Incorporating Clinical Domain Knowledge into Causality Deep Learning

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

Jia Li

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

Prof. Vipin Kumar, Advisor
Prof. Gyorgy Simon, Co-advisor

January, 2023

Acknowledgements

First of all, I want to express my appreciation for all who supported and guided me in my thesis work and in my graduate career

Foremost, I am extremely grateful to my thesis advisor Prof. Vipin Kumar, who accepted my appeal to change research direction in my second year, and gave me the most precious opportunity in my life, to step into the palace of data science research. Once, being a scientist was only a dream of little me. But in our research lab, I gradually realized that my enthusiasm and intelligence are respected and my dreams are supported. Especially in the last year of my program, Vipin gave full recognition of my whimsical ideas, and continuously supported me to finish the work - This experience changed my mind about my career, let me bravely follow my heart and pursue an academic position, which I had never imagined before.

I am also grateful to all of my thesis committee members: my co-advisor, Prof. Gyorgy Simon, Prof. Shashi Shekhar, and my instructor Dr. Michael Steinbach for their valuable suggestions in guiding my Ph.D. research and polishing my thesis.

Furthermore, I am very fortunate to have so many brilliant colleagues in the research lab to support me, Haoyu Yang, Xiaowei Jia, Xiang Li, Shaoming Xu, and Mengdie Wang.

Finally, I am very grateful to my parents, because as the single daughter in the family, none of my accomplishments would have been possible without their tolerance and understanding.

Dedication

To my parents, my best friends Qiman Gao and Yujie Zhao, my cat Echo, and this wonderful world.

Abstract

Applying Deep Learning (DL) models to graphical causal learning has brought outstanding effectiveness and efficiency. But it seems odd that their relative applications are still far from widespread practical use in domain sciences like health informatics. As the author of this thesis, I focus on the research of causality hidden in healthcare data for years during my Ph.D. program, at the same time, with a strong interest in seeking the underlying reason behind the "gap" between statistical analytics and advanced Machine Learning (ML).

In the research on EHR (Electronic Healthcare Records), I realize some confounding bias inherently exists in data, such that DL cannot automatically detect or adjust it. The underlying reason eventually traces back to the *Geometric Meaning of Causal Graphs* - Causal graphs can be *multi-dimensional* because of the relatively independent timelines. Various individual-level features can lead to inter-timeline inconsistency and ultimately result in confounding bias. This bias is initially named Causal Representation Bias (CRB) in this thesis. However, its existence has rarely been noticed.

More importantly, such a *blind spot* has formed the fundamental obstacle for cross-applications over the two fields. DL models tend to overlook CRBs and thus stop in front of model *Generalizability*, while statistical methods lack a geometric global view and are blinded to the inter-timeline-transformation defined model *Individualization*.

This thesis focuses on investigating CRB's existence and its underlying theoretical scheme. It starts with an introduction (Chapter 1), including a discussion of the dilemma situation in causal research, a geometric explanation of CRB, and a proposal of Causal Representation Learning (CRL). Next, Chapter 2 reviews causal learning methods in classical statistics and state-of-the-art ML. The following three chapters demonstrate my relative works, ordered by their relevance: Work in Chapter 3 aims to experimentally verify the existence of "invisible" CRB to DL models; Chapter 4 focuses on recovering the causally missed data values considering the potentially existing CRBs; Chapter 5 analyzes an instance of CRB from the health informatics perspective. In the end, in Chapter 6, I will share my thinking flow of investigating CRB, conclude this work's value, and summarize directions for future work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Learning causality is an important topic throughout the health sciences, as well as in many other research fields, like meteorology, biology, epidemiology, social sciences, and policy-making [1–3]. It is particularly vital in medical analytics, a field that solves problems like finding adverse drug events or risk factors for disease. Machine Learning (ML) methods have shown that we can make significant progress in understanding the causal factors underlying data, such as Electronic Health Records (EHR). Since EHR data is an excellent source for causal inferences and analysis, considerable effort has been expended to implement comparative effectiveness analysis to evaluate known treatments or discover new treatments for different diseases in real-world environments. However, with commonly used ML models, obtaining unbiased estimates of causal effects from observational EHR data can still be challenging.

During my Ph.D. program, I was particularly interested in investigating the *theoretical gaps* between advanced causality machine learning methods and traditional statistical causal analytics regarding EHR data mining. Such a research interest was triggered by a curious phenomenon based on my long-term observations - Even with the successful theoretical development of interdisciplinary applications across the two areas in recent years, the causal learning methodologies from computer science barely impacted the domain empirically. This is not only due to the exceptional sensitivity of making changes in the healthcare field. It can also point to the lack of enthusiasm for the domain scientists to do so, although they widely recognize the advantages of machine learning. There is a growing consensus that machine learning can perform

powerful modeling while statistics explicitly provide model interpretations. However, it is not intuitive to understand why effectiveness and interpretability can form a paradox. Specifically, the interpretable causal deep learning [4,5] has been oriented for a few years, which aims to construct neural networks according to the known causal relations, but why its general success can hardly be reached [6]? Furthermore, why is it hindering Deep Learning (DL) but not statistical modeling if some barrier exists?

By summarizing the existing causal learning framework, I found that machine learning and statistics are restricted by their perspectives - One is data-driven, while another is knowledge-driven, which forms the mismatching between learnings and interests. In other words, some gap exists between the causal relations we generally learn from observational data, and the existing domain knowledge people have understood.

Remarkably, in deep learning applications on EHR data, I realize some confounding bias inherently exists and cannot be automatically adjusted. The underlying reason traces back to *Multi-Dimensionality* of the acyclic causal graph, which is a particular property of causality problems (distinguished from correlations) but has rarely been noticed. In such a graph, confounding bias may happen in a subspace of the graph space, while the learning process occurs in another and makes the bias unobservable. In my research problems, the 3-dimensional DAG (Directed Acyclic Graph) space is defined by multiple timelines, including absolute and relative ones. However, different timelines are not distinguishable from the view of general deep learning because the input data matrices can only reflect a single timeline. On the other side, statistical analytics focuses more on element-wise inferences, e.g., the local impact of confounding relationships, thus lacking a global structural sight to realize the dimensionality of the causal graph. In this thesis, I initially define such an invisible confounding bias as **Causal Representation Bias** (CRB), where I will describe its mechanism and critical influences in detail.

The key to a solution is to realize the causal learning process and the confounding bias adjustment in the same subspace. Thus creating a 2-dimensional latent space by extracting causal variables' representations can be a good way. And that leads to my proposed solution, the **Causal Representation Learning** (CRL) framework. It is like building up the desired causal relations in a representation latent space, instead of the original data space. Since the latent space is generally accessible for all DL-based models, any learning process or model calibration can be appropriately performed in

the space and preserve their effectiveness for further access. I subsequently realized that the CRL framework is not only for solving the possibly existent invisible biases, i.e., the CRBs. It can be generally helpful for making causal models generalizable, especially DL-based ones.

It is worth mentioning that the term "CRL" was initially raised in another work [7] very recently, where the author Scholkopf et al. focus on knowledge generalization problems, and proposed "causal representation learning" as a promising new direction to deal with them. In this paper, they point out that "most work in causality starts from the premise that the causal variables are given", and describes three critical challenges we are facing: 1) The robustness of modeling is still weak; 2) The learned knowledge lacks reusable mechanisms that allow it to be generalized; 3) We do not have effective ways to interpret deep learning results to form new knowledge. However, they were targeting a specific computer vision problem and aimed to make "low-level knowledge" reusable when learning "high-level knowledge" in their practice, but not mean to define a general framework, like the proposed CRL in this thesis.

Their work inspired me to realize that the invisible bias problem I want to solve is essentially attributable to the generalizability of causal models, because the adjustment can be seen as a model calibration process, independent from the causal model establishing process. For causal models, generalizability means that the causal knowledge learned from one set of observations can be utilized when modeling causality on other relative observations. From the learning perspective, the purpose is to achieve sufficient flexibility in causal modeling: A graphical causal relation can be modeled structurally step-by-step, separately by partitions, or hierarchically by changing granularities, such that the established models (or part of them) can be used when extending the causal graph or be modified to fit another dataset. Such flexible extension and modification can be easily realized in a latent space, where all causal variables can be adequately represented as feature vectors, and causal learning can be appropriately performed among them. In summary, the proposed CRL framework provides a new way to establish generalizable structural causal models; specifically, for multi-dimensional causality (like the one in EHR data), it can help to solve the invisible biases that possibly happened in other subspaces.

I will begin this chapter's content (in 1.1) by analyzing the causality research status

and summarizing the current causal learning framework. Next, section 1.2 concentrates on theoretical contributions, mainly focusing on the initially raised *Geometric Meaning of Causal Graphs*, based on which I will demonstrate the inherently existing bias Causal Representation Bias (CRB) and its significant influences. Section 1.3 will introduce the proposed CRL (Causal Representation Learning) framework and the theoretical hypothesis it is based on as a general resolution for CRB and other model generalization problems. In the last section, 1.4, the organization of this thesis will be presented, which is also an overview of my related works in the past few years to discover and verify the existence of CRB initially. In addition, for the realization of CRL and relative experiments, one can refer to my other paper [8], which is not included in this thesis due to space limitations.

1.1 Status of Causal Research

1.1.1 Two Sides of Causality

Causal analytics on top of observational data through causal graphs has wide applications in many fields, such as medicine, epidemiology, econometrics, biology, social science, and even policymaking. Better extracted causal information from observational data can help us to answer questions such as "what treatments are most effective against complications of patients", "what can be done to control disease spread", or "which long-term factors determine economic growth".

The study of causality originated from classical statistics, which has developed for decades and contributed critical causal understanding to establishing knowledge and theories in many areas. Technological improvement in collecting high-quality data has brought fundamental challenges to statistical causal analytics: learning graphical causality from data to be Bayesian networks is NP-hard. Meanwhile, machine learning techniques show great potential in causal learning; specifically, it has become a burgeoning topic in the Artificial Intelligence (AI) field as part of bedrock [9]. So, it is not surprising that in recent years causal learning has been a burgeoning topic in Machine Learning (ML).

Conventionally, causal analytics begins with the model specification with prior knowledge, assuming the underlying distributions, thus may lead to model approximation

errors (often called model selection bias in statistics). A significant advantage of such classical modeling is to keep it interpretable, thus making it easy to leverage domain knowledge [10]. On the other hand, Deep Learning (DL) applications bring outstanding effectiveness to causal learning. They usually do not require model specifications but reduce modeling approximation errors automatically by optimizing neural network connections. Some recent works greatly enhanced their model efficiency by approximating the discrete structure constraint as a continuous optimization function [4, 11], thus bypassing the NP-hard combinational searching scheme. However, in the meantime, the black-box nature of DL diminishes the modeling interpretability.

A common understanding is that ML significantly improves causal modeling over statistical methods. A Structure Causal Model (SCM) can capture interventional distributions and computes potential outcomes accordingly [7]. In contrast, statistical models can only allow some specified causal reasoning. But the biggest problem is that we instinctively assume what we learn from the observational data includes everything we are interested in. In other words, some inherent gaps exist between the data-driven perspective of computer scientists and the knowledge-driven understanding of the domain scientists. The gaps can be noticed as an odd exclusivity between applications of the two fields. Since the significant theories of health informatics are founded on comprehensive causal analytics, this field naturally demands causal modeling improvement. But in reality, the fundamental role of statistical models has little changed in empirical applications of this domain, compared to the flourishing development of cross-disciplinary research works.

1.1.2 Current Framework

Suppose two variables X and Y exist; then they are said to be causally related if some event of Y must follow some changing of X. [12]. There are two types of causal questions. The first deals with questions such as "Do cell phones cause brain cancer?" In this case, we want to know the causal effect of X on Y. In medical research language, X is referred to as an Exposure (or Intervention), and Y is the Outcome. The challenges are finding a parameter θ that characterizes the causal influence of X on Y and a way to estimate θ. This is usually what we mean when we refer to Causal Inference. The second question is: Given a set of variables, determine the causal relationship between

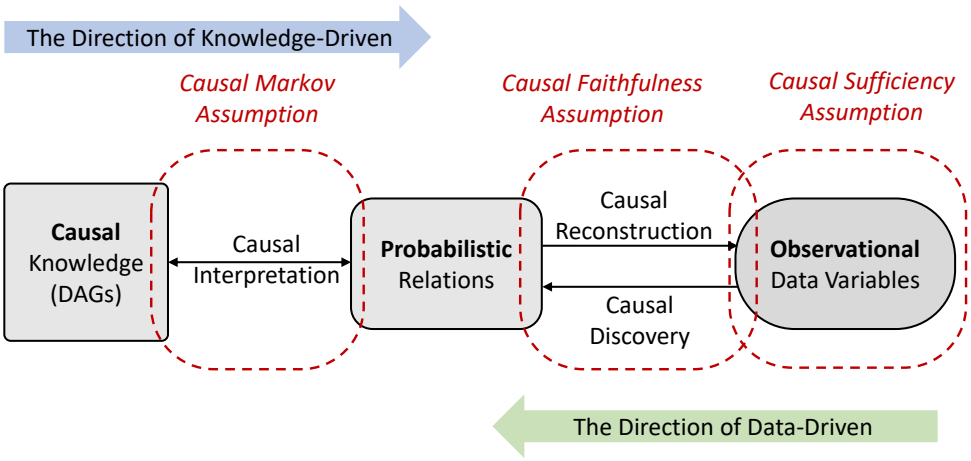


Figure 1.1: The current causal learning framework and three standard causal assumptions

the variables. The set of all causal relationships is referred to as the Causal Structure of these variables. This is called Causal Discovery. Briefly, Causal discovery is responsible for analyzing and creating models that illustrate the relationships inherent in the data, while Causal inference aims to study the possible effects of altering variable values in a given system [13]

Traditional statistical techniques focus on causal inference tasks. One may perform interventions or manipulations as part of a randomized experiment to understand causal relations. This is costly, sometimes impossible due to ethical concerns, and unrealistic or impractical at a large scale. Accordingly, statisticians created many empirical methods to manually extract "randomized" groups from observational data and then estimate causal effects using them to avoid selection bias (i.e., the imbalance between study groups). A more detailed account of modern statistical techniques for causal inference is provided in Section 2.2.

On the other side, the data-driven causal-structure discovery methods were initially developed in the ML community while built on statistical approaches for testing independence between variables. The discovered conditional-independence structure is described as a graph, e.g., in the form of a Bayesian Network (BN), which can be interpreted as a structural causal graph subsequentially.

I summarise the current causal learning framework as Fig 1.1, to describe all types of

learning processes between what we have known (i.e., Causal Knowledge) and what we can see (i.e., Data Variables), and mark the instinct directions of data-driven learning and knowledge-driven learning. Statistical analytics and ML methods can be seen as different ways of realization for these processes but consistent in the same purpose. Here the term Causal Reconstruction includes but is not limited to the modeling part of causal inference defined above. The word "reconstruction" refers to all types of applications to observational data that aim to reflect or have to rely on some priorly known relations. For example, estimating the causal effects from X to Y belongs to causal inference, whose modeling process can be seen as reconstructing the known causal relation $X \to Y$ on data. Furthermore, reconstruction can refer to some other applications not belonging to inference, like establishing a synthetic data generator according to a given DAG from domain knowledge.

From the interpretation view, a causal relation is equal to a correlation with certain temporal order for the events, as $X \to Y$ means X has to happen before Y. But from the view of models, the order of events does not have to be meaningful, and even though it is, they can be ordered in any reasoning, like spatially, but not necessarily to reflect a timeline. In other words, the dependent relations we learned from data unnecessarily indicate causations but are highly likely to be hybrid reasoning in practice, even though the training data have been arranged in chronological order. A typical hybrid scenario is a causation with hidden confounders: for a pair of independent variables, if their common causes exist but are unobservable, the learned pairwise association between them can only be some meaningful correlation, which can conclude in a pseudo-causation and violate the independence. Unfortunately, such a pseudo-relationship can only be identified manually, relying on some prior knowledge. In other words, the manual interpretation process is indispensable for data-driven machine learning on causality. So it is not strange that the black-box nature of DL disables causal interpretation and impedes its popularity in domain science applications. Specifically, the neural networks often provide no meaningful insights into their learned dependencies, although with accurate outputs.

As emphasized in Figure 1.1, the causal discovery is essentially detecting probabilistic dependencies, which are unnecessary to be causations. But in practice applications, we usually neglect that causal discovery consists of two processes, discovery and making

causal. Many ML methods have been well-developed for the discovery process, such as Structural Equation Models (SEMs) or Functional Causal Models (FCMs) [13,14]. The discovered causal knowledge is often described by Directed Acyclic Graphs (DAGs), although it can also be represented by other types of graphs. Any well-constructed knowledge in a field has been developed with great effort by domain scientists, often over a long period. By leveraging such knowledge, data-driven learning models can discover unrevealed knowledge besides contributing to improving modeling performances.

From a philosophical perspective, any interpreted causal relations are not certain to be true, because we have naturally assumed the connection from distributional dependency to causal interpretation in a human-understandable way. Indeed, the causality we are talking about is more likely to be "something to believe in for now" rather than a causal truth, which is called the causal Markov assumption. But from a data science perspective, it is out of the scope, and our concern majorly focuses on the other two standard assumptions about causality.

The standard causal assumptions include the following:

1. Causal Markov Assumption: All the relevant probabilistic information about a variable (ignoring its effects) that can be obtained from a system is contained in its direct causes.

2. Faithfulness Assumption: Any population generated by a causal graph (satisfying Causally Markov) has the exact independence relations entailed by the causal graph.

3. Sufficiency Assumption: The set of measured variables in a causal graph contains all of the common causes among these variables.

As shown in Figure 1.1, the standard assumptions are applied in different processes within the framework: causal Markov assumption is about interpretation, faithfulness assumption acts as the premise of modeling observational data, while the sufficiency assumption asserts the data contains all common causes and thus is cautiously used in practice.

The general Markov Condition is the medium for connecting the human-understandable DAGs with the observable probabilistic distributions, regardless of causality. Additionally, when DAGs are interpreted causally, the Markov condition is assumed to be the

connection between the causal structure and probabilistic independence and is specifically named Causal Markov Condition. As mentioned above, violation of the causal Markov condition (or assumption) is out of data science scope. Thus, we do not need to worry about such situations.

Faithfulness assumes that any independence occurring in the population arises from the causal structure generating it, rather than coincidence. Although it seems arbitrary to eliminate all possible coincidences, this assumption is very reasonable in practice and has been widely embraced in domain sciences [15]. It does not mean to assert reality, but codifies the preference of taking all observed regularities into account when modeling causal relations.

Satisfying causal Sufficiency means that no common hidden causes (i.e., confounders) exist, which rarely seems reasonable in practice. But not assuming sufficiency on observational data does not require revealing hidden confounders. Given prior knowledge, the existence or the role of unobserved confounders may have been known and provides a reference for modeling the observed variables. On the other hand, inappropriately assuming sufficiency may lead to modeling bias, or an empirical violation of the other prior knowledge-based assumptions. For example, two variables have been known to be independent, but neglecting some confounders may make their observations unfaithfully dependent.

1.2 Blindspot in Causality ML

In statistical analytics, the specified models usually reflect the best guesses from domain scientists about the underlying causality based on the assumption that the variables in observational data can fully represent the model's variables, which is usually referred to as *causal sufficiency*, as one of the standard assumptions of causal learning. In other words, sufficiency is to assume the consistency between knowledge and data, but in reality, it usually can hardly be true.

Figure 1.2 illustrates the scenario that causal sufficiency is violated, according to the current framework in Figure 1.1. Two colors identify two differently driven learning processes, and the unmatching between them leads to the violation. The reconstruction aims to build up the causality on data, which is given by prior knowledge, but the

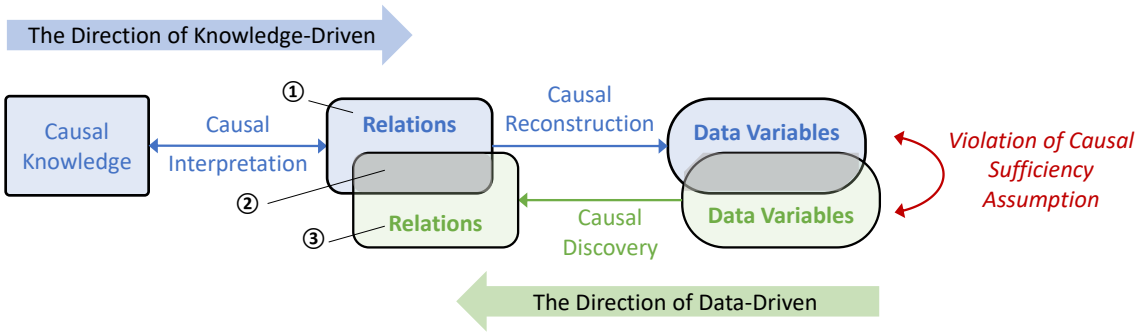


Figure 1.2: Sources of Causal-Learning Biases

reconstructed causal variables unmatch the observed ones from the data. For example, a causal effect estimation model may output biased predictions compared to observational measurements. This bias may come from improper model specification or optimization, and can also come from unknown confounding relations or so. On the other hand, since the discovery process is data-driven, the relations it discovers unmatch the current knowledge because they may contain some unrevealed knowledge or mislead by some unknown noises in data.

It is worth mentioning that people may conclude such unmatching as a violation of *faithfulness* assumption, say that the knowledge is certainly to be true but the causal relations from it cannot be correctly modeled on the data. Interestingly, sometimes it can also be seen as a violation of *sufficiency* since, from the angle of observation, the given knowledge may be too ideal and exclude some practical factors that impact the data generation. Referring to the initial intention of assuming *faithfulness* [10], it is to describe the specified models are entirely accurate, and no modeling error exists, which means we can ignore the empirical bias brought by modeling processes, but regardless of how the data generated. Therefore, by definition, they can be distinguished as: the faithfulness violation describes the bias between model and knowledge, while the sufficiency violation means the one between knowledge and data.

However, deliberately differentiating them is usually not meaningful, because they are hybrid revealed as "modeling errors" altogether, and the appropriate classification (if exists) entirely depends on practice. For example, some deterministic factors of

data generation are overlooked by people and lead to modeling biases; then, it can be explained as either unrevealed knowledge or incomplete models. And somehow, identifying the biases is precisely the purpose of causal learning.

Referring to Figure 1.2, it is easy to understand that the motivation of Machine Learning (ML) on causality is to increase the (2) area and, accordingly, to reduce (1) and (3) areas. Such as, improving causal models is to increase the (2) area by reducing modeling errors for both directions: An improved causal discovery can "move" the data-driven relations (Relations in green) closer to the knowledge-driven ones (Relations in blue); while better reconstructed causal relations (Data Variables in blue) will be closer to the reality (Data Variables in green). However, such modeling improvements have their upper limits, not able to shrink the (1) area to vanish, which can include three different types of components:

1. Reducible modeling biases (equal to faithfulness violations), including approximation errors (named "model selection bias" in statistics) and estimation errors.

2. Some causal relations are known and expected to be but not observable in current data. They can only be solved by changing expectations.

3. Some causal relations are known and observable in current data; they are expected to be but not adequately represented in the causal learning process, which is defined as Causal Representation Bias (CRB). The ignorance of CRB forms *blind spot* in causality machine learning.

The concept of CRB is initially proposed in this thesis. It is also demonstrated in my recent work [8] under the context of EHR (Electronic Healthcare Records) data causal learning. The fundamental condition of generating CRB is that the underlying causal graph is *Multi-Dimensional* because of multiple *relatively independent timelines*.

We often use Directed Acyclic Graph (DAG) to describe causal relations, but the *geometric meaning* of causal DAG has not been systematically discussed yet. Given a causal graph G that includes a set of variables V as its vertices, the relations over V must be learned along a timeline. For example, we need successive observations over months to learn how the economic situation influences stock movements. The causal model must be established on the economic situation relative variables observed within

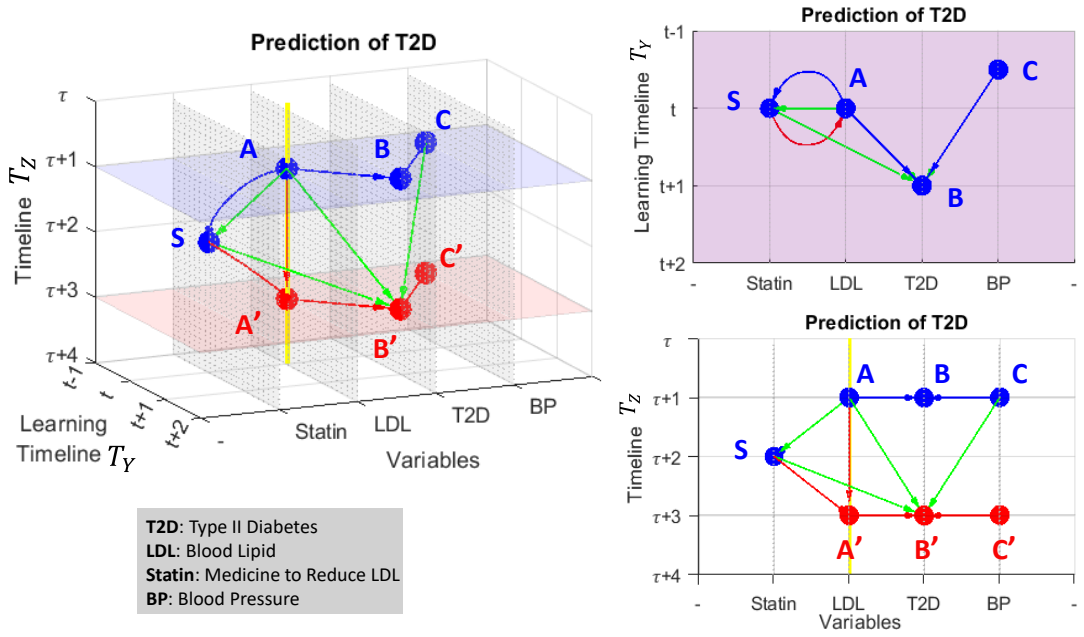


Figure 1.3: A 3-dimensional causal graph within two timelines. The learning target is to predict the risk of developing T2D given values of BP and LDL, which can be causally impacted by using Statin. ($A \to S$ and $S \to A'$ are intensionally curved to display overlapped edges.)

a specific period (e.g., a month) ahead, and the natural months' timeline is what we are learning along. For convenience, in the following, the timeline representing the target causal relations is named *Learning Timeline*, and the plane spanned by the two axes, defined by set V, and the learning timeline, is called *Learning Plane*.

Notably, the learning timeline can be either the natural timeline (such as dates of the calendar, timing hours, etc.) or any reasonable relative timeline, e.g., to model the monthly effect of a medication, which can happen on patients in any nature month of any year. In short, a timeline in a causal graph has to represent some meaningful causal effects and carry the relations' subgraph accordingly. Two timelines are called to be *Relatively Independent* if they represent different causal relations (regardless if they have the same time step length). Two relatively independent timelines can be joined because of shared causal variables.

Figure 1.3 displays a typical joint-timelines scenario of medical effects estimation in EHR data, having two independent timelines T_Y and T_Z, where T_Y in natural dates is

the learning timeline that represents the causally changed performances of patients, including the measurements of blood lipid LDL and blood pressure BP, as well as the estimated probability of developing type II diabetes, T2D. The relative timeline T_Z describes the medical effects of Statin medication, whose usage is determined by physicians based on the criteria of LDL and directly reduces the patients' LDL values. The medical effects will subsequentially decrease the risk of T2D via the causation from LDL to T2D. Here the learning task is to predict T2D risk given the patient's LDL and BP values in the last time step and his/her Stantin using history accordingly.

The right upper part of Figure 1.3 shows the learning plane defined by T_Y, from which view the variable Statin is observable but does not belong to this plane. Thus the time stamp t is not meaningful to the node S, and the causal effects between S and A turn out to be a circle from such a view. While the circle no longer exists by involving the relative timeline T_Z, the nodes A, B, and C can be split by two to differentiate the roles before and after taking Statin, respectively, as shown at the right bottom of Figure 1.3. The green lines indicate the causal relations we aim to model. An appropriate learning process should be performed in an eligible learning plane. In other words, it is necessary to determine whether the green-colored causal subgraph can define a 2-dimensional plane.

Regularly, when applying machine learning models (including deep neural networks) to causality problems, validation of the learning plane is assumed to be *True* by default since it is the premise of modeling. However, this condition may not be satisfied when the causal relations cross multiple independent timelines.

Theorem 1. *A valid causal learning plane must satisfy that the underlying causation of any samples in the data can be represented as a linear transformation of the causal graph in this plane.*

Corollary 1. *If a causal graph across multiple independent timelines does not contain any confounding relations, then it is in a valid causal learning plane.*

Theorem 1 defines the geometric meaning of causal graphs. In the simplest case, modeling a pair-wise causal relation from observations is equivalent to estimating the correlation between the two variables, such that the deviations of all the data samples can be minimized to normally distributed. In other words, the underlying correlation

in any data sample can be seen as a linear transformation of the estimated average correlation. By extension, modeling a causal graph is also the process of finding such an estimation. All the pair-wise causal relations (i.e., edges) in this graph must be recovered simultaneously from the estimation by a *single* linear transformation for any data samples. Since two vectors can define a 2-dimensional plane, the causal graph described in Corollary 1 can certainly form a valid hyperplane for causal learning.

In Figure 1.3, the green-colored target learning relations contain a confounding relation over nodes A, S, and B', which requires further discussion about its validation of being a learning plane. Considering the two edges $A \to S$ and $A \to B'$, according to Corollary 1, the plane they spanned is valid for causal learning. Because $A \to B'$ can be decomposed as two successive vectors, $A \to A'$ along T_Z and $A' \to B'$ along T_Y, this plane has crossed two timelines; Which implies that the relative changing of T_Z and T_Y within the single linear transformation have been determined. And accordingly, the transformation of another edge, $S \to B'$, connecting two nodes belonging to T_Z and T_Y individually, has also been determined but in a passive way. Under this circumstance, assuming that the passive transformation of edge $S \to B'$ can *adequately represent* the observational relation in the data coincidentally is arbitrary. The risk of leading to bias is commonly ignored in causality machine learning. It is, therefore, necessary to propose the name Causal Representation Bias (CRB) here to refer to it explicitly.

Interestingly, such a linear transformation defines the data points' *Individualized features*. In recent healthcare studies, capturing the personal level characteristics of patients has become a common goal. Once there exist confounding relations across independent timelines, assuming all causal variables in the graph are adequately represented in the learning process means that all the pair-wise causal relations are consistently transformed. Specifically, in Figure 1.3, that means if we learned edges $A \to S$, $S \to B'$, and $A \to B'$ individually, their three models can output the same results as if we learned the three edges simultaneously in one process. In other words, for any available data point (in the data space), we can always find a complete triangular to correctly describe its underlying causal relation, within the two timelines T_Z and T_Y. This clearly indicates all data samples should be from the same patient, who runs a consistent causal system in the body. However, obtaining enough samples from a single patient to perform causal learning in EHR data is almost impossible. And that is the reason people

find identifying individual-level patient features very challenging. The geometric explanation of causal learning above provides a straightforward manner to instance it - An individualized feature can be represented as a distinct linear transformation matrix.

This thesis focus on confirming the existence of Causal Representation Bias (CRB). Mainly, the work described in Chapter 3 uses comprehensive experiments to verify this point successfully. In my most recent work [8], a general solution for CRB is proposed, called Causal Representation Learning (CRL). To be brief, CRL aims to intentionally create a valid learning plane for any causal graphs, which is also referred to as the latent *representation feature space* [8]. This can also be seen as the resolution of how to make causal models generalizable, especially deep-learning-based ones.

In addition, the geometric meaning of causal graphs provides a golden rule of differentiating *Casality* problems away from the regular *Correlation* problems. Suppose the causal graph can initially define a valid learning plane, like the case raised in Corollary 1. In that case, its causal relations are equal to correlations, whether with or without causal interpretation. On the other hand, learning causality has to face the risk of CRB brought by independent timelines, like the example in Figure 1.3 with regular EHR data. The CRB adjustment must be made through *asynchronous processes* based on prior knowledge, like pre-process or post-process, independent from the target causal learning - Since CRB is the confounding bias in other timelines, out of the learning plane, it cannot be spontaneously detected or adjusted by the causal learning process. The work in Chapter 3 shows a practical example of "asynchronous CRB adjustment" under the EHR data context.

1.2.1 From the Data-Driven Perspective

In machine learning, we are accustomed to focusing on modeling performance. One can hardly find reasons to be concerned about something out of the learning scope since, usually, being observable means being learnable. But interestingly, this is not always the case in causality problems.

Keeping the causal models interpretable by leveraging prior knowledge can be greatly helpful for avoiding potential representation biases and preventing the possible blind spot mentioned above. Nowadays, the interdisciplinary applications of domain knowledge to enhance machine learning models have been a thriving area in health informatics.

Knowledge about underlying causality can often deliver extraneous information about unobservable confounders in the data. One can consciously adjust confounding biases as a pre-processing or post-processing, to produce de-confounded pseudo datasets or adjusted learning models. In other words, some causal knowledge may be out of the learning scope but can be utilized in other asynchronous assistant processes.

Causal knowledge can be used in many ways, like building a reasonable model structure, selecting proper training samples, calibrating established causal models, etc. But we rarely notice the distinction between *Synchronous* and *Asynchronous* utilizations. For example, we can perform resampling as a pre-process to construct a balanced training dataset from the observations known to be imbalanced, which is an asynchronous process; besides, maybe we can also reconstruct the model to eliminate extra samples at input layer consciously, and directly use the raw observational data for training, which is a synchronous process. However, for the blind spot mentioned above, the aim of adjusting CRBs can *only* be achieved in the former but not later.

Fortunately, I have the opportunity to extensively explore two different causal data with typical causality and correlation problems, respectively. One is the EHR data, whose causality is supported by comprehensive healthcare knowledge. Another is from hydrology, which is to record how environmental factors (including temperature, solar radiation, precipitation, and so on) impact the volume of streamflow months later in a particular district, whose underlying causality can be roughly explained by some physical conversion formulas [8, 16]. Interestingly, I noticed that, unlike EHR data, the causality in the hydrology data is only "two-dimensional" by considering the natural timeline and the sequence of all variables as two axes. Because all records are about the same district, the underlying physical mechanism identically determines all variables' changing in this dataset. Thus, any pair-wise causation in its causal graph consistently happens along the same timeline. In this circumstance, the timeline only means a particular order of events, which is unnecessary to be temporally meaningful. For the model, it is the same as any regular dataset with other reasoning orders, e.g., some records ordered by spatial locations, or even random order - Which indicates the interested relations can be considered as just *correlation*, but not necessary to be *causation*.

It can be easily understood by analogy: if we consider one particular district in hydrology data as equal to one particular patient in EHR data, it is straightforward to

know the EHR data has *one more dimension* than the hydrology data because EHR contains numerous patients daily visiting the hospital. However, selecting records of a single patient from EHR data is meaningless since there may be only $2 \sim 3$ records, which is impossible modeling.

Remarkably, it does not imply that the relationship from the same data source is restricted to one type, either causation or correlation. Indeed, it depends on the particular questions one aims to solve. Suppose we are interested in the standard physical regulation on the earth instead of some local impact. In that case, we can collect the hydrological records over many different districts to comprise the dataset, in which the causal relations are highly likely to be multi-dimensional, because different physical factors may vary along different relative timelines. On the other side, the data from clinical trials (consistently recording certain patients in the long-term) can be seen as a correlation, more superficial than the complex relationship in EHR data, although their interpretations are from the same knowledge system. This is because the intentional randomization of the patient cohort eliminates the individual-level variances. In other words, it removes the natural timeline axis to make the causations equivalent to correlations in a two-dimensional graph space. And notably, to some degree, this can also be seen as an asynchronous pre-process.

A fascinating phenomenon can demonstrate the existing troubles that multi-dimensional causality brings to machine learning: People have made great efforts to realize *interpretable DL causal models* by leveraging domain knowledge but can hardly achieve general success (please refer to section 2.1 for detailed reviews). For example, a recent attempt is to construct neural network architecture according to priorly known causal structures and simultaneously optimize the network connections [6]. However, the conclusion can hardly support that involving this knowledge can significantly improve performance; on the contrary, it is more likely the performance and the interpretable structures form a dilemma strangely. The geometric interpretation of multi-dimensional causality can help to explain the underlying reason - The involved causal knowledge is certainly correct, but some causations may happen along timelines other than the learning one, thus can hardly be informative for the learning process.

1.2.2 From the Knowledge-Driven Perspective

From the knowledge-driven view, as in classic health informatics, model specifications are fundamentally based on domain knowledge to realize specified-only causal reasoning within statistical models. From the data-driven machine learning's view, the specified-only implies them to be "limited," but the key advantage of this way is ensuring that the model always serves causal interpretations rather than in the opposite way. For this reason, the validation of the causal modeling plane has never been a concern from the domain's perspective since, the established knowledge has guaranteed it. However, their learning scope is correspondingly highly restricted, such that they can only deal with the causal relations in a single timeline but cannot capture the geometric meaning of transformations between the relatively independent multiple timelines. More importantly, in current society, the need for identifying personalized features keeps increasing in many fields of applications; But such features may be mainly defined by the relative relations among independent timelines, e.g., each patient has indistinct reactions to complex medications, whose effects belong to complex joint timelines. The traditional models naturally lack the global view of causal graphs, and thus can hardly handle such problems.

This does not mean the interpretable statistical models have little concern about potential representation biases. In these models, the causal variables are risky to be under-represented or distorted because of such models' natural limitation on computational capability. It often happens that people have been aware of some minor representation biases but leave them over to keep the models intuitively interpretable.

For example, suppose $A \to B \to C$ is given knowledge, assuming that A and C are independent given B. However, in practice, the observations of A accidentally capture some minor features of C and lead to a weak correlation between A and C. Because in A the extraneous features are strongly associated with its major features and can hardly be split away linearly, such small representation bias has to be ignored and remains in the established causal models based on linear regression. From the larger scale perspective, along with knowledge development, numerous small biases can be gradually accumulated in models or datasets and result in essential problems, like adversely affecting modeling performances or misleading knowledge discovery.

On the other hand, since models have to be supported by knowledge, they usually

cannot proactively figure out new knowledge, but commonly rely on manually adjusting hypotheses and detections by chance. This situation does happen in reality. For instance, let's refer to a diabetes-related work [17] published in the 2020 Nature genetics journal. LDL cholesterol is one of the vital signs to diagnose T2D (Type II Diabetes), and statin is commonly used as a first-line medicine to reduce LDL. By substitution, the causal effect that A (statin) reduces B (LDL) is known, and it is also acknowledged that elevated B (LDL) often indicates a higher risk of C (T2D). Accordingly, all other methods in comparisons [17] confirmed the positive effects from B (LDL) to C (T2D) by given A (in clinical trials of statins), represented as the causation $A \to (-B) \to (-C)$. But by observational data from trials, larger A (using statin) shows an increased risk of C (T2D), i.e., $A \to (-B) \to C$. In other words, some hidden correlation between A and C impacts the data distribution to show a paradox. The clinical hypothesis is that statins and their target enzyme affect multiple systems. Thus, part of their effects on T2D is possibly mediated by other mechanisms, i.e., A has some association with C independent of B. In this work, Morrison et al. contribute a new method to conclude that 13% of LDL acts through some other factor, i.e., 13% causal effects of A on C come from the hidden correlation.

This real case is a vivid example of how people manually adjust the hidden confounding bias by referring to clinical knowledge. In practice, such biases and unidentified causal relations commonly exist. So far, they may not seriously affect the development and progress of traditional statistical analytics, but under the current background of information and data explosion, knowledge expansion and development have greatly sped up, which can make biases accumulate rapidly and lead to more significant problems than before.

1.3 Causal Representation Learning

Among the thriving interdisciplinary applications combining machine learning and health statistical methods, many works are, essentially, realizing an adjustment of some known representation biases, leveraging domain knowledge. From this angle, the existence of CRB has been recognized for a long time, but in the way of case-by-case in various scientific fields. Previously, there may not be a strong demand for extracting a generic

concept for it and proposing general resolutions accordingly. The case-by-case adjustment works well mostly, and making the models interpretable has reached the goal of modeling, usually without concerns about how to generalize the model further.

But in the era of information explosion, and given the increasing interest in Deep Learning (DL) applications, causality learning has been on a much larger scale and facing more general problems. Commonly, the DL models take advantage of handling numerous variables from a global view but provide no meaningful insights about the learned relations, which makes manually adjusting models very inconvenient. Therefore, it does make sense that many DL works implicitly assume the adequate representation of observed causal variables to avoid the demand for more detailed interpretation. But such a strategy brought more fundamental problems - These causal models can be badly generalizable. The Causal Representation Bias (CRB) crossing independent timelines may have distorted the models; Since such biases come from the inconsistent individualized features included by the data, the correspondingly established models have been "specialized" somehow instead of representing the most general underlying causal knowledge.

On the other hand, in the research area of domain knowledge, the capability of knowledge-driven statistical models has approximately reached its ceiling. Investigating new knowledge increasingly relies on machine learning with relatively more global perspectives. Typically, in health informatics, individual-level patient features are critical references for medical decision-making, but statistical methods cannot investigate them crossing independent timelines. Without a geometric understanding of the multi-dimensional causal graph, the desired personalized features can hardly be represented.

Therefore, the dilemma has been formed, such that the advanced DL models tend to conceal CRBs and hence are stopped in front of *generalizability*; While statistical methods are restricted intra-timeline and blinded to the inter-timelines transformation defined *individualization*. There has been a demand to break the "wall" between the two sides of causality research. This motivated me to seek a general resolution and eventually propose **Causal Representation Learning** (CRL) in this thesis. In brief, CRL is a general method of finding a valid learning plane for any causal graphs, providing a reference frame for generalizing graphical models and also extracting individualized features.

The term *Causal Representation* was first raised by Scholkopf et al. [7] last year (2021). In this paper, the author points out that "most work in causality starts from the premise that the causal variables are given", and describes three critical challenges we are facing on the way toward applicable Artificial Intelligence (AI): 1) The robustness of modeling is still weak; 2) The learned knowledge lacks reusable mechanisms that allow it to be generalized; 3) We do not have effective ways to refine the discovered causality using existing knowledge. Scholkopf et al. [7] mainly focus on knowledge generalization (making low-level knowledge reusable when learning high-level knowledge) in a specific field of computer vision. Thus they raise the concept of "causal representation learning" but do not concentrate on its geometric meaning in general.

Once the causal variables and relations can be accurately represented in a latent feature space, AKA a valid learning plane, the modeling process can be performed on feature vectors instead of the unchangeable observational data variables. I expect that generalizing the concept of Causal Representation as a general machine learning method would have more significant benefits than resolving the three underlying problems mentioned above. It can considerably expand causal learning's potential applications and maybe even start a new trend of causality research in the near future. The roughly summarized advantages of CRL applications are as below:

- Firstly, it can be helpful to reduce the adjustable part of CRB on variables in causal learning and, consequently, be conducive to causal knowledge refinement. Without the restriction on observed data variables, the theoretical causal relations can be freely set up on appropriately designed representations, which can accurately represent the desired causal variables or factors. In the example mentioned in section **??**, one can extract a representation feature vector to purely represent the causal variable A without extra correlation with C (if the correlation is estimable). Then the causation $A \to B \to C$ can be accurately learned by building models on these representation vectors instead of the data variables.

- Secondly, it may provide a fundamental solution for the DL model generalization problem. Suppose we aim to find a general causal relationship over a set of causal variables, which applies to all datasets collected under different conditions. By

representing the causal variables as a group of feature vectors, the desired causation can be described as SEMs (or FCMs) over these vectors in the appropriately designed feature space. Although training on a particular dataset may lead to over-fitting, after training through all available datasets, the causal knowledge commonly shared among them could be refined and learned by the SEMs, such that the desired causal information has been preserved in the feature space. In the form of representations, the learned causal relations saved in this space could generally be accessible to all deep learning-based models.

- Finally, it can provide a novel perspective on model interpretability. Under this new framework, DL would be released from reproducing structural causal knowledge as it is supposed to be and only focus on adequately representing causal variables and their functions. The representations extracted by DL do not need to be interpretable but can be reconstructed to form conventionally interpretable models and their corresponding variables. In short, DL acts as causal information storage, while the interpretation is still attributed to conventional models given existing knowledge.

1.3.1 Proposed Framework

Under the current causal learning framework (as shown in Fig 1.1), the existence of CRB can easily be overlooked since it assumes the validation of *learning plane* by default. Compared to traditional statistical analytics, Deep Learning (DL) is more advanced in performance but tends to downplay the interpretation of modeling due to its black-box nature. However, being aware of CRB usually depends on how well the prior causal knowledge has been applied. Specifically, the known causal relations must be identifiable from the model and can be applied to check possible biases. Without any structural interpretation, even though a DL model occasionally "understands" the given causal knowledge, its generalization will still be a problem.

Another trouble with DL on causality is that the global optimization strategy brings high efficiency but probably increases the overall biases simultaneously. Unlike traditional heuristics, which can deal with the local confounding bias in each step regardless of other variables, the global DL models wrap up the optimizations of all (pairwise)

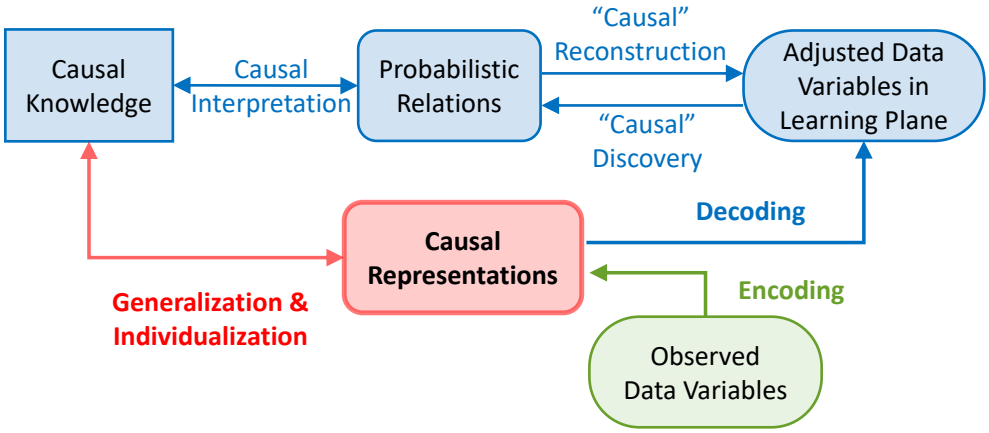


Figure 1.4: The Proposed Framework: Causal Representation Learning (CRL). The red color indicates the causal representation feature space; The green part is the initially observed data space; And the blue means being in the valid learning plane(s) where the known CRBs have been adjusted, and prior causal knowledge from conventional causal analytics can be accurately interpretable.

causations in the graph in one run, possibly letting the local confounding bias impact the global performance.

In recent years, the demands for interpretable causal models significantly boosted cross-disciplinary applications, including the numerous efforts to make domain knowledge utilizable to DL models. However, these efforts are often case-by-case for practical problems and can not be generalized easily. For example, in my previous work [**?**], we extract a linearly computable estimator to partially adjust confounding bias in medical effects, but with the premise that the known causal structure is luckily to be derivable.

The Causal Representation Learning (CRL) framework is shown in Figure 1.4, which aims to find valid learning planes for general causal learning problems and accordingly realize generalizable adjustment for the inherently existing CRBs in observed data. The basic principle is to separate the representation adjustment, and knowledge-guided learning as independent processes finished in two different spaces. This way can make DL concentrate on the learning processes at latent space and let the existing causal analytics or knowledge navigate model interpretation. Specifically, we aim to realize an autoencoder to extract representation for each pair-wise causal relation, such that

causal variables can be represented as feature vectors and the correlations among them can be modeled in the latent space. Accordingly, regular learning tasks can also be performed in such a latent space, including causal discoveries and reconstructions.

From the angle of DL models, this framework can explicitly realize their structural interpretations by leveraging domain knowledge, thus, can be widely used to realize DL model generalization. And such applications can be either on causality or correlation problems, with and without meaningful causal interpretation.

1.3.2 Theoretical Hypothesis

Causal Representation Learning (CRL) uses the DL-based data representation extraction method to construct a latent feature space for any graphical causal relationship as its valid learning plane.

The concept of the *learning plane* is raised in section 1.1.3, which comprises two axes, the set of causal variables and the learning timeline. A valid one must carry the target causal graph. In other words, all pair-wise causal relations in the graph must be consistent along the learning timeline. Since it is easy to ensure the timeline for a pair-wise correlation manually, intuitively, we can go through all the edges to establish pair-wise models one by one in the latent space, such that all causal variables from the data space are appropriately represented as individual feature vectors in this space. And this latent space plays the role of the "axis" all causal variables lying independently.

The representation vectors can be extracted using autoencoders. The balanced initialized autoencoders learn the subspace spanned by the top principal components of the data [18–20], which is well known as that autoencoders perform the unsupervised principal component analysis (PCA).

Theorem 2. *Suppose matrix X is augmented by column vectors of all variables in a causal graph G, then the representation feature space of G must have a dimension larger than $rank(X)$.*

Variables in the causal graph G can have various dimensionalities, but they will be represented as samely dimensional vectors in the latent space. Since we want the variables to be represented distinctively and use their distances to represent correlations, such a space should simultaneously provide enough degree of freedom for all

pair-wise causal modelings. A reasonable dimensionality of the latent space should be between the rank of X and the column number of X, i.e., the sum of all variables' dimensions. (suppose the number of time steps, i.e., the row number of X, is much larger than its column number.) Thus, we can basically expect each pair-wise model to be a *higher-dimensional* representation extraction, unlike the regular lower-dimensional autoencoders. Please refer to my recent work [8] for details of the initially designed autoencoder architecture.

It can be seen that each causal variable needs to be over-represented by a redundant dimensional vector. The principle is based on the fact that *over-parameterized* autoencoders learn contractive solutions around the training examples in the latent space [21–23], which can be seen as a manifestation of overfitting. But such overfitting is necessary for free space to allow the two vectors "move" toward their training samples, respectively, in each pair-wise causal learning, such that the distance between them can let us model their conditional distribution, AKA the pair-wise causal effect.

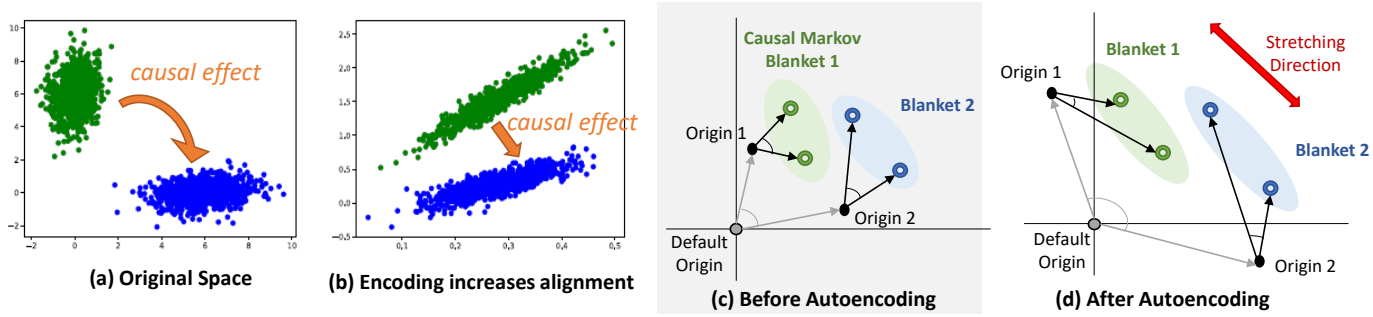


Figure 1.5: Examples for the hypothesis in Corollary 2.

Corollary 2 (Hypothesis). *A lower-rank representation feature space can emphasize more significant causal relations in graph G.*

Additionally, I hypothesize that for causal graph G, its corresponding causal relationship G' in the representation feature space can become more representative than G regarding causal inferences or causal discoveries. This hypothesis is inspired by a proven proposition that "autoencoders can produce aligned latent spaces by stretching along the top left singular vectors of the data" [23] based on the PCA natural of autoencoders. This principle has been widely used in many downstream learning tasks, such as finding

meaningful semantic directions to identifying word analogies [24, 25], and embedding gene expression data [26, 27], but not been referred to for solving causality learning problems yet.

According to the principle, the causal variables can also become more aligned in the latent space. Suppose this representation feature space is spanned by selected top principal components of X instead of fully ranked X. The causally related representation vectors will be stretched along their principal commonality. Then, the differences among the vectors can represent the causal effects more particularly, as shown in the example Figure 1.5 (a) and (b).

Furthermore, from a structural view, embedding G into such a latent space can potentially emphasize more significant relations and make the causal discovery easier. For example, in Figure 1.5 (c), to discover the two separate causal Markov blankets, we will want to reduce the similarity between them and increase the internal similarity for each of them. Suppose we use cosine similarity to evaluate distance in causal discovery. It is promising to find a stretching direction via SVD (singular values decomposition) to realize this purpose, as shown in (d).

However, this corollary is still in the hypothesis stage and needs further rigorous demonstration in future work.

1.4 Thesis Organization

This thesis will start with a literature review of causal inference methodologies, including an analysis of challenges in causal learning from the perspective of healthcare informatics. The major content of my research works will be introduced in Chapters 3-5. In the end, it will be wrapped up by a conclusion chapter with summarizing discussion.

Chapter 3-5 contains three individual works, which were finished in different stages of my program, and are ordered by their relevance to Causal Representation Bias (CRB). Basically, the work in Chapter 3 focuses on how to adjust this bias for deep learning-based models from the view of variables; The work in Chapter 4 is about how to produce a de-biased dataset for general learning, which could be considered as an adjustment from the view of data sampling; And the last work in Chapter 5 investigates, from a domain point of view, a paradox result from this bias.

Chapter 3: Improving Deep Learning Predictive Performance by Incorporating Causal Effects Domain Knowledge.

In this work, we incorporate statistically learned causal effects into the deep learning architecture, to improve its causally predictive performance on time series. The interesting causal effects consist of the relative effects of diabetes medicine, are evaluated based on 10-year follow-up Electronic Health Record (EHR) data obtained from the Mayo Clinic with research consent.

Given the increasing interest in applying deep learning to healthcare data to achieve outstanding predictive performance, we were curious about whether, in addition to the high accuracy, deep learning can automatically capture clinically meaningful causal information from time series. In this work, we designed comprehensive experiments, to compare the performance differences of RNN models, whether involving prior causal knowledge or not. We concluded that deep learning cannot fully extract causal medical effects automatically from EHR time series but needs additional support from domain knowledge.

The relevant medicine effects are calculated using a conventional statistical method, named Propensity Score Matching (PMS), for patients with diabetes and related complications, including heart disease, kidney disease, and stroke. We implemented several different ways to incorporate these effects into deep learning architectures, and also

reported conclusions on various RNN mechanisms.

A theoretical proof, which is also provided in this chapter, explains how prior knowledge provides causal information gain to deep learning, and corroborates that the observable variables in EHR data do not fully provide relevant causal information. Since deep learning does not require model selections, this work also emphasizes that the existence of CRB is independent of model selection bias. Furthermore, in this work, we first described the role of Causal Representation explicitly and demonstrated its importance in causal learning.

Chapter 4: Dealing with the Causally Informative Missing Values in EHR Data.

This is my earliest work investigating the influence of CRB in machine learning. It focuses on informative missing values in EHR data, using a group-specific learning methodology to consider the prediction task separately. It successfully verified the hypothesis that the missingness was primarily formed causally rather than randomly.

To be specific, we had the idea that a particular missing pattern (a certain combination of variables with missing values) may be a result of some underlying reason, which acts as an unobserved confounder and creates data selection bias in the patient cohort. Thus, the predictive performance of the overall cohort must be influenced by the bias and can be potentially improved. Group-specific learning methodology considers the cohort as multiple subgroups, according to their different missingness patterns, and performs predictions separately in each subgroup. By assuming the existence of the hidden confounder, we adjust the selection bias accordingly for each pattern and use the adjusted subgroups to perform predictions respectively. By integrating the predicted results from subgroups, we find that the overall predictive performance has been successfully improved as expected, which verified our hypothesis of causally formed missingness.

It is worth noting that this work intended to investigate the existence of hidden confounding and how to de-biased the data, but not for the purpose of creating better-performing prediction models. Thus, the method proposed here does not necessarily have high efficiency, and did not consider any higher-order nonlinear causation. However, it was nonetheless useful to design an imputation method that imputes missing

values by patterns, leverages the pattern-wise causal information for adjustment, and outputs a completed de-biased dataset that can be efficiently modeled in general.

Chapter 5: Assessment of the Association Paradox between BMI and All-Cause Mortality.

This is a statistical assessment work completed early, which aims to investigate a phenomenon essentially resulting from the CRBs of the observable data variables - the association paradox of BMI (body mass index) versus Mortality.

The paradox, which is controversial, is that under some conditions, higher BMI is predictive of less possible mortality. In health informatics, the potential explanations of this paradox were described in relation to the nature of the clinical condition, or methodological biases. However, these biases are not present consistently in all the studies. The aim of this study is to use a risk model based on different levels of co-morbidities, to define the baseline mortality risk and analyze the association between it and BMI. The conclusion shows that baseline comorbidities, through the risk they confer on patients, are the primary driver of all-cause mortality, while BMI only plays a secondary role.

From the view of clinical statistics, this work aims to interpret how clinical characteristics influence the probability of mortality. And with respect to modeling techniques, it implies that the underlying causation cannot be described by any single variable but attributed to some complex association among multiple variables. Furthermore, since it can hardly be described as a specified parametric model, the causal relation of interest is more likely to be nonlinear.

Inherently, the paradox implies that BMI cannot adequately represent the desired causation. This work is looking for a better Causal Representation from the associated observable variables but fails to resort to specific models due to the complexity of underlying nonlinear relations. Doing this work brought some requirements regarding Causal Representation to the surface: 1) the representation has to be identifiable regardless of the modeling method, and 2) it has to be able to describe nonlinear associations among the variables.

Chapter 2

Reviews

As we briefly described in the Introduction, unlike regular associations, causality answers the question, "What will happen to Y if we change X?" The traditional way of detecting causal effects is through carefully designed experiments, such as randomized clinical trials in health care, but such experiments are usually expensive, impractical, and sometimes even impossible. Thus, much attention was drawn to the question, "How do we reveal causal information from purely observational data?" In other words, is it possible to figure out the exact causal relationship among different components of the complex system only by its observed outputs?

As a classical causal model, Bayesian Networks (BNs), represent the conditional dependency relationships among variables using Directed Acyclic Graphs (DAGs). This method as well as the DAG causal representation have been widely used to address causal questions in many applications, such as estimating the effect of medication, predicting storms by maritime meteorological observation, and so on. However, learning such a network from observational data is NP-hard [6, 28]. To address this challenge, a great variety of causal models were developed given different applications in the last decades to discover the causal relationship among observable variables. We discuss generic (application-independent) causal analytics methods in Section 2.1. and then in Section 2.2, we focus on casual analytics specifically in healthcare.

2.1 Deep Learning in Causality

In recent years, Deep Learning (DL) has been applied to realize highly efficient graphical causal models. The key novelty is to turn the discrete possible-DAGs-space searching problem into a smooth function optimization question [6]. Specifically, these methods quantify the Acyclic-ness of the graph's adjacency matrix to be a continuous constraint, which established an elegant mathematical connection between the classical combinatorial algorithms and the continuous optimization techniques of machine learning [6]. For instance, Lachapelle et al. described how to convert the DAG learning into a pure neural network learning one in 2019 [5]; and Zheng et al. transformed it into a generalized nonparametric problem that requires no modeling assumptions in 2020 [4]. In short, this methodology established an elegant mathematical connection between the classical combinatorial algorithms and the continuous optimization techniques of machine learning.

Based on these realizations, people have made efforts toward the interpretable DL causal models. One way is to figure out conditional independence by finding zero derivatives of the graphical causal functions [5]. Another way is to learn it from the weights of the neural networks [4]: among all Markov causal paths that lead the direction from X to Y, if at least one of them has weight 0, then Y does not depend on X. Meanwhile, the adjacency matrix was optimized together with the DAG constraint by maximizing the log-likelihood using the augmented Lagrangian method. Such a weights-detecting methodology is also called a Neural Architecture Search (NAS).

However, this strategy can hardly achieve general success. It is still very difficult to determine how well the knowledge has been utilized during the networking optimization or why it has not been. The work in Chapter 3 has confirmed that networking optimization cannot be automatically guided by causal knowledge in some practical occasions. The fundamental reason has been discussed in the geometric meaning analysis of causal graphs.

On the other hand, the concept of representation learning is not particularly for DL methods, but was proposed about 100 years ago, including the unsupervised principal component analysis (PCA) in 1901 and supervised linear discriminant analysis (LDA) in 1936. In recent years, the deep network architectures for representation learning

have turned out to be widely applied [29]. Data representation describes how much the explanatory factors of variation behind the data have been revealed [30], and thus plays an important role in determining the success of machine learning methods. In the quest for applicable Artificial Intelligence (AI), more powerful methods are motivated to be designed in this direction.

2.2 Causal Analytic Methods

The evolution of causal analytic techniques looks back on decades of development and several textbooks have been published on the topic [31–33]. Reviewing all causal inference and causal discovery methods is thus unnecessary, and impractical. In this section, we mainly focus on causal discovery techniques that most closely relate to the present thesis. We discuss these techniques in three categories based on the principle they operate on. A more detailed account of causal inference methods will be provided in the context of healthcare in Section 2.3.

2.2.1 Conventional Heuristic Methods

Usually, people utilize the independency relationships among variables from the data to discover the underlying causal structures. To avoid a combinatorial explosion in the search space (of all possible DAG solutions), most methods approximate the solutions by using constraints or score-based heuristics [6].

The typical constraint-based methods include PC and Fast Causal Inference (FCI) [34]. PC provides a search architecture for the selected statistical procedure to decide conditional dependence, such as some background knowledge-directed hypothesis tests, with the assumption of no unobserved confounder. FCI is an important variation of PC, supplemented with a heuristic guideline, which tolerates unobserved confounders to some degree. Such approaches are widely applicable if given reliable conditional independence testing. Still, they do not necessarily discover full causal information because multiple causal structures may satisfy the same conditional independence constraint.

A score-based method, e.g., Greedy Equivalence Search (GES) [35], directs its searching using a score-function as a heuristic, but is still computationally costly due to the combinatorial optimization nature and inefficient searching strategy.

2.2.2 Directed Graphical Models

Directed graphical causal models are commonly presented as "Structural Equation models" (SEMs), or "Functional Causal Models" (FCMs) [13, 14]. They assume the value of each variable to be a deterministic function of its direct causes in the DAG, along with some unmeasured extraneous influences. The algorithm applies constraint on the global joint probability distribution over all variables, so as to decompose it into a product of simpler factors, which is taken as a representation of the conditional independence relationships in data [36–38].

However, this class of models does not necessarily have a causal interpretation. In practice, the causal sufficiency assumption is often violated, and for some causalities with multiple independent timelines, Causal Representation Bias (CRB) is risky to happen. Due to the existence of CRB, the DAG structure only represents what we have known, so the separation procedure applied to the DAG can only provide necessary that are not sufficient conditions for identifying the variable-level conditional independence relationship. There may exist additional relations of interest but not entailed by the separation.

2.3 Causal Learning of Healthcare Data

Although Machine Learning (ML) methods have already successfully improved predictions on EHR data, such as modeling disease trajectories, forecasting disease deterioration, risk factors, and mortality rate, as well as predicting treatment responses, the demand for model interpretation is more fundamental than accuracy [39]. For the purpose of helping domain scientists understand the best way to act and give treatment, ML models must be able to interpret causal relationships in EHR data. In this context, machine learning methods for causal inference can be used to learn the effects of treatments from observational data and subsequently provide clinicians with actionable intelligence for making treatment decisions.

In this section, we will review current state-of-the-art causal inference methods applied to EHR observational data, in the longitudinal setting for learning temporal causality.

2.3.1 Causality in EHR

As part of the American Recovery and Reinvestment Act of 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted, to promote the adoption and meaningful use of health information technology. Close to $30 billion dollars was spent on creating a nationwide repository of electronic Health Records (EHRs), which consists of patient attributes such as demographics, diagnosis codes, laboratory test results, and vital information. Furthermore, in the future, all such information about patient healthcare attributes will be directly collected via EHRs.

In the past decade, the established EHR repository has significantly contributed to healthcare studies, because it facilitated the secure, efficient and effective sharing and using of electronic health information when and where it is needed. Specifically, EHRs can help health care providers recommend treatments that are better tailored to an individual's preferences, genetics, and concurrent treatments; it can help individuals make better treatment decisions and health-impacting decisions outside of the care delivery system; and can help reduce care delivery redundancy and cost by allowing test results to be reused while supporting analyses to pinpoint waste. These applications of these studies have been important contributors to improving health outcomes, improving health care quality, and lowering health care costs – the three overarching aims that the U.S. is striving to achieve.

From the data science point of view, EHR data has enabled better documentation of existing treatments and medical interventions. As a data platform, it helps the data researchers efficiently investigate underlying information and realize more solid applications. For example, a better understanding of existent disease trajectories can give rise to more accurate disease forecasting, and correctly evaluating the existent treatment effects can provide important references for physicians. These applications could promote care coordination between service providers and patients' healthcare outcomes, and reduce health disparities, thereby improving the overall quality of the healthcare management system.

One of the factors making EHRs valuable is that they contain rich causal information obtained from actual clinical practice. Causality is an important concept throughout health science development, and it is particularly vital for clinical informatics work, such as finding adverse drug events or risk factors for disease by using EHR data to

support clinical decisions. There have been works discussing how we can discover or estimate causality from large-scale observational data with high accuracy, in the context of machine learning methodology. But essentially, learning from EHRs is to enable reasonable bioinformatic inferences, and thus, any conclusions need to be interpretable. In this work, we aim to draw clinically meaningful conclusions from EHR data by identifying and understanding the hidden causal effects.

2.3.2 Challenges of Causal Learning

Before delving into the question of how to find causality, we need to understand why we need causes at all, and why associations cannot be used instead. The primary uses of machine learning methods for causal analytics are predictions and explanations.

Predictions on EHR data are usually aimed at determining the probability that a health event is to happen given certain conditions, such as new medications used or a change in the living environment. Predictive models are frequently constructed based on associations alone, as in some clinical informatics studies [40]. But this can be problematic as we do not know why the associations hold, thereby cannot know how and when they will stop working. For example, we may be able to predict the respiratory disease rate in a city is based on the daily automobile exhaust volume, but the correspondence between city traffic and respiratory health is indirect. In addition to daily commutes, many variables may affect air quality. Once we have data on the daily air quality index, information about vehicle traffic volume becomes redundant. With seasonal flu outbreaks, the effect of air quality may only be a weak factor. More generally, black box models based on associations may have redundant or unstable variables, leading to biased estimation or unnecessary medical tests.

Another primary use of causal inference is interpretation. We are seeking the influence of particular events, such as how much the concentration of particulate matter in the air influences human lung health; and also, the relationship between two phenomena, e.g., how an adverse drug event is associated with an underlying disease and medical history. In the former case, we generally want explanations for causal inferences and predictive rules, referred to as causal effects. For the purpose of medication evaluation for patients with different characteristics, we are interested in Individual Treatment Effects (ITE) estimation, an important research branch in clinical studies, especially when

such studies are going to be used for real services, such as clinical decision support. In the case of relationship interpretation, we aim to describe how the elements cause each other or are commonly caused by what element, which is usually considered to be causal structure discovery.

EHR data contains much clinical information about heterogeneous patients and their treatment responses. To estimate ITE from EHRs, machine learning models are trained to estimate potential outcomes (i.e., the unobserved patient outcomes under different treatment options). These models are conditioned on each patient's clinical state, which is based on how other patients have responded to the treatment in historical records. However, the complex causality in her data poses several challenges and requires strong assumptions that must be rigorously assessed before modeling.

The first challenge is the complex confounded relationship. In observational data, the treatment is often highly dependent on the patient's clinical conditions. For instance, cancer patients with more aggressive tumors are more likely to receive more extensive treatments, but despite the more aggressive treatment, they typically still have worse outcomes. This phenomenon may induce biased results that, without model adjustment, can incorrectly lead to interpreting the dependence as causation and concluding that the aggressive treatment is harmful to patients. Based on clinical knowledge, one can consciously de-confound the known relations when modeling. But unknown confounders, due to incomplete knowledge or coincidental confounding in data, probably still exist and can lead to unknown biases.

The second challenge is data heterogeneity. As real-world observations, the patients recorded in EHRs are highly heterogeneous in both individual characteristics and treatments, which leads to a covariate shift in the training data distribution and consequentially may result in model selection bias. For instance, the outcomes of sicker patients receiving extensive treatments would distribute differently from those with mild diseases and ordinary treatments. Thus, the overall supervised learning models can be biased and cannot be generalized well over various patient cohorts.

The third challenge is the inevitability of improper causal assumptions. Even with appropriate model adjustment given prior knowledge, once the learning model has been selected, we still have to assume that no additional hidden confounders exist. In other words, the causal *Suffeciency* has to be assumed for the adjusted model. In

particular, the prior knowledge, including the underlying causal structure (i.e., causal graph), usually comes from consultations with doctors or key opinion leaders, who pay more attention to model interpretability than to improved predictive accuracy. Thus, the specified models probably contain unknown bias, which empirically leads to unknown dependencies that may violate the given knowledge or mislead the results.

In summary, due to the nature of EHR data, causal discovery and inference create significant challenges for machine learning. Whereas many methods have been developed to handle the possible model selection biases in various applications, from the general view, most challenges are still unsolvable. Knowledge from domain experts can theoretically describe the fundamental causation, but cannot fully explain the empirical distribution of observations, while the inevitable modeling bias makes it hard for data-driven methods to generalize and integrate the existing knowledge. So far, we rarely find works that aim to systemically discuss or evaluate the information gap between theoretical knowledge and its practical representation in data. There is still a long way for us to go toward interpretable causal modeling.

2.3.3 Treatment Effect Estimations

Conventional statistical methods to estimate individualized treatment effects are usually applied in a cross-sectional study. In other words, a single treatment decision is made based on the one-time measurement of patient features. Advanced statistical techniques to model causal relationships in complex longitudinal data (such as structural marginal models and G-estimators for complex longitudinal data [33]) exist, but extending to address the causal representation bias is highly non-trivial. For this reason, we chiefly focus on machine learning methods. Machine learning methods allow modeling time series to perform causal inference, which can extract actionable intelligence from patient historical records, and goes beyond standard supervised learning methods for prediction. In the following paragraphs, I will describe what is possible to achieve using current state-of-the-art machine learning methods for treatment effect estimation.

For ML, the training data can include all observational records collected before the treatment decision is made, which contains patient demographic data, such as age, sex, and race, and also clinical outcomes, such as diagnosis history, comorbidities, symptoms, and laboratory measurements. Conditioned on these data, ML methods can be used to

estimate the effects of single or multiple treatments on various patient outcomes.

But compared to the cross-sectional setting, it becomes more challenging to estimate treatment effects on time series, which involve temporal information, and the presence of time-dependent confounders: patient outcomes are affected by past treatments, and these treatments are covariates that influence future treatments and outcomes. For instance, for Type II Diabetes (T2D) patients, we need to estimate the effect of an antiglycemic drug on the risk of experiencing a cardiac event. (An effective antiglycemic drug will lower blood glucose levels.) Sustained high blood glucose levels for a patient will result in an escalation of the treatment, while also leading to an increased risk of cardiac events. Thus, blood glucose levels are confounders that cause both treatment and outcomes. Because the blood glucose levels are only affected by past treatments, this is considered time-varying confounding.

The way to adjust these confounding biases is to involve conditions that lead to these treatment decisions, which is also named the de-confounding process. Therefore, in addition to the observable patient trajectories, we also need the counterfactual values of outcomes. In other words, we need the observational data about the patients who have received the treatment, as well as those who did not receive (but could have been treated), and their corresponding outcomes. In conclusion, this de-confounding method works based on two assumptions: 1) The unconfoundedness assumption in the sequential setting, which means that at each time step, we observe all variables affecting the patient's treatment and outcome; 2) the treatment decision conditions remain consistent along the timeline. To be specific, at any time step we assume that each treatment option has a certain non-zero probability of being administered. The second assumption can be assessed by computing the probability of a patient receiving each possible treatment at each time step. Moreover, validating the unconfoundedness assumption requires domain expert knowledge, performing sensitivity analysis [41, 42], or estimating substitutes for the hidden confounders [43].

2.3.4 Deep Learning on EHR Predictions

During the last few years, Deep Learning (DL) techniques have achieved great success in EHR data analysis, due to their ability to effectively construct deep hierarchical features and capture long-range dependencies in data [44]. As one branch of machine learning

methods, deep learning has been widely used in EHR data analysis in multiple types of predictive modeling tasks, such as patient diagnosis prediction, features forecasting, risk estimation, and so on. Despite these regular prediction tasks that machine learning is usually used on, DL approaches have been used target more challenging problems. For example, Deep Patient [45] predicts multi-outcomes of patients, Deeper [46] targets hospital re-admissions, and Doctor AI [47] was proposed for heart failure prediction. Apart from achieving higher prediction accuracy, DL models have also been employed to abstract information, such as the representation of medical concepts [48, 49], and patient features [45].

In time series EHR data analysis, i.e., in the sequential setting, RNNs have been extensively adopted to yield better performance than traditional time-series models [50,51]. The two most popular RNN mechanisms, Gated Recurrent Units (GRU) [47,52] and Long Short-term Memory (LSTM) [53,54], were developed to leverage the temporal clinical information.

However, the challenges brought by the special characteristics of EHR data still hold for DL methods. Generally, convolutional neural networks assume a certain structure of variables and consider each variable consistently in a global view. In RNNs, the events input at each time step are explained by a unique group of attributes. But in clinical practice, the same events can have different effects on different variables, according to the different disease severity, which may be unobserved. But current neural networks models are mostly based on the no hidden confounders assumption, which may lead to the loss of such relationships, and cannot guarantee the learned models to be clinically interpretable.

In my literature review of the state-of-the-art works, people mostly deal with causal effects in EHR data in two ways: 1) simply focus on a single disease, and assume the clinical events, like medications, to be uniquely effective [55, 56], and 2) perform causal discovery to describe the complex relationship, based on standard causal assumptions [36, 57]

On the other hand, causal inference combined with DL approaches has shown increased interest, including combining DL methods with causal discovery, in both the DL and the causality communities [58]. These works can be classified into two categories: 1) learning causal model structure, and 2) estimating the causal effects of a particular

input. In the first category, a neural network is viewed as a Functional Causal Model (FCM) [36–38], where each directed edge represents the causal relation between two variables and is described by a function. In the second category, people aim to answer the question "How much would perturbing a particular input variable affect the output?" [59], where the causal structure is left as a black box [60], or assumed to be well-known [58, 61]. As mentioned ahead, these causality inference works are usually based on standard causal assumptions [15].

For the underlying causality of EHR data, the standard assumptions may be unrealistic. A simple counterexample is that unmeasurable causes usually exist, e.g., the experience of doctors. Thus, we have a basic hypothesis that although DL models are potentially able to enhance the predictive accuracy for treatment effect estimation, they are still under the restrictions of causal representation bias in data.

Chapter 3

Improving Deep Learning Predictive Performance by Incorporating Causal Effects Domain Knowledge

3.1 Introduction

Given the wide adoption of Electronic Health Records (EHR) systems in the US, there is a surge of interest in building EHR data-driven models for predicting clinical events. These models aim to estimate the risk for disease development, which enables better diagnosis, prognosis and treatment selection. Their predictive performance is crucial for healthcare quality, especially in the context of the real-time tools available at the bedside (e.g. precision medicine [62]).

It is known that disease development is determined by underlying causal mechanisms, such as functional genetic variants, environmental factors, and sequential effects through organs. Particularly, for chronic diseases, the risk factors are shared over multiple comorbidities, which leads to complex developing trajectories that are causally connected. Thus, causal mechanism discovery is essential to establish interpretable and accurate predictive models.

Recently, deep learning (DL) techniques have drawn much attention from healthcare investigators because of their demonstrated superior performance in predicting disease risks using EHR data [44, 63]. Containing a wealth of real-world data, EHR holds promise for the discovery of underlying causal mechanisms, yet the existing DL applications fall short of leveraging them due to the special characteristics of EHR data, which has restricted their performance and interpretability.

In this paper, we propose the establishment of a new DL-based predictive model by leveraging clinical knowledge that has been accumulated over many years. In particular, we use a classic method to estimate the causal effects of clinical treatments from EHR data and incorporate them into the DL model. The applied domain knowledge brings complementary causal information, provides more accurate predictions, and also makes the DL models to be more interpretable. The conclusion has corroberated our initial hypothesis that DL models are unable to automatically discover clinically interpretable causal effects from EHR data.

We will start from the analysis of underlying causal relations in EHR data (Section 3.2); and propose our methodology (Section 3.3). For experiments, we firstly describe the experimental designs (Section 3.4), and then fully discuss the experimental results (Section 3.5), followed by the summary of conclusions (Section 3.6).

3.2 Analysis of Clinical Causation

In this section, we will discuss the characteristics of clinically explained causal relations hidden in EHR data, and use specific examples to illustrate its violations of the standard assumptions.

3.2.1 Underlying Causal Structure

Consider an EHR dataset that spans T time steps. At each time step t, the observed events consist of the patient features X (e.g. lab results, vital signs, and demographics), the received treatments M (such as medications and procedures), and the assigned diagnosis codes Y (e.g. ICD-10 code for diabetes).

Fig. 3.1 displays two different causal relationships: the left one is obtained only based on observations, and the right one is the underlying (but unobserved) causality

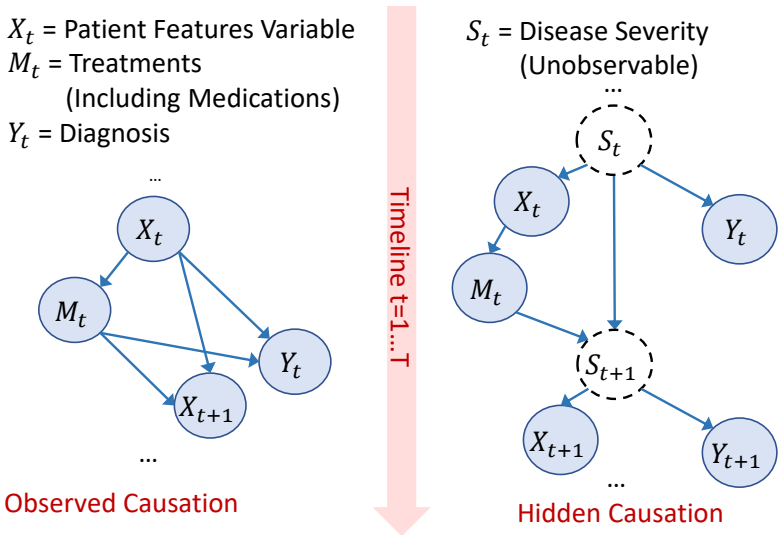


Figure 3.1: Comparison of observed relation and the true causation hidden in EHR. The solid and filled circle represents observable events, while the dashed unfilled one means unobservable.

in general EHR data, manually identified with domain knowledge.

It is reasonable to assume that a hidden variable S_t exists, which fully represents the patient's *Disease Severity* at time t. In other words, it consists of all necessary features to be the representation, while X_t only comprises observable ones. Thus, an ideal S_t could be more (or at least the same) informative than X_t. For instance, a medicine may hurt renal function as a side effect, and will reflect on glucose (a lab result) after 5 time steps. S_t should capture the functional injury right after the medicine is taken, even though X_t has no reflection. Then, if observed, S_t will be predictive of the future effects on X_{t+5}.

However, as far as we know, no instrument exists to precisely scale "organ damage". Thus the disease severity is not directly observable, but able to be inferred from relevant lab results, or different medication levels, which are observable features. Therefore, the conventional methods usually consider the left side of Fig. 3.1 as the causation, which is assuming X_t to be S_t, along with some domain-specific scales.

3.2.2 Violation of Causal Assumptions

Replacing unobservable S_t with observed X_t works mostly in practice, but the nature of EHR data is unchangeable. Distorted causal mechanism leads to violations of standard

assumptions in clinical analysis.

In the following, we will describe particular characteristics of EHR data from three aspects, with specific examples.

Unreliable Criterion

The disease criterion is usually based on lab results. But it is not reliable because the sign can be triggered by other factors and lead to false positives. To solve this Causal Sufficiency violation problem, professional instructions should be input as additional variables.

[Example A]. For diabetes patients, high Blood Pressure (BP) is a sign of danger. But in cardiovascular surgery (which may be caused by severe diabetes), increasing BP
by anticoagulant drugs is mandatory to prevent blood clotting.

Misunderstood Medications

For chronic diseases, higher level medicine use is considered a criterion of less healthy. But the reality may be counter-intuitive. To fully understand medical effects, more information is in demand. Put differently, the Causal Sufficiency assumption is violated.

[Example B]. Increasing insulin use is a significant sign of worse diabetes. But after a kidney transplant surgery (which may be caused by severe diabetes), this increasing conversely means a more functional body.

Incomplete Info

The Causal Markov Condition assumption is invalid on EHR data because patients' current status and medication are insufficient to determine severity. Their current personal treatment effects may help, if well estimated.

[Example C]. A pregnant woman who just developed gestational Hypertension (HTN) has very similar lab results with an HTN patient who just got pregnant, but differentiate on personal effects to anti-hypertension drugs.

3.3 Proposed Methodology

Our objective is to find a treatment effect estimator to augment the EHR data, which extracts the underlying causal information without any additional data. We integrate the estimator into modeling to maximally rationalize the standard assumptions, and eventually improve the predictive performance of DL models.

The proposed effect estimator is based on well-developed clinical knowledge. In this section, we will firstly introduce its definition (subsection 3.3.1), then the calculation method will be presented (subsection 3.3.2), and finally followed by the theoretical analysis (subsection 3.3.3).

3.3.1 Treatment Effect Estimator Δ

We represent the patients' observable individual-level features as variable X, such as lab results value and vital signs; and use M to denote the desired treatments applied to patients, including medications, surgeries, and so on.

Let vector $M = (m_1, m_2, \ldots, m_n), m_i \in [0, 1]$, where m_i indicates whether the ith treatment (n treatments in total) happened ($m_i = 1$) or not ($m_i = 0$). We use M^{i+} and M^{i-} to represent a pair of treatments that only have different values on $\{m_i\}$, and keep identical on the other treatments. In specific, M^{i-} means $(m_1, m_2, \ldots, m_i = 0 \ldots, m_n)$, while M^{i+} means $(m_1, m_2, \ldots, m_i = 1 \ldots, m_n)$.

Assume that a patient is known to have treatment M^{i-} at time $t-1$, denoted as M^{i-}_{t-1}, and at the next time step t, he/she can be possibly added $\{m_i\}$ to become M^{i+}_t, or keeps in M^{i-}_t steadily. To be brief, we use X^{i+}_{t+1} to represent the patient's features at time $t+1$ if $\{m_i\}$ is added at t, $X^{i+}_{t+1} = (X_{t+1}|M^{i+}_t)$, and accordingly, $X^{i-}_{t+1} = (X_{t+1}|M^{i-}_t)$ indicates his/her features if another possibility happened. Then,

Definition 1. *With M^{i-}_{t-1} known, Treatment Effect Estimator Δ^i_{t+1} about treatments $\{m_i\}$ at time t, is defined as*

$$\Delta^i_{t+1} = \begin{cases} (X^{i+}_{t+1} - X^{i-}_{t+1}), & \text{if } M^{i+}_t \text{happened} \\ 0, & \text{otherwise} \end{cases}$$

Here $\{m_i\}$ could be considered as any desired treatment combination, as a valid subset of $\{m_1, m_2, \ldots, m_n\}$, i.e. $\{m_i\} = \{m_{i_1}, m_{i_2}, \ldots, m_{i_j}\} \subseteq \{m_1, m_2, \ldots, m_n\}$

Since M_t^{i-} and M_t^{i+} are exclusive to each other, only one of X_{t+1}^{i-} and X_{t+1}^{i+} could be observable. By definition, Δ_{t+1} only has meaningful value for the patients who have M_t^{i+} truly happened at t. Therefore, we call the observed X_{t+1}^{i+} as *Factual Value* and the unobserved X_{t+1}^{i-} as *Counterfactual Value* of X at time $t+1$.

The proposed estimator Δ is inspired by a well-grounded concept in clinical study: *Individual Treatment Effect* (ITE) [64]. ITE is usually defined as the individual-level difference between the factual and counterfactual values of the outcome for each patient, such as some diagnosis or mortality risk.

To estimate a meaningful Δ, in this work, we adopted one of the conventional clinical statistical methods for ITE estimation.

3.3.2 Estimation of Δ Value

As shown in Fig. 3.1, because of confounding, the effect of M is mixed with other causal effects. Here we employed *Propensity Score Matching* (PSM), a widely used method, to isolate the treatment effects and pursue unbiased estimation.

PSM aims to create a de-confounded pseudo-population, on which the counterfactual value X_{t+1}^{i-} could be estimated without bias. This pseudo-population consists of a number of patient pairs, in which the two patients are comparable.

Definition 2. *In prediction task of treatment M_t, a pair of patients (r_1, r_2) is called Comparable iff they have M_t^{i+} and M_t^{i-} respectively, and have the X_t values to be approximately same predictive about M_t. i.e. $\mathbb{P}(M_t|X_{r_1,t}) \approx \mathbb{P}(M_t|X_{r_2,t})$. And $\mathbb{P}(M_t^{i+}|X_{r,t})$ is defined as patient r's Propensity Score.*

Specifically, for the patients who have known M_{t-1}^{i-}, we perform logistic regression to calculate their Propensity Scores, $\mathbb{P}(M_t^{i+}|X_t)$, i.e. the probability of having M_t^{i+} by given X_t. Then we split them into two groups: One is *Treated* group with known M_t^{i+}, and another is *Untreated* group with known M_t^{i-}. Comparable patients are selected from the two groups respectively, whose propensity scores differ within a small caliper. This process is called *Matching*, and the matched patient pairs comprise the target pseudo-population.

Because the value of X_t has been de-confounded with M_t in this population, we can easily estimate the unconfounded counterfactual value as a function $X_{t+1}^{i-} = f(X_t, M_t^{i-})$.

3.3.3 Theoretical Functional Analysis

In this subsection, we will analyze the role of Δ.

Firstly, Δ is informative

Let's restate the conception of S_t, the Disease Severity about Time t, by the essential assumptions below. The temporal causal relationship we are talking about is described as a DAG:

Asm 1. Time Series Assumption*: Given S_t as parents, its child nodes can only happen at the same or a later time, i.e. for any $X_{t'}$ caused by S_t, $t' >= t$.*

Asm 2. Info Sufficient Assumption*: S_t is informative enough that given S_t as parents, for any child node of it (including S_{t+1}), all the common causes in the DAG have been included in the DAG.*

Asm 3. Linear Progress Assumption*: About any time t, S_t exists and linearly causes S_{t+1}. In other words, S_{t+1} can be described as a linear transformation of S_t for any t.*

Because M_t can only be one of M_t^{i+} and M_t^{i-}, S_{t+1} can be rewrote as $S_{t+1} = [(S_{t+1}|M_t^{i+}) - (S_{t+1}|M_t^{i-})]\mathbb{1}(M_t^{i+}) + (S_{t+1}|M_t^{i-})$. Let's define its two terms separately as below:

Definition 3*. With M_{t-1}^{i-} known, the Treatment Action $\mathcal{M}_t^i$ at time t is define as $\mathcal{M}_t^i = [M_t^{i+} - M_t^{i-}]\mathbb{1}(M_t^{i+})$, representing the treatment changing that happened at time t.*

Definition 4*. The Naturally Progressed severity at time $t+1$ is defined as $S_{t+1}^N = (S_{t+1}|M_t^{i-})$.*

Definition 5*. The M-Changed part of severity at time $t+1$ is defined as $S_{t+1}^M = [(S_{t+1}|M_t^{i+}) - (S_{t+1}|M_t^{i-})]\mathbb{1}(M_t^{i+})$*
$= [(S_{t+1}|M_t^{i+}) - (S_{t+1}|M_t^{i-})]\mathbb{1}(\mathcal{M}_t^i = 0)$.

We also split X_{i+1} into two terms as $X_{i+1} = X_{t+1}^M + X_{t+1}^N$, and let $X_{t+1}^N = (X_{t+1}|M_t^{i-})$, and $X_{t+1}^M = [(X_{t+1}|M_t^{i+}) - (X_{t+1}|M_t^{i-})]\mathbb{1}(\mathcal{M}_t^i = 0)$. Then we have property as below:

Property*: $S_{t+1}^N \perp\!\!\!\perp S_{t+1}^M$ and $X_{t+1}^M \perp\!\!\!\perp S_{t+1}^N$, given S_t.*

Proof: By given S_t, value of S_{t+1} is determined. Then we have S^N_{t+1} as constant, and S^M_{t+1} to be a variable only about $\mathcal{M}^i_t$, thus $S^N_{t+1} \perp\!\!\!\perp S^M_{t+1}$. Suppose $X^M_{t+1} = f_1(S^M_{t+1}) + f_2(S^N_{t+1})$ where $f_1()$ and $f_2()$ are general causal mapping functions. By definition, $X^M_{t+1} = 0$ iff $S^M_{t+1} = 0$, so we have $f_2(S^N_{t+1}) = -f_1(0)$ to be constant, thus $X^M_{t+1} \perp\!\!\!\perp S^N_{t+1}$.

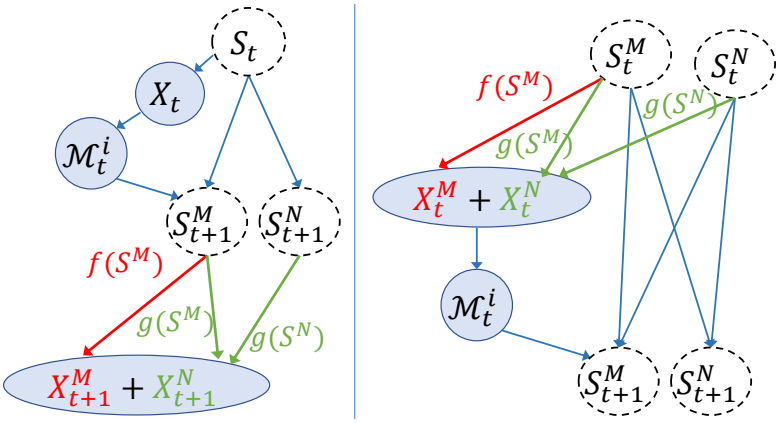


Figure 3.2: Two Partial DAGs for disassembled causality. The solid circle means observable variables, while the dashed one is not.

Theorem 1. *X^M_t can be calculated w/o estimation of S_t.*

Proof: Without loss of generality, suppose $X^M_{t+1} = f(S_{t+1})$, and $X^N_{t+1} = g(S_{t+1})$, where $f()$ and $g()$ are general functions that respectively describe a causal mapping relation, from the disease severity space to the data feature space. Then given Property we have $X^M_{t+1} = f(S^M_{t+1} + S^N_{t+1}) = f(S^M_{t+1})$, and $X^N_{t+1} = g(S^M_{t+1} + S^N_{t+1}) = g(S^M_{t+1}) + g(S^N_{t+1})$. The relationship of function $f()$ and $g()$ are illustrated in Fig.3.2. Substituting the former expression into the later one, we have:

$$X^N_{t+1} = g(f^{-1}(X^M_{t+1})) + g(S^N_{t+1}) \tag{3.1}$$

The right side of Fig.3.2 displays the preceding causal relations of S^M_{t+1} and S^N_{t+1}. Given the observed $\mathcal{M}^i_t$, we have S^M_{t+1} to be independent with X_t, where $X_t = X^M_t + X^N_t$. We also have S^M_t independent with S^N_t. Therefore, S^M_{t+1} can be expressed as the sum of three independent terms:

$$S^M_{t+1} = \mathcal{M}^i_t + \alpha S^M_t + \beta S^N_t \tag{3.2}$$

Where α and β are two linear transformation matrices that describe the linear causal relations $S_t^M \to S_{t+1}^M$ and $S_t^N \to S_{t+1}^M$ respectively. Then from $X_{t+1}^M = f(S_{t+1}^M)$ we have:

$$\begin{aligned} X_{t+1}^M &= f(\mathcal{M}_t^i + \alpha S_t^M + \beta S_t^N) \\ &= C + \alpha X_t^M + \beta f(S_t^N) \end{aligned} \tag{3.3}$$

Where C indicates a constant variable. Then from Eq.(3.3) we have an expression about S_t^N as: $S_t^N = \frac{1}{\beta} f^{-1}(X_{t+1}^M - \alpha X_t^M - C)$. Additionally, because of the isomorphism of t and $t+1$, from Eq.(3.1) we can deduct another expression about S_t^N as $S_t^N = g^{-1}(X_t^N) - f^{-1}(X_t^M)$. By integrating them we have:

$$X_{t+1}^M = (\alpha - \beta) X_t^M + \beta f(g^{-1}(X_t^N)) + C. \tag{3.4}$$

By Eq.(3.4), given mapping functions $f(), g(), \alpha$ and β, X_{t+1}^M can be expressed without severity values, so proof is done.

Theorem 2. *The proposed estimator Δ_t can reveal hidden information from unobserved severity S_t.*

Proof: By definition we have expression about Δ_{t+1} as $\Delta_{t+1} = [(X_{t+1}|M_t^{i+}) - (X_{t+1}|M_t^{i-})]\mathbb{1}(\mathcal{M}_t^i = 0)$, thus $\Delta_t = X_t^M$. Since X_t is observable, having Δ_t is equivalent to having X_t^M and X_t^N both become known.

With known X_{t+1}^M, X_t^M and X_t^N, by assuming proper values of α and β, from Eq.(3.4) we can learn the nested mapping function $f(g^{-1}(\cdot))$ as:

$$f(g^{-1}(X_t^N)) = \frac{1}{\beta} X_{t+1}^M + (1 - \frac{\alpha}{\beta}) X_t^M - \frac{C}{\beta}. \tag{3.5}$$

Because function $f()$ and $g()$ integrally represent the causal mapping from unobserved S_t to observed X_t, proof is done.

Theorem 3. *Incorporating Δ_t can improve predictions.*

Proof: From Theorem 2 we know that given Δ_t and X_t, value of X_{t+1}^M could be predictable, with proper setting of α and β. Furthermore, from Eq.(3.1) we know that with known $g(S_{t+1}^M) = f(g^{-1}(X_{+1}^M))$ and $g(S_{t+1}^N) = X_{t+1}^N - g(S_{t+1}^M)$, the relation between $g(S_{t+1}^M)$ and $g(S_{t+1}^N)$ can also be learned. Thus both X_{t+1}^M and X_{t+1}^N become predictable with Δ_t given.

If no prior knowledge is applicable to set up α and β, they will be randomly assigned reasonable values in the learning process of any model. However, compared to linear regressions, DL methods can be improved more on predictive performances. Because linear models imply linear assumptions: 1) $g(S_{t+1}^M)$ and $g(S_{t+1}^N)$ are linear related, thus $g()$ is linear; 2) $f(g^{-1}())$ is a linear transform, in other words, $f()$ and $g()$ are linearly associated, thus $g()$ is also linear. On the contrary, DL is capable to learn nonlinear relations, with $f()$ and $g()$ both being nonlinear, which is mostly the actual circumstance.

Secondly, Δ can guide NNs attention

The effect estimator Δ_{t+1} is initialized as a zero vector, whose digits can only be assigned non-zero values when the changing of treatment, i.e. the changing from M^{i-} to M^{i+}, happening at time t can significantly impact the immediate subsequent disease severity.

Despite large possible treatment combinations for $\{m_i\}$ in theory, the number of interesting treatment changes is often limited in practice. Therefore, a selection of treatment changes can be used prior to generating sparse Δ values, which helps dramatically reduce the computational cost.

In DL training process, it is highly efficient to introduce a sparse matrix (or vector) as an input to guide the neural networks' attention. Because the zero values can avoid numerous meaningless back-propagation computing.

Furthermore, Δ can be set as manually controllable, to provide a new interface for physicians to input domain knowledge. e.g. By manually zeroing out the less important digits, we can make the output to be more clinically interpretable.

Since Δ is able to reveal the hidden disease severity, it can contribute to correcting the standard assumption violations, like the feasible solutions of the aforementioned examples:

[Example A]. At surgery time, disable the BP relative digits in Δ to avoid wrong attention.

[Example B]. The positive effect reflected on Δ will distinguish this special case.

[Example C]. The individual-level effects Δ will differentiate the two conditions.

Table 3.1: Study Cohort Description

Variable	**Median**	**Interquartile Range**
Age [years]	45	31, 59
Male [%]	43.4	
LDL [mg/dL]	105	83, 130
TG [mg/dL]	130	91, 187
HDL [mg/dL]	50	41, 61
SBP [Hg mm]	120	108, 132
DBP [Hg mm]	70	60, 78
FPG [mg/dL]	101	92, 117
Follow-up [years]	10	6, 12
	Percent	**Number of patients**
Antihypertensive medication	42.4	30968
Antilipemic medication	32.8	23925
Progressed to DM	16.9	12367
Progressed to CAD	17.9	13065
Progressed to CKD	7.2	5267
Progressed to Stroke	1.1	779

3.4 Experimental Design

In experiments, we aim to verify that Δ is informative enough to significantly improve DL predictive performance.

For convincing conclusions, we employed two types of datasets to perform experiments: (1) a real-world EHR dataset of type-II diabetes patients; (2) multiple synthetic datasets generated with underlying causation, determined by random DAGs.

Due to the practical limitation of real data (e.g. missing values), the advantage of Δ cannot be fully revealed on it. That's why synthetic data is necessary.

In this section, we will introduce the two types of datasets and the experimental tasks on them respectively. The synthetic data generation method will be illustrated and the source code has been published for free downloading.

3.4.1 Real EHR Data

Mayo Clinic (MC) provides primary care to residents of Olmsted County, Minnesota, and has an integrated electronic health record system including diagnoses, medications, laboratory results, and clinical notes.

We used a retrospective cohort of de-identified data from 73,045 primary care patients at Mayo Clinic, Rochester, MN with research consent. The cohort consists of patients aged $\geq$18 and $\leq$ 89 at baseline on Jan. 1st, 2005, having at least one visit before and after baseline. These patients were followed until 2017 (the median follow-up time is 10 years). We extracted patient demographics, and diagnoses (ICD-9), laboratory results, vital signs, and medications longitudinally for six non-overlapping time windows: before-2004, 2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-after. For the outcomes that happened in time window t, we make predictions based on observations from previous time windows, i.e., t-1, t-2, etc. In each time window, the latest observed values are taken for the predictors.

Cohort Details

As the description in Table 3.1.

Medications

All referred medications for T2DM and related comorbidities have been rolled up to National Drug File Reference Terminology NDF-RT pharmaceutical subclasses.

In this dataset, We have three medicine classes: T2DM (DM) drugs, Hyperlipidemia (HL) drugs, and Hypertension (HTN) drugs. Table 3.2 displays all three classes in increasing levels, where a higher level is commonly for higher severity. Each class is encoded as a binary vector with one digit representing one level, and we finally concatenate all three vectors to comprise the treatment indicator M.

Prediction Tasks

We comprehensively set up two types of tasks. One type is binary classification, and another one is continuous outcome regression, as below:

Table 3.2: The treatment levels for medication classes

Treatment Levels	Medication Classes		
	DM	HL	HTN
Level-1	1st non-insulin drug	statin as the 1st drug	1st drug
Level-2	2nd or more non-insulin drug	non-statin drug as the 1st drug	2nd drug
Level-3	insulin added as the 1st drug	non-statin drug added to statin	more drugs
Level-4	insulin added to non-insulin drug(s)	non-statin drug added to non-statin drug(s)	-
Level-5	non-insulin drug added to insulin	-	-

Task 1: Disease diagnosis predictions. The outcomes in this task include the disease diabetes (DM) and its three complications (CAD, CKD, and Stroke). We aim to predict whether the disease developed in the subsequent time window, using predictors from the immediately preceding time windows.

For a brief, we refer to the number of preceding time windows as *Time Step* or *Step*. e.g. Step=1 means the predictors from the preceding time window (2011-2012); Step=2 means from the preceding two time windows (2009-2010 and 2011-2012).

We perform four groups of experiments, with Time Step being $1, 2, 3, 4$ respectively. **Task 2: Lab results forecasting.** The outcomes in this task are continuous values, including three critical lab results in the immediately subsequent time window. Same as Task 1, four groups of experiments have Time Step $= 1, 2, 3, 4$ respectively.

Compared Models

We implemented two RNN architectures in a different ways of using Δ: one is simply augmenting Δ as input, and another is pre-training with Δ, where the latter one is expected to be more efficient to utilize Δ's information. RNN-GRU and RNN-LSTM mechanisms are used respectively. Besides, as a comparison with RNNs, two linear

models are implemented as baseline methods.

For each model, we obtain a pair of performances: with and without Δ input, whose difference reflects Δ's effect.

Model 1: Augmentation Architecture RNN. For each time window t, we concatenate X_t with Δ_t as new predictors, then use such augmented data as input to RNN model.

Model 2: Unsupervised Pre-training RNN. An alternative approach to use Δ to guide the learning process of RNN model: We pre-train the RNN such that the model predicts Δ_{t+1} from X_t at each t. Since the pre-training does not require labels, i.e. unsupervised, test data can also be used.

Model 3: Logistic Regression Model. Δ_t variables are simply augmented with X_t as new predictors.

Model 4: Linear Mixed-Effects Model. A common method in EHR data analysis, as patients are usually categorized as subgroups, and a mixed-effect model is convenient to capture the group-level effect. Δ is also augmented as new predictors.

3.4.2 Synthetic EHR Data

Causal data Generation Method

The synthetic EHR data generation comprises three stages: I) Initialization, II) Disease Assignment, III) Treatment Assignment.

The three stages output three different views of the data matrix, X_{init}, X_{dise} and X_{trea} one by one, where everyone is based on the preceding one. But only the last matrix X_{trea} will be exported as the synthetic EHR data, and the front two are hidden ground truth. Δ matrix generated along with X_{trea}.

Stage I: Initialization. Randomly set up the healthy values for each lab result variable by the normal distribution, then initialize all patients' healthy data X_{init} accordingly.

Stage II: Disease Assignment. Randomly set up diseases and the DAG causal graph of them, then build up causality function from diseases to lab results, i.e. diseases influence lab value. Generate all patients' diseased data X_{dise} accordingly.

Stage III: Treatment Assignment. Randomly set up multiple medicine lines for each disease, and build up the causality function from medicines to the responded disease, i.e.

Table 3.3: Notations in Causal Data Generation Algorithm

$t = 1 \dots T$	T time windows
$P = \{p\}_{N_p} \mid p \in \mathcal{P}$	The set of N_p patients in space $\mathcal{P}$
$L = \{l\}_{N_l} \mid l \in \mathcal{L}$	The set of N_l lab results in space $\mathcal{L}$
$D = \{d\}_{N_d} \mid d \in \mathcal{D}$	The set of N_d diseases in space $\mathcal{D}$
$M = \{m\}_{N_m} \mid m \in \mathcal{M}$	The set of N_m medicine in space $\mathcal{M}$
$e(d_1, d_2) \mid d_1, d_2 \in D$	A causal relation from d_1 to d_2.
$G(D, \{e\})$	The causal graph of diseases D.
$V(G) = D$	Function of getting vertices of G.
$\mathcal{L} = \mathbf{F}_d(\mathcal{D})$	The causation mapping from $\mathcal{D}$ to $\mathcal{L}$.
$\mathcal{D} = \mathbf{F}_m(\mathcal{M})$	The causation mapping from $\mathcal{M}$ to $\mathcal{D}$.
X and Δ	Data matrix and Δ matrix.

treatments affect disease severity. Update all patients' treated data X_{trea} accordingly, and in each iteration, record the difference after and before updating to be the value of Δ.

The details of the three stages are described in Algorithms 1, 2, and 3, and all used notations are listed in Table 3.3.

As the example displayed in Fig. 3.3, for patient "No.3", the three views of his values on "Lab No.6" are colored by blue, red, and green respectively, and all of them keep

Algorithm 1: (Stage I) Initialization

Result: $(N_p \times T \times N_l)$ data matrix X_{init}
for $p = (1 \dots N_p)$ **do**
  `// For each patient` p
  for $l = (1 \dots N_l)$ **do**
    `// For each lab result` l
    Random (μ, σ) ;
    Random $(x^1, \dots, x^T)$ as $x^t \sim \mathcal{N}(\mu, \sigma), t = 1 \dots T$;
    $X_{init}[p, :, l] = (x^1, \dots, x^T)$;
  end
end

Algorithm 2: (Stage II) Disease Assignment

Result: $(N_p \times T \times N_l)$ data matrix X_{dise}
Initialize $X_{dise} = X_{init}$;
Random graph $G = (D, \{e\})$; `// Diseases causal graph`
Random $\mathcal{L} = \mathbf{F}_d(\mathcal{D})$; `// Causation from` $\mathcal{D}$ `to` $\mathcal{L}$
for $p = (1 \dots N_p)$ **do**
    `// For each patient` p
    Random path g from G ;
    for $d \in V(g)$ **do**
        `// For each` d `in path`
        Update L by $L = \mathbf{F}_d(d)$; `// Update lab results`
        for $l \in L$ **do**
            Update $X_{dise}[p, :, l]$; `// Update time series`
        end
    end
end

Algorithm 3: (Stage III) Treatment Assignment

Result: $(N_p \times T \times N_l)$ data matrix X_{trea} and Δ
Initialize $X_{trea} = X_{dise}$;
Initialize Δ as zeros;
Random $\mathcal{D} = \mathbf{F}_m(\mathcal{M})$; `// Causation from` $\mathcal{M}$ `to` $\mathcal{D}$
for $p = (1 \dots N_p)$ **do**
    `// For each patient` p
    Random medicine set M from $\mathcal{M}$;
    for $m \in M$ **do**
        Update D by $D = \mathbf{F}_m(m)$; `// Update diseases`
        for $d \in D$ **do**
            Update L by $L = \mathbf{F}_d(d)$; `// Update labs`
            for $l \in L$ **do**
                `// Update time series`
                Record $x_0 = X_{trea}[p, :, l]$;
                Update $X_{trea}[p, :, l]$;
                Update $\Delta[p, :, l] = X_{trea}[p, :, l] - x_0$;
            end
        end
    end
end

changing along the timeline. This patient developed a sequence of diseases (not shown in the figure) in a causal order, whose influence on "Lab No.6" is reflected as the value changing of the red curve. To cure him, the treatments, i.e. medicines, are assigned based on the observed values of red curve (in accordance to pre-determined criterion), and the treatment effects are reflected as the value changing of the green curve. Each disease has a sequence of applicable medicines, ordered by increasing levels, and the ith level is named as "ith Line".

The introduced causal data generator has been published for free downloading[1] , implemented on R platform.

As a reminder, the generator is for the purpose of chronic diseases simulation, who develops as a group of comorbidities, such as diabetes, hypertension, and heart disease. Particularly, the disease severity $d \in D$ naturally progresses non-decreasing along the timeline; And all medicines in M are in persistent use, i.e. no stop once start taken. It is not applicable to generating pandemic diseases.

Synthetic Setting

Particularly, in our experiments, we set up the observational timeline as 40 years, i.e. 40 time steps in total, and varying numbers of synthetic patients are sampled by demand (from 500 to 5000).

There are 20 lab results, whose values are always under observation along the timeline, and 10 diseases (denoted as "$d_1, \ldots, d_{10}$") that each disease can possibly influence $1 \sim 4$ labs value. These 10 diseases form a group of comorbidities, and with a pre-determined causal graph G, there exists at most 20 possible disease-developing paths (i.e. disease trajectories).

We set up 10 classes of medicines corresponding to the 10 diseases. Each class contains $1 \sim 3$ medicine levels, ordered with increasing impacts on labs value, including both positive normal effects and negative side effects.

[1] https://github.com/kflijia/Causal-EHR-Generator

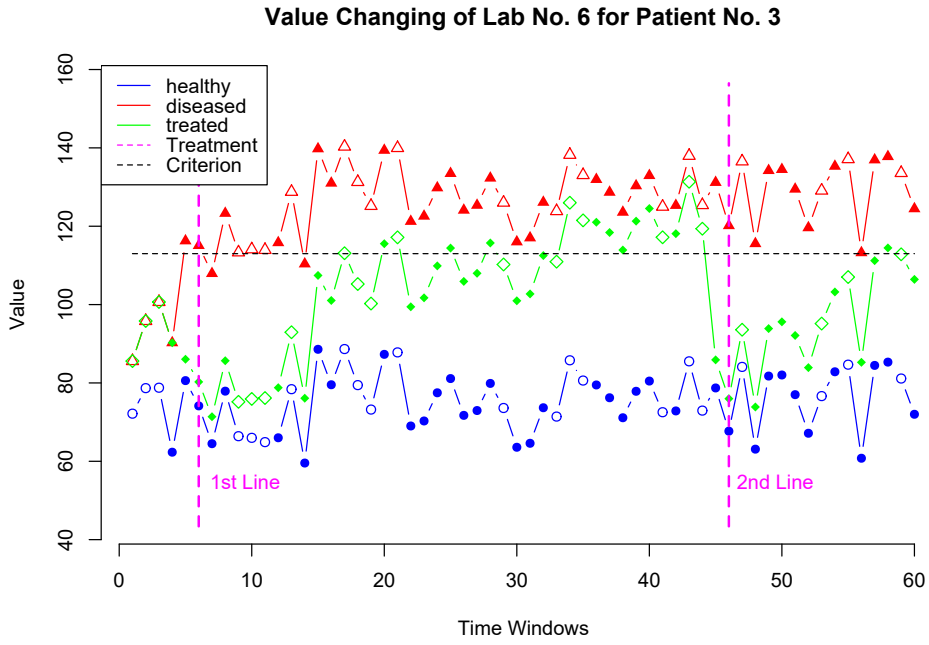


Figure 3.3: Example of synthetic data. The filled and unfilled markers indicate the random hospital visiting and absence, simulated by Poisson distribution.

Prediction Tasks

The experiments on synthetic data aim to exhibit the predictive performance changing, by varying hyper-parameter settings. This includes: 1) Time Step (i.e. the number of observed preceding years); 2) the sample size of training data (i.e. the number of synthetic patients).

From these results, we can confirm the conditions, in which the proposed Δ can be most effective, i.e. the most helpful to enhance the predictive performance. Additionally, we can observe whether involving Δ would influence the computing efficiency, i.e. running time of the training process.

For well concentrating, unlike in the real data experiments, here we only perform the binary classification tasks: To predict the 5-years-risk of the synthetic disease diagnosis, i.e. whether the patient will develop the disease within future 5 years.

We adopt AUC (Area under the ROC Curve) as the performance evaluation, and

for each model, we compare the two AUCs with and without Δ input. Additionally, their training time of them is also recorded and compared at the end.

Compared Models

- GLM, i.e. Logistic Linear Regression Model.
- RNN-GRU in Unsupervised Pre-training Architecture.
- RNN-LSTM in Unsupervised Pre-training Architecture.

Here we dropped the Linear Mixed-Effects Model and the Δ-Augmentation RNNs, because they did not show out-performance compared to their brother methods.

Our expectations include 1) RNN-GRU and RNN-LSTM can outperform the linear model GLM in most of the tasks, especially with a large sample size; 2) adding Δ input can significantly enhance predictive performance for both linear model and RNN model; 3) The training time is expected to be longer with Δ added, but not dramatically.

3.5 Experimental Results

For real EHR data, all experiments are conducted for 20 times, with 20 randomly splitting of training and test. For synthetic EHR data, we independently generate 5 different data, and perform all experiments for 4 times on each one, with 4 randomly splitting.

All shown results are on averaged values over these independent multiple runs, and the corresponding conference intervals (CI) are also displayed, as the shadowed areas in the figures.

The splitting rate is $0.7 : 0.3$ for training and test.

3.5.1 Results on Real Diabetes EHR Data

Table 3.4 and 3.5 display the performances of binary diagnosis predictions (in AUC) and continuous labs value forecasts (in MSE) respectively.

Each row in the table represents an individual task with a specified outcome and Time Step. The columns indicate applied methods, and the bold text in each row

Table 3.4: Disease Diagnosis Prediction Performance of 4 Outcomes on Real EHR Dataset

AUC Performance											
Models		GLM		GLMER		LSTM			GRU		
Methods		No Δ	Δ Augment	No Δ	Δ Augment	No Δ	Δ Augment	Δ PreTrain	No Δ	Δ Augment	Δ PreTrain
T2DM	Step=1	0.7783	0.7806**	0.7778	0.7803**	0.7829	0.7005	0.8067***	0.7846	0.7019	**0.8069*** **
	Step=2	0.8000	0.8041***	0.7488	0.7529***	0.8325	0.7387	0.8448***	0.8437	0.7459	**0.8588*** **
	Step=3	0.6867	0.6942***	0.6483	0.6533***	0.6888	0.6782	**0.7083*** **	0.6778	0.6665	0.7054***
	Step=4	0.6722	0.6756***	0.6712	0.6758***	0.6559	0.6845***	0.6861***	0.6564	0.6722*	**0.6893*** **
CAD	Step=1	0.7059	**0.7202*** **	0.6998	0.7156***	0.6868	0.6577	0.7061***	0.6874	0.6607	0.7104***
	Step=2	0.7127	0.7233***	0.6805	0.6887***	0.7122	0.6707	0.7213***	0.7169	0.6692	**0.7292*** **
	Step=3	0.6376	0.6418***	0.6259	0.6280***	0.6121	0.6310***	0.6337***	0.6049	0.6320***	**0.6372*** **
	Step=4	0.6475	0.6527***	0.6487	**0.6538*** **	0.6079	0.6418***	0.6454***	0.6059	0.6494***	0.6486***
CKD	Step=1	0.8154	0.8159	0.8154	**0.8159**	0.7685	0.6470	0.7877***	0.7684	0.6469	0.7942***
	Step=2	0.8182	**0.8200*** **	0.7792	0.7814***	0.7990	0.6632	0.8094***	0.7899	0.6603	0.8147***
	Step=3	0.7816	**0.7829***	0.7456	0.7469*	0.7352	0.6539	0.7596***	0.7327	0.6661	0.7660***
	Step=4	0.7687	0.7698*	0.7720	**0.7731***	0.6802	0.6814	0.7299***	0.6775	0.6816	0.7332***
Stroke	Step=1	0.6409	0.6395	0.6409	0.6394	0.6165	0.6143	**0.6754*** **	0.6115	0.6242	0.6675***
	Step=2	0.6393	0.6388	0.6454	0.6482	0.6053	0.6118	**0.6642*** **	0.6011	0.6113*	0.6634***
	Step=3	0.6343	0.6443***	0.6519	0.6412	0.6024	0.6233*	**0.6614*** **	0.6059	0.6104	0.6528***
	Step=4	0.6491	0.6565**	0.6453	0.6550**	0.6123	0.6532***	**0.7013*** **	0.6044	0.6588***	0.6882***

Table 3.5: Labs Value Forecast Performance of 3 Different Outcomes on Real EHR Dataset

MSE Performance											
Models		LM		LMER		LSTM			GRU		
Methods		No Δ	Δ Augment	No Δ	Δ Augment	No Δ	Δ Augment	Δ PreTrain	No Δ	Δ Augment	Δ PreTrain
LDL	Step=1	969.53	908.32***	991.06	932.49***	1006.04	917.85***	**749.58*****	983.1	903.26***	750.38***
	Step=2	964.51	885.84***	970.12	893.03***	987.93	941.46**	714.05***	986.62	902.53**	**546.27***
	Step=3	946.47	792.52***	955.84	810.55***	1104.67	903.83***	729.19***	996.01	1012.31	**516.05*****
	Step=4	827.67	681.55***	827.66	**681.41*****	866.8	836.14***	689.75***	1092.29	818.17*	863.58***
SBP	Step=1	466.74	457.74***	468.02	460.37***	512.51	479.64***	433.21***	512.98	484.75***	**421.41*****
	Step=2	432.93	418.55***	430.61	418.09***	488.42	458.56***	**396.06*****	488.7	473.07*	418.13***
	Step=3	385.72	367.77***	385.2	368.41***	444.67	419.25**	**343.38*****	459.5	464.11	375.98***
	Step=4	358.91	**337.74****	358.91	337.74**	479.9	460.52**	361.04**	507.64	445.96***	361.31***
TG	Step=1	5262.75	4754.05***	5732.52	5425.32***	5465.9	4732.7***	4230.93***	5428.54	4661.68***	**4090.07*****
	Step=2	6105.08	5339.08***	6556.5	5807.74***	6349.99	5365.36***	4812.35***	6506.7	5300.33***	**4523.19*****
	Step=3	4571.69	3971.84***	4538.45	4026.53***	4864.27	4497.26***	**3487.98****	4745.23	4373.88***	4478.10***
	Step=4	4102.99	3342.96***	4105.74	3347.37***	4761.85	4217.16***	**2223.00****	5654.48	5128.49*	3593.77*

Table 3.6: P-values of t-test on AUC vectors pair

$10^{-3}*$		Sample Size				
Models	Time Steps	250	500	1500	2500	5000
GLM	Step=5	3.2903	1.1429	0.6845	0.7187	0.7336
	Step=10	0.2065	0.9230	0.0682	0.3762	0.0130
	Step=15	1.5441	2.6313	1.0823	2.4657	0.8256
GRU	Step=5	0.0000	0.0000	0.0015	0.0000	0.1455
	Step=10	0.0021	0.0046	0.0347	0.2035	0.0486
	Step=15	0.3203	0.4357	0.1109	0.6493	4.0279
LSTM	Step=5	0.0005	0.0000	0.0008	0.0003	0.0001
	Step=10	0.0229	0.4461	1.1835	1.2984	8.7623
	Step=15	0.0651	2.6749	9.9602	12.3447	39.8459

indicates the best performer. The star symbols stand for the significance level of the improvement with Δ's participation, compared to the no Δ version. The improvement is verified by one-side paired t-tests, and significance levels are defined as p-value intervals: $***$ for $(-\infty, 0.001]$, $**$ for $(0.001, 0.01]$, and $*$ for $(0.01, 0.05]$.

In table 3.4 and 3.5, the RNN methods with Δ pre-training architecture, including both RNN-GRU and RNN-LSTM (but especially LSTM), outperformed the others mostly. To be specific, they won 10 times out of 16 binary diagnosis prediction tasks, and 10 times out of 12 continuous lab value forecasting tasks. On the other hand, RNN methods without Δ input never outperformed the pre-training ones.

Except for RNNs, Δ can also benefit linear models. Most of the time, with Δ augmented, liner models performed better, but not as significantly as RNNs.

3.5.2 Results On Synthetic EHR Data

Figure 3.4 shows AUC performances of the three compared models. The predicted outcomes are five randomly selected synthetic diseases, whose performances are equally weighted.

For each model, with and without Δ performances are in the same color but distinguished by solid and dashed line style. With varying Time Step=$5, 10, 15$, the sample

size increases from 250 to 5000 gradually. The shadows in the figure represent CI.

As we expected, the performance with Δ is always better than the one without Δ input. And both RNN-GRU and RNN-LSTM consistently perform better than GLM when the sample size is large enough (in our case $>= 500$).

The hyper-parameters (Time Step and sample size) obviously influence the performance. With larger Time Step, all models tend to perform better, due to more historical information observed; With increasing sample size, RNN models present an upward trend on performance, but level off when sample size has been large enough. On the other hand, GLM model slightly performs the upward trend but is much weaker than RNNs, especially with small Time Step $= 5$.

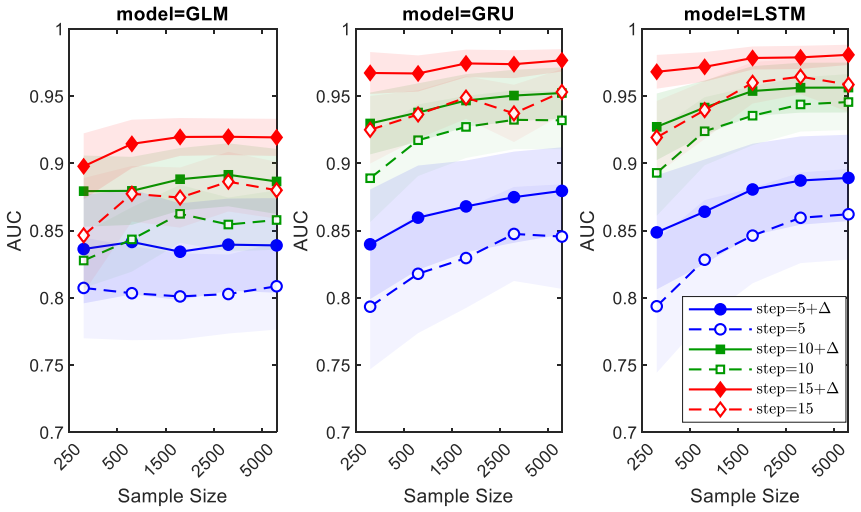


Figure 3.4: Comparison of the predictive performances with increasing numbers of samples, for different Time Steps.

The effect of adding Δ variables is compared in Figure 3.5, which shows the AUC gain of involving Δ, i.e. the difference of with-Δ-AUC minus without-Δ-AUC, and keeps consistent layout with Figure 3.4. With a shrinking sample size (from 5000 to 250), RNN-LSTM clearly presents the trend that Δ becomes more helpful in improving predictions; RNN-GRU trends the same but less obviously; This pattern has almost vanished in GLM model. On the other hand, increasing Time Step from 5 to 15 did not result in any monotonic tendency.

For more convincing conclusions, we performed paired t-test on each pair of AUC

vectors. Each vector consists of the performances of all synthetic disease prediction tasks. In our case, the vector has a length of 80. The p-values are listed in Table 3.6, and all turn out to be significant. Particularly, RNN-LSTM shows a clear pattern that when the sample size is comparatively small, Δ can be more certain to be helpful, which is consistent with the conclusion from Figure 3.5.

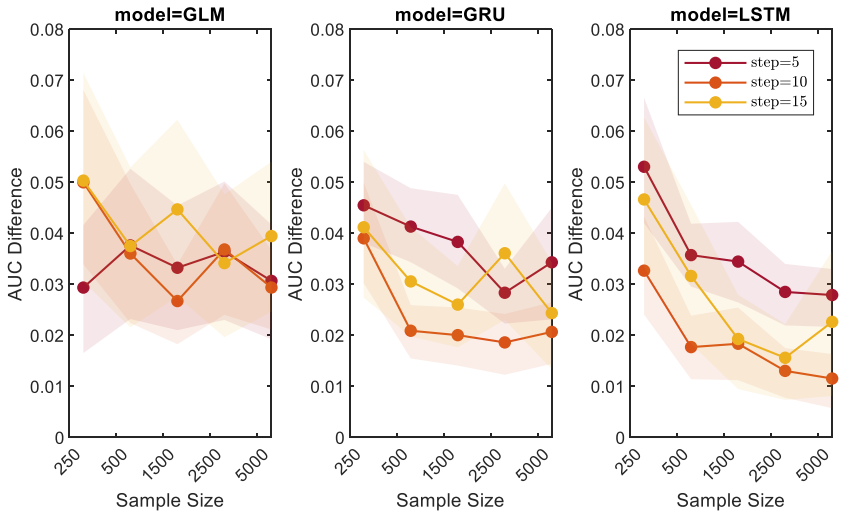


Figure 3.5: Performance gains by adding Δ for three methods: GLM, GRU, and LSTM.

Among the three models, RNN-LSTM acts the most stably to present Δ's effect changing. Because LSTM has the most complex architecture, and the largest number of parameters, it is more able to capture enough information from data than the other two. In contrast, the simplest model GLM failed to fully utilize the new information brought by Δ.

On the side of training efficiency, adding Δ does not dramatically influence it. For GLM, augmentation of Δ increased training time by around 80%, e.g. with sample size=5000 and Time Step=15, training without Δ runs 10.8 secs by average, and adding Δ increases it to 18.82 secs. RNN models spend hundreds of secs on the training process, and after pre-training using Δ, their training time was reduced by around 1/3 because of prior knowledge. However, the entire time cost is almost doubled compared to the version without pre-training.

3.6 Conclusion

In this paper, we proposed the treatment effect estimator Δ, which leverages clinical knowledge to explicitly bring new causal information to RNN.

With comprehensive experiments, we demonstrated that Δ can significantly enhance the predictive performance of RNN models, especially when the sample size is comparatively small. This also confirmed our initial hypothesis that RNN is unable to automatically discover treatment effects from EHR data. It is because the observed variables in causal data inherently cannot represent all underlying causal factors, which include both naturally progressed severity and medical changed severity in our case.

As in the background of temporal EHR data analysis, we proposed a generally applicable methodology to effectively adjust the causal representation bias in causal data, and eventually helpful for reducing estimation bias of general modeling.

Chapter 4

Dealing with the Causally Informative Missing Values in EHR Data

4.1 Introduction

Missing values pose a significant challenge in data analysis since accurate information is usually required for unbiased and precise inference. Particularly, in clinical studies, the information accuracy is critical but missing data are a common occurrence [65].

There are numerous ways in which data can be missing. When diseases have multiple etiologies, such as infections, which can have multiple sources, the diagnostic test can depend on the etiology, i.e. among the many possible tests, only the tests most appropriate for the infection source will be performed, while the others are left unobserved (missing). Also, when easily observable vital signs all indicate the absence of infection, lab tests for infection may not be performed, which is another source of missing values. A third source, longitudinally missing data, is related to repeated measures: when a patient drops out of a study, all future observations will be missing. In this study, we are focusing on data elements that are completely missing. In particular, the much simpler case longitudinal missingness is out of scope.

Missing values in clinical studies are common, and traditional methods for handling

them can introduce unpredictable bias [66]. For example, missing values can suggest favorable outcomes, when the disease they are meant to diagnose can be ruled out without performing the test. They can suggest also unfavorable outcomes when they are missing simply because there was not enough time to safely perform the tests. When the missing value pattern is driven by disease etiology, it can impact diagnostic accuracy, prognosis, and overall outcomes.

The most common method of addressing missing values in clinical studies is to simply discard records with missing data. Since the missingness itself can be associated with outcomes, simply discarding these records introduces bias and also reduces the sample size [67].

An alternative is to impute missing values. Most existing missing value imputation methods only yield unbiased results when the data are missing at random: whether the data element is observed or not is independent of its unobserved value conditional on the other observed values in the record. In the above examples, this assumption often does not hold true and these values are missing-not-at-random (MNAR). Some Maximum-Likelihood-based methods such as Pattern Mixture [68], were designed to perform imputation under the MNAR assumption. Instead of assuming that the data is missing at random, they explicitly model the missingness mechanism; in other words, they replace the randomness assumption with model assumptions.

In this paper, we propose a methodology, Pattern-Wise Analysis, inspired by pattern mixture models, to carry out modeling without imputation. Instead, separate models are constructed for selected patterns of missingness. We will demonstrate that even the simplest scheme for Pattern-Wise Analysis outperforms the commonly used imputation methods.

Pattern-Wise Analysis has its own challenges: the number of patterns can be too numerous, many patterns may have an insufficient sample size, and the data distribution in a pattern can be different from the data distribution outside the pattern. We propose methods to overcome these challenges from three complementary perspectives: (i) From a model selection perspective, where model selection is used to determine which patterns we should model; (ii) A distributional perspective, where the training data set is expanded with observations in a distribution-preserving fashion so that modeling this pattern becomes possible; (iii) From a causal perspective, where a causal structure for

the MNAR mechanism is assumed and exploited to convert the problem from MNAR to MAR. We will demonstrate that each of these perspectives improves predictive performance.

We performed our evaluation on synthetic MNAR data, and also in the context of sepsis using real electronic health records (EHR) data collected from a tertiary care provider. Sepsis (blood poisoning, blood infection) is a severe form of infection, where the infection enters the bloodstream, spreads to, and damages multiple organs. Sepsis is highly morbid with a 25-50% percent mortality rate [69]. We aim to predict whether patients who present with infection in the emergency room and are subsequently hospitalized are going to progress to sepsis. Sepsis is an excellent use case due to the multitude of infection sources many of which have their own unique tests [70]. This leads to a high rate of potentially non-randomly missing data.

The rest of this paper is organized as follows: Section 2 describes related works. Section 3 presents the Pattern-Wise Analysis methodology. In Section 4, we present the experimental evaluation and Section 5 concludes our work.

Contributions

Our specific contributions are as follows:

- We develop a Pattern-Wise Analysis methodology that performs modeling without imputation, discarding records, or making restrictive assumptions about the missingness mechanism
- We demonstrate that even the simplest Pattern-Wise Analysis algorithm outperforms the commonly used imputation methods.
- We present specific algorithms to implement Pattern-Wise Analysis from three complementary perspectives, isolate the effect of these improvements and identify scenarios where they work best.

4.2 Related Work

In this work, we assume that the outcome is fully observable. We define *missing data* as measurements that are not observed at any point in time; the longitudinal missingness is out of scope.

4.2.1 Missing Data Mechanisms

Based on the taxonomy developed by Rubin in 1976 [71] and Little and Rubin in 2002 [72], missing data mechanisms are classified as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR).

Throughout the following sections, let X denote an $n \times k$ covariate matrix that may contain missing values and Y the fully observed outcome vector, where n is the number of instances, and k is the number of covariates. We define the missing data indicator matrix $M = (m_{ij})$, such that $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ if x_{ij} is present. Let X_{obs} denote the observed components or entries of X, and X_{mis} the missing components.

Data is said to be missing **completely at random** (MCAR) if the missingness does not depend on any data, either observed or missing

$$Pr(M|X_{obs}, X_{mis}) = Pr(M). \tag{4.1}$$

This means that the observed components X_{obs} are just a random sample of the full matrix X, missing for example as a result of accidental omission.

Data is **missing at random** (MAR), when the missingness M is independent of the missing values X_{mis} given the observed covariates (X_{obs})

$$Pr(M|X_{obs}, X_{mis}) = Pr(M|X_{obs}). \tag{4.2}$$

For example, in cancer clinical trials, the size of a primary tumor is often missing, and it may depend on the type of the primary tumor that is often fully observed. If the missingness of the size only depends on the type, then it is considered MAR.

From an analysis perspective, MAR is a relevant mechanism, because it characterizes the (unobservable) distribution of the missing values, which is the basis for imputation [67].

Data is **missing not at random** (MNAR) if it is neither MCAR nor MAR

$$Pr(M|X_{obs}, X_{mis}) \neq Pr(M|X_{obs}). \tag{4.3}$$

MNAR is usually called informative or nonignorable missingness, which implies that the missing values can no longer be inferred from the observed values.

Note that from the incomplete dataset itself, we cannot even determine whether missingness is MNAR or MAR. The growing consensus is to perform sensitivity analysis to determine whether the MAR assumption is appropriate [73].

4.2.2 Three ways of Handling Missing Values

As the most straightforward way of dealing with missing values, Complete Cases Analysis is the most popular method. In Complete Case Analysis the incomplete records are omitted and only the complete cases are analyzed. This leads to two problems: (i) when missingness is MNAR the analysis can become biased; and (ii) when a great number of records are omitted, we can incur substantial information loss.

An alternative is an imputation. The goal of imputation is to complete the data matrix with estimates for the missing elements, which differs from our goal of constructing the best model for the outcome. Given a large number of imputation methods in existence (see [74] for a textbook on imputation), reviewing them comprehensively is outside the scope of this manuscript. We will review the most popular methods in Section 4.2.3 and review ways to correct bias which are useful for our proposed Pattern-Wise Analysis in Section 4.2.4.

A third method of handling missing values is to construct models without imputation. Decision trees have built-in methods for directly handling missing values either by implicit imputation or through surrogate variables (a variable that is correlated with the variable that has a missing value and is observed) [75]. Similarly, Expectation-Maximization-based methods can construct a model without explicitly completing the data matrix [76]. However, all of these methods still estimate the unobserved values either through the surrogate or through the expectation, thus they perform imputation implicitly.

4.2.3 Imputation in Practice

While the imputation literature is very rich, the methods that are used in practice are few. In a recent survey of 140 clinical studies of sepsis between 2010 and 2016, we found that most studies did not even acknowledge the problem of missing values. The vast majority of the studies that acknowledged missing values used the default method of the analysis software (which is complete case analysis in the case of the R software). Only 65 papers explained explicitly how they handled missing values, and of these 65 papers, 17 used complete case analysis (discarded records with missing values), 11 used mean (or zero) imputation, 19 used Multivariate Imputation via Chained Equations (MICE), 4

used maximum likelihood imputation, and 3 used pattern mixture models; the remaining studies used manual imputation, discarded variables with too many missing values and used a combination of the above depending on the missing rate.

Below, we briefly describe these methods. **Single Value Imputation** operates under the MCAR assumption and missing data is imputed column-wise with a single value such as zero or the mean of the observed values in the column (mean imputation). **Multivariate Imputation via Chained Equations** (MICE or MI) assumes the MAR mechanism and imputes the missing values through a series of regression analysis [77]. **Pattern Mixture** (Mix) models operate under MNAR and model the missingness mechanism and the data jointly [78]. An imputation model is constructed for each missing value pattern (combination of variables with missing values). Both Mix and MI require the correct specification of an imputation model.

4.2.4 Bias Correction

The key problem with missing values is that they can bias the results. Many strategies have been devised to correct these biases. The simplest method is to use imputation in conjunction with a bias-correcting indicator variable, which simply denotes whether a particular element was imputed or observed.

A more sophisticated approach to correcting for bias is to model the missing mechanism and the data jointly. One such set of techniques is the Pattern Mixture models from Section 4.2.3, while another is the inverse propensity (of missingness) weighing (IPW) [79]. In IPW, observation weights, proportional to the observation's propensity to be missing, are used in the imputation model. This approach is sensitive to the correct specification of both the imputation and the missing mechanism models. In contrast, Mix "only" requires that we specify the imputation model correctly and that the pattern has sufficient observations to estimate the imputation model. Both techniques work best when the number of patterns is small. As the number of patterns increases and the number of observations in each pattern decreases, both the imputation model and (in the case of IPW) the missingness mechanism model become inaccurate. For this reason, these methods are most often used in repeated measures (longitudinal) studies [80], where the main reason for missing data is patients dropping out of the study and thus the number of missingness patterns is limited.

4.3 Pattern-Wise Analysis

The goal of the analysis is to learn a model from partially observed data for an outcome Y assuming that the missingness in the data is MNAR. We build this model in a pattern-wise fashion, constructing a sub-model for each pattern of missingness. A missingness pattern is a set of variables that are missing together and a set of variables that are present together. We assume that observations belonging to each pattern are generated by the same process (e.g. the same disease etiology).

Modeling a population in a per-pattern fashion raises a number of questions. **Q1**: Can the method handle MNAR patterns? **Q2**: Can the method handle a large number of patterns? **Q3**: Can the method model small patterns (i.e. small number of observations in a pattern)?

In the following paragraphs, we will first formalize the Pattern-Wise Analysis problem, then present an overview of the various methods under the Pattern-Wise Analysis umbrella and describe the details of these methods.

4.3.1 Definitions

Let X be an $n \times k$ covariate matrix with rows denoting the n observations and columns corresponding to the k predictors.

Let a **missingness pattern** (or pattern for short) p be a binary vector of length k. If the jth component of p is 1, then the value of the jth variable is missing; if it is 0, the value is present. Let M be a binary matrix of the same dimensions as X such that the missingness pattern for observation i is M_i.

The pattern p defines a **subpopulation**

$$\mathcal{S}^p = \{x_i \in X | M_i = p\}$$

that has exactly the same variables observed as those observed in p and exactly the same variables missing that are missing in p.

Let $\mathcal{P}$ be the set of all missingness patterns existing in the data set. Let R be an $n \times |\mathcal{P}|$ binary matrix, called the **pattern matrix**, where the rows are observations, the columns are patterns, and $R_{ij} = 1$ if and only if the ith observation has a missingness pattern that coincides with the jth pattern in $\mathcal{P}$. R shows which missingness pattern each observation follows.

A pattern $\widehat{p}$ is a **super-pattern** of pattern p if the set of observed variables in $\widehat{p}$ is a super-set of the set of observed variables in p. In other words, all variables observed in p are also observed in $\widehat{p}$, but $\widehat{p}$ has additional observed variables. If we wish to build a model for the subpopulation defined by p, we can use observations from the subpopulation defined by $\widehat{p}$.

Let P denote the set of all super-pattern of p and let $\mathcal{E}^P$ denote the subpopulation

$$\mathcal{E}^P = \bigcup_{\widehat{p} \in P} \mathcal{S}^{\widehat{p}}.$$

We call $\mathcal{E}^P$ the **exten*sion* subpopulation** of p as it can extend $\mathcal{S}^p$ to a set of observations such that all variables observed in p are also observed in $\mathcal{E}^P$. The resulting subpopulation is the **extend*ed* subpopulation** $\mathcal{E}^P \cup \mathcal{S}^p$. Note that $\mathcal{E}^P$ does not contain $\mathcal{S}^p$. The extended subpopulation is the largest subpopulation that has all observed variables in p observed and is thus the largest set of observations that can be potentially used to build a model for pattern p.

We call a **sub-model** for pattern p the model built on the subpopulation $\mathcal{S}^p$. An **extended sub-model** for pattern p is model built on some or all observations in the extended subpopulation $\mathcal{E}^P \cup \mathcal{S}^p$.

4.3.2 Overview of the Pattern-Wise Analysis Models

The baseline model in the Pattern-Wise Analysis methodology is **Per-Pattern Model (PPM)**. We simply extract all patterns of missingness from the data and build a sub-model for each pattern. This model can correct for biases stemming from MNAR, and can capture differences in the effect of predictors on the outcome cross the patterns. However, if the pattern consists of too few observations, the model may fail.

The first variant of PPM we consider is one that performs model selection (**PPM+MS**). PPM+MS can select subpopulations for which to build sub-models. It does not have the ability to build a better model for these (small) subpopulations but has the ability to avoid the decrease in predictive performance due to overfitting the sub-model.

To build better models for each pattern, we consider extending the training set by borrowing observations from the extension subpopulation. **PPM+Ext** uses the entire extended subpopulation ($\mathcal{S}^p \cup \mathcal{E}^P$) of each pattern p. When the distribution of the data

in $\mathcal{E}^P$ matches that in $\mathcal{S}^p$, this approach leads to a better model; otherwise the extended sub-model can misrepresent the relationship between X and Y.

To ensure that the relationship between X and Y is preserved, we only use a portion of the extension subpopulation where the joint distribution of (X, Y) matches that of $\mathcal{S}^p$. The resultant model is the distribution matched PPM (**PPM+DM**). Since the subpopulation that needs extension is small, matching the distribution is non-trivial.

As an alternative to PPM+DM, we can use a generic causal structure to describe how missingness relates to the unobserved values of X as well as the outcome Y. The resultant model is PPM with causal matching (**PPM+CM**).

The final model is actually an imputation-based model, which is positioned between pure imputation-based models and per-pattern models. This model, which we call Bias Corrected Imputation (**PPM+BCI**), applies per-pattern bias correction to a mean imputed model.

Whether can handle it?	**Q1**: MNAR	**Q2**: Many Patterns	**Q3**: Small Pattern
PPM+BCI and PPM+MS	√	√	×
PPM	√	×	×
PPM+Ext	×	×	√
PPM+CM and PPM+DM	√	×	√

Table 4.1: Whether the PPM-family methods satisfy challenging questions. Check mark means Yes, and cross means No.

Table 4.1 provides a summary of which questions these various models can answer. In the following sections, we describe these models in detail.

4.3.3 Per-Pattern Model

The construction of the baseline Per-Pattern Model (PPM) is rather straightforward. A set $\mathcal{P}$ of all missingness patterns are extracted from the data and for each pattern, p in $\mathcal{P}$ a sub-model is constructed. This sub-model is a Lasso-penalized logistic regression model. While the Lasso penalty may set many of the coefficients in some of the small sub-models to 0, all sub-models have at least an intercept, thus the PPM methodology does not perform model selection (it performs variable selection). The optimal penalty

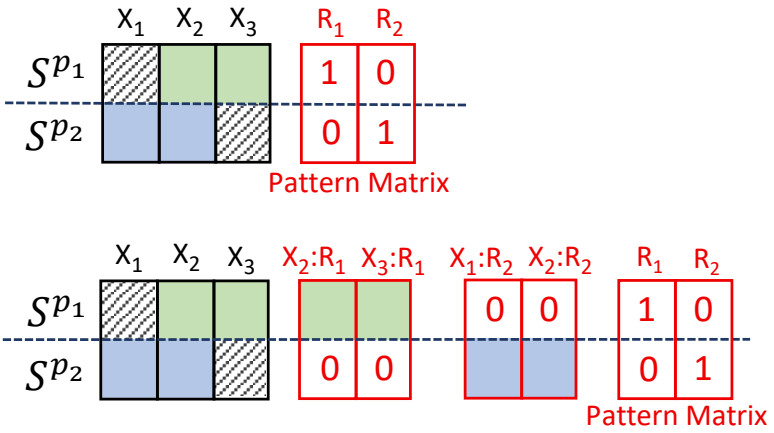


Figure 4.1: Example of PPM+BIC (top figure) and PPM+MS (bottom figure). the green rectangle represents $\mathcal{S}^{\mathbf{p}_1}$, the observed part of the sub-population with missing pattern $\mathbf{p}_1$, and blue rectangle represents $\mathcal{S}^{\mathbf{p}_2}$. The shaded area represents missing values in the original data matrix and is mean-imputed here.

was determined via cross-validation.

4.3.4 Model Selection

The submodels for Per-Pattern Model with Model Selection (PPM+MS) and Bias Corrected Imputation (PPM+BCI) are constructed very similarly, so we discuss them together.

For PPM+BCI, the partially observed covariate matrix X is completed using mean imputation, then juxtaposed with the pattern matrix R, and a penalized logistic regression model is constructed. Recall that the pattern matrix R is an indicator matrix showing which pattern each observation follows, thus the R matrix provides a per-pattern intercept to a single global model.

In the case of the PPM+MS model, the mean-imputed covariate matrix is extended by $(X : R)$, a matrix representing the interaction between the pattern matrix R and the observation matrix X. All columns containing unobserved components are dropped from $(X : R)$.

Figure 4.1 illustrates the construction of the PPM+MS covariate matrix. There are two missing patterns: $\mathbf{p}_1$ with $\{X_2, X_3\}$ observed and $\mathbf{p}_2$ with $\{X_1, X_2\}$ observed.

The pattern matrix R has two columns R_1 and R_2. In method PPM+BCI, R is simply juxtaposed to mean-imputed X; while in the PPM+MS method, $(X\colon R) = \{X_2\colon R_1, X_3\colon R_1, X_1\colon R_2, X_2\colon R_2\}$ are added as well. Note that $(X\colon R_1) = \{X_1\colon R_1, X_2\colon R_1, X_3\colon R_1\}$ initially, but $\{X_1\colon R_1\}$ was dropped because X_1 is missing in $\mathbf{p}_1$; and so did $\{X_3\colon R_2\}$. The coefficients of $(X\colon R)$ are the coefficient of the sub-model for each pattern $\mathbf{p}$, and the coefficient of R is the intercept of the sub-model.

Technically, the PPM+MS model is built on a mean-imputed observation matrix, but the sub-model of each pattern is built on its respective observed data (the $(X\colon R)$ matrix), hence the imputation is not necessary. Given the presence of the R matrix, missing values in X could be filled with any number (e.g. 0) without affecting the model; or the X matrix can also be omitted entirely. The presence of X merely helps model selection: if a sub-model is not required, all of its coefficients can be set to zero, suggesting the model coincides with a global model (which is the PPM+BCI model).

PPM+MS is applicable even when the number of patterns is very high because it has the ability to ignore smaller patterns that do not affect the predictive performance. However, some of these ignored patterns could be clinically interesting.

4.3.5 Extended Sub-Models

The simplest model that uses extended sub-models is the PPM+Ext. Recall that for a pattern p, P represents the set of all super-patterns, patterns in which the set of observed variables is a superset of the set of observed variables in p. The collection of all observations that follow any of the super-patterns in P is the exten*sion* subpopulation $\mathcal{E}^P$ and in the case of PPM+Ext, an extended sub-model is built on the extend*ed* subpopulation $\mathcal{S}^p \cup \mathcal{E}^P$ of each pattern p in $\mathcal{P}$ (set of all patterns in the data). These models are lasso-penalized logistic regression models. No model selection is applied.

This method solves the small pattern problem, but it actually operates under the MAR assumption, which implies $\mathcal{S}^p$ and $\mathcal{E}^P$ are assumed independent and identically distributed (i.i.d.). Below, we present two variants that address this limitation from two perspectives: a causal and a distributional perspective.

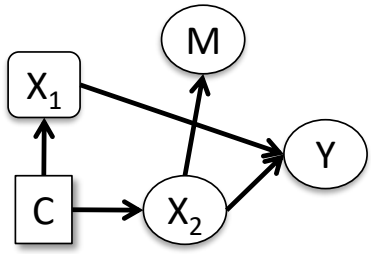


Figure 4.2: Causal graph representing informative missingness through the MNAR mechanism. The node in a rectangle is unobservable, the node in a rounded rectangle is partially observable and nodes in circles are fully observable. Edges indicate causation. The causal graph is a priori known.

Causal Matching

Consider the causal graph in Figure 4.2. X_1 and X_2 are covariates, Y is the outcome, and M indicates whether X_1 is missing. An unobservable extraneous mechanism C triggers changes in X_1 and X_2. X_2 is always observable and it can suggest values of X_1. As a result, X_1 may not be observed ($M = 1$). For example, C can be an immune response to an infection (Y), which causes abnormal vital signs X_2 (such as fever or high heart rate) and also causes elevated white blood cell count (WBC) X_1. When the vital signs are normal, the physicians may rule out infection and elect not to measure WBC.

This example represents informative missingness through the MNAR mechanism: although the unobserved normal value of X_1 did not cause it to be missing (a physician's decision based on observing X_2 caused it), normal values of X_1 are predictive of both the missingness (M) of X_1 and the outcome Y (no infection). Since the missing values of X_1 are predictive of missingness, this is MNAR.If we balance the distribution of X_2 across the levels of M through, for example, propensity score matching, we can remove the $X_2 \rightarrow M$ edge and can render the problem MAR, because X_1 is no longer associated with M via X_2.

Given two groups, "treated" and "untreated", propensity score matching attempts to construct a new sample consisting of matched observations from these two groups: for each "treated" observation, K "untreated" observations are selected. In the desired sample, the two groups are identical and only differ in "treatment" and outcome. This

Algorithm 4: Causal Matching

Data: $\mathcal{S}^p$, $\mathcal{E}^P$, K, ϵ
Result: $\mathcal{E}^P_{mtc}$, L^p
$\mathcal{E}^P_{mtc} = \{\}$;
Learn $PS(\mathcal{S}^p \cup \mathcal{E}^P)$;
for $x \in \mathcal{S}^p$ **do**
  Sort $\mathcal{E}^P$ by $|PS(\mathcal{E}^P) - PS(x)|$ increasingly;
  for $i = 1 \dots K$ **do**
    $\widehat{x} = \mathcal{E}^P(i)$;
    if $\widehat{x} \notin \mathcal{E}^P_{mtc}$ *and* $|PS(\widehat{x})\text{-}PS(x)| < \epsilon$ **then**
      $\mathcal{E}^P_{mtc} = \{\mathcal{E}^P_{mtc}, \widehat{x}\}$;
    end
  end
end
Train L^p on $\mathcal{S}^p \cup \mathcal{E}^P_{mtc}$;

implies that the covariate distributions (relevant to treatment) across the two groups are identical. This is achieved by matching observations that have the same propensity (probability) of belonging to the "treated" group. In our application, "treatment" is whether an observation follows a particular pattern and the goal is to balance the observed covariates between $\mathcal{S}^p$ (the "treated" group) and $\mathcal{E}^P$ (the "untreated" group), where p is a pattern and P are the set of all super-patterns of p. In other words, we are balancing the observed covariates between the subpopulation that follows p and the extension subpopulation of p.

For each observation in $\mathcal{E}^P \cup \mathcal{S}^p$, we compute the propensity (probability) of following p through logistic regression. The propensity score is the linear prediction for this model. For each observation in $\mathcal{S}^p$, K candidates from $\mathcal{E}^P$ are selected (matched) without replacement. Two observations form a match if their propensity scores are within a caliper ϵ. The details are given in Algorithm 4, in which $\mathcal{E}^P_{mtc}$ is the matched observations from $\mathcal{E}^P$, L^p is the learned extended sub-model on $\mathcal{E}^P_{mtc} \cup \mathcal{S}^p$, and $PS(\cdot)$ means "propensity score of".

The CM method has multiple advantages: Firstly, it can extend the small patterns in a distribution-preserving way (**Q3** satisfied). Secondly, by comparing its performance

Algorithm 5: Distribution Matching

Data: $\mathcal{S}^p+$, $\mathcal{E}^P+$, ϵ
Result: $\mathcal{E}^P_{mtc}+$, L^p
$\mathcal{E}^P_{mtc}+ = \{\}$;
Learn $PS(\mathcal{S}^p + \cup\ \mathcal{E}^P+)$;
for $\widehat{x} \in \mathcal{E}^P+$ **do**
  $d = min($ |PS($\widehat{x}$) - PS($\mathcal{S}^p+$)|) ;
  if $d < \epsilon$ **then**
    $\mathcal{E}^P_{mtc}+ = \{\mathcal{E}^P_{mtc}+, \widehat{x}\}$;
  end
end
Train L^p on $\mathcal{S}^p \cup \mathcal{E}^P_{mtc}+$;

with the PPM+Ext method for each pattern, one can easily find which patterns are MNAR (**Q1** satisfied). It can achieve high accuracy, especially when missing patterns are mostly MNAR.

Distribution Matching (DM)

The problem with $\mathcal{S}^p$ being differently distributed from $\mathcal{E}^P$ is that the relationship between X and Y may differ between $\mathcal{S}^p$ and $\mathcal{E}^P$. To ensure that this is not the case, we aim to identify a subset of $\mathcal{E}^P$, in which the joint distribution of (X, Y) is identical to that in $\mathcal{S}^p$. Given that $\mathcal{S}^p$ is small, this is a non-trivial task.

In the DM method, we utilize k-Nearest Neighbor matching: for each observation in $\mathcal{S}^p$, we find all observations within ϵ caliper (distance) from $\mathcal{E}^P$. The similarity metric used in the nearest neighbor computation is the difference between two observations' propensity to belong to pattern p. The procedure is given in Algorithm 5, in which "+" means that the subpopulation is on joint (X, Y) instead of only X.

DM has an apparent advantage that small patterns can almost always be extended to a learnable scale (**Q3** satisfied), but it can lead to overfitting accordingly. The major concern for DM is that because it matches on the outcome as well, the inherent relationship between covariates and outcome could be broken. Besides, DM is able to find MNAR patterns in the same way as CM (**Q1** satisfied).

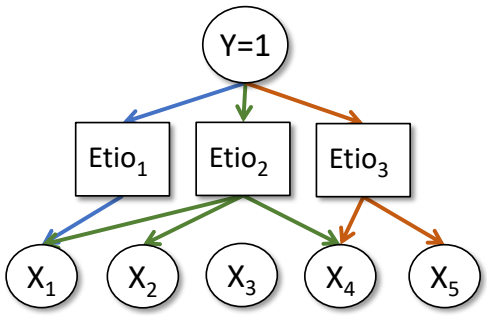


Figure 4.3: An example of generating synthetic EHR data. Consistently with Fig 4.2, the node shapes (rectangle and circles) stand for unknown and fully known respectively.

4.4 Experimental Evaluations

We present two experiments: one on synthetic EHR datasets with MNAR missingness to study the methods' behavior as the missing rate increases and to illustrate how some of the most successful methods achieve their high performance; and the second one on real EHR data predicting sepsis to provide a real-world application. We compared the proposed PPM-family methods (PPM+BCI, PPM+MS, PPM, PPM+Ext, PPM+CM, and PPM+DM) on both data sets with state-of-the-art imputation methods: Mean Imputation (**Mean**) for MCAR data, Multiple Imputation (**MI**) for MAR, and Pattern Mixture Imputation (**Mix**) for MNAR. In all experiments, the Lasso-penalized logistic regression was used with Area under the ROC curve (AUC) as the evaluation criterion. All methods were implemented in R; we used the "MICE" package with appropriate (Gaussian) settings for MI and "glmnet" for the regression models.

In the synthetic data experiment, we generate 10 datasets and report the average (10-fold) cross-validated performance over the 10 runs; and on the real sepsis data, we also perform 10 runs of 10-fold cross-validation with partitioning. We report the average performance of the runs and the 95% empirical Confidential Interval (CI).

4.4.1 Experiments on Synthetic EHR Data

Data Generation

Our data generation follows the missingness mechanism assumed for Figure 4.2 and a Bayes Network-based methodology is used for inducing non-random missingness [81]. At

a high level, we first generate data for a disease-free population. Then, for a randomly selected set of patients, we adjust the covariate distributions to simulate the presence of disease ($Y = 1$). Such systematic changes to the covariates impose a dependence structure, which we learn through BN structure learning and exploit to introduce non-random missingness.

First, we model a disease-free population. Covariates $X_1 \dots X_6$ are initialized as a multivariate normal distribution with non-zero covariances, representing, for example, the normal values of vital signs or lab results. Next, we select a random sample set from this cohort as patients with the disease ($Y = 1$). Diseases can have multiple etiologies (they can develop in different ways), affecting different vital signs or lab results. For each disease patient, we randomly select (a latent) etiology (the rectangles in Figure 4.3) and modify the corresponding covariates. For example, etiology 3 affects X_4 and X_5, so the joint distributions of (X_4, X_5) are shifted to have larger mean values than healthy patients. X_1, which can be affected by either etiology 1 or 2, will follow different distributions in patients assigned to etiology 1 and 2, and will also be different from healthy patients.

Next, we introduce MNAR missingness. The change we made to the covariate distributions to transform disease-free patients into diseased patients imposes a correlation structure that is different from the original disease-free structure. We use BN to learn this new structure [81]. The stronger the relationship is between two covariates, the more likely that we will make one of the two missings (delete it). For example, if the patient's condition can be deduced from a set of lab tests, certain other tests may not be necessary to perform (they will be missing). Note that in such cases, much of the "missing" information is redundant and the outcome can be predicted.

The "bnlearn" R package is used to automatically learn BN.

Results on Synthetic Data (Experiment 1)

We generate 10 different synthetic datasets with 6 covariates and 5,000 observations each and test all methods on them (except PPM+DM, which we will discuss later). We report the performances in Figure 4.4. The horizontal axis corresponds to increasing missing data rates and the vertical axis shows the average AUC (over 10 data sets, 10 runs, cross-validated) and its confidence interval (as error bars). For better visibility,

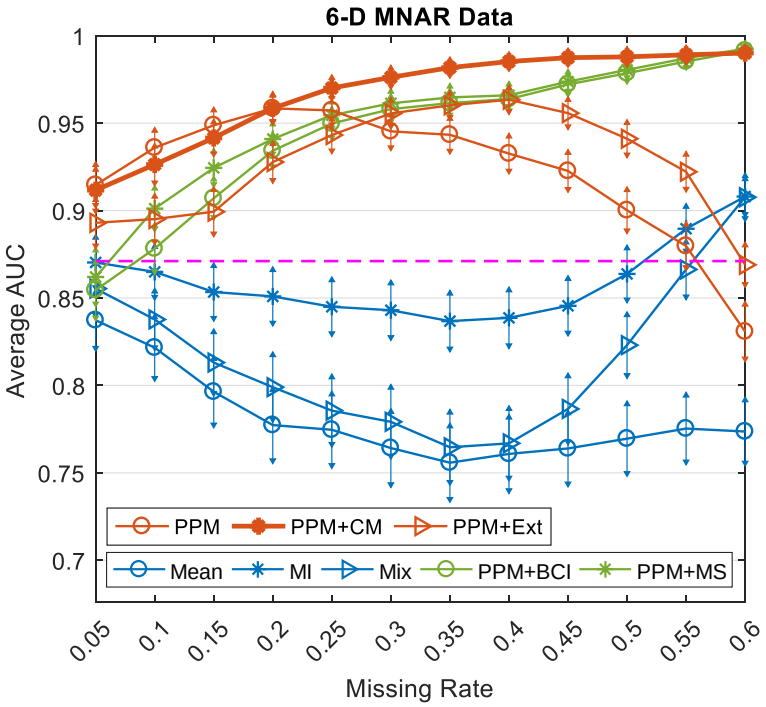


Figure 4.4: Performance Comparison on 6-Dimensional Synthetic Data with MNAR missingness. The vertical error bars are 95% Confidence Intervals (CI) for AUC, averaged over the 10 runs. The magenta dashed line indicates performance on complete data

we use blue color for imputation methods, green for PPM+BCI and PPM+MS, and red for PPM, PPM+CM, and PPM+Ext. Since the simulated covariates are Gaussian, a linear Gaussian model estimate is used in both MI and Mix. Our proposed methods are in green and red.

The proposed methods all outperformed the imputation methods (blue lines) in all cases except the (base) PPM and PPM+Ext when the missingness rate was in excess of 0.55. Overall, PPM+CM had the highest AUC (and also the narrowest CI), with PPM+MS and PPM+BCI being a close second. As the missing rate increased, the underlying MNAR missing mechanism became more obvious, which negatively impacted some of the PPM methods.

To understand the performance differences, we need to consider two phenomena.

First, as the missingness rate increases, the number of missingness patterns increases, and the number of per-pattern observations decrease (because the patterns are disjoint

Average SMD over the 10 Runs				
X	missing rate 0.2		missing rate 0.5	
	Before	After	Before	After
X_1	0.0479	0.0397	0.0062	0.0048
X_2	-0.0790	-0.0214	-0.1099	-0.0430
X_3	0.0569	0.0104	-0.1424	-0.0831
X_4	-0.0646	0.0013	-0.2648	-0.0136
X_5	0.1844	0.0973	0.1453	0.0613
X_6	0.1483	0.0847	0.3304	0.0111

Number of Observations Per Pattern				
Runs	missing rate 0.2		missing rate 0.5	
	Before	After	Before	After
run 1	250	683	125	153
run 2	185	494	98	115
run 3	139	904	104	187
run 4	147	832	125	208
run 5	217	651	143	187
...	...	...	...	...

Table 4.2: Comparison before and after Causal Matching

in terms of the observations). As the number of observations decreases, at some point, modeling the subpopulation becomes more challenging. For PPM, this point is clearly visible and appears to be at missingness rates between 0.15 and 0.25: beyond this point, its performance degrades. Methods that can perform well at higher missingness rates can either expand their training set or perform model selection (and avoid building models that would perform poorly).

The second phenomenon is that the number of patterns does not grow indefinitely; it peaks at a missingness rate of 0.45. Increasing the missingness rate beyond this point creates empty patterns (patterns with zero observations), thus the number of patterns decreases, making the modeling task easier. Although missing data implies information loss, much of the missing data is actually redundant, and hence PPM+CM (and PPM+BCI and PPM+MS, as well) is managed to achieve very high performance despite the high missing rate. PPM+Ext, however, did not benefit from this.

To illustrate the difference between PPM+CM and PPM+Ext, we created Table 4.2, which shows the Standardized Mean Difference (SMD) in the covariates and the number of observations between $\mathcal{S}^p$ and $\mathcal{E}^P$ before and after causal matching. SMD is the difference between the means of X_i, in terms of the pooled standard deviations and it is a standard metric to evaluate the quality of propensity matching. Small absolute values of SMD indicate high similarity between the two compared distributions. Table 4.2 shows that the extended subpopulation $\mathcal{S}^p \cup \mathcal{E}^P$ without matching has a distribution that is significantly different from $\mathcal{S}^p$ and hence it is not appropriate to build a sub-model for $\mathcal{S}^p$. On the other hand, PPM+CM, which performs matching, used observations from $\mathcal{E}^P$ that collectively match the distribution in $\mathcal{S}^p$ and thus the extended subpopulation after matching is appropriate for modeling $\mathcal{S}^p$. The distributional differences between $\mathcal{S}^p$ and $\mathcal{E}^P$ prevented PPM+Ext from achieving higher performance. This is a direct consequence of non-random missingness.

Despite the reduced performance, both PPM and PPM+Ext still (mostly) work better than imputations.

The three imputation methods perform substantially worse than all of the proposed methods (except the base PPM and PPM+Ext at very high missing rates). Although Pattern Mixture Imputation (**Mix**) is specifically designed for MNAR mechanism, it performs worse than even Multiple Imputation (**MI**). As an imputation method, Mix

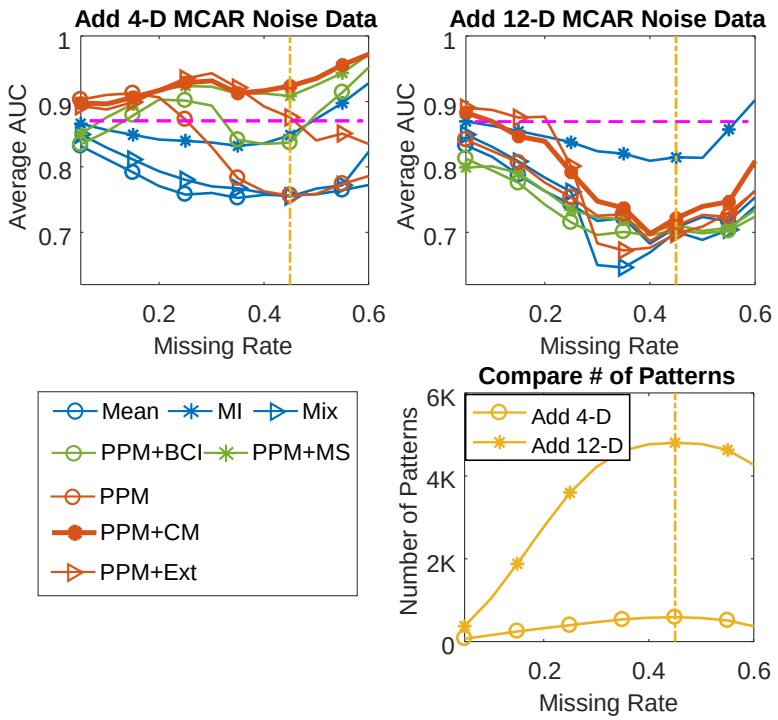


Figure 4.5: Performance Comparison When Adding Different Dimensional Noise Data with MCAR, and Their Changing of Number of Patterns. The vertical dashed yellow line indicates the peak of the number of patterns.

models the missing variables rather than the outcome, which introduces unnecessary model assumptions and errors. MI reduces imputation errors through resampling and learning from multiple runs. Resampling comes at a cost: MI has the longest running time among all methods, whose maximal is 0.3 minute (at a rate of 0.4), which is three times as long as PPM+CM (running time data is not shown for brevity).

Results on Synthetic Data (Experiment 2)

We demonstrated that PPM+CM can effectively handle MNAR when the number of the missing patterns is relatively small. When the missing patterns no longer correspond to meaningful disease etiologies and become random, their number can increase exponentially. Having too many patterns can be mitigated in a straightforward way by building models only for the largest patterns or for the least random patterns. Investigating such

strategies is outside the scope of this paper, however, understanding how and when the algorithms fail is valuable.

Adding MCAR noise constitutes the worst-case scenario because it increases the number of missingness patterns exponentially. Figure 4.5 displays the performances of adding 4 and 12 Dimensional MCAR noise data, respectively, and can be interpreted analogously to Figure 4.4. Our results show that PPM+CM outperforms the other methods until 8-D noise data added (not shown). 8-dimensional noise could produce as many as 3,000 missing patterns, which is very high compared with the 5,000 observations. When the added noise dimension exceeds 12, the number of patterns reaches 5,000 implying that only one observation exists in each pattern. Statistical tests to detect MCAR missingness exist and appropriate imputation methods can be used to impute MCAR patterns as a preprocessing before applying the PPM-family methods to model data with non-random missingness.

Computation Cost. The computation cost of PPM-family methods depends on the number of missing patterns, rather than data dimensions. In reality, the number of disease etiologies is limited, thus the number of meaningful patterns is also limited. The above extreme scenario (5000 patterns within 18 variables) is only meant to illustrate the limitation of the PPMs methods and is not a practical scenario.

Furthermore, it is worth mentioning that in our experiments, MI always had the longest running time. In the presence of 12-D noise data, MI runs in 4.5 minutes, while PPM+CM is completed in 2.1 minutes.

4.4.2 Experiments on Real EHR Data

Sepsis Dataset

We use the clinical data repository of a large health care system situated in the Midwestern United States. In a retrospective cohort study of 26,009 adult patients who presented at the emergency department (ED) between 2010 and 2015 with infection and subsequently got admitted to the hospital, 7,175 developed sepsis, and 18,834 did not progress to sepsis. For this cohort, we extracted 21 covariates (vital signs, lab results, and medication administrations) from the ED visit to predict whether patients have sepsis during the hospitalization. Infection, sepsis, severe sepsis, and septic shock

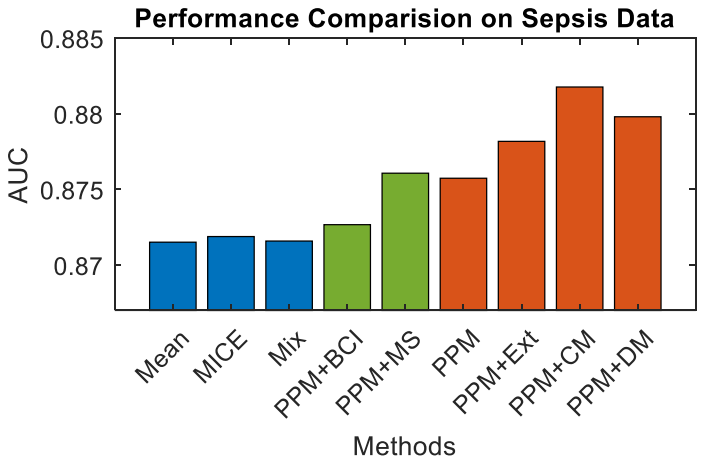


Figure 4.6: Performance Comparison on sepsis dataset: Blue, green and red bars are for different display groups.

were identified by their ICD-9 and/or ICD-10 codes. Multiple results for the same lab or vital signs were summarized as the latest result in the ED before admission to the hospital.

$59,103$ of $26,009 \times 21$ entries are unobserved, corresponding to an average missing rate of 9.88%. There are 137 missing patterns in total, 99 of them contains fewer than 16 observations (patients), while the largest pattern has $6,002$ observations, i.e. 23.1% of the entire population. We further checked the proportion of sepsis observations in the missing patterns. For the entire population, this is 27.59%, but for 137 missing patterns, this is much higher, on average 50%.

Results on Sepsis Data

Figure 4.6 displays the performance comparison of all methods in terms of the average AUC over the 10 cross-validation runs. Pattern-Wise methods, even the simplest PPM, outperformed all imputation methods. Among the PPM-family methods, PPM+CM performs best.

Due to model selection, PPM+MS performs better than both PPM+BCI and PPM (without model selection). Also, PPM+Ext works better than PPM (without extension) because MAR missingness dominates the missing patterns in this data, thus including

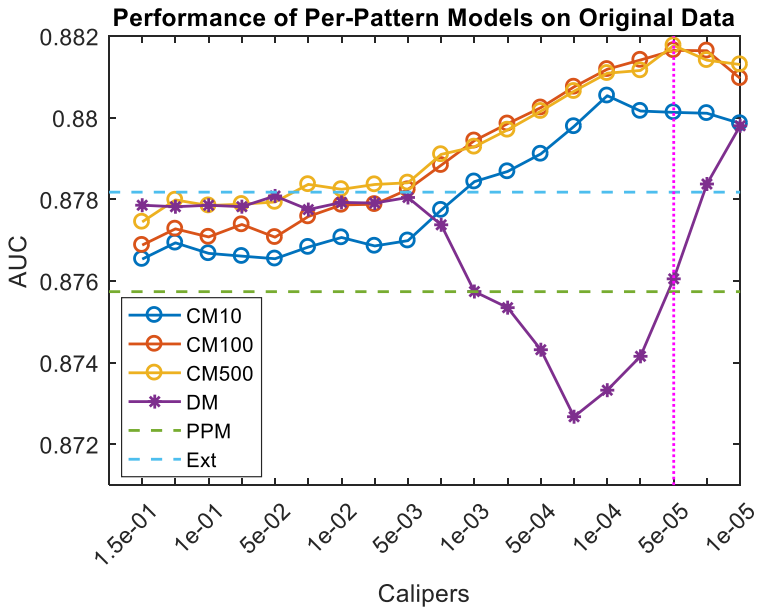


Figure 4.7: Performance of PPM-family models with decreasing calipers

all observations from the extension subpopulation is advantageous.

Figure 4.7 displays the performance of PPM+CM and PPM+DM as a function of caliper size. Recall that caliper controls how close the propensity scores must be for a match. A caliper size of 0 requires an exact match: an observation in $\mathcal{S}^p$ matches an observation in $\mathcal{E}^P$ only if they are identical. To put the effect of the caliper size into context, the figure also displays the performance of PPM and PPM+Ext. Since these methods do not use a caliper, their performance is constant and are hence represented as horizontal dashed lines. PPM+Ext is the top line and PPM is the bottom one.

Our expectation is that as the caliper decreases, the matched observations are increasingly similar to the observations in the subpopulation, making the distributions of $\mathcal{E}^P_{mtc}$ and $\mathcal{S}^p$ increasingly similar. As the two distributions become increasingly similar, the problem is gradually converted from MNAR to MAR, resulting in improved performance. As the caliper becomes too small, the number of successful matches decreases, and the training set becomes too small to build a good model. The performance here drops.

The behavior of PPM+CM is congruent with this expectation. We tried several K

(required number of matches in $\mathcal{E}^P$ for each observation in $\mathcal{S}^p$). The figure shows 10, 100, and 500, labeled as CM10, CM100, and CM500 in the figure, respectively. Training set expansion is most advantageous when $\mathcal{S}^p$ is small. With 16 patients in some of the small patterns, estimating even an intercept is difficult. As we expand the training set (essentially) 10 times, we can reliably estimate not only the intercept but also multiple (observed) coefficients. Increasing K further to 100 approximately increases the training set size by a factor of 100. Naturally, not all observations in $\mathcal{S}^p$ will have matches, but we expect the resultant training set to be much larger than at $K = 10$. With a maximum of 21 features, increasing K from 100 to 500 had a minimal benefit: the training set size at $K = 100$ was sufficiently large to estimate all coefficients.

The behavior of PPM+DM was surprising. As the caliper decreases from 1 towards 0, the training sample gradually changes from $\mathcal{S}^p \cup \mathcal{E}^P$ to $\mathcal{S}^p$, i.e. PPM+DM changes from PPM+Ext to PPM. But at some caliper value, the performance sharply drops even below that of PPM. Recall that PPM+DM matches on the joint probability of (X, Y) not just on X as PPM+CM does. When the data is MNAR, the missingness pattern itself is predictive of Y; and conversely, Y will be a good predictor of the propensity of an observation belonging to a particular pattern. For some MNAR patterns, the outcome Y becomes the dominant predictor and ultimately, we match chiefly on the outcome. This can distort the relationship between X and Y as the "matched" observations in the extension subpopulation can have very different distributions in terms of X. This explains the drop in performance. As the caliper becomes tighter, the matching is increasingly based on both the covariates and the outcome, making the AUC rise back and ultimately surpass the PPM+Ext method.

Although PPM+DM can work well with a good caliper choice, we do not recommend using it in general, because there is always the risk of skewing the X distribution. In contrast, PPM+CM can consistently outperform all baseline methods for a relatively wide range of calipers.

MNAR Patterns Analysis on Sepsis Data

To illustrate that PPM+CM is capable of accurately preserving the distribution of small subpopulations, we analyze some small MNAR patterns. Table 4.3 shows two such patterns. The first one is lactate in patients with missing respiratory rate and

temperature. Lactate is a critical measure of severe sepsis and elevated lactate is strongly associated with mortality. The second one is glucose. In sepsis, low values of glucose are associated with malnutrition, which in turn is associated with increased mortality.

In pattern #1, the missing variables are Respiratory Rate (Resp) and Temperature (Temp), while lactate is observed and is the dominant variable. We compare the lactate distribution in the training samples used for PPM, PPM+CM, and PPM+Ext methods to the one from the mean imputed complete data through the Kolmogorov-Smirnov (KS) test, whose statistics and p-value are listed. Patients in PPM training samples, $\mathcal{S}^p$, have lower lactate than the entire population and are thus at lower motility risk. PPM+CM preserves this characteristic well, despite extending the subpopulation from 17 to 90 observations. In contrast, the PPM+Ext method over-extends the subpopulation to 315, resulting in a lactate distribution that is more similar to the entire (imputed) population (The KS statistic decreased). PPM+CM has a much better AUC evaluation on this pattern compared to PPM+Ext. PPM method fails to construct a reasonable model on such few observations.

In pattern #2, where in addition to Resp and Temp, lactate is also missing, we are observing the distribution of glucose. We performed the KS test, which shows that the patients in pattern #2 have higher glucose than the entire population, i.e. lower risk of malnutrition. PPM+CM preserves this feature well, while PPM+Ext almost loses it by adding too many (9400+) observations (the KS statistic is an order of magnitude lower).

4.5 Conclusions

In this manuscript, we proposed a general methodology, Pattern-Wise Analysis, for predictive modeling in the presence of non-randomly missing values. At a high level, Patter-Wise Analysis constructs a model for each missingness pattern.

We have demonstrated that even the simplest patter-wise method, the baseline Per-Pattern Modeling (PPM), outperforms the most popular imputation methods. There are two interrelated challenges that PPM faces: there can be too many patterns (and thus too many models to construct) and each pattern may contain too few observations.

Among the many variants of PPM, we recommend Causal Matching (PPM+CM)

Pattern #1: missing (Resp, Temp)				
		PPM	PPM+CM	PPM+Ext
KS test on Lactate	stats	0.5179	0.5196	0.4015
	p-value	0.00011	1.88e-25	0.00e+00
Number of Observations		17	90	315
AUC		NULL	0.885	0.687

Pattern #2: missing (Lactate, Resp, Temp)				
		PPM	PPM+CM	PPM+Ext
KS test on Glucose	stats	0.2462	0.2015	0.0288
	p-value	0.0208	8.92e-06	9.78e-06
Number of Observations		32	144	9503
AUC		0.663	0.762	0.634

Table 4.3: Significant parameters analysis on two MNAR patterns

for MNAR missingness in EHR data analysis. The experimental results demonstrate that regardless of whether missingness is MAR or MNAR, CM method can outperform the basic imputation methods, and also most other proposed methods. This conclusion confirms that our causal model for the informative missingness is reasonable. While PPM+DM can achieve marginally higher performance than PPM+CM, PPM+CM is a safer choice because it performs well for a much wider range of parameter settings.

Out methodology assumes a small number of meaningful missingness patterns, which could, for example, correspond to disease etiologies. We have demonstrated on a real-world sepsis data set, that the method can handle a practical (but unknown) mixture of MAR, MCAR, and MNAR missingness without having to take any special measures.

However, especially when MAR or MCAR missingness dominates, the number of missingness patterns can become exponentially large, and building a separate model for each pattern is intractable. There are methods that can identify random patterns, and statistical tests to identify MAR/MCAR missingness, so the number of missingness patterns can be reduced by imputing MAR/MCAR missing data and only using the PPM methodology for the MNAR patterns. Alternatively to imputation, one can apply

the PPM methodology only to the largest or least random missingness patterns.,

In summary, our suggestion is that if the missingness is likely to be MNAR, do not make MAR assumptions and apply imputations just for simplicity. The PPM+CM or PPM+MS are safe and high-performing choices. If the number of missingness patterns appears intractably large, one can either eliminate the MAR/MCAR patterns (via imputation) or can choose to model only the largest or least random patterns.

Chapter 5

Assessment of the Association Paradox between BMI and All-Cause Mortality

5.1 Abstract

Objective

The association between body mass index (BMI) and all-cause mortality is controversial, frequently referred to as a paradox. Whether the cause is metabolic factors or statistical biases is still controversial. We assessed the association between BMI and all-cause mortality considering a wide range of comorbidities and baseline mortality risk.

Methods

A retrospective cohort study of Olmsted County residents with at least one BMI measurement between 2000–2005, clinical data in the electronic health record, and minimum 8-year followup or death within this time. The cohort was categorized based on baseline mortality risk: Low, Medium, Medium-high, High, and Very-high. All-cause mortality was assessed for BMI intervals of 5 and $0.5 Kg/m^2$.

Results

Of $39,739$ subjects (average age 52.6, range 18–89; 38.1% male) 11.86% died during 8-year follow-up. The 8-year all-cause mortality risk had a "U" shape with a flat nadir in all the risk groups. Extreme BMI showed higher risk (BMI $< 15 = 36.4\%, 15 to < 20 = 15.4\% and \geq 45 = 13.7\%$), while intermediate BMI categories showed a plateau between 10.6 and 12.5%. The increased risk attributed to baseline risk and comorbidities was more obvious than the risk based on BMI increase within the same risk groups.

Conclusions

It is confirmed that there exists non-linear causality or complex association from BMI to all-cause mortality when evaluated including comorbidities and baseline mortality risk. To be specific, from the view of conventional statistic linear modeling, comorbidities are better predictors of mortality risk most time, but, the extreme BMIs replace comorbidities to be the dominating predictor. Therefore, to keep the model conventional and interpretable, we suggest to add up BMI variable as the nonlinear condition (e.g. learned by multi-layers perceptron), based on the well-defined linear regression from comorbidities to mortality.

In patients with no or few comorbidities, BMI seems to better define mortality risk. Aggressive management of comorbidities may provide better survival outcomes for patients with body mass between normal and moderate obesity.

5.2 Introduction

The association between body weight distribution is defined by body mass index (BMI= Kg/m^2) and all-cause mortality is a controversial topic. Multiple studies, including several systematic reviews and meta-analyses, have attempted to explain this association and found different results [82–89]. The general association of BMI and all-cause mortality follows a U or J curve, with very high mortality among people with very low BMI ($<$ 18.5) and very high (BMI$>$ 40). However, the most common unexpected finding is that people defined as having a normal or ideal weight with BMI of 18.5 to 25 do not necessarily have the best survival. In many cases, overweight people (BMI 25 to 30), and those with mild to moderate obesity, (BMI of 30 to 35 and 35 to 40), show

the best survival. This phenomenon has been described as the "obesity paradox" and it is the subject of intense review due to the potential and very significant impact on many aspects of routine clinical practice and the healthcare in general [90–94]. The obesity paradox has been described not only in the general population [82] but also in multiple cohorts of people with highly prevalent medical conditions including diabetes, heart disease [91, 95–97] , kidney disease [98, 99], cancer [100, 101], stoke [102, 103] , and rheumatoid arthritis [104], among others. The potential explanations are described sometimes in relation to the nature of the clinical condition, and sometimes attributed to methodological biases. For example, in patients with cancer and on chemotherapy, having more metabolic reserves may improve their chances of survival. However, in other conditions such as diabetes and congestive heart failure, obesity associated with better survival cannot be easily explained. Many possible methodological biases have been proposed, including reverse causation, confounding factors, and several biases such as selection, survival, and treatment biases [105]. However, these biases are not present consistently in all the studies. Many discussions also center on the definition of normal weight and obesity by using BMI [106, 107]. Experts agree that BMI is an imperfect measure of body fat and may be influenced by many factors, including body composition of muscle mass, fat distribution, visceral vs. subcutaneous fat, and ectopic fat. Even more important, physical fitness and nutritional status may play a more important role than BMI in predicting overall health and risk of mortality [108]. However, one concept seems to be well accepted, obesity is a well-established and important risk factor for many metabolic and cardiorespiratory conditions, and these conditions have been associated with reduced survival [109]. Now, the question raised by this paradox is: In the presence of one or more clinical conditions, and once modern therapeutic interventions have been taken to manage these underlying metabolic consequences of obesity, is BMI still a main determinant of survival? Are there other factors in the complex relationship between diseases, therapy , and follow-up that become more important than the BMI? Based on these observations, the aim of this study was to use a new approach to assess the relationship between BMI and all-cause mortality by using a risk model based on multiple clinical characteristics, including comorbidities, to define baseline mortality risk and analyze the association between this mortality risk and BMI using the BMI in categories and as a continuous variable.

5.3 Methodology

5.3.1 Study setting and participants

A retrospective cohort study was conducted using clinical data from the Mayo Clinic electronic medical records and related databases including the Rochester Epidemiology Project (REP). Mayo Clinic is a large integrated Medical Center located in Rochester, Minnesota, that provides care to national and international patients but also to the Rochester community. It has an integrated, inpatient and outpatient, electronic medical record (Epic, Madison, WI) and comprehensive electronic clinical data dating from the 1990s The REP is a research data repository approved for medical research. It links together almost all the medical records of the residents of Olmsted County, MN, over several decades [110,111]. From this repository, electronic searches identified a cohort of $52,148$ Olmsted county residents seen at Mayo Clinic in Rochester, Minnesota, between January$1, 2000$ and January $1, 2005$. This cohort included adults aged 18 or older, alive on $2005-01-01$, and with at least one BMI measurement between 2000 and 2005. The Median follow-up was 9.2 years. Subjects that dropped out without an event within 8 years and those for whom 8-year mortality could not be calculated were excluded leaving $39,739$ subjects in the study cohort. According to Minnesota state law (Minnesota state privacy law, Statute 144.335), subjects without research authorization were excluded. The study followed best practice to protect the confidentiality of the medical data, including those related to the Health Insurance Portability and Accountability Act (HIPAA). This study was reviewed and approved by Olmsted Medical Center and Mayo Clinic Institutional Review Boards.

5.3.2 Data collection

Data were collected using an electronic search of the REP database and the Mayo Clinic electronic health record including inpatient and outpatient visits. The baseline data was collected between January$1, 2000$, and January $1, 2005$, and included demographic data, vital signs, laboratory results, medications, and diagnoses. We collected height, and weight, and calculated BMI that were measured during clinic visits and documented in the electronic medical records. Multiple BMI observations were aggregated by computing their average. Prescriptions between 2003–2005 were collected and aggregated using

NDF- RT therapeutic classes (antidiabetic, antihypertensive, and antihyperlipidemic drugs). We used the Elixhauser Comorbidity list to categorize the comorbidities of the patients using the International Classification of Diseases (ICD) diagnosis codes found in the EHR. Each comorbidity category was considered either present or not present (dichotomous) [112]. The primary outcome was all-cause mortality. Mortality status was collected using the REP database during the entire follow-up period and allowed to calculate 8-year mortality rates [111].

5.3.3 Mortality risk model and risk groups

We constructed a risk model to assess the individual 8-year all-cause mortality risk at baseline, excluding the direct effect of obesity for all the subjects in the cohort. A logistic regression model was used with independent variables including age, sex, Exlihauser Comorbidities except for obesity, and additional variables including hyperlipidemia, coronary artery disease, chronic kidney disease, and stroke. There were no missing values and backward elimination was used for feature selection. Only hyperlipidemia showed the opposite effect due to collinearity with coronary artery disease. The risk model achieved an area under the receiver operating characteristic curve (AUC) of 0.91.

The cohort was categorized into five risk groups using percentiles of the patient's baseline mortality risks derived from the model:

- Low: patients in the lower half of the risk scores.
- Medium: patients between $50th$ and $75th$ percentile of the risk scores.
- Medium-high: patients between $75th$ and $90th$ percentile.
- High: patients between the $90th$ and $95th$ percentile.
- Very-high: patients above the 95% percentile.

Higher risk groups had much more comorbidities than lower risk groups. The low-risk group had on average 0.8 comorbid conditions, the medium-risk group had 2, the medium-high had 3.6, the high-risk group had 5.1 and the very-high risk group had 7.8 and an approximately 80%8-year mortality rate.

5.3.4 Body mass index

The average BMI for each individual at baseline was used in the study. BMI categories were defined every 5 points starting at < 15 and ending at ≥ 45. These categories differ from those defined by the National Institutes of Health and the World Health Organization in the categories with a cutoff of 18.5, underweight < 18.5 and normal 18.5 to < 25. These categories allowed the analysis to avoid preconceptions related to what should be normal or abnormal (underweight, normal, and obesity) [106, 113]. The BMI was also used as a continuous variable with intervals of $0.5Kg/m^2$.

5.3.5 Statistical analysis

The unadjusted mortality rate was calculated as the fraction of patients who died within 8 years among patients in the given risk group and BMI category. To better understand the effect of BMI on mortality, BMI was also used as a continuous variable at $0.5kg/m^2$ resolution. For each BMI level, between 15 and 55, we computed the unadjusted mortality rate using kernel estimation with a Gaussian kernel. The empirical 95% confidence interval was obtained through 200 iterations of bootstrap resampling. To adjust for baseline risk and to assess whether some of the excess mortality risks could be attributed to BMI, we considered residuals. For a patient, the residual was the difference between the patient's predicted risk and the actual outcome. It could be interpreted as the patient's excess risk of mortality not explained by the baseline comorbidities. As before, for each BMI level between 15 and 55 we computed the average residual (excess risk of mortality) using kernel estimation with a Gaussian kernel and bandwidth of $0.5kg/m^2$. The empirical 95% confidence interval was obtained through 200 iterations of bootstrap resampling.

5.4 Experimental Results

The cohort of $39,785$ adult subjects (average age 52.6, range 18–89; 38.1% male) had an overall all-cause mortality of $11.9\%(4,713$ deaths) in 8-year follow-up. The mortality risk model was used to define five risk groups (Low, Medium, Medium-high, High and Very-high) as defined in the methodology. The table in Fig 5.1shows the baseline characteristics of the entire cohort and the risk groups. The mean age and the prevalence of

Baseline comorbidities and mortality risk model.

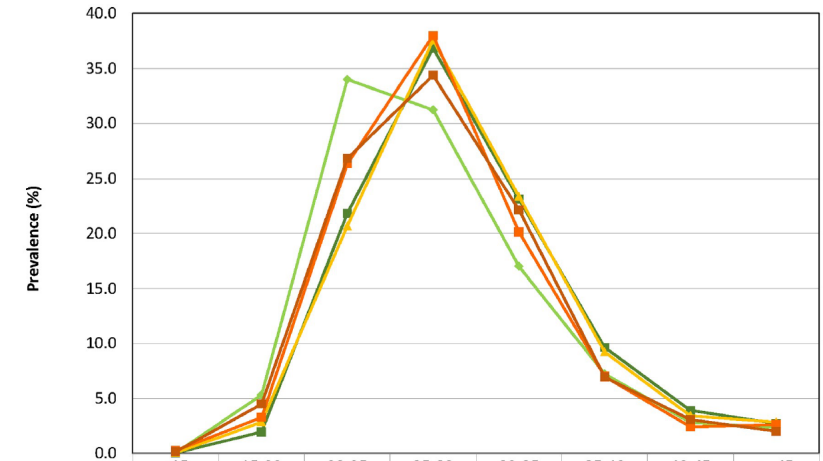
Comorbidities	Prevalence of Comorbidities in the Risk Groups (%)						Mortality Risk Model	
	Low	Medium	Medium High	High	Very High	Entire Cohort	Coefficient	p-value
N	19865	9939	5961	1987	1987	39739	-	-
(Intercept)	-	-	-	-	-	-	-8.112	<0.001
Age (years, mean)	38.1	58.1	72.0	80.9	83.0	52.6	0.082	<0.001
Male (%)	31.8	43.0	45.9	43.3	47.1	38.1	0.231	<0.001
Hyperlipidemia	24.8	49.3	62.2	57.9	53.5	39.6	-0.584	<0.001
Hypertension	13.6	44.3	70.6	82.4	89.3	37.1	0.132	0.007
Chronic lung disease	13.7	20.4	26.2	33.9	46.5	19.9	0.452	<0.001
Deficiency anemia	6.1	11.2	23.1	36.5	61.4	14.2	0.232	<0.001
Psychoses	9.2	11.9	9.9	13.7	27.1	11.1	0.372	<0.001
Fluid and electrolytes disorder	3.0	6.2	16.2	33.7	65.4	10.5	0.555	<0.001
Coronary artery disease	1.4	11.6	31.5	46.3	62.0	13.7	0.342	<0.001
Valvular disease	2.7	7.2	17.7	29.1	46.0	9.5	0.196	0
Other neurological disorders	2.2	6.1	11.9	21.9	42.0	7.6	0.754	<0.001
Diabetes mellitus	2.7	11.7	22.0	26.1	33.8	10.5	0.442	<0.001
Peripheral vascular disorder	0.7	3.8	13.7	24.0	37.3	6.4	0.410	<0.001
Solid tumor without metastasis	1.4	6.3	16.1	23.0	30.0	7.4	0.400	<0.001
Weight loss	1.4	2.3	4.7	9.6	21.5	3.6	0.562	<0.001
Congestive heart disease	0.1	0.7	5.2	19.3	52.6	4.6	0.860	<0.001
Alcohol abuse	2.2	4.3	4.9	4.1	5.6	3.4	0.622	<0.001
Chronic kidney disease	0.2	0.8	3.2	7.4	21.9	2.2	0.612	<0.001
Pulmonary circulation disorder	0.4	1.0	4.1	8.5	18.8	2.4	0.291	0.002
Coagulopathy	0.8	1.6	3.9	6.5	14.9	2.4	0.485	<0.001
Liver disease	1.0	2.6	3.7	3.4	4.8	2.1	0.507	<0.001
Stroke	0.1	0.7	3.2	7.1	15.7	1.8	0.251	0.019
Drug abuse	1.6	2.2	2.0	2.1	2.1	1.8	0.755	<0.001
Metastatic cancer	0.0	0.9	3.4	7.0	12.4	1.7	1.212	<0.001
Paralysis	0.3	1.0	2.3	4.0	11.1	1.5	0.744	<0.001
Lymphoma	0.2	0.6	1.9	2.4	3.9	0.8	0.588	0
HIV/AIDS	0.0	0.0	0.2	0.2	0.1	0.0	1.810	0.003

Comorbidities with prevalence less than 1% were omitted in this table.

Figure 5.1: Baseline comorbidities and mortality risk model.

the comorbidities increased with the increase in the mortality risk in all the risk groups.

The distribution of BMI within each risk group was very similar except for the low-risk group that had slightly more subjects with BMI of 20 to < 25 and slightly less with BMI of 25 to < 35 (Fig 5.2). The category with the most subjects was BMI between 20 and 35 with $13,543$ subjects. The number of subjects with BMI < 15 was small ($n = 22$), followed by BMI ≥ 45 (n = 1000). The rest of the BMI categories had between $1,259$ and $13,543$ subjects (Table in Fig 5.3).



	<15	15-20	20-25	25-30	30-35	35-40	40-45	>=45
Low	0.0	5.3	34.0	31.2	17.0	7.2	2.8	2.4
Medium	0.0	2.0	21.8	36.8	23.1	9.6	3.9	2.7
Medium-high	0.1	2.9	20.7	37.6	23.4	9.2	3.4	2.8
High	0.3	3.3	26.3	38.0	20.1	7.0	2.4	2.6
Very-high	0.2	4.5	26.8	34.4	22.1	6.9	3.1	2.0

Figure 5.2: Distribution of the prevalence (%) of the eight BMI categories among the five mortality risk groups: Low, medium, medium-high, high and very-high risk.

Mortality risk among the risk groups based on BMI

The table in Fig 5.3 shows the 8-year mortality rates in each BMI category for the entire cohort and each risk group. The categories with very low BMI and very high BMI showed the highest mortality rates for all the risk groups. The intermediate categories showed lower risk, picturing the traditional "U" distribution for all the risk groups but with a wide and plateau nadir. The changes in mortality rates between risk groups were more significant compared to changes between BMI categories within the same risk group. For example, for BMI 25 to < 30 the mortality rate for the Low, Medium, Medium-high, and High risk groups is $0.012, 0.043, 0.222 and 0.514$ respectively. However, in the Medium-high risk group the mortality rate for the BMI of 25 to $< 30, 30$ to < 35 and 35 to < 40 was $0.222, 0.226$ and 0.239 respectively. These findings are better appreciated in Fig 5.4 where each risk group and the entire cohort are represented as their unadjusted mortality risk and 95% confidence interval, using BMI as a continuous variable. The mortality risk increases significantly between the risk groups, with relatively small changes attributable to changes in BMI, except for very low and very high BMI.

Raw 8-year mortality rates in the entire cohort and the risk groups by BMI categories.

BMI	Entire Cohort			Mortality Risk Groups				
	N = 39739	Mortality N = 4713	Mortality Rate (0.1186)	Low	Medium	Medium High	High	Very High
<15	22	8	0.364	0.000	0.000	0.333	0.800	1.000
15–<20	1576	243	0.154	0.009	0.108	0.453	0.785	0.944
20–<25	11212	1193	0.106	0.006	0.040	0.278	0.581	0.784
25–<30	13543	1652	0.122	0.012	0.043	0.222	0.514	0.782
30–<35	7909	989	0.125	0.011	0.044	0.226	0.513	0.752
35–<40	3218	349	0.108	0.006	0.039	0.239	0.496	0.746
40–<45	1259	142	0.113	0.014	0.057	0.225	0.521	0.672
>=45	1000	137	0.137	0.021	0.074	0.266	0.577	0.800

BMI = body mass index, Kg/m^2.

Figure 5.3: Raw 8-year mortality rates in the entire cohort and the risk groups by BMI categories.

In these extreme BMI categories, the mortality risk seems to have a noticeable increase but with wide confidence intervals. The nadir for each group is a plateau, except for the high and Very-high risk groups that have fluctuations in mortality risk in the BMI over 35 and 40. Mortality seems to be low in the Very-high risk group with BMI close to 40. These values are difficult to interpret and likely are a statistical anomaly evident by the wide confidence intervals that denote the small and likely heterogeneous group of subjects.

Effect of body mass index on mortality adjusted by comorbidities

Fig 5.5 shows the excess risk of mortality as a function of BMI as a continuous variable ($0.5 kg/m^2$ between 15 and 55) for the entire cohort and each risk group. The residual mortality risk that is not explained by the comorbidities but attributed to BMI has again a "U" distribution with a plateau nadir for the entire cohort and all the risk groups. For all the groups, low BMI and very high BMI were associated with mortality higher than the average of the risk group (residuals are consistently above the zero line). Conversely, intermediate BMIs were associated with lower mortality than the average of the risk group (residuals were consistently below the zero line). This distribution was easy to appreciate in the entire cohort and the Low, Medium, and Medium-high groups. In the High and Very-high groups this trend is less clear, in part due to the low number of subjects and the consequent wide confidence intervals. Also, the low risk does not seem to correspond to the traditional definitions of normal, overweight, and obesity. For

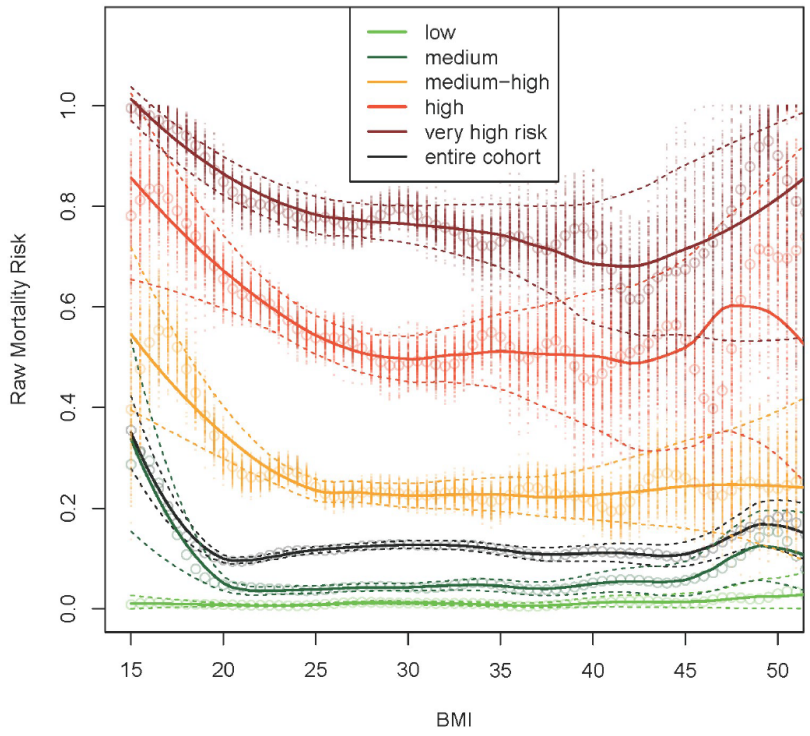


Figure 5.4: Raw 8-year mortality rates at each $0.5Kg/m^2$ BMI intervals between 15 and 55 in the entire population (black line) and the five mortality risk groups. Dots represent the point estimates of the mortality rate in the 200 bootstrap iterations, circles are their mean, the solid line is a smooth curve representing the averages, and the dashed lines are the 95% empirical confidence interval.

example, for the entire cohort, BMI approximately between 20 and 40 showed similarly residual mortality risk when adjusted by comorbidities.

5.5 Discussion and Conclusion

Our results show that the main driver of all-cause mortality risk is comorbidities, assessed in our study by a risk model mainly based on baseline comorbidities and the categorization of the risk in five mortality risk groups. While moving to a higher comorbidity-based risk group (by developing new comorbidities) increases the mortality risk by approximately 20–30% (except between low and medium risk groups), the

interquartile variation within the risk groups is much smaller when we exclude extreme low and extreme high BMI. In general, within each risk group, the distribution of BMI was close to a normal distribution and very similar among all the groups. The association between BMI and all-cause mortality displayed a "U" distribution regardless of the baseline mortality risk. Our data did not show a nadir in a specific BMI category; instead, the nadir was in general a plateau including what is defined as normal, overweight, and obesity with some variations across the different risk groups. These findings suggest that the mortality risk is essentially flat across these BMI categories. Patients with extreme BMI (underweight and extreme obesity) show an increased risk of mortality, independent of the baseline comorbidity risk, but the (interquartile) variation within a risk group is much smaller than the risk difference between groups. Subjects in the low and medium risk groups, which comprise 75% of the population, have very few or no comorbidities. In the absence of serious comorbidities in these patients, mortality depends primarily on BMI. In particular, BMI in excess of $38 kg/m^2$ is associated with increased mortality. In low risk patients, the increase in mortality between 25 and $40 kg/m^2$ is a relative 65% (from 0.8% to 1.10%). Similarly, in medium-risk subjects, the increase in mortality risk between 25 and $40 kg/m^2$ is 68% (from 3.6% to 5.3%). In the absence of significant comorbidities, BMI is an important driver of mortality; however, the variation of mortality risk based on BMI changes within each group is much smaller than the risk difference between the groups. We also encountered some paradoxical findings, but only in the top 5% of the population with the most baseline comorbidities. These patients have 78% mortality in 8 years on average. In these subjects, and only in these subjects, obesity with BMI < 42 may be protective, but BMI in excess of $42 kg/m^2$ is still associated with increased mortality. However, these data should be interpreted cautiously, given the small number of subjects and wide heterogeneity represented by the wide confidence intervals. The stark contrast between the lack of correlation between all-cause mortality and BMI categories defined as normal, overweight, and obesity, and the strong association between the baseline comorbidities and all-cause mortality have significant clinical relevance. Obesity is an important risk factor for many chronic and common clinical conditions; however, once the conditions are overt, the mortality risk seems to be more closely correlated with the prevalence and management of these conditions than with the body weight except for extreme values. Obesity, measured by BMI,

is a readily available clinical indicator used across the health care system triggering additional clinical interventions. However, our data showed that the BMI distribution is very similar among all the risk groups, which should reinforce the premise to screen and manage comorbidities regardless of the BMI. The obesity paradox has been observed and reported in several common conditions including diabetes, coronary artery disease, hypertension, etc., for which BMI is an important known risk factor [114]. The cause of these observations requires an additional evaluation that needs to be specific to each condition, considering not only additional comorbidities but also temporal metabolic and therapeutic changes that could potentially lend a survival advantage, as long as the BMI is not extreme. The purpose of this work is neither to disprove the existence of such an effect nor to dismiss BMI as a clinical risk factor, but to convey the need for aggressive management of comorbidities regardless of BMI. Patients with low BMI may look healthier, which in turn may lead to being less aggressive in screening for or treating comorbidities. For example, high BMI has been identified as an independent predictor of recognition and management of prediabetes while low BMI predicted lack of recognition [115]. The development of new comorbidities imparts a far greater impact on mortality risk than the variation of BMI; thus, prevention, recognition, and management of comorbidities should be of high priority, regardless of BMI. Our findings summarize a complex relationship between BMI and all-cause mortality and may not solve the current controversy. However, our findings are in keeping with the results of many other studies that describe this complex relationship that cannot be attributed to any single statistical bias or limitation of the BMI as a definition of obesity. We must become critical of the current definitions of ideal body mass and what is or is not abnormal or related to health risks. Only by reevaluating our current premises, we will be able to find evidence (no expert opinions) that can be implemented in daily clinical practice.

considered in our study. The data collection was done using an electronic search which allowed the use of a large data set but inevitably decreases the accuracy and details of individual variables. One relevant variable that was a casualty of our electronic searches was smoking status. Smoking is associated to lower BMI and higher mortality and it is well known as a risk factor for several comorbidities including cancer, cardiovascular and lung diseases, which were included in our prediction model. Unfortunately,

smoking data are usually recorded as text in the clinical notes but lack standardization and reliability. The limited structured data available in the medical record were contradictory and found to be too inaccurate to be used in the model. However, for other relevant variables, we used data collected during routine clinical care and stored in the EHR and other data repositories intended for research. Still, our study uses data from a defined cohort seeking care in one specific medical center which may limit generalizability and our findings should be subjected to validation. This is not a population-based study and likely healthy individuals are underrepresented. Our results should be used in the context of clinical care, especially in subjects with comorbidities. In conclusion, our study shows that baseline comorbidities, through the risk they confer on patients, are the primary driver of all-cause mortality, while BMI only plays a secondary role. In general, across all baseline comorbidity-based risk groups, the association between BMI and all-cause mortality displayed U-shaped distribution. The nadir was a plateau including what is defined as normal, overweight, and obesity with variations across the different risk groups. Patients with extreme BMIs within each risk group showed an increased risk of mortality, but the variation of mortality risk within each group was smaller than between groups. This suggests that patients should be treated aggressively to prevent or manage comorbidities, regardless of BMI.

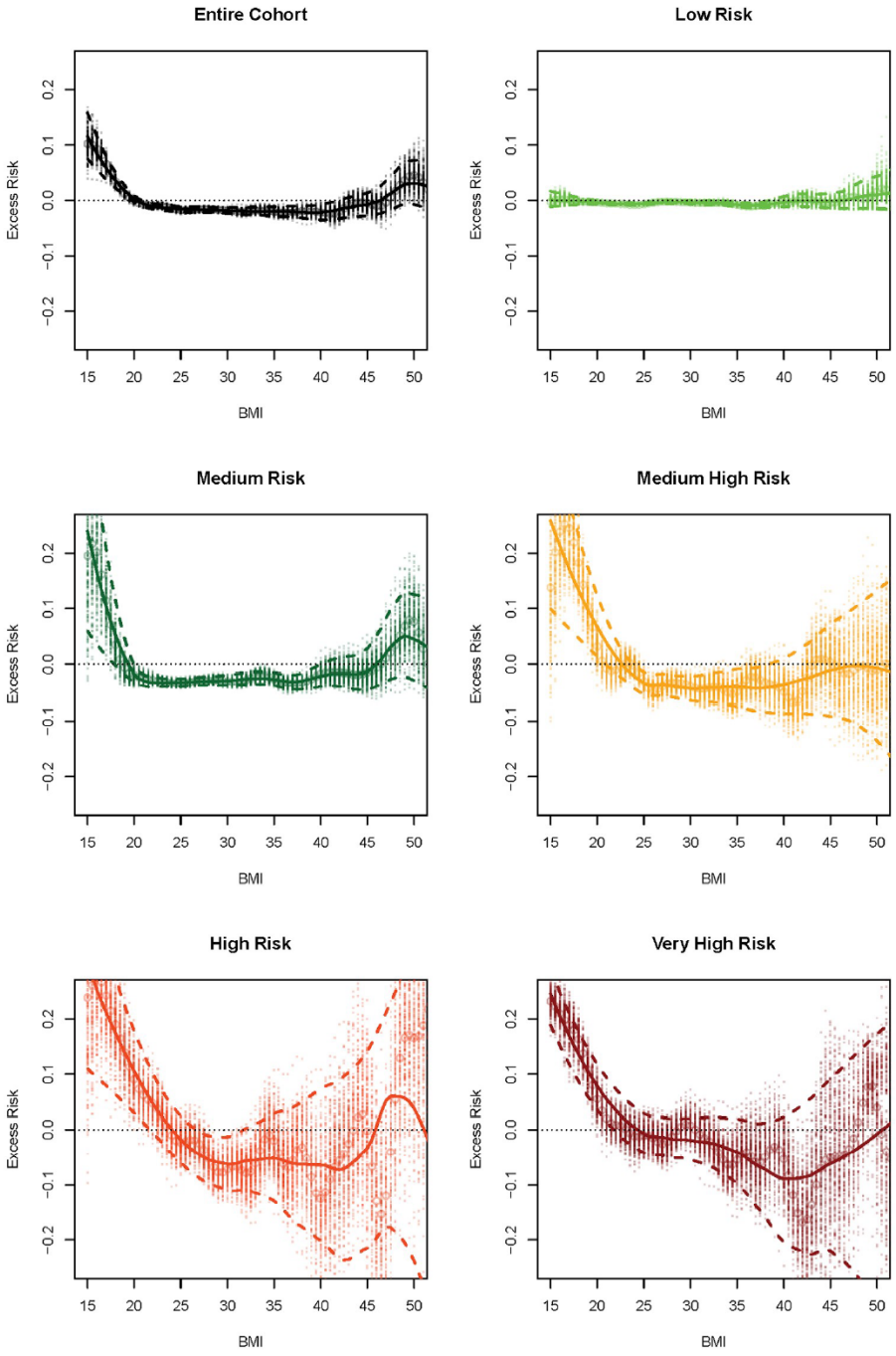


Figure 5.5: Excess risk of all-cause mortality attributed to the BMI not explained by the baseline comorbidities (residual risk) after adjusting for baseline mortality risk. The residual risk was estimated for each $0.5 Kg/m^2$ BMI interval between 15 and 55. The dots represent the point estimates of the mortality rate in the 200 bootstrap iterations, circles are the mean, the solid line is a smooth curve representing the averages and the dashed lines are the 95% empirical confidence interval. Values above zero represent increased risk and values below zero represent decreased risk compared to the group average.

Chapter 6

Conclusions and Future Work

This thesis introduced three causality-related research works I finished sequentially in my program. It was a long process to discover the existence and then investigate the influences of Causal Representation Bias from EHR data, an original idea that I refined from some commonly encountered modeling biases during causal learning.

Initially, this bias appeared as "incomplete causal information". After encountering this phenomenon multiple times, a question lingered: since all information is learned from the observations, including the complementary information I successfully figured out, why didn't the models fully learn it in one-pass modeling, but instead require side processes to complete it? It was hard to find the answer because what I was looking for is supposed to be irrelevant to model selection, but it can only be reflected in some causal modeling practices. However, the causal modeling biases are usually complex and one can hardly identify certain reasons that lead to these biases. For example, some bias may come from inappropriate model parameter settings, but some have been created in the model selection process, or even earlier, in the data sampling process. While my purpose is to identify the bias inherently hidden in the data that was caused by the referred information gap.

This question motivated me to complete the deep learning modeling work introduced in Chapter 3. Since deep learning doesn't require model specification, it can help to reduce empirical biases as much as possible and may eventually expose the inherent biases contained in data. The results of this work turned out to be very illustrative: The "causal information shortage" still exists, and dealing with that issue allowed me

to improve deep learning's performance significantly.

With further research on causal inference theory, I noticed that there often exists a "mismatch" between what the data can show and what people have known, especially in the domain sciences that have many years of development. Practically, in my works, complementary information usually comes from domain knowledge rather than observations. Accordingly, my works were essentially an undertaking to transform such information into a form utilizable by models.

On the other hand, the phenomenon of "causal information redundancy" could possibly exist, where some observed variables are assumed to represent certain factors as a convention, but in practice, they may contain additional factors beyond the desired factors only. Such extraneous information can eventually be shown to be modeling selection bias, i.e., the current model cannot selectively utilize only the desired information and eliminate the extra information. We used to pay major attention to the modeling process and thus may have neglected that such a bias is from the data, and that bias in data could accumulate in the subsequent processes.

In short, such a bias can be described as "How well can the data variables represent the causal knowledge?" Or in other words, it is about the degree of representativeness of the data variables. Although this bias has nothing to do with the modeling process, it can become very noticeable in the form of modeling biases, or empirical violations of causal assumptions.

Indeed, the existence of the Causal Representation Bias may be much more common than previously thought. In practical applications, many works for improvement are essentially trying to address this bias, but doing so on a case-by-case, and field-by-field basis. For example, leveraging knowledge works as an adjustment for this bias, to help models utilize knowledge better; and interpretable modeling transfers data into a form as close to existing knowledge as possible.

It is instinctive to build up models on observed data variables because there seems to be no other choice. But nowadays, deep learning techniques provide the possibility of creating variables that are what we want, rather than what we see, but it raises another big challenge: the created variable is no longer directly interpretable by us. What we can only know from deep learning is that it has learned the information per our requirements. Such a variable is more properly called a representation, and the

corresponding entity is a vector in feature space without the need to be explainable.

In simple terms, the deep learning-based representation extraction method has the promise of realizing the true "causal variables", which are expected to be much purer representations than the data variables. However, I found an interesting truth that the idea of "causal representation" has actually been used by people since a long time ago [7, 116], even before the birth of deep learning.

In the work of Chapter 3, we defined "disease severity" as a variable S, which is an abstract conception borrowed from classical healthcare informatics. It is a virtual entity that represents all information needed to describe the patient's illness, which helps people to picture the underlying causal graph more easily, just like the one shown in Figure 3.2. And accordingly, the proposed estimator Δ can be seen as a realization of some specific part of S. Fortunately, we found a way of numerically expressing S partially by referring to the observed variable X in this particular work. In general, the true causal variables may still remain abstract descriptions. For example, in the work of Chapter 5, "reason of mortality" is the variable(s) that physicians eventually aim to describe. Still, for now, it can only be a profile existent in a statement rather than explicit models.

Since we have raised the causal representation as a generic conception, can we find a general realization? The deep learning-based representation can satisfy our need to make every representation vector identifiable in the feature space without specifying models, and deep learning is good at handling nonlinear relations. As for model interpretation, we can still rely on conventional models, except that the variables they model could come from data variables reconstructed by deep learning from the causal representation space, instead of the observational data only. Therefore, I was motivated in the direction of establish a new causal-learning framework to realize these ideas (as shown in Figure 1.4).

As mentioned in section 1.3, the DL-based general realization is in my ongoing work and is expected to be in progress for years. Recently, the biggest development is that the most significant technical challenge, a reversible neural network architecture for representation extraction and recovery, has been basically completed. Based on that, I completely realized a causal discovery process in the representation feature space, and also built up the corresponding SEM model on a demo dataset. At the same time, I am still working on the demo of CRB adjustment on representations. However, even

with the core functions, there is still a long way to go in improving the technique and exploring its applications. This is impossible for a single individual but will require a lot of teamwork. For now, the most promising application is to provide non-model-based de-confounded causal data simulation as complementary support to existing causal inference.

From the perspective of machine learning, I highly expect that causal representation learning would start a new direction for causality research, where more and more works will be produced in the future that make substantial contributions, to improve the robustness and interpretability of causal modeling, as well as rendering immense support for the development of domain knowledge.

References

[1] Christopher J Wood and Robert W Spekkens. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of bell-inequality violations require fine-tuning. *New Journal of Physics*, 17(3):033002, 2015.

[2] Matej Vuković and Stefan Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

[3] Mohammed Ombadi, Phu Nguyen, Soroosh Sorooshian, and Kuo-lin Hsu. Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resources Research*, 56(7):e2020WR027251, 2020.

[4] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[5] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[6] Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427, 2020.

[7] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[8] Jia Li, Xiang Li, Xiaowei Jia, Michael Steinbach, and Vipin Kumar. Realization of causal representation learning to adjust confounding bias in latent space. *arXiv preprint arXiv:2211.08573*, 2022.

[9] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33:9180–9192, 2020.

[10] Richard Scheines. An introduction to causal inference. 1997.

[11] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

[13] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[14] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020.

[15] Pearl Judea. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.

[16] Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan S Read, Jacob A Zwart, Michael Steinbach, and Vipin Kumar. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3):1–26, 2021.

[17] Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics*, 52(7):740–747, 2020.

[18] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[19] Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

[20] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016.

[21] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117(44):27162–27170, 2020.

[22] Adityanarayanan Radhakrishnan, Karren Yang, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. *arXiv preprint arXiv:1810.10333*, 2018.

[23] Saachi Jain, Adityanarayanan Radhakrishnan, and Caroline Uhler. A mechanism for producing aligned latent spaces with autoencoders. *arXiv preprint arXiv:2106.15456*, 2021.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[26] Anastasiya Belyaeva, Louis Cammarata, Adityanarayanan Radhakrishnan, Chandler Squires, Karren Dai Yang, GV Shivashankar, and Caroline Uhler. Causal network models of sars-cov-2 expression and aging to identify candidates for drug repurposing. *Nature communications*, 12(1):1024, 2021.

[27] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[28] Xiao Xie, Fan Du, and Yingcai Wu. A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1448–1458, 2020.

[29] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016.

[30] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1(2665):2012, 2012.

[31] Judea Pearl. *Causality*. Cambridge university press, 2009.

[32] Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

[33] Miguel A Hernán and James M Robins. Causal inference, 2010.

[34] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

[35] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.

[36] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.

[37] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pages 39–80. Springer, 2018.

[38] Andrew R Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *arXiv preprint arXiv:2104.08043*, 2021.

[39] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[40] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.

[41] Daniel Scharfstein, Aidan McDermott, William Olson, and Frank Wiegand. Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Statistics in Biopharmaceutical Research*, 6(4):338–348, 2014.

[42] AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.

[43] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

[44] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[45] 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(1):1–10, 2016.

[46] Nilmini Wickramasinghe. Deepr: a convolutional net for medical records. 2017.

[47] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[48] 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*, pages 1495–1504, 2016.

[49] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 30–41. Springer, 2016.

[50] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 93–101. IEEE, 2016.

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

[52] Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *arXiv preprint arXiv:1608.05745*, 2016.

[53] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

[54] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[55] Laila Rasmy, Yonghui Wu, Ningtao Wang, Xin Geng, W Jim Zheng, Fei Wang, Hulin Wu, Hua Xu, and Degui Zhi. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84:11–16, 2018.

[56] 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*, 24(2):361–370, 2017.

[57] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[58] Jonathan D Young, Bryan Andrews, Gregory F Cooper, and Xinghua Lu. Learning latent causal structures with a redundant input neural network. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 62–91. PMLR, 2020.

[59] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR, 2019.

[60] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*, 2020.

[61] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

[62] Hans-Peter Deigner and Matthias Kohl. *Precision medicine: tools and quantitative approaches*. Academic Press, 2018.

[63] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[64] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[65] Geert Molenberghs and Michael Kenward. *Missing data in clinical studies*, volume 61. John Wiley & Sons, 2007.

[66] National Research Council et al. *The prevention and treatment of missing data in clinical trials*. National Academies Press, 2011.

[67] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[68] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[69] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehr): A survey. *Department of Computer Science and Engineering*, 2015.

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

[71] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[72] Donald B Rubin and Roderick JA Little. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*, 2002.

[73] James R Carpenter, Michael G Kenward, and Ian R White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical methods in medical research*, 16(3):259–275, 2007.

[74] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010.

[75] Shichao Zhang, Zhenxing Qin, Charles X Ling, and Shengli Sheng. " missing is useful": missing values in cost-sensitive decision trees. *IEEE transactions on knowledge and data engineering*, 17(12):1689–1693, 2005.

[76] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[77] Yang C Yuan. Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49:1–11, 2010.

[78] Donald Hedeker and Robert D Gibbons. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods*, 2(1):64, 1997.

[79] Andrea Rotnitzky. Inverse probability weighted methods. *Longitudinal data analysis*, pages 453–476, 2009.

[80] Jean Mundahl Engels and Paula Diehr. Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, 56(10):968–976, 2003.

[81] Alexina Mason, Nicky Best, Sylvia Richardson, and IAN PLEWIS. Strategy for modelling non-random missing data mechanisms in observational studies using bayesian methods. 2010.

[82] Katherine M Flegal, Brian K Kit, Heather Orpana, and Barry I Graubard. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *Jama*, 309(1):71–82, 2013.

[83] Emanuele Di Angelantonio, Shilpa N Bhupathiraju, David Wormser, Pei Gao, Stephen Kaptoge, Amy Berrington De Gonzalez, Benjamin J Cairns, Rachel Huxley, Chandra L Jackson, Grace Joshy, et al. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *The Lancet*, 388(10046):776–786, 2016.

[84] Jufen Zhang, Aine Begley, Ruth Jackson, Michael Harrison, Pierpaolo Pellicori, Andrew L Clark, and John GF Cleland. Body mass index and all-cause mortality

in heart failure patients with normal and reduced ventricular ejection fraction: a dose–response meta-analysis. *Clinical Research in Cardiology*, 108(2):119–132, 2019.

[85] JE Winter, RJ MacInnis, and CA Nowson. The influence of age on the bmi and all-cause mortality association: A meta-analysis. *The journal of nutrition, health & aging*, 21(10):1254–1258, 2017.

[86] Abhishek Sharma, Carl J Lavie, Jeffrey S Borer, Ajay Vallakati, Sunny Goel, Francisco Lopez-Jimenez, Armin Arbab-Zadeh, Debabrata Mukherjee, and Jason M Lazar. Meta-analysis of the relation of body mass index to all-cause and cardiovascular mortality and hospitalization in patients with chronic heart failure. *The American journal of cardiology*, 115(10):1428–1434, 2015.

[87] Xin Liu, Lixia Xie, Wengen Zhu, and Yue Zhou. Association of body mass index and all-cause mortality in patients after cardiac surgery: A dose-response meta-analysis. *Nutrition*, 72:110696, 2020.

[88] Dagfinn Aune, Abhijit Sen, Manya Prasad, Teresa Norat, Imre Janszky, Serena Tonstad, Pål Romundstad, and Lars J Vatten. Bmi and all cause mortality: systematic review and non-linear dose-response meta-analysis of 230 cohort studies with 3.74 million deaths among 30.3 million participants. *bmj*, 353, 2016.

[89] Francesco Zaccardi, Nafeesa N Dhalwani, Dimitris Papamargaritis, David R Webb, Gavin J Murphy, Melanie J Davies, and Kamlesh Khunti. Nonlinear association of bmi with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies. *Diabetologia*, 60(2):240–248, 2017.

[90] Erwin Fleischmann, Nancy Teal, John Dudley, Warren May, John D Bower, and Abdulla K Salahudeen. Influence of excess weight on mortality and hospital stay in 1346 hemodialysis patients. *Kidney international*, 55(4):1560–1567, 1999.

[91] Xiao Liu, Linjuan Guo, Kaiwen Xiao, Wengen Zhu, Menglu Liu, Rong Wan, and Kui Hong. The obesity paradox for outcomes in atrial fibrillation: Evidence from

an exposure-effect analysis of prospective studies. *Obesity Reviews*, 21(3):e12970, 2020.

[92] Reka Hidvegi, Christian Puelacher, Danielle M Gualandro, Andreas Lampart, Giovanna Lurati Buse, Angelika Hammerer-Lerchner, Joan Walter, Marcel Liffert, Daniel Bolliger, Luzius Steiner, et al. Obesity paradox and perioperative myocardial infarction/injury in non-cardiac surgery. *Clinical Research in Cardiology*, 109(9):1140–1147, 2020.

[93] JE Dvorak, ELW Lester, PJ Maluso, L Tatebe, V Schlanser, M Kaminsky, T Messer, AJ Dennis, F Starr, and F Bokhari. The obesity paradox in the trauma patient: normal may not be better. *World journal of surgery*, 44(6):1817–1823, 2020.

[94] Audrey De Jong, Boris Jung, Gérald Chanques, Samir Jaber, and Nicolas Molinari. Obesity and mortality in critically ill patients: another case of the simpson paradox? *Chest*, 141(6):1637–1638, 2012.

[95] Jonathan Y Xia, Donald M Lloyd-Jones, and Sadiya S Khan. Association of body mass index with mortality in cardiovascular disease: new insights into the obesity paradox from multiple perspectives. *Trends in cardiovascular medicine*, 29(4):220–225, 2019.

[96] Gjin Ndrepepa and Adnan Kastrati. Obesity paradox after percutaneous coronary intervention—closing in on the truth behind the phenomenon. *EuroIntervention*, 15(13):1120–2, 2020.

[97] Marijana Tadic and Cesare Cuspidi. Obesity and heart failure with preserved ejection fraction: a paradox or something else? *Heart failure reviews*, 24(3):379–385, 2019.

[98] Reza Mohebi, Amirali Simforoosh, Maryam Tohidi, Fereidoun Azizi, and Farzad Hadaegh. Obesity paradox and risk of mortality events in chronic kidney disease patients: a decade of follow-up in tehran lipid and glucose study. *Journal of Renal Nutrition*, 25(4):345–350, 2015.

[99] Ramzi Vareldzis, Mihran Naljayan, and Efrain Reisin. The incidence and pathophysiology of the obesity paradox: should peritoneal dialysis and kidney transplant be offered to patients with obesity and end-stage renal disease? *Current hypertension reports*, 20(10):1–8, 2018.

[100] Beata Ujvari, Camille Jacqueline, Dorothée Missé, Valentin Amar, Jay C Fitzpatrick, Geordie Jennings, Christa Beckmann, Sophie Rome, Peter A Biro, Robert Gatenby, et al. Obesity paradox in cancer: Is bigger really better? *Evolutionary applications*, 12(6):1092–1095, 2019.

[101] Dong Hoon Lee and Edward L Giovannucci. The obesity paradox in cancer: epidemiologic insights and perspectives. *Current nutrition reports*, 8(3):175–181, 2019.

[102] Emilio Rodríguez-Castro, Manuel Rodríguez-Yáñez, Susana Arias-Rivas, María Santamaría-Cadavid, Iria López-Dequidt, Pablo Hervella, Miguel López, Francisco Campos, Tomás Sobrino, and José Castillo. Obesity paradox in ischemic stroke: clinical and molecular insights. *Translational stroke research*, 10(6):639–649, 2019.

[103] Haydn Hoffman, Muhammad S Jalal, Taylor Furst, and Lawrence S Chin. The obesity paradox in spontaneous intracerebral hemorrhage: results from a retrospective analysis of the nationwide inpatient sample. *Neurocritical care*, 32(3):765–774, 2020.

[104] Joshua F Baker, Erica Billig, Kaleb Michaud, Said Ibrahim, Liron Caplan, Grant W Cannon, Andrew Stokes, Vikas Majithia, and Ted R Mikuls. Weight loss, the obesity paradox, and the risk of death in rheumatoid arthritis. *Arthritis & Rheumatology*, 67(7):1711–1717, 2015.

[105] Christina-Maria Kastorini and Demosthenes B Panagiotakos. The obesity paradox: methodological considerations based on epidemiological and clinical evidence—new insights. *Maturitas*, 72(3):220–224, 2012.

[106] National Heart, Lung, Blood Institute, and North American Association for the Study of Obesity. *The practical guide: identification, evaluation, and treatment of overweight and obesity in adults*. The Institute, 2000.

[107] David Berrigan, Richard P Troiano, and Barry I Graubard. Bmi and mortality: the limits of epidemiological evidence. *The Lancet*, 388(10046):734–736, 2016.

[108] Lara L Roberson, Ehimen C Aneni, Wasim Maziak, Arthur Agatston, Theodore Feldman, Maribeth Rouseff, Thinh Tran, Michael J Blaha, Raul D Santos, Andrei Sposito, et al. Beyond bmi: The "metabolically healthy obese" phenotype & its association with clinical/subclinical cardiovascular disease and all-cause mortality–a systematic review. *BMC public health*, 14(1):1–12, 2014.

[109] Aviva Must, Jennifer Spadano, Eugenie H Coakley, Alison E Field, Graham Colditz, and William H Dietz. The disease burden associated with overweight and obesity. *Jama*, 282(16):1523–1529, 1999.

[110] 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. In *Mayo Clinic proceedings*, volume 87, pages 1202–1213. Elsevier, 2012.

[111] Jennifer L St Sauver, Brandon R Grossardt, Barbara P Yawn, L Joseph Melton III, 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.

[112] Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, pages 8–27, 1998.

[113] Centers for Disease Control, Prevention, et al. Overweight and obesity. centers for disease control and prevention website, 2010.

[114] Seth Uretsky, Franz H Messerli, Sripal Bangalore, Annette Champion, Rhonda M Cooper-DeHoff, Qian Zhou, and Carl J Pepine. Obesity paradox in patients with hypertension and coronary artery disease. *The American journal of medicine*, 120(10):863–870, 2007.

[115] et al. Njeru, Jane W. Clinical recognition and management of patients with prediabetes. *Endocrine Practice 25.6 (2019): 545-553.*, 2019.

[116] Thomas A Medsger, Stefano Bombardieri, László Czirják, Raffaella Scorza, A Della Rossa, and Walter Bencivelli. Assessment of disease severity and prognosis. *Clinical and experimental rheumatology*, 21(3; SUPP/29):S42–S46, 2003.