trajectory that is commonly used for patient education. We also observe the most likely trajectory is far more likely than the others. Counting the number of patients (Table 5.5) who show no evidence of following a different trajectory confirms that the vast majority of the patients follow this trajectory. This satisfies our definition of typical trajectory.

5.5.2 Atypical trajectories

There is evidence in our data that patients can follow trajectories different from the typical one (i.e., p(HLD → HTN → IFG → T2D)). In our definition, a patient is said to

follow a trajectory $p(\text{HLD} \rightarrow \text{HTN} \rightarrow \text{IFG} \rightarrow \text{T2D})$, if and only if his sequence of comorbidities are constant with that trajectory. Formally if the patient has k comorbidities, these have to coincide with the first k diseases along the trajectory. A patient follows the typical trajectory if his comorbidities are consistent with the typical trajectory; and follows an atypical trajectory if he shows evidence to contrary. For example a patient with comorbidities {HLD, HTN} follows the typical trajectory we identified, while a patients with comorbidities {HLD, IFG} shows evidence to contrary, i.e. skipped "HTN", because HTN preceded IFG along the trajectory. A patient who has no comorbidity is assumed to follow the typical trajectory. We enumerate the atypical trajectories based on which conditions of the typical trajectory are "skipped".

Table 5.5 shows the typical and atypical trajectories with detailed information. Each row in the table corresponds to a patient group, presenting with a set of comorbidities at baseline. Patients with T2D at baseline are excluded, so we omit T2D from the table. The column "No." is simply an arbitrary identifier assigned to the group. We also show the total number of patients in this group and the number of cases, patients who developed T2D during the follow-up period. We assign these patient groups to trajectories, which we also show in the table. For instance, we assign No. 7 group to the atypical trajectory. Patients in group No. 7 present with HLD and IFG but not with HTN. We do not know whether they developed IFG or HLD first, but it is inconsequential: the fact that they have IFG and no HTN offers evidence that they did not follow the typical trajectory. In the typical trajectory, patients develop HTN before they develop IFG; thus, every patient with IFG should also present with HTN.

The number of patients who followed atypical trajectories is substantial. From the table, we can see that 6,626 of the 43,509 patients (15%) followed atypical trajectories; and more importantly, 1,300 of 4,795 (27%) cases (patients who developed T2D) followed atypical trajectories.

5.5.3 Atypical trajectories and the risk of developing type 2 diabetes

To study whether the trajectory influences the patients' risk of progression to diabetes, we have built a regression model for incident diabetes which, besides the usual comorbidities, also includes the trajectory as an independent variable. Table 5.6 shows the predictors and their coefficient estimates. The predictors describing HLD, HTN, IFG

Table 5.6: Predictors and coefficient estimates from the type 2 diabetes predictive model.

Variable	Coefficient estimate (SE)	p-value
(Intercept)	-13.55 (0.44)	<.001
age	0.01 (0.00)	<.001
male	-0.16 (0.39)	<.001
gluc	0.11 (0.00)	<.001
hld.untx	0.38 (0.08)	<.001
hld.tx	0.29 (0.09)	<.001
htn.untx	0.19 (0.07)	.005
htn.tx	0.26 (0.06)	<.001
ifg.pre1	0.20 (0.07)	.005
ifg.pre2	0.00 (0.12)	.978
obese.ovrwght	0.10 (0.05)	.048
obese.obese	0.45 (0.05)	<.001
trajskip.htn	0.24 (0.08)	.002
trajskip.hl	-0.06 (0.13)	.650
trajskip.both	-0.54 (0.16)	<.001

and obesity are ordinal; their levels are ordered: "no" (no sign of disease) is less severe than "untx" (no treatment needed) and "untx" is less severe than "tx" (treated). The effect of each level is measured relative to the next lower level. For instance, the effect of hld.tx is measured relative to hld.untx: requiring treatment for hyperlipidemia (hld.tx) increases the log odds of progression to diabetes by .29 relative to patients who do not require treatment for hyperlipidemia (hld.untx).

Two of the atypical trajectories are significant. The atypical trajectory where patients skip HTN (patients with HLD and IFG but without HTN) increased the log odds of developing T2D by 0.24 compared to the typical trajectory. At first, this appears as if the lack of HTN increased the risk. The risk of T2D depends on the deterioration of the underlying metabolic health and the comorbidities, including HTN, are imperfect indicators of the deterioration of the metabolic health. One probable explanation is that the metabolic health of the patients with HLD and IFG (but without HTN) has deteriorated just as far as that of the patient with HTN, but their blood pressure has not yet increased sufficiently to meet the HTN diagnosis criteria. In such patients, the deterioration of the underlying condition, which typically manifests itself in the HTN disease, cannot contribute to the diabetes risk through the HTN variable, but its

detrimental effect is captured through the trajectory variable.

The atypical trajectory where patients skip both HLD and HTN altered the log odds of developing T2D by -0.54 (i.e. decrease it by .54) as compared to the typical trajectory. In patients following the typical trajectory, IFG increases the (log odds of the) risk of diabetes by .20 or by .20+.00=.20 depending how far the patient has progressed. However, in the absence of both HLD and HTN, these .20 overestimate the patients' actual risk, thus the trajectory adjusts the risk (downwards). In other words, for patients who neither present with HLD nor with HTN, elevated fasting glucose is not as damaging (with regard to diabetes) as we would expect assuming that IFG is independent of these conditions.

5.6 Discussion

In this work, we studied a novel approach to infer disease progression from Electronic Health Records (EHR). EHR with their large sample size and long follow-up time are becoming increasingly popular for population-based disease progression studies. However, unreliable diagnostic codes in the EHR data combined with the slow onset of many of the chronic diseases make it virtually impossible for us to directly observe trajectories, sequences in which the diseases develop. In this work, we described methods to sidestep or overcome these issues and discover interesting, previously unknown knowledge.

Specifically, we overcame the problem of unreliable diagnostic codes through phenotyping. Phenotyping refers to the combined use of diagnosis codes, lab results and medications to determine whether a patient presents with a condition at a given time. As phenotypes, we created an ordinal variable for each condition of interest, which, besides indicating the presence of a condition, also encoded its severity.

Solving the issue of onset dates is more challenging and we sidestepped it by simply assuming that the onset date occurred before the earliest recording date. Even if we managed to estimate the onset dates accurately, the pre-baseline period (and the 13 years of follow-up in general) was insufficient to observe entire trajectories. Instead of directly observing, we inferred the trajectories from snapshots. We used likelihood estimation to find a typical trajectory, which coincided with the trajectory that is commonly used for diabetes patient education.

We found that in the context of diabetes, some atypical trajectories had significant effect on the risk of progression to T2D. We observed that "skipping" HTN increased the risk of T2D by approximately the same amount as HTN itself; and we also observed that having high blood glucose without HTN or HLD is not as damaging as one would expect under the assumption that these conditions affect the risk of T2D independently. These are novel findings that were previously not known and not even studied.

Given the popularity of EHR data as a research platform, we expect larger sample sizes and longer follow-up times in the future. With the explosive growth of wearable health devices providing real-time physiological measurements, we may be able to infer the onset dates with better accuracy. Unfortunately, these improvements will not be able to completely eliminate the issues addressed in this work. The need for using historic data will remain and along with it the uncertainty in the historic data will remain, as well. Methods such as the ones proposed in this paper will still be required to help unlock the full potential of historic data.

Chapter 6

Discovery of multiple disease trajectories

6.1 Introduction

Disease trajectories[11], the order in which diseases develop, is an important problem in medical research. Multiple ongoing initiatives encourage the inclusion of a broad range of patient information into clinical decision making. Disease trajectories are an example of such currently underutilized but useful information. Recent studies showed that patients who share the same risk factors at an encounter, can experience substantially different outcomes[114, 115, 10] and these differences could be partially explained by the patient's progression patterns[11, 116] before the encounter in question. While this suggests that trajectories are valuable for the effective prevention and management of diseases[117, 118], currently there are no algorithms specifically to extract disease trajectories from electronic health records (EHR) data.

Computational methods that can be adapted to incorporate sequential information into clinical studies exist[35, 36, 37, 38, 39]. Sequential pattern mining (SPM)[26] is the most promising such method. The goal of SPM is to extract all sequences and their subsequences from a database of events. While SPM can extract sequences of clinical events (e.g. diseases, phenotypes, disease severity indicators[119, 120]), due to its fragmented nature, EHR data does not contain full trajectories; we can only extract partial trajectories. SPM offers no facility to combine partial trajectories into

full trajectories. In addition, SPM suffers from a tendency to yield an exponentially large number of trajectories. Although each trajectory is self-explanatory, the sheer number of them can ultimately render the results uninterpretable.

Causal structure discovery (CSD)[27, 28] offers an alternative approach to trajectory inference. Given a set of observations described by features (variables), the goal of CSD is to construct a partially directed graph with nodes representing features and edges representing causal relationships between the nodes they connect. The edges are oriented in the direction of cause to effect if this direction can be inferred. The problem with this approach for learning trajectories is twofold. First, it does not give us trajectories; it gives us a causal graph, from which we have to extract trajectories. Second, causality is neither required nor sufficient for explaining sequences of clinical events. For example, multiple diseases can have a latent common cause and these diseases can represent different stages of deterioration. These diseases do not cause each other but are associated, have a definite temporal ordering, and can be predictive of the next disease the patient is going to develop.

In this work, we focus on the development of a novel computational method specifically for learning disease trajectories from EHR data. These trajectories are sequences of clinical events (phenotypes or disease diagnoses) that have a definite temporal ordering along the trajectory and are associated with each other (but do not have to cause each other). The resulting set of trajectories have to be able to explain the observed clinical events.

Specifically, we have three key contributions. First, we propose a trajectory extraction algorithm. Inspired by sequential pattern mining, the proposed extraction algorithm extracts partial sequences from EHR data and combines them into full trajectories. Second, we develop filtering criteria based on the association and precedence of pairs of successive clinical events. Third, we propose a likelihood function for trajectories assessing the risk of developing a set of outcomes given a set of trajectories. The likelihood function is used to evaluate the goodness of fit of a set of disease trajectories.

We applied the proposed method in the context of Type 2 Diabetes (T2D) and its complications. T2D is a growing public concern in the United States[1]. 9.4% of the US population was reported to have T2D in 2015[2]. It is the 7th leading cause of death in the US[1] with accompanying severe chronic diseases including macro and

microvascular diseases[1, 5]. Since T2D and associated complications are progressive and non-reversible[3, 4], early identification and appropriate interventions are important[6, 7]. This makes T2D and its complications an ideal application to demonstrate our proposed methods.

EHR data of 53,409 patients at Mayo Clinic[83] were used as a development cohort. Diabetes trajectories were extracted and internally validated in terms of their ability to explain the observed partial trajectories. External validation was then performed on EHR data of 59,686 patients from an independent health system, M Health Fairview.

The remainder of this manuscript is organized as follows: in Section 6.2, we describe the data that are used in this study, present the proposed methods, and describe how we evaluated the models; in Section 6.3, we demonstrate and evaluate our proposed approaches on real EHR data; and in Section 6.4, we discuss the results.

6.2 Materials and Methods

6.2.1 Study design

Table 6.1: The baseline and its follow-up characteristics and phenotypes.

	Mayo Clinic		Fairview Health Services	
	2005∼2007	2012∼2014	2005∼2007	2012∼2014
Demographics, vitals and labs				
Age	46.2	–	48	–
Male	42.0 %	–	32.7 %	–
Death	0.0 %	2.6 %	0.0 %	1.4 %
Body Mass Index (BMI)	28.5	29.0	23.3	22.9
Systolic Blood Pressure (SBP)	122.4 mmHg	123.8 mmHg	120.9 mmHg	123.3 mmHg
Diastolic Blood Pressure (DBP)	73.0 mmHg	74.1 mmHg	73.8 mmHg	74.3 mmHg
High-Density Lipoprotein (HDL)	56.8 mg/dL	55.6 mg/dL	51.2 mg/dL	53.2 mg/dL
Low-Density Lipoprotein (LDL)	111.7 mg/dL	102.8 mg/dL	114.6 mg/dL	106.4 mg/dL
Triglycerides (TG)	138.1 mg/dL	136.0 mg/dL	135.2 mg/dL	136.9 mg/dL
Hemoglobin A1C (HbA1C)	6.0 %	6.1 %	6.2 %	6.2 %
Fasting Blood Glucose (FBG)	102.0 mg/dL	104.8 mg/dL	–	110.6 mg/dL
Random Blood Glucose (RBG)	117.6 mg/dL	–	102.7 mg/dL	105.7 mg/dL
Estimated Glomerular Filtration Rate (eGFR)	80.8	82.0	80.9	84.1
Phenotype				
Obesity (OB)	35.0 %	45.1 %	28.8 %	34.6 %
Hyperlipidemia (HLD)	34.2 %	47.6 %	37.3 %	55.5 %
Hypertension (HTN)	43.4 %	57.6 %	42.7 %	66.0 %
Impaired Fasting Glucose (IFG)	23.8 %	37.9 %	13.7 %	27.8 %
Type 2 Diabetes (T2D)	6.2 %	10.9 %	8.0 %	14.4 %
Chronic Renal Failure (CRF)	5.9 %	11.2 %	6.0 %	13.8 %
Peripheral Vascular Disease (PVD)	2.1 %	4.9 %	1.3 %	5.6 %
Coronary Artery Disease (CAD)	5.9 %	10.7 %	5.1 %	12.8 %
Myocardial Infarction (MI)	1.6 %	3.6 %	0.6 %	4.1 %
Cerebrovascular Accident (CVA)	1.8 %	5.2 %	1.9 %	7.6 %
Congestive Heart Failure (CHF)	0.2 %	2.4 %	1.0 %	4.5 %

A retrospective observational study was conducted. We utilized two large cohorts of 53,409 Mayo Clinic and 59,686 M Health Fairview primary care patients aged 18 or over in 2005. We collected patients' medical histories, including demographics, diagnosis codes, vital signs, laboratory test results, and prescriptions. Patients' medical history is segmented into two distinct three-year windows (2005-2007 and 2012-2014) with a four-year inter-window gap. We use the first window as the baseline of our study, and the second window as a follow-up period to see the disease progressions. Phenotypes were identified within each window in a cumulative manner: chronic conditions present in the baseline window are carried forward to the follow-up window. Table A.1 in the Appendix lists the precise definition of the phenotypes. Table 6.1 describes the cohorts in each window.

6.2.2 Trajectory discovery

Let $V = \{v_1, \ldots, v_N\}$ be a set of all events in the EHR data where each event v represents a phenotype or a disease from Table 6.1, and N is the number of total phenotypes. In the followings, we use the terms events, diseases, and phenotypes interchangeably. A trajectory τ is a sequence of events

$$\tau = v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow v_l$$

where $\rightarrow$ denotes temporal succession: the event on the left-hand side is followed by the event on the right-hand side. We assume that each patient follows a single trajectory.

Let $x_i^{(t)}$ be an observation vector of all events in V for a patient i in time window t. Specifically, $x_{i,j}^{(t)} = 1$ indicates whether the patient i has developed the event (disease) v_j during or before the time window t while $x_{i,j}^{(t)} = 0$ indicates that the patient i has had no indication of the disease v_j during or before the time window t. For convenience, we denote $x_{i,j}^{(t)} = 1$ as $x_{i,j}^{(t)}$ and $x_{i,j}^{(t)} = 0$ as $\overline{x}_{i,j}^{(t)}$ respectively. Since all diseases are progressive and non-reversible, $x_{i,j}^{(t)} = 1$ implies $x_{i,j}^{(t+1)} = 1$ for all t.

Patient i follows a trajectory τ if and only if for all time t during the study period, the set $s = \left\{ v_j \mid x_{i,j}^t \right\}$ of events that patients had developed by time t is identical to the first $\|s\|$ diseases on the trajectory τ.

Algorithm for extracting disease trajectories from EHR data

The proposed extraction algorithm obtains a trajectory set by concatenating progression pairs. A progression pair ρ is an ordered pair of a (potentially empty) set s of events and a single event v, $\rho = (s, v)$. In our application, s represents the set of diseases the patient has developed in or before the baseline window and v is a disease newly developed in the follow-up window. In our application, s represents the set of diseases the patient has developed in or before the baseline window and v is a disease newly developed in the follow-up window. Patients can develop multiple new diseases in which case multiple progression pairs are produced.

Trajectories are built up iteratively, concatenating two progression pairs in each iteration. Two progression pairs, $\rho_1 = (s_1, v_1)$ and $\rho_2 = (s_2, v_2)$, can be concatenated into $s_1 \to v_1 \to v_2$ if and only if $s_1 \cup \{v_1\} = s_2$. Initially, we consider progression pairs $\rho = (\phi, v_1)$ where $v_1 \in V$ and ϕ represents an empty set (patient has no diseases in the baseline window). Next, we extend the trajectory using the progression pair $(\{v_1\}, v_2)$ yielding a trajectory $v_1 \to v_2$; then we further extend it using the progression pair $(\{v_1, v_2\}, v_3)$ into $v_1 \to v_2 \to v_3$, continuing until no applicable progression pair can be found. Algorithm 6.1 formally describes the process.

Criteria for filtering disease trajectories

The objective of the filtering criteria is to introduce the constraints that specify the relationship between two adjacent events on a disease trajectory. Consider a trajectory $\tau = v_1 \to \cdots \to v_m \to v_n \to \ldots$, and two temporally ordered events, v_m and v_n, v_n follows v_m. We introduce three heuristic-based criteria that capture the relationship between two successive events as follows:

- *Association*: v_m and v_n are temporally associated given τ if and only if developing v_n is more likely in the presence of v_m than in its absence:

$$\Pr\left(v_m^{(1)}, v_n^{(1)} \mid v_m^{(0)}, \overline{v}_n^{(0)}, \tau\right) > \Pr\left(\overline{v}_m^{(1)}, v_n^{(1)} \mid \overline{v}_m^{(0)}, \overline{v}_n^{(0)}, \tau\right).$$

- *Strict precedence*: v_m precedes v_n given τ if and only if among patients with both

Algorithm 6.1: Extracting disease trajectories from EHR data

Input: $X^{(0)}$, $X^{(1)}$, *threshold*
Output: T

Function `ExtractProgressionPairs`($X^{(0)}$, $X^{(1)}$, *threshold*):
  $S :=$ distinct events on $X^{(0)}$
  for $s \in S$ **do**
    for $v \in V$ **do**
      $\rho := (s, v)$
      if $\sigma(\rho, X^{(0)}, X^{(1)}) > threshold$ **then**
        $\mathrm{P} := \mathrm{P} \cup \{\rho\}$
      end
    end
  end
  return P

Function `GrowTrajectories`(τ, P, T):
  $\mathrm{P}' := \{\forall \rho = (s, v) \mid (\rho \in \mathrm{P}) \wedge (s = \{v \mid v \in \tau\})\}$
  if $\mathrm{P}' \neq \phi$ **then**
    for $\rho \in \mathrm{P}'$ **do**
      $\tau' := \underbrace{\ldots}_{\tau} \rightarrow v$ where v is in $\rho = (s, v)$
      $\mathrm{T} :=$ `GrowTrajectories`(τ',P,T)
    end
  else
    $\mathrm{T} := \mathrm{T} \cup \{\tau\}$
  end
  return T

Function `ExtractTrajectories`($X^{(0)}$, $X^{(1)}$, *threshold*):
  $\mathrm{P} :=$ `ExtractProgressionPairs`($X^{(0)}$, $X^{(1)}$, *threshold*)
  $\mathrm{T} := \phi$
  for $v \in V$ **do**
    if $(\phi, v) \in \mathrm{P}$ **then**
      $\tau := \phi \rightarrow v$
      $\mathrm{T} :=$ `GrowTrajectories`(τ,P,T)
    end
  end
  return T

v_m and v_n, v_n is more likely than v_m to be the newly developed disease

$$\Pr\left(v_m^{(0)}, \overline{v}_n^{(0)} \mid v_m^{(1)}, v_n^{(1)}, \tau\right) > \Pr\left(\overline{v}_m^{(0)}, v_n^{(0)} \mid v_m^{(1)}, v_n^{(1)}, \tau\right).$$

- *Partial precedence*: v_m partially precedes v_n given τ if and only if

$$\Pr\left(v_m^{(0)}, \overline{v}_n^{(0)} \mid v_m^{(1)}, v_n^{(1)}, \tau\right) \geq \Pr\left(\overline{v}_m^{(0)}, v_n^{(0)} \mid v_m^{(1)}, v_n^{(1)}, \tau\right).$$

Partial precedence differs from strict precedence in that it allows for the two probabilities to be statistically equal, meaning that v_m and v_n developed concurrently.

A trajectory is filtered out if any two successive events on the trajectory do not satisfy the criteria. One or more criteria can be applied at the same time. Specifically, we will consider the following combinations: *association (A)*, *strict precedence (P_S)*, *partial precedence (P_P)*, *association and strict precedence ($A\&P_S$)*, *association and partial precedence ($A\&P_P$)*. Note that we do not apply strict precedence and partial precedence at the same time.

Likelihood function for assessing the risk of developing a set of outcomes along a trajectory

The (partial) likelihood, the joint probability of observing the outcomes (presence or absence of all diseases in the follow-up window) given the baseline diseases and a set of trajectories, is key to assessing the goodness-of-fit of the trajectory set and for ranking the trajectories. The need to define a new likelihood arises because (i) outcomes are not independent of each other, (ii) the dependence structure among the outcomes depends on which trajectories the patient follows and (iii) we may not be able to uniquely determine at baseline which trajectory the patient follows, and (iv) the dependence structure is simple since the trajectories are sequences.

Partial likelihood Let $x_i^{(0)}$ and $x_i^{(1)}$ denote the state of all diseases (present or absent) for patient i in the baseline window and in the follow-up window, respectively. We can partition $x_i^{(1)}$ based on when the patient developed or is expected to develop the

corresponding diseases

$$x_i^{(1)} = \{p_i, n_i, \bar{f}_i, \bar{u}_i, e_i\},$$

where p_i is the set of diseases that patient i developed in or before the baseline window (**p**re-existing), n_i is the set of diseases patient i newly developed in the follow-up window (**n**ew), f_i is the set of diseases the patient is likely to develop along the trajectory in the future (**f**uture), u_i is the set of diseases that the patient is not expected to develop along this trajectory (**u**nlikely), and e_i is the set of diseases that the patient was not supposed to develop following trajectory τ (**e**rror). The bar above f_i and u_i indicates that none of these diseases is present in the follow-up window. If the patient indeed follows τ, e_i is empty.

The probability of observing the set of follow-up diseases (described by $x_i^{(1)}$) given the baseline disease states $x_i^{(0)}$ while following trajectory τ is

$$\begin{aligned}
&\Pr(x_i^{(1)} | x_i^{(0)}, \tau) \\
&= \Pr(p_i, n_i, \bar{f}_i, \bar{u}_i \mid x_i^{(0)}, \tau) \\
&= \Pr(p_i \mid n_i, \bar{f}_i, \bar{u}_i, x_i^{(0)}, \tau) \times \Pr(n_i \mid \bar{f}_i, \bar{u}_i, x_i^{(0)}, \tau) \times \qquad (6.1) \\
&\quad \Pr(\bar{f}_i \mid \bar{u}_i, x_i^{(0)}, \tau) \times \Pr(\bar{u}_i \mid x_i^{(0)}, \tau) \\
&= \Pr(p_i \mid x_i^{(0)}) \times \Pr(n_i \mid x_i^{(0)}, \tau) \times \qquad (6.2) \\
&\quad \Pr(\bar{f}_i \mid n_i, x_i^{(0)}, \tau) \times \Pr(u_i \mid \tau) \\
&= \Pr(n_i \mid \bar{f}_i, x_i^{(0)}, \tau) \times \Pr(\bar{f}_i \mid x_i^{(0)}, \tau) \qquad (6.3)
\end{aligned}$$

Eq. (6.1) follows from the chain rule of probabilities and assumes that e_i is empty. If e_i is not empty, we handle it later as an error condition. Eq. (6.2) holds true for the following reasons. First, p_i, the set of pre-existing diseases, only depends on the baseline disease states $x_i^{(0)}$ (in the absence of errors). Since these are chronic diseases, which can be controlled but not cured, $\Pr(p_i \mid x_i^{(1)}) = 1$, which we use in Eq. (6.3). As patients progress from the baseline diseases along the trajectory τ, they develop the new diseases n_i during the follow-up, leaving $\bar{f}_i$ to develop later. Accordingly, n_i depends on the trajectory τ, the baseline disease states $x_i^{(1)}$ and $\bar{f}_i$ showing how far along τ the patient progressed. Once we account for the dependence between n_i and $\bar{f}_i$, $\bar{f}_i$ only depends on τ: the set of diseases that are possible along τ. Similarly, $\bar{u}_i$, the set of

diseases that are not possible along τ only depend on the trajectory and $\Pr(\bar{u}_i|\tau) = 1$, which we use in Eq. (6.3).

The two terms in Eq. (6.3) can be computed as follows. With $n_i = \{n_{i,1}, n_{i,2}, \ldots, n_{i,k}\}$, where diseases are indexed in the same order as they appear on the trajectory τ,

$$\begin{aligned}
&\Pr(n_i \mid \bar{f}_i, x_i^{(0)}, \tau) \\
&= \Pr(n_{i,1} \mid \bar{f}_i, x_i^{(0)}, \tau) \times \\
&\quad \Pr(n_{i,2} \mid n_{i,1}, \bar{f}_i, x_i^{(0)}, \tau) \times \cdots \times \\
&\quad \Pr(n_{i,k} \mid n_{i,1}, \ldots, n_{i,k-1}, \bar{f}_i, x_i^{(0)}, \tau).
\end{aligned}$$

The first new disease, $\Pr(n_{i,1} \mid \bar{f}_i, x_i^{(0)}, \tau)$ can be directly observed from data. When patients develop multiple new diseases, subsequent new disease probabilities $\Pr(n_{i,k} \mid n_{i,1}, \ldots, n_{i,k-1}, \bar{f}_i, x_i^{(0)}, \tau)$ $(k \geq 2)$ are approximated as if the patient had already developed $n_{i,1}, \ldots, n_{i,k-1}$ by baseline.

Finally, the last term in Eq. (6.3) is computed as

$$\Pr(\bar{f}_i \mid n_i, x_i^{(0)}, \tau) = \Pr(\bar{f}_1 \mid n_i, x_i^{(0)}, \tau),$$

where f_1 is the first disease in f_i along τ. If the patient has not progressed to f_1 along τ, then $\Pr(\bar{f}_j \mid \bar{f}_1, \tau) = 1$ for all subsequent $(j \geq 2)$ diseases along the trajectory.

Errors in trajectory assignment When the patient is assigned an incorrect trajectory, we may find that e_i is not empty or that some of the diseases along the trajectory are "skipped". In such cases, we set $\Pr(x_i^{(1)} \mid x_i^{(0)}, \tau)$ to be close to 0, indicating that the follow-up disease states are impossible given the trajectory.

Under-determined trajectory assignment Multiple trajectories may share a common prefix therefore it may not be possible to assign the patient to a unique trajectory based on the baseline disease states. Let T denote the set of all trajectories.

$$\begin{aligned}
&\Pr(x_i^{(1)} \mid x_i^{(0)}, T) \\
&= \sum\nolimits_{\tau \in T} \Pr(x_i^{(1)} \mid x_i^{(0)}, \tau) \times \Pr(\tau \mid x_i^{(0)}),
\end{aligned} \tag{6.4}$$

where

$$\Pr(x_i^{(1)} \mid x_i^{(0)}, \tau)$$
$$= \begin{cases} 0 & \text{if patient does not follow } \tau \text{ at baseline} \\ 0 & \text{if assignment of } \tau \text{ to patient } i \text{ is in error} \\ \Pr(x_i^{(1)} \mid x_i^{(0)}, \tau) & \text{from Eq. (6.3) otherwise} \end{cases}$$

and $\Pr(\tau \mid x_i^{(0)})$ is determined from the training data as the proportion of patients who end up following τ in the follow-up window among those who share the same baseline disease state $x_i^{(0)}$. In computing $\Pr(\tau \mid x_i^{(0)})$, patients for whom the trajectory cannot be uniquely determined even in the follow-up window are ignored.

Likelihood The likelihood of the trajectory set T given the baseline and follow-up disease states in the population is computed as

$$\mathcal{L}(T) = \prod_i \Pr(x_i^{(1)} \mid x_i^{(0)}, T) \tag{6.5}$$

using eq. (6.4). Note that this is a *partial* likelihood since we do not compute $\Pr(x_i^{(0)})$. We chose not to compute it because the trajectories are based on observed changes in disease states and such changes cannot be reliably observed before the baseline.

Ranking Algorithm We rank disease trajectories using the forward selection method. We begin with an empty set of trajectories, and repeat, including the most significant trajectory, i.e., the trajectory that can maximize the likelihood most, into the set until there is no trajectory available. We rank in the order of inclusion.

6.2.3 Competing trajectory extraction methods

Although ours is the first method specifically for extracting disease trajectories from EHR records, for comparative evaluation, we propose three adaptations of Bayes network structure discovery algorithms to trajectory discovery.

- *BN-I (Causal inference-based method)*: The causal network (graph) is learned by

the max-min hill-climbing (MMHC) algorithm[121] and trajectories are extracted by a depth-first traversal of the graph.

- *DBN-I (Causal inference-based method)*: We apply a Dynamic Bayes Network (DBN) structure learning algorithm to the data and extract trajectories through depth-first traversal of the unrolled graph[122, 123].

- *DBN-G (Generative model-based method)*: We discover the causal network from the original data using the DBN discovery algorithm (as we did for DBN-I). Then we generate synthetic data from the DBN[124] containing full trajectories, and finally, we extract full disease trajectories directly from synthetic sequential data.

6.2.4 Experimental setup

We evaluate the trajectory sets based on their ability to explain the diseases observed in the follow-up window (*follow-up diseases*) given the diseases in the baseline window (*baseline diseases*) and the set of trajectories. We use the log-likelihood as the metric of how well a trajectory set explains the follow-up diseases. If a trajectory is correct and can "predict" the subsequent diseases (i.e. follow-up diseases) correctly, the probability of observing the set of follow-up diseases is high, resulting in a small negative log-likelihood; otherwise, the follow-up disease set is unexpected and the probability is low (resulting in a large negative log-likelihood).

Evaluation method

Internal evaluation. We generated 1,000 datasets via bootstrap resampling on the Mayo Clinic (MC) data. Each bootstrapped dataset consists of a training set and an out-of-bag test set. Trajectories were extracted from the training set and evaluated on the out-of-bag test set. **External evaluation**. Trajectories were extracted from the 1,000 bootstrap replicas of the MC data set and evaluated on the entire FHS data set. Reported are the mean of the 1,000 log-likelihood estimates and its empirical 95% confidence interval.

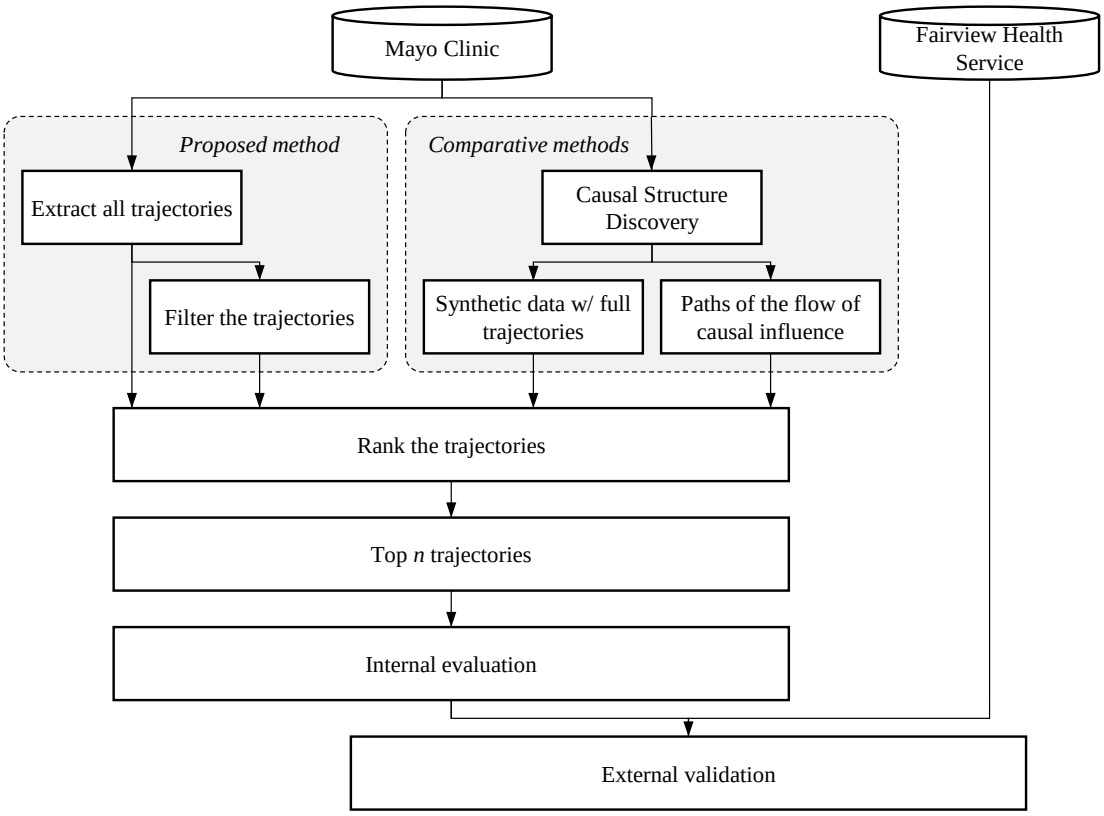


Figure 6.1: An overview of the experimental setup.

Experiments

We conduct three experiments in total. First, we evaluated the Proposed method (without filtering) and the three comparison algorithms on extracting disease trajectories from EHR data. This experiment results in four sets of trajectories, each set named after the algorithm that produced it: *Proposed*, *BN-I*, *DBN-I*, and *DBN-G*. We evaluate the four trajectory sets using log-likelihood as the metric internally and externally as described previously.

These algorithms can potentially produce large trajectory sets, impeding interpretation. Next, we use the proposed ranking algorithm to select the top n trajectories ($n = 1, \ldots, 20$) and evaluate the resulting n trajectory sets (for each algorithm) as described above. Only the training data is used for ranking.

Third, we evaluate the proposed filtering criteria. Six trajectory sets are obtained

after applying combinations of filtering methods to the disease trajectories produced by the Proposed method as follows: *unfiltered (U)*, *association (A)*, *strict precedence (P_S)*, *partial precedence (P_P)*, *association and strict precedence ($A\&P_S$)*, *association and partial precedence ($A\&P_P$)*. We follow the same evaluation process as above. Figure 6.1 describes an overview of the experimental setup.

6.3 Results

6.3.1 Evaluation of the algorithms for extracting disease trajectories

Table 6.2: The number of disease trajectories obtained by four extraction methods.

Method	The number of trajectories	Log-likelihood	
		Mayo Clinic	Fairview Health Services
Proposed	5,005	-4.20 (-4.30, -4.10)	-5.73 (-5.85, -5.60)
BN-I	76	-20.95 (-21.13, -20.71)	-22.62 (-22.80, -22.47)
DBN-I	79	-20.49 (-20.68, -20.25)	-22.93 (-23.10, -22.76)
DBN-G	63	-9.14 (-9.31, -8.99)	-11.43 (-11.61, -11.24)

First, we evaluated the four algorithms (*Proposed*, *BN-I*, *DBN-I*, and *DBN-G*) for extracting disease trajectories from EHR data. Table 6.2 shows the number of trajectories in each trajectory set extracted from the Mayo Clinic (MC) data, the log-likelihood of the trajectory sets along with the 95% confidence interval on the MC data and externally on the M Health Fairview (FHS) data. Table 6.2 reveals that Proposed shows the highest log-likelihood, followed by *DBN-G*, and *BN-I* and *DBN-I*.

Some of the algorithms discovered very large sets of trajectories, hindering expert interpretation. We evaluated the four algorithms based on their performance on a smaller, more readily interpretable, set of trajectories. Figure 6.2 depicts the log-likelihood of the top n trajectories, $n = 1, \cdots, 20$. Each color and line type represents a different extraction algorithm and filtering criterion. The lists of the top 20 disease trajectories obtained by the four extraction algorithms can be found in Table A.2 to A.5 in the Appendix. In Figure 6.2 (left), *Proposed* shows the highest log-likelihood over the entire range of n on the MC data. *DBN-G* shows the same log-likelihood as the *Proposed*

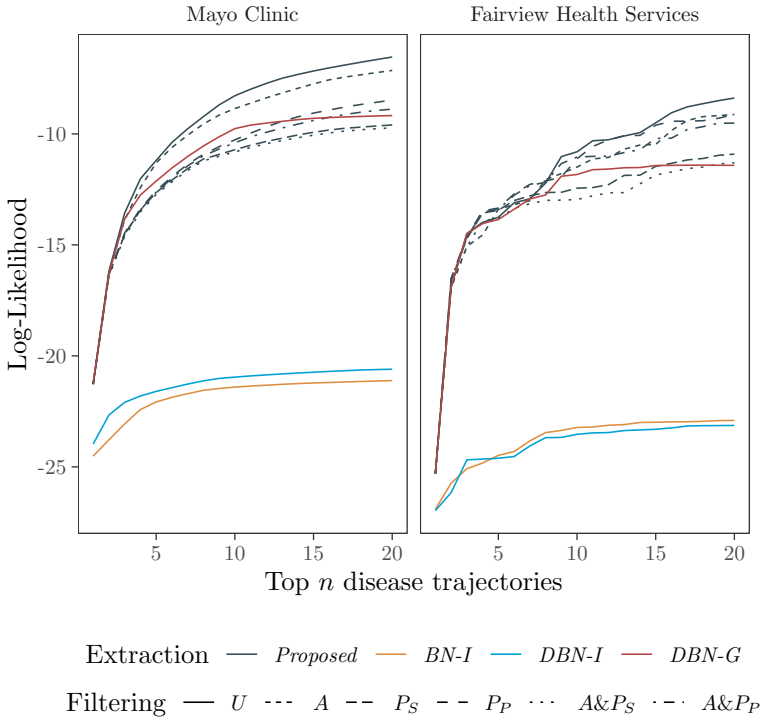


Figure 6.2: The log-likelihood of top n disease trajectories explaining the disease progressions most.

on the top (one) trajectory while the difference in log-likelihood between *Proposed* and *DBN-G* keeps increasing as we include more trajectories. *BN-I* and *DBN-I* show the lowest log-likelihood over the entire range. The top (one) disease trajectory on *BN-I* and *DBN-I* shows approximately 2.7 to 3.3 lower log-likelihood ($e^3 = 20$ times lower likelihood) than *Proposed*.

The slopes of *BN-I*, *DBN-I*, and *DBN-G* become flat after the top 4 to 12 trajectories while the slope of *Proposed* remains steeper, indicating that additional trajectories can help *Proposed* but not the three competing methods. Specifically, the increase in the log-likelihood as a result of adding an extra trajectory becomes less than 1.0% after the top 4 trajectories for *BN-I* and *DBN-I*, and the top 12 trajectories for *DBN-G*, while for *Proposed*, the change in log-likelihood is greater than 1.0% even after the top 20 trajectories. Using the top 20 trajectories, *Proposed* shows approximately 2.6 to 14.6

higher log-likelihood than *BN-I*, *DBN-I*, and *DBN-G*.

6.3.2 Evaluation of the filtering criteria

Table 6.3: The number of disease trajectories obtained after applying combinations of filtering methods to the *Proposed*.

Method	The number of trajectories	Log-likelihood	
		Mayo Clinic	Fairview Health Services
U	5,005	-4.20 (-4.30, -4.10)	-5.73 (-5.85, -5.60)
A	334	-5.42 (-5.54, -5.29)	-6.89 (-7.02, -6.76)
P_S	31	-9.08 (-9.24, -8.92)	-10.50 (-10.68, -10.33)
P_P	778	-6.93 (-7.08, -6.79)	-7.67 (-7.80, -7.51)
$A\&P_S$	26	-9.40 (-9.57, -9.23)	-11.01 (-11.20, -10.84)
$A\&P_P$	43	-8.11 (-8.27, -7.95)	-8.69 (-8.84, -8.52)

Rather than taking the top n trajectories, a more flexible way of reducing the number of trajectories and thus improving interpretability is to use the proposed filtering criteria. Six combinations of the filtering criteria (U the set of unfiltered trajectories, A, P_S, P_P, $A\&P_S$, and $A\&P_P$ as defined in Section 6.2.2) yielded six trajectory sets. For each trajectory set, Table 6.3 shows the number of trajectories, the log-likelihood with its 95% confidence interval internally on the MC data and externally on the FHS data. The use of filtering criteria achieved a substantial reduction in the number of trajectories (80% to 100-fold reduction), at the cost of reducing the log-likelihood by only 1.2 to 5.2. Among the filtered set of trajectories, $A\&P_S$ has the lowest (worst) log-likelihood, which is still 10 higher than *BN-I* or *DBN-I*. The log-likelihood of $A\&P_S$ is comparable to the third competing method, *DBN-G*, but achieves this similar performance (-9.4 vs -9.1 on MC) with 55% fewer trajectories (26 vs 63).

6.3.3 Trade-off

Figure 6.2 shows that an increasing number of trajectories typically yields better log-likelihood but impedes interpretation. Therefore, a tradeoff exists between interpretability and performance, which is controlled through the number of trajectories. In Figure

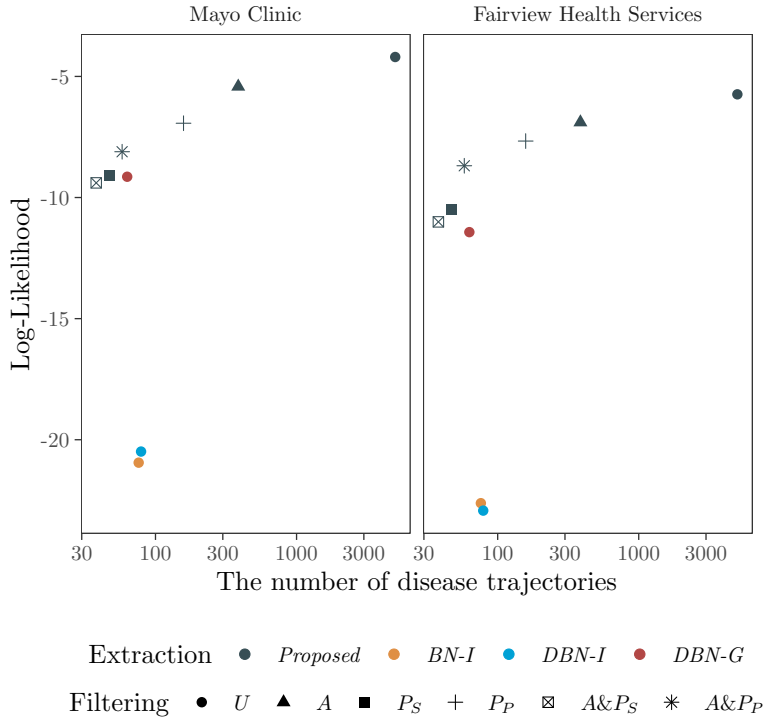


Figure 6.3: The log-likelihood of the full model, i.e., complete list of disease trajectories we obtained.

6.3, we visualize this tradeoff. Figure 6.3 shows the number of disease trajectories obtained by the extraction algorithms with/without filtering criteria on the horizontal axis and the corresponding log-likelihood on the vertical axis evaluated internally on the MC data (left pane) and externally on the FHS data (right pane). Again, each color and line type represents different extraction algorithms and filtering criteria.

6.3.4 External validation

Results of external validation are shown in the right pane of Figures 6.2 and 6.3. The observations from the internal validation are largely mirrored in the external validation: (i) higher number of trajectories typically lead to better log-likelihood; (ii) the Proposed method without filtering consistently outperforms the competing methods; (iii) there are filtering methods (P_S and $A\&P_P$) that achieve higher likelihood with fewer trajectories

than the competing methods. Small inconsistencies exist in Figure 6.2. For example, association-based filtering A outperforms the unfiltered set U for the top 5 trajectories. This suggests that the ranking of the trajectories is slightly different between MC and FHS, which is unsurprising given the significantly different patient populations that these two health systems serve.

6.4 Discussion

Ongoing initiatives in health care including evidence-based medicine, precision medicine, and learning health systems emphasize the use of all available data for decision making. One important type of information, which is almost completely ignored in today's clinical practice, is disease trajectories. Trajectories are useful in medical sciences for understanding disease progression patterns, in clinical practice for risk stratification, and in health care delivery for focusing preventive efforts on the diseases the patient is most likely to develop next along the trajectory. However, high performance in explaining follow-up outcomes, internal and external validity, interpretability, and correct ordering of diseases along each trajectory are critical for these applications.

Definitions of trajectories Trajectories can be defined in many different ways. Trajectories have been viewed before for example as time series of lab results[114, 115, 125]. In diabetes, trajectories of HbA1c have been shown to be predictive of outcomes decades before the onset of those outcomes[114, 115]. Lab results trajectories are trivially visualizable and interpretable, and can be integrated into risk models[125]. However, these trajectories only look at a single aspect of the patient's health, namely the condition that is described by the lab result in question.

In our work, we defined trajectories more broadly, providing a more comprehensive view of the multiple facets of complex conditions like diabetes. Our trajectories are sequences of clinical events (phenotypes) representing the worsening of the metabolic syndrome. Trajectories defined in this way are self-explanatory, offering a straightforward interpretation as long as the number of trajectories is not overwhelming. Since the clinical events can describe multiple facets of health, these trajectories can simultaneously encompass many aspects of a disease or a syndrome. We showed that trajectories

defined in this way and extracted through our novel algorithm can explain observed outcomes in two independent health systems equally well and substantially better than competing approaches.

High performance Log-likelihood assesses the ability of a set of trajectories to explain the observed outcomes (follow-up disease set). We have shown that our proposed method achieves a higher likelihood than the competing methods, but likelihood by itself is not intuitive making it difficult to assess how well the trajectory set predicts the outcomes. Over the first 10 trajectories, the probability of the observing the next outcome along the trajectory is on average 46.9% (IQR 29.4% to 63.1%) higher than assuming that the outcomes develop independently of each other (given the set of diseases that the patient had at baseline) and subsequent outcomes have even higher probability. This means that the trajectories enrich the patient subpopulation that follows them with positive outcomes, even given the baseline diseases. Since the baseline diseases are given, this enrichment is possible because diseases develop in a number of typical sequences and our algorithm successfully extracted these sequences.

Internal and external validity Our experimental setup through using out-of-bag samples ensures internal validity; and we have shown that the key conclusions remained valid through external validation.

Interpretability A set of 5,000 trajectories is not easily interpretable by a human expert. We presented two ways to reduce the number of trajectories. First, we considered the top n trajectories. Using a single trajectory, the log-likelihood is approximately -20 and using the entire trajectory set, the log-likelihood is -4.2, giving a range of 15.8 that can be explained by additional trajectories. We developed a ranking method that could select a small set of trajectories that could explain outcomes almost as well as the full set. Using $n = 10$ trajectories improves the likelihood from -20 (for a single trajectory) to -8, which is 76% of the possible improvement range; and using 20 trajectories improves the likelihood to -5, which is 94.9% of the possible improvement. A set of 10 or 20 trajectories are small enough to be interpretable, yet they explain outcomes with only 5% to 24% percent loss of log-likelihood relative to the full set of 5,000 trajectories.

The second way of reducing the trajectories was to use the proposed filtering criteria and we discuss their effect below.

Correct ordering of diseases The true set of trajectories that patients follow are not known. In the absence of gold standard knowledge, assessing the correctness of the trajectory set is difficult, and we present the trajectory sets in the Appendix for the interested readers. In lieu of a formal evaluation, we discuss two ordering errors that the competing methods made but our method avoided: developing complications in the absence of their risk factors; and developing risk factors *after* the onset of their complications. While these are both possible, they are highly atypical.

First, we look at developing complications without their risk factors. For instance, it is well established in the literature that hypertension (HTN) is a risk factor of chronic renal failure (CRF)[126], yet, the competing methods captured CRF without the development of HTN. This happened, because the causal inference-based methods identified HL as a common cause of HTN and CRF and attributed all the risk of HTN on CRF to HL. This could be true, but it caused the causal-based method to miss the fact that HTN is a risk factor of and precedes CRF. The proposed method avoided such mistakes because the association-based filtering criterion discards trajectories where associated risk factors such as HTN for CRF are missing.

Second, risk factors and complications are in the reverse order. As an example, obesity (OB) is known as one of the main risk factors of type 2 diabetes (T2D), and indeed, in our data sets 70% of diabetic patients have OB. However, trajectory sets obtained by causal inference-based methods include multiple trajectories in which T2D precedes OB. The *strict/partial precedence* criteria filter out such trajectories.

Generalizability . We presented our algorithm on the use-case of diabetes due to the availability of data for external validation and our clinical expertise with this disease. However, the proposed methods are more generally applicable to diseases where phenotype-based events can be defined, and their temporal ordering represents a deterioration of health.

6.5 Conclusion

Trajectories from the proposed algorithm explain disease progressions best among the compared algorithms. Precedence and/or association-based filtering reduced the number of trajectories significantly while preserving their ability to explain disease progressions in both cohorts and helped the proposed method order diseases along the trajectory more correctly than competing methods. Trajectories themselves are self-explanatory and the filtered sets are sufficiently concise to maintain interpretability by clinicians.

Chapter 7

Contributions to Science

7.1 Contribution to Health Informatics

In this thesis, we have developed a number of methodologies that directly help discover trajectories, and many of them are also more generally applicable to other domains.

First, our main contribution is new methods to discover disease trajectories from EHRs data. EHR data poses several key challenges to extract true trajectories due to inaccurate timestamps and limited follow-up. In order to address these challenges, we proposed trajectory discovery methods with two different approaches. The first approach uses likelihood estimation to find a typical trajectory and enumerates the atypical trajectories based on which conditions of the typical trajectory are "skipped". The second approach extracts a comprehensive list of trajectories from EHR data and selects meaningful trajectories using filtering criteria based on association and precedence of two succeeding diseases on trajectories. We claim that our two methods can overcome above challenges.

Second, we proposed a new knowledge-driven data representation for clinical data mining, including trajectory mining. A data representation for trajectory mining needs to be high interpretable representation and must help achieve high performance on analytics based on binary or event input data. We demonstrated that our proposed representation is interpretable, as the variables it uses follow clinical reasoning, and it offers the overall highest performance among competitive representations on association analysis. High performance on association analysis is particularly important as our

trajectory mining algorithm will largely be based on association analysis algorithms.

Third, we proposed a novel approach to estimate the onset time of disease from intermittent observations with informative censoring. Many research questions rely on the accurate estimation of onset times for diseases. Because of the inherent limitations of EHR data, we cannot always directly extract this information from EHRs as they contain the recorded time for diseases rather than the onset time. We demonstrated the effectiveness of the proposed method through a concrete application, in which we estimate the onset time of T2D based on the intermittently observed trajectory of HbA1c to optimize the time when a patient needs to come back for diabetes screening.

7.2 Anticipated contributions to Medicine

Diabetes is heterogeneous, its manifestation can differ from patient to patient; patients belonging to separate subtypes of diabetes show different progression over time. Understanding the heterogeneity needs to be a key part of providing personalized diabetes care. However, the majority of existing studies show limited capabilities in differentiating patients. Existing studies only assess the patient's conditions at a specific point in time but do not fully utilize all information such as patient's medical history. Therefore, these studies show insufficient abilities of explicitly modeling population heterogeneity or identifying groups that benefit for an intervention. In this thesis, we aim to extract Type 2 Diabetes (T2D) trajectories, temporal sequences of events towards T2D. The followings are anticipated contributions to medicine.

First, T2D trajectories allow us a precise way to quantify the risk of T2D from subpopulations. Assessing the risk correctly is especially important for improving outcomes. The overestimation of T2D risk can lead to unnecessary medical expenses due to aggressive intervention including shortening follow-up periods, and the underestimation can lead to the failures of providing proper intervention in a timely manner resulting in the worsening of the patient's condition.

Second, T2D trajectories allow us to assess a likelihood of upcoming events given patient's current conditions. T2D, its comorbidities and complications progress slowly over time, but progression is not reversible in general. Therefore, it is important to forecast the impending complications so that we can proactively help patients avoid or

delay progression. Each disease has its own risk models and guidelines, yet, there is no risk model or guideline that can cover these diseases comprehensively.

Third, T2D trajectories allow us to define the intervention group precisely. One of the common clinical practice issues is that aggressive intervention without a comprehensive understanding of the patient's health can cause adverse events[127]. On the other hand, aggressive intervention without a comprehensive understanding of the patient's health can lead to a waste of medical resources. We expect that our proposed trajectories can help physicians and other health professionals select only patients who need treatment. This can improve the outcome and minimize adverse events as well.

Finally, T2D trajectories can be integrated into clinical decision support systems (CDSS). The CDSS can trace patient's medical history in EHRs and can infer trajectories that are likely to follow. Therefore, we expect that the CDSS can assist physicians and other health professionals with clinical decision tasks mentioned above efficiently and effectively in real time.

Chapter 8

Summary and Conclusion

In this thesis, we have conducted several studies aimed to discover type 2 diabetes (T2D) trajectories from Electronic Health Records (EHR) data.

The first study proposed a new clinical representation to make data more clinically understandable and to enrich it with clinical knowledge to support personalized healthcare. The proposed representation summarized numerous facets of a disease into a single hierarchical variable where the hierarchy replicates the clinical reasoning steps of determining the severity of a certain disease. We demonstrated that our proposed representation is interpretable, as the variables it uses follow clinical reasoning, and it offers the overall highest performance among competitive representations on two common analytic tasks, regression and association analysis.

The second study aimed to develop a predictive model to estimate onset time of slow onset diseases primarily based on their defining laboratory results. In this work, we modeled the patients' HbA1c trajectories through Bayesian networks (BN) to estimate the onset time of diabetes. The BN was applied to describe dependency relationships among variables, enabling the separation of the (unobservable) actual and observed HbA1c level. We demonstrated that the proposed model reflects the actual changes in HbA1c level well, and we also showed that the model has the ability to accurately estimate the time to the onset of diabetes.

The third study aimed to focus on whether the trajectory influences the risk of T2D. Maximum likelihood via the chain rule was used to find a typical trajectory among potential trajectories. We identified a typical trajectory that most people follow, which is a

sequence of diseases from hyperlipidemia (HLD) through hypertension (HTN), impaired fasting glucose (IFG) to T2D. Further, we found that the sequence of comorbidities that a patient followed was predictive of his/her T2D risk even after adjusting for the severities of the comorbidities.

In the fourth study, we proposed a new computational method for learning disease trajectories from EHR data. The proposed method consists of three parts: first, an algorithm for extracting trajectories from EHR data, second, three criteria for filtering trajectories, and third, a likelihood function for assessing the risk of developing a set of outcomes given a trajectory set. We confirmed that the proposed algorithm can extract the most comprehensive disease trajectory set available on data, and the use of the proposed filtering criteria selects a small subset of disease trajectories that are highly interpretable with a minimal loss of the ability to explain disease progressions in a cohort.

In conclusion, we demonstrated that the proposed methodologies can overcome the limitations of EHR data for trajectory mining, and we successfully developed a methodology that extracts clinically meaningful disease trajectories that can explain the observation well.

References

[1] Centers for Disease Control and Prevention. Diabetes Report Card 2014. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2014.

[2] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2017. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2017.

[3] Christopher D. Saudek. Can Diabetes Be Cured? *JAMA*, 301(15):1588, apr 2009.

[4] John B. Buse, Sonia Caprio, William T. Cefalu, Antonio Ceriello, Stefano Del Prato, Silvio E. Inzucchi, Sue McLaughlin, Gordon L. Phillips, R. Paul Robertson, Francesco Rubino, Richard Kahn, and M. Sue Kirkman. How Do We Define Cure of Diabetes? *Diabetes Care*, 32(11):2133–2135, nov 2009.

[5] Josephine M. Forbes and Mark E. Cooper. Mechanisms of diabetic complications. *Physiological reviews*, 93(1):137–88, jan 2013.

[6] Jaakko Tuomilehto, Jaana Lindström, Johan G. Eriksson, Timo T. Valle, Helena Hämäläinen, Pirjo Ilanne-Parikka, Sirkka Keinänen-Kiukaanniemi, Mauri Laakso, Anne Louheranta, Merja Rastas, Virpi Salminen, Sirkka Aunola, Zygimantas Cepaitis, Vladislav Moltchanov, Martti Hakumäki, Marjo Mannelin, Vesa Martikkala, Jouko Sundvall, and Matti Uusitupa. Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. *New England Journal of Medicine*, 344(18):1343–1350, may 2001.

[7] Diabetes Prevention Program Research Group. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *New England Journal of Medicine*, 346(6):393–403, feb 2002.

[8] 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, dec 2011.

[9] Ali Abbasi, Linda M. Peelen, Eva Corpeleijn, Yvonne T. van der Schouw, Ronald P. Stolk, Annemieke M. W. Spijkerman, Daphne L. van der A, Karel G. M. Moons, Gerjan Navis, Stephan J. L. Bakker, and Joline W. J. Beulens. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ*, 345(sep18 2):e5900–e5900, sep 2012.

[10] Gerben Hulsegge, Annemieke M. W. Spijkerman, Y T van der Schouw, Stephan J. L. Bakker, Ronald T. Gansevoort, Henriette A. Smit, and Wilhelmina M. M. Verschuren. Trajectories of metabolic risk factors and biochemical markers prior to the onset of type 2 diabetes: the population-based longitudinal Doetinchem study. *Nutrition & Diabetes*, 7(5):e270–e270, may 2017.

[11] Wonsuk Oh, Era Kim, M. Regina Castro, Pedro J. Caraballo, Vipin Kumar, Michael S. Steinbach, and Gyorgy J. Simon. Type 2 Diabetes Mellitus Trajectories and Associated Risks. *Big Data*, 4(1):25–30, mar 2016.

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

[13] JaWanna Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. Technical report, Office of the National Coordinator for Health Information Technology, Washington, D.C., 2016.

[14] Era Kim, Wonsuk Oh, David S. Pieczkiewicz, M. Regina Castro, Pedro J. Caraballo, and Gyorgy J. Simon. Divisive Hierarchical Clustering towards Identifying

Clinically Significant Pre-Diabetes Subpopulations. In *Proceedings of the 2014 AMIA Annual Symposium (AMIA'14)*, 2014.

[15] Pranjul Yadav, Lisiane Pruinelli, Andrew Hangsleben, Sanjoy Dey, Katherine Hauwiller, Bonnie L. Westra, Connie W. Delaney, Vipin Kumar, Michael S. Steinbach, and Gyorgy J. Simon. Modeling Trajectories for Diabetes Complications. In *Proceedings of the 4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining (DMMH-SDM'15)*, 2015.

[16] Li Li, Wei-Yi Cheng, Benjamin S. Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P. Bottinger, and Joel T. Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, 7(311):311ra174–311ra174, oct 2015.

[17] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 292(3):344–350, 2016.

[18] Naren Ramakrishnan, David A. Hanauer, and Benjamin J. Keller. Mining Electronic Health Records. *Computer*, 43(10):77–81, oct 2010.

[19] Krzysztof J. Cios and G. William Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24, sep 2002.

[20] Craig K. Enders. A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research. *Psychosomatic Medicine*, 68(3):427–436, may 2006.

[21] Stuart L. Silverman. From Randomized Controlled Trials to Observational Studies. *The American Journal of Medicine*, 122(2):114–120, feb 2009.

[22] Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. Resolving the Bias in Electronic Medical Records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pages 2171–2180, New York, New York, USA, 2017. ACM Press.

[23] Simon Lebech Cichosz, Mette Dencker Johansen, and Ole Hejlesen. Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. *Journal of Diabetes Science and Technology*, 10(1):27–34, 2016.

[24] Era Kim, Pedro J Caraballo, M Regina Castro, David S Pieczkiewicz, and Gyorgy J Simon. Towards more Accessible Precision Medicine : Building a more Transferable Machine Learning Model to Support Prognostic Decisions for Micro- and Macrovascular Complications of Type 2 Diabetes Mellitus. 2019.

[25] Andrew L. Beam, Arjun K. Manrai, and Marzyeh Ghassemi. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*, pages 6–7, jan 2020.

[26] Tian-Rui Li, Yang Xu, Da Ruan, and Wu-ming Pan. Sequential Pattern Mining*. In *Intelligent Data Mining*, pages 103–122. Springer Berlin Heidelberg, jul 2005.

[27] Yang Zhou. Structure Learning of Probabilistic Graphical Models: A Comprehensive Survey. nov 2011, 1111.6925.

[28] Kalia Orphanou, Athena Stassopoulou, and Elpida Keravnou. Temporal abstraction and temporal Bayesian networks in clinical domains: A survey. *Artificial Intelligence in Medicine*, 60(3):133–149, 2014.

[29] Paul Dagum and Adam Galper. Forecasting sleep apnea with dynamic network models. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pages 64–71, 1993.

[30] Sathyakama Sandilya and R. Bharat Rao. Continuous-time Bayesian modeling of clinical data. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM'04)*, pages 512–516, Philadelphia, PA, apr 2004. Society for Industrial and Applied Mathematics.

[31] Gustavo Arroyo-Figueroa and Luis Enrique Sucar. Temporal Bayesian Network of Events for Diagnosis and Prediction in Dynamic Domains. *Applied Intelligence*, 23(2):77–86, oct 2005.

[32] Huidong Jin, Jie Chen, Hongxing He, Graham J. Williams, Chris Kelman, and Christine M. O'Keefe. Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions. *IEEE Transactions on Information Technology in Biomedicine*, 12(4):488–500, jul 2008.

[33] Theodore Charitos, Linda C. van der Gaag, Stefan Visscher, Karin A. M. Schurink, and Peter J. F. Lucas. A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Systems with Applications*, 36(2):1249–1258, 2009.

[34] R. Marshall Austin, Agnieszka Onisko, and Marek J. Druzdzel. The Pittsburgh Cervical Cancer Screening Model: a risk assessment tool. *Archives of Pathology & Laboratory Medicine*, 134(5):744–750, 2010.

[35] Huidong Jin, Jie Chen, Hongxing He, Chris Kelman, Damien McAullay, and Christine M. O'Keefe. Signaling Potential Adverse Drug Reactions from Administrative Health Databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):839–853, jun 2010.

[36] G. Niklas Norén, Johan Hopstadius, Andrew Bate, Kristina Star, and I. Ralph Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361–387, may 2010.

[37] Elena Gatti, Davide Luciani, and Fabio Stella. A continuous time Bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, 24(4):496–515, dec 2012.

[38] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 705–714, New York, New York, USA, 2015. ACM Press.

[39] Fei Wang, Chuanren Liu, Yajuan Wang, Jianying Hu, and Guoqiang Yu. A Graph Based Methodology for Temporal Signature Identification from EHR. In

Proceedings of the 2015 AMIA Annual Symposium (AMIA'15), pages 1269–78, 2015.

[40] Serguei S.V. Pakhomov, Harry Hemingway, Susan A. Weston, Steven J. Jacobsen, Richard Rodeheffer, and Véronique L. Roger. Epidemiology of angina pectoris: Role of natural language processing of the medical record. *American Heart Journal*, 153(4):666–673, apr 2007.

[41] Li-wei Lehman, Mohammed Saeed, William Long, Joon Lee, and Roger Mark. Risk Stratification of ICU Patients Using Topic Models Inferred from Unstructured Progress Notes. In *Proceedings of the 2012 AMIA Annual Symposium (AMIA'12)*, volume 2012, pages 505–11, 2012.

[42] Wonsuk Oh, Pranjul Yadav, Vipin Kumar, Pedro J. Caraballo, M. Regina Castro, Michael S. Steinbach, and Gyorgy J. Simon. Estimating Disease Onset Time by Modeling Lab Result Trajectories via Bayes Networks. In *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI'17)*, pages 374–379, 2017.

[43] Fabienne C. Bourgeois, Karen L. Olson, and Kenneth D. Mandl. Patients Treated at Multiple Acute Health Care Facilities. *Archives of Internal Medicine*, 170(22):1989, dec 2010.

[44] Wei-Qi Wei, Cynthia L. Leibson, Jeanine E. Ransom, Abel N. Kho, Pedro J. Caraballo, High Seng Chai, Barbara P. Yawn, Jennifer A. Pacheco, and Christopher G. Chute. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association*, 19(2):219–224, mar 2012.

[45] Alan R. Saltiel and C. Ronald Kahn. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, dec 2001.

[46] Peter W. F. Wilson, James B. Meigs, Lisa Sullivan, Caroline S. Fox, David M. Nathan, and Ralph B. D'Agostino. Prediction of incident diabetes mellitus

in middle-aged adults: the Framingham Offspring Study. *Archives of Internal Medicine*, 167(10):1068–74, may 2007.

[47] Kurt George Matthew Mayer Alberti and Paul Z. Zimmet. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation. *Diabetic Medicine*, 15(7):539–553, jul 1998.

[48] Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 25(Supplement 1):S5–S20, jan 2002.

[49] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 27(Supplement 1):S5–S10, jan 2004.

[50] Daphne Sl Gardner and E. Shyong Tai. Clinical features and treatment of maturity onset diabetes of the young (MODY). *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 5:101–8, jan 2012.

[51] 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):1084–94, mar 2014.

[52] Centers for Disease Control and Prevention. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2014.

[53] Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011, 2011.

[54] James P. Boyle, Amanda A. Honeycutt, K. M. Venkat Narayan, Thomas J. Hoerger, Linda S. Geiss, Hong Chen, and Theodore J. Thompson. Projection of Diabetes Burden Through 2050: Impact of changing demography and disease prevalence in the U.S. *Diabetes Care*, 24(11):1936–1940, nov 2001.

[55] James P. Boyle, Theodore J. Thompson, Edward W. Gregg, Lawrence E. Barker, and David F. Williamson. Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Population Health Metrics*, 8(1):29, 2010.

[56] Wenya Yang, Timothy M. Dall, Pragna Halder, Paul Gallo, Stacey L. Kowal, Paul F. Hogan, and Matthew P. Petersen. Economic Costs of Diabetes in the U.S. in 2012. *Diabetes Care*, 36(4):1033–1046, apr 2013.

[57] Pang-Ning Tan, Michael S. Steinbach, and Vipin Kumar. Data. In *Introduction to Data Mining*. Pearson, 2006.

[58] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 35, pages 1798–1828, aug 2013, 1206.5538.

[59] Joerg O W Pelz, Alexander Stojadinovic, Aviram Nissan, Werner Hohenberger, and Jesus Esquivel. Evaluation of a peritoneal surface disease severity score in patients with colon cancer with peritoneal carcinomatosis. *Journal of Surgical Oncology*, 99(1):9–15, 2009.

[60] Thomas A. Medsger, Stefano Bombardieri, László Czirják, Raffaella Scorza, Alessandra Della Rossa, and Walter Bencivelli. Assessment of disease severity and prognosis. *Clinical and Experimental Rheumatology*, 21(3 Suppl 29):S42–6, 2003.

[61] Patrick S. Kamath, Russell H. Wiesner, Michael Malinchoc, Walter Kremers, Terry M. Therneau, Catherine L. Kosberg, Gennaro D'amico, E. Rolland Dickson, and W. Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.

[62] Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C. Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383, jan 1987.

[63] Jonathon Shlens. A Tutorial on Principal Component Analysis. apr 2014, 1404.1100.

[64] Joyce C. Ho, Joydeep Ghosh, Steven R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, 2013.

[65] Yejin Kim, Robert El-Kareh, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. Discriminative and Distinct Phenotyping by Constrained Tensor Factorization. *Scientific Reports*, 7(1):1114, 2017.

[66] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[67] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, jul 2006.

[68] Geoffrey E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11(10):428–434, 2007, 1102.0183.

[69] Cédric Rose, Cherif Smaili, and François Charpillet. A dynamic Bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*. IEEE, 2005.

[70] Pranjul Yadav, Michael S. Steinbach, Lisiane Pruinelli, Bonnie L. Westra, Connie W. Delaney, Vipin Kumar, and Gyorgy J. Simon. Forensic Style Analysis with Survival Trajectories. In *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM'15)*, pages 1069–1074. IEEE, nov 2015.

[71] Logan Dumitrescu, Marylyn D. Ritchie, Joshua C. Denny, Nihal M. El Rouby, Caitrin W. McDonough, Yuki Bradford, Andrea H. Ramirez, Suzette J. Bielinski, Melissa A. Basford, High Seng Chai, Peggy L. Peissig, David S. Carrell, Jyotishman Pathak, Luke V. Rasmussen, Xiaoming Wang, Jennifer A. Pacheco, Abel N. Kho, M. Geoffrey Hayes, Martha Matsumoto, Maureen E. Smith, Rongling Li, Rhonda M. Cooper-DeHoff, Iftikhar J. Kullo, Christopher G. Chute, Rex L. Chisholm, Gail P. Jarvik, Eric B. Larson, David Carey, Catherine A. McCarty, Marc S. Williams, Dan M. Roden, Erwin P. Bottinger, Julie A. Johnson, Mariza

de Andrade, and Dana C. Crawford. Genome-wide study of resistant hypertension identified from electronic health records. *PLOS ONE*, 12(2):e0171745, feb 2017.

[72] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining Association in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*, pages 207–216, 1993.

[73] Juliana Tolles and William J. Meurer. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, 316(5):762–774, 2016.

[74] David S. Moore, George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. W. H. Freeman, 6th edition, 2007.

[75] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008, 0701907v3.

[76] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pages 1495–1504, New York, New York, USA, 2016. ACM Press, 1602.05568.

[77] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(April):26094, may 2016.

[78] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017, 1607.07519.

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

[80] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noémie Elhadad, Stephen B. Johnson, and Albert M. Lai. A review of approaches to

identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, mar 2014.

[81] Centers for Disease Control and Prevention. Diabetes Report Card 2012: National and State Profile of Diabetes and Its Complications. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2012.

[82] Massimo Pietropaolo, Emma Barinas-Mitchell, and Lewis H. Kuller. The heterogeneity of diabetes: unraveling a dispute: is systemic inflammation related to islet autoimmunity? *Diabetes*, 56(5):1189–97, may 2007.

[83] Jennifer L. St Sauver, Brandon R. Grossardt, Barbara P. Yawn, L. Joseph Melton, Joshua J. Pankratz, Scott M. Brue, and Walter A. Rocca. Data resource profile: The rochester epidemiology project (REP) medical records-linkage system. *International Journal of Epidemiology*, 41(6):1614–1624, 2012.

[84] Andrew J. Landgraf and Yoonkyung Lee. Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. (1999), 2015, 1510.06112.

[85] Bing Liu, Wynne Hsu, Yiming Ma, and Blwhy Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 80–86, New York, NY, 1998.

[86] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care*, 41(Supplement 1):S13–S27, 2018.

[87] World Health Organization. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. Technical report, 2011.

[88] American Diabetes Association. Standards of Medical Care in Diabetes-2016. *Diabetes Care*, 39(Supplement_1), 2016.

[89] Clare Bradley and Jane Speight. Patient perceptions of diabetes and diabetes therapy: assessing quality of life. *Diabetes/Metabolism Research and Reviews*, 18(S3):S64–S69, sep 2002.

[90] Elbert S. Huang, Sydney E. S. Brown, Bernard G. Ewigman, Edward C. Foley, and David O. Meltzer. Patient Perceptions of Quality of Life With Diabetes-Related Complications and Treatments. *Diabetes Care*, 30(10):2478–2483, oct 2007.

[91] Edward W. Gregg, Haiying Chen, Lynne E. Wagenknecht, Jeanne M. Clark, Linda M. Delahanty, John Bantle, Henry J. Pownall, Karen C. Johnson, Monika M. Safford, Abbas E. Kitabchi, F. Xavier Pi-Sunyer, Rena R. Wing, and Alain G. Bertoni. Association of an Intensive Lifestyle Intervention With Remission of Type 2 Diabetes. *JAMA*, 308(23):2489, dec 2012.

[92] Alan J. Garber, Martin J. Abrahamson, Joshua I. Barzilay, Lawrence Blonde, Zachary T. Bloomgarden, Michael A. Bush, Samuel Dagogo-Jack, Ralph A. DeFronzo, Daniel Einhorn, Vivian A. Fonseca, Jeffrey R. Garber, W. Timothy Garvey, George Grunberger, Yehuda Handelsman, Robert R. Henry, Irl B. Hirsch, Paul S. Jellinger, Janet B. McGill, Jeffrey I. Mechanick, Paul D. Rosenblit, and Guillermo E. Umpierrez. Consensus statement by the American association of clinical endocrinologists and American college of endocrinology on the comprehensive type 2 diabetes management algorithm - 2016 executive summary. *Endocrine Practice*, 22(1):84–113, jan 2016.

[93] Janice A. Kolberg, T. Jorgensen, Robert W. Gerwien, Sarah Hamren, Michael P. McKenna, Edward Moler, Michael W. Rowe, Mickey S. Urdea, Xiaomei M. Xu, Torben Hansen, Oluf Pedersen, and Knut Borch-Johnsen. Development of a Type 2 Diabetes Risk Model From a Panel of Serum Biomarkers From the Inter99 Cohort. *Diabetes Care*, 32(7):1207–1212, jul 2009.

[94] Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):130–141, jan 2015.

[95] Richard Kahn, Peter Alperin, David M. Eddy, Knut Borch-Johnsen, John B. Buse, Justin Feigelman, Edward W. Gregg, Rury R. Holman, M. Sue Kirkman, Michael P. Stern, Jaakko Tuomilehto, and Nicholas J. Wareham. Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis. *The Lancet*, 375(9723):1365–1374, apr 2010.

[96] Gregory A. Nichols, Teresa A. Hillier, and Jonathan B. Brown. Progression From Newly Acquired Impaired Fasting Glusose to Type 2 Diabetes. *Diabetes Care*, 30(2):228–233, feb 2007.

[97] Peter J. F. Lucas, Linda C. van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214, 2004.

[98] David Heckerman. A Tutorial on Learning with Bayesian Networks. In *Innovations in Bayesian Networks*, pages 33–82. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[99] Nicholas J. Wareham and Stephen O'Rahilly. The changing classification and diagnosis of diabetes. *BMJ*, 317(7155):359–360, aug 1998.

[100] Monica Chan, Poh Lian Lim, Angela Chow, Mar Kyaw Win, and Timothy M. Barkham. Surveillance for Clostridium difficile infection: ICD-9 coding has poor sensitivity compared to laboratory diagnosis in hospital patients, Singapore. *PLoS ONE*, 6(1):8–11, 2011.

[101] Rachel L. Richesson, Shelley A. Rusincovitch, Douglas Wixted, Bryan C. Batch, Mark N. Feinglos, Marie Lynn Miranda, W. Ed Hammond, Robert M. Califf, and Susan E. Spratt. A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association*, 20(e2):e319–e326, dec 2013.

[102] Pornthep Tanpowpong, Sarabeth Broder-Fingert, Joshua C. Obuch, David O. Rahni, Aubrey J. Katz, Daniel A. Leffler, Ciaran P. Kelly, and Carlos A. Camargo. Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease. *Annals of Epidemiology*, 23(3):136–142, 2013.

[103] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, 2003.

[104] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[105] Mei Chen, Richard N. Bergman, and Daniel Porte. Insulin Resistance and β-Cell Dysfunction in Aging: The Importance of Dietary Carbohydrate. *The Journal of Clinical Endocrinology & Metabolism*, 67(5):951–957, nov 1988.

[106] Steven E. Kahn, Mark E. Cooper, and Stefano Del Prato. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *The Lancet*, 383(9922):1068–1083, mar 2014.

[107] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, feb 1996.

[108] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15:104–116, 2017.

[109] Pranjul Yadav, Michael S. Steinbach, Vipin Kumar, and Gyorgy J. Simon. Mining Electronic Health Records (EHRs): A Survey. *ACM Computing Surveys*, 50(6):1–40, jan 2018, 1702.03222.

[110] David M. Eddy and Leonard Schlessinger. Archimedes: A trial-validated model of diabetes. *Diabetes Care*, 26(11):3093–3101, nov 2003.

[111] Philip M. Clarke, Alistair M. Gray, Andrew Briggs, Andrew J. Farmer, Paul Fenn, Richard J. Stevens, David R. Matthews, Irene M. Stratton, and Rury R. Holman. A model to estimate the lifetime health outcomes of patients with Type 2 diabetes: The United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). *Diabetologia*, 47(68):1747–1759, 2004.

[112] Walter A. Rocca, Barbara P. Yawn, Jennifer L. St Sauver, Brandon R. Grossardt, and L. Joseph Melton. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clinic proceedings*, 87(12):1202–13, dec 2012.

[113] American Diabetes Association. Standards of Medical Care in Diabetes-2014. *Diabetes Care*, 37(Supplement_1):S14–S80, jan 2014.

[114] Iris Walraven, M. Ruth Mast, Trynke Hoekstra, A. P. Danielle Jansen, Amber A. W. A. van der Heijden, Simone P. Rauh, Femke Rutters, Esther van 't Riet, Petra J. M. Elders, Annette C. Moll, Bettine C. P. Polak, Jacqueline M. Dekker, and Giel Nijpels. Distinct HbA1c trajectories in a type 2 diabetes cohort. *Acta Diabetologica*, 52(2):267–275, apr 2015.

[115] Mette K. Beck, Anders Boeck Jensen, Annelaura Bach Nielsen, Anders Perner, Pope L. Moseley, and Søren Brunak. Diagnosis trajectories of prior multimorbidity predict sepsis mortality. *Scientific Reports*, 6(November):1–9, 2016.

[116] John D. Piette and Eve A. Kerr. The Impact of Comorbid Chronic Conditions on Diabetes Care. *Diabetes Care*, 29(3):725–731, mar 2006.

[117] Kristine Færch, Daniel R. Witte, Adam G. Tabák, Leigh Perreault, Christian Herder, Eric J. Brunner, Mika Kivimäki, and Dorte Vistisen. Trajectories of cardiometabolic risk factors before diagnosis of three subtypes of type 2 diabetes: a post-hoc analysis of the longitudinal Whitehall II cohort study. *The Lancet Diabetes & Endocrinology*, 1(1):43–51, sep 2013.

[118] Sanket S. Dhruva, Chenxi Huang, Erica S. Spatz, Andreas C. Coppi, Frederick Warner, Shu-Xia Li, Haiqun Lin, Xiao Xu, Curt D. Furberg, Barry R. Davis, Sara L. Pressel, Ronald R. Coifman, and Harlan M. Krumholz. Heterogeneity in Early Responses in ALLHAT (Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial). *Hypertension*, 70(1):94–102, jul 2017.

[119] Wonsuk Oh, Michael S. Steinbach, M. Regina Castro, Kevin A. Peterson, Vipin Kumar, Pedro J. Caraballo, and Gyorgy J. Simon. Evaluating the Impact of Data

Representation on EHR-based Analytic Tasks. In *Proceedings of the 17th World Congress of Medical and Health Informatics (MedInfo2019)*, 2019.

[120] Che Ngufor, Pedro J. Caraballo, Thomas J. O'Byrne, David Chen, Nilay D. Shah, Lisiane Pruinelli, Michael Steinbach, and Gyorgy Simon. Development and Validation of a Risk Stratification Model Using Disease Severity Hierarchy for Mortality or Major Cardiovascular Event. *JAMA Network Open*, 3(7):e208270, jul 2020.

[121] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[122] Seth Lloyd. Causality and Information Flow. In *Information Dynamics. NATO ASI Series (Series B: Physics)*, pages 131–142. 1991.

[123] Gregory F. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. In *Computation, Causation, and Discovery*, chapter 1, pages 3–62. 1999.

[124] Jim Young, Patrick Graham, and Richard Penny. Using bayesian networks to create synthetic data. *Journal of Official Statistics*, 25(4):549–567, 2009.

[125] Gyorgy J. Simon, Kevin A. Peterson, M. Regina Castro, Michael S. Steinbach, Vipin Kumar, and Pedro J. Caraballo. Predicting diabetes clinical outcomes using longitudinal risk factor trajectories. *BMC Medical Informatics and Decision Making*, 20(1):6, dec 2020.

[126] Suzanne Oparil, Maria Czarina Acelajado, George L. Bakris, Dan R. Berlowitz, Renata Cífková, Anna F. Dominiczak, Guido Grassi, Jens Jordan, Neil R. Poulter, Anthony Rodgers, and Paul K. Whelton. Hypertension. *Nature Reviews Disease Primers*, 4(1):18014, jun 2018.

[127] Michal Ozery-Flato, Liat Ein-Dor, Naama Parush-Shear-Yashuv, Ranit Aharonov, Hani Neuvirth, Martin S. Kohn, and Jianying Hu. Identifying and investigating unexpected response to treatment: A diabetes case study. *Big Data*, 4(3):148–159, 2016.

Appendix A

Supplementary

Table A.1: The baseline and its follow-up characteristics and phenotypes.

Abbreviation	Description	Definition
OB	Obesity	One or more BMI $\geq$ 30
HL	Hyperlipidemia	Any two or more Dx, LDL $\geq$ 130 mg/dL, Rx
HTN	Hypertension	Any two or more Dx, [SBP $\geq$ 140 mg/dL or DBP $\geq$ 90 mg/dL] Rx
IFG	Impaired fasting glucose	Any two or more Dx, [A1c $\geq$ 5.7 % or FG $\geq$ 100 mg/dL or RG $\geq$ 140 mg/dL], Metformin
DM	Type 2 diabetes	Any two or more Dx [A1c $\geq$ 6.5 or FG $\geq$ 125 mg/dL or RG $\geq$ 200 mg/dL], Rx
CRF	Chronic renal failure	[One or more Dx and one or more GFR 60 ml/min/1.73 m2] or [Three or more GFR 60 ml/min/1.73 m2]
PVD	Peripheral vascular disease	One or more Dx
CAD	Coronary artery disease	One or more Dx
MI	Myocardial infarction	One or more Dx
CVA	Cerebrovascular accident	One or more Dx
CHF	Congestive heart failure	One or more Dx

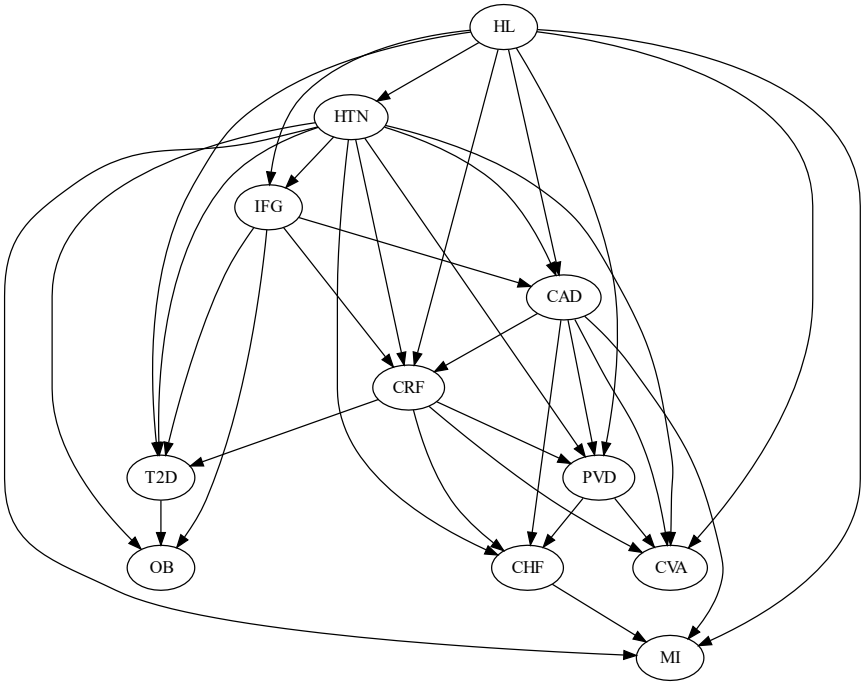
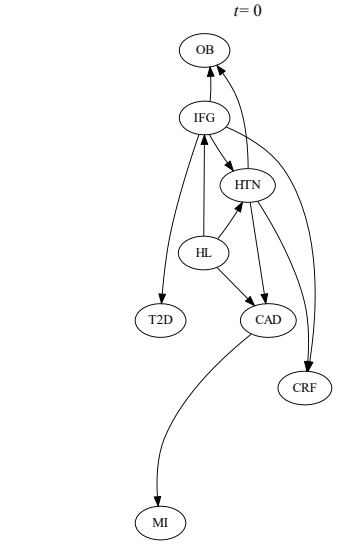
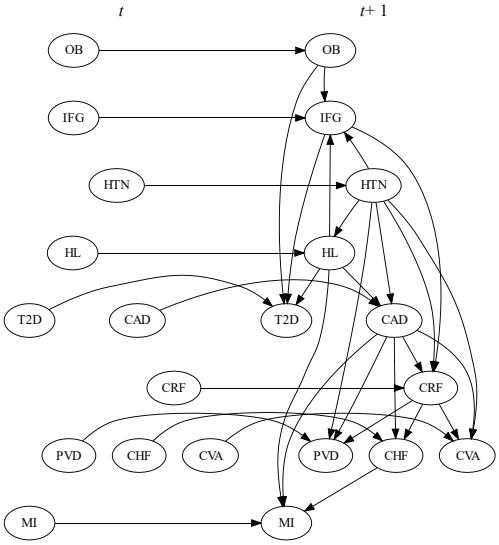


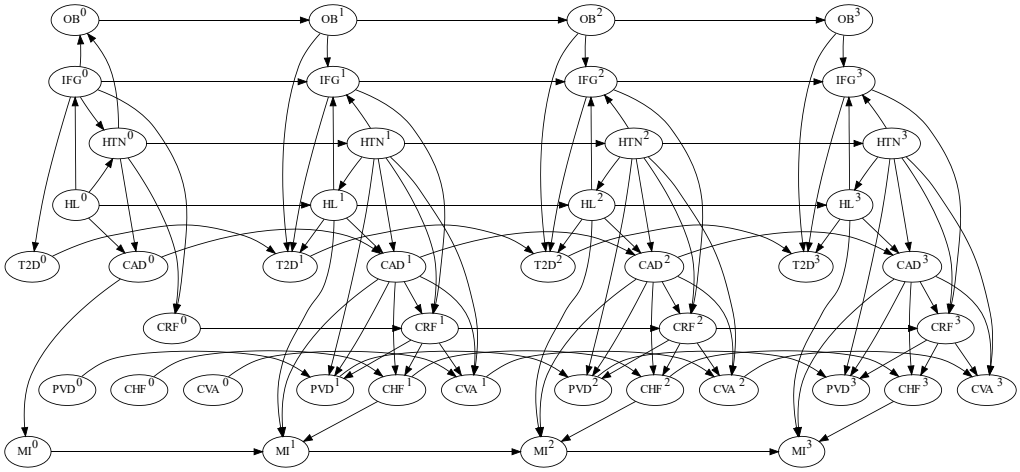
Figure A.1: Bayesian network.



(i) Initial network



(ii) Transfer network



(iii) Unrolled DBN for the process duration

Figure A.2: dynamic Bayesian network.

Table A.2: The top 20 disease trajectories obtained by the proposed disease trajectories extraction method.

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CAD → MI
02.	No condition → HTN → HL → IFG → T2D → OB → CAD → MI → PVD
03.	No condition → HL → OB → IFG → HTN → CRF → CAD → MI → CHF
04.	No condition → IFG → HTN → OB → HL → CAD → MI → T2D → CHF
05.	No condition → OB → IFG → HTN → T2D → HL → PVD → CAD
06.	No condition → HL → IFG → HTN → CRF → T2D → OB → CVA
07.	No condition → HTN → OB → IFG → CRF → HL → T2D → CAD → PVD
08.	No condition → OB → HL → HTN → CAD → IFG → T2D → CRF → CHF
09.	No condition → HL → HTN → CAD → IFG → CRF → PVD → CHF
10.	No condition → HTN → IFG → HL → OB → CVA → T2D
11.	No condition → HTN → CRF → HL → OB → IFG → PVD
12.	No condition → OB → IFG → HL → T2D → HTN → CRF → PVD
13.	No condition → HTN → HL → OB → IFG → PVD → CAD → T2D
14.	No condition → CAD → HL → HTN → MI → IFG → OB → CVA
15.	No condition → IFG → HL → T2D → HTN → CAD → CRF → PVD
16.	No condition → IFG → T2D → HTN → HL → CRF → OB → CHF
17.	No condition → HL → HTN → CRF → IFG → OB → CVA
18.	No condition → CVA → HTN → HL → IFG
19.	No condition → OB → HTN → CRF → HL → CAD → IFG → PVD
20.	No condition → HTN → PVD → HL → IFG → T2D → CRF

Table A.3: The top 20 disease trajectories obtained by the causal inference-based method applied with a Bayesian network (BN).

Rank	Trajectories
01.	No condition → HL → HTN → IFG → OB
02.	No condition → HL → HTN → IFG → T2D → OB
03.	No condition → HL → HTN → OB
04.	No condition → HL → IFG → OB
05.	No condition → HL → HTN → IFG → CRF → T2D → OB
06.	No condition → HL → HTN → IFG → CAD → CRF → T2D → OB
07.	No condition → HL → HTN → CAD → MI
08.	No condition → HL → HTN → CRF → CVA
09.	No condition → HL → IFG → T2D → OB
10.	No condition → HL → HTN → CVA
11.	No condition → HL → HTN → IFG → CAD → MI
12.	No condition → HL → HTN → PVD → CVA
13.	No condition → HL → CAD → MI
14.	No condition → HL → HTN → CAD → CRF → PVD → CHF → MI
15.	No condition → HL → HTN → IFG → CRF → PVD → CVA
16.	No condition → HL → HTN → IFG → CAD → PVD → CVA
17.	No condition → HL → HTN → IFG → CAD → CRF → PVD → CHF → MI
18.	No condition → HL → CVA
19.	No condition → HL → CRF → PVD → CHF → MI
20.	No condition → HL → HTN → IFG → CRF → CVA

Table A.4: The top 20 disease trajectories obtained by the causal inference-based method applied with a dynamic Bayesian network (DBN).

Rank	Trajectories
01.	No condition → HL → HTN → OB → IFG → T2D
02.	No condition → HL → IFG → HTN → CRF → PVD
03.	No condition → HL → HTN → IFG → T2D
04.	No condition → HL → IFG → OB → T2D
05.	No condition → HL → HTN → OB → IFG → CRF → PVD
06.	No condition → HL → IFG → HTN → CAD → CRF → PVD
07.	No condition → HL → HTN → CAD → MI
08.	No condition → HL → HTN → CRF → CVA
09.	No condition → HL → IFG → HTN → OB → T2D
10.	No condition → HL → HTN → CVA
11.	No condition → HL → HTN → IFG → CRF → CVA
12.	No condition → HL → IFG → HTN → CAD → MI
13.	No condition → HL → IFG → T2D
14.	No condition → HL → IFG → HTN → CVA
15.	No condition → HL → CAD → MI
16.	No condition → HL → HTN → PVD
17.	No condition → HL → HTN → CAD → CRF → PVD
18.	No condition → HL → IFG → HTN → PVD
19.	No condition → HL → HTN → OB → IFG → CRF → CVA
20.	No condition → HL → IFG → HTN → CAD → PVD

Table A.5: The top 20 disease trajectories obtained by the generative model-based method applied with a DBN.

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CAD → CHF
02.	No condition → HTN → HL → IFG → T2D → OB → CRF → CVA
03.	No condition → HL → OB → IFG → HTN → T2D → CRF → PVD
04.	No condition → HTN → IFG → OB → HL → T2D → CRF → PVD
05.	No condition → OB → IFG → T2D → HTN → HL → CRF → PVD
06.	No condition → HL → IFG → HTN → T2D → OB → CRF → CVA
07.	No condition → OB → HL → IFG → T2D → HTN → CAD → MI → CRF
08.	No condition → HTN → OB → HL → IFG → CRF → T2D → PVD
09.	No condition → HL → HTN → OB → IFG → T2D → CRF → CVA
10.	No condition → OB → HTN → IFG → T2D → HL → CRF → PVD
11.	No condition → OB → HL → HTN → IFG → T2D → CRF → CAD
12.	No condition → HTN → HL → IFG → OB → T2D → CAD → MI → CRF
13.	No condition → HTN → PVD → OB → HL → IFG → T2D
14.	No condition → OB → IFG → HTN → HL → T2D → CRF → CVA
15.	No condition → HL → IFG → T2D → OB → HTN → CRF → CVA
16.	No condition → OB → HTN → HL → IFG → T2D → CVA
17.	No condition → OB → HTN → HL → IFG → CRF → CVA
18.	No condition → HTN → OB → HL → IFG → T2D → CAD → PVD
19.	No condition → HL → IFG → OB → T2D → HTN → CRF → PVD
20.	No condition → OB → HTN → HL → IFG → T2D → PVD

Table A.6: Trajectories after applying the association-based filtering method.

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CVA
02.	No condition → HTN → HL → IFG → T2D → OB → CAD → MI
03.	No condition → HL → IFG → OB → HTN → CRF → CAD → MI → CHF
04.	No condition → IFG → HTN → OB → T2D → HL → CRF → CHF
05.	No condition → OB → HL → IFG → HTN → CAD → MI
06.	No condition → OB → IFG → HTN → HL → CVA → CRF
07.	No condition → HTN → OB → IFG → HL → CRF → PVD
08.	No condition → HL → HTN → IFG → CRF
09.	No condition → HTN → IFG → HL → CAD → MI
10.	No condition → CAD → HL → HTN → OB → IFG → T2D → CRF → CHF
11.	No condition → OB → HL → HTN → CRF → CHF
12.	No condition → IFG → HL → HTN → OB → T2D → PVD
13.	No condition → CRF → HTN → HL → CAD → PVD
14.	No condition → HL → HTN → OB → IFG → CRF → CVA
15.	No condition → OB → IFG → HL → T2D → HTN → CVA
16.	No condition → IFG → T2D → HTN → HL → CRF → CVA
17.	No condition → HL → IFG → T2D → HTN → CAD → CRF
18.	No condition → HTN → CAD → HL → IFG → PVD → CRF → CHF
19.	No condition → CRF → HTN → OB → IFG → HL → CAD → MI → CHF
20.	No condition → HTN → OB → HL → IFG → CAD → PVD

Table A.7: Trajectories after applying the precedence-based filtering method (precedence: succeeding events need to precede to preceding events).

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CVA
02.	No condition → HTN → HL → IFG → T2D → CRF
03.	No condition → HL → IFG → T2D
04.	No condition → OB → HL → IFG → T2D
05.	No condition → OB → HTN → IFG → T2D → CAD
06.	No condition → HTN → IFG → CRF
07.	No condition → IFG → T2D
08.	No condition → OB → IFG → T2D
09.	No condition → OB → HTN → HL → IFG → CAD → MI
10.	No condition → HTN → HL → IFG → CRF → CVA
11.	No condition → OB → HTN → HL → IFG → CRF → PVD
12.	No condition → HTN → HL → IFG → CAD → MI
13.	No condition → HTN → HL → CRF → CVA
14.	No condition → OB → HTN → HL → IFG → T2D → CAD → MI → PVD
15.	No condition → HTN → HL → CAD → CHF
16.	No condition → HTN → CRF
17.	No condition → OB → HTN → HL → CRF → CHF
18.	No condition → OB → HTN → CRF
19.	No condition → HTN → IFG → T2D
20.	No condition → HTN → HL → IFG → CVA

Table A.8: Trajectories after applying the precedence-based filtering method (precedence: succeeding events need to precede to preceding events or happen at the same time).

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CAD → MI
02.	No condition → HTN → HL → IFG → T2D → CRF → CAD → PVD
03.	No condition → HL → IFG → T2D
04.	No condition → OB → HL → IFG → T2D
05.	No condition → OB → HTN → IFG → T2D → CRF → PVD
06.	No condition → IFG → T2D → HTN → HL → CAD → MI
07.	No condition → HTN → IFG → CRF
08.	No condition → OB → IFG → T2D → HL
09.	No condition → HTN → CRF → IFG → OB → HL → CAD → MI
10.	No condition → HL → HTN → CVA
11.	No condition → HTN → HL → CAD → MI → IFG → OB → T2D → PVD
12.	No condition → OB → HL → HTN → CAD → MI
13.	No condition → HTN → HL → IFG → CRF → CAD → PVD → CHF
14.	No condition → OB → HTN → HL → IFG → CAD → MI → CRF → CHF
15.	No condition → HTN → HL → CRF → CAD → IFG → CVA
16.	No condition → HTN → HL → IFG → CAD → T2D → CRF → PVD
17.	No condition → OB → HTN → HL → IFG → T2D → CAD → CRF → PVD
18.	No condition → OB → HTN → HL → IFG → CRF → CAD → T2D → CHF
19.	No condition → HTN → HL → PVD → IFG → T2D → CRF
20.	No condition → OB → HTN → HL → CRF → CAD → CHF

Table A.9: Trajectories after applying both association and precedence-based filtering method (precedence: succeeding events need to precede to preceding events).

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CVA
02.	No condition → HTN → HL → IFG → T2D
03.	No condition → HL → IFG → T2D
04.	No condition → OB → HL → IFG → T2D
05.	No condition → OB → HTN → IFG → T2D → CAD
06.	No condition → HTN → IFG → CRF
07.	No condition → IFG → T2D
08.	No condition → OB → IFG → T2D
09.	No condition → OB → HTN → HL → IFG → CAD → MI
10.	No condition → OB → HTN → HL → IFG → CRF → PVD
11.	No condition → HTN → HL → IFG → CRF
12.	No condition → HTN → HL → IFG → CAD → MI
13.	No condition → OB → HTN → HL → IFG → T2D → CAD → MI
14.	No condition → HTN → HL → CRF → CHF
15.	No condition → HTN → HL → CAD → CHF
16.	No condition → HTN → CRF
17.	No condition → OB → HTN → HL → CRF → CHF
18.	No condition → OB → HTN → CRF
19.	No condition → HTN → IFG → T2D
20.	No condition → HTN → CVA

Table A.10: Trajectories after applying both association and precedence-based filtering method (precedence: succeeding events need to precede to preceding events or happen at the same time).

Rank	Trajectories
01.	No condition → OB → HTN → HL → IFG → T2D → CRF → CVA
02.	No condition → HTN → HL → IFG → T2D → CRF → CVA
03.	No condition → HL → IFG → T2D
04.	No condition → OB → HL → IFG → T2D
05.	No condition → OB → HTN → IFG → T2D → CAD
06.	No condition → IFG → T2D → HTN → HL → CAD → MI
07.	No condition → HTN → IFG → CRF
08.	No condition → OB → IFG → T2D → HL
09.	No condition → HL → HTN → CVA
10.	No condition → OB → HL → HTN → CAD → MI
11.	No condition → OB → HTN → HL → IFG → CRF → CAD → MI → CHF
12.	No condition → OB → HTN → HL → IFG → CAD → MI
13.	No condition → HTN → HL → IFG → CRF
14.	No condition → HTN → HL → IFG → CAD → MI
15.	No condition → HTN → HL → CAD → MI
16.	No condition → OB → HTN → HL → IFG → T2D → CAD → MI
17.	No condition → HTN → HL → CRF → CHF
18.	No condition → HTN → CRF
19.	No condition → OB → HTN → HL → CRF → CHF
20.	No condition → OB → HTN → IFG → CRF