ADVANCING CAUSAL ANALYTICS USING BIOMEDICAL DATA

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

Xinpeng Shen

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

Dr. Gyorgy J. Simon, Advisor
Dr. Sisi Ma, Advisor

January 2024
Acknowledgements

I would like to express my deepest gratitude to my advisors, Professor Gyorgy Simon and Professor Sisi Ma. Their invaluable guidance, patience, and expertise have guided me throughout my research journey. Without their support, I cannot imagine having reached this stage of my academic career. I am also deeply thankful to Professor Prashanthi Vemuri for her insightful suggestions and advice during my Ph.D. studies. Their passion for sciences has been truly inspiring and I have grown tremendously.

Besides this, I also want to thank my committee members, Professor Lisiane Pruinelli and Professor Steven Johnson. Their assistance during my preliminary exam and dissertation has been invaluable. The peers and colleagues at the Institute for Health Informatics at University of Minnesota has been instrumental in my success. The welcoming and supportive environment of the department has significantly enriched my experience.

In the end, I would also like to express my appreciation to my family and friends. Their endless support and encouragement have been my strength throughout this journey. Studying abroad presents its own set of challenges, and it was their unwavering belief in me that made this possible.

This journey would not have been possible without each every one of you.
Abstract

Understanding the distinction between association and causality is crucial in the fields of health informatics and biomedicine, as causality allows for the modeling of manipulable relationships. Various statistical methods have been developed as alternatives to randomized clinical trials, which often are impractical due to ethical or cost considerations. The process of causal analysis typically involves two steps: causal structure discovery and causal effect estimation (also known as causal inference). The process of extracting causal structure from observational data, known as computational causal structure discovery (CSD), is an emerging field that has garnered considerable attention in recent years. Once the causal structure is known, or partially understood, the estimation of specific causal effects can be undertaken using causal inference (CI) methods.

As vast biomedical data repositories continue to emerge, understanding how to effectively process causal structure discovery and causal effect estimation has become crucial for better utilization of observational biomedical data. This thesis aims to contribute by exploring existing causal discovery methods and developing new methodologies to address practical challenges in discovering causal relationships from biomedical datasets. The work is composed of four chapters that explore the use of existing methodologies to discover causal relationships among biomarkers related to Alzheimer's disease (AD), estimate causal effects related to Alzheimer's disease, and propose two methods to address data challenges inherent in electronic health records (EHRs) datasets.
Ultimately, the research presented in this thesis offers practical examples of applying CSD and causal inference methods to address biomedical problems. It also proposes two novel methods to navigate prevalent data challenges, which are crucial for effectively utilizing EHR data and extracting meaningful causal relationships.
Table of Contents

Acknowledgements ........................................................................................................... i

Abstract ............................................................................................................................ ii

List of Tables .................................................................................................................... v

List of Figures .................................................................................................................. vi

Introduction ....................................................................................................................... 1

1. Chapter I: Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer’s Pathophysiology .................................................................................. 11

2. Chapter II: Causal structure discovery identifies risk factors and early brain markers related to evolution of white matter hyperintensities ............................................................................ 36


Discussion and Conclusion ............................................................................................... 103

Reference .......................................................................................................................... 110
List of Tables

Table 1.1 Characteristics for Continuous and Categorical Variables........................................ 19

Table 1.2 Recovery rate of edges....................................................................................................... 30

Table 1.3 Typical problems and solutions.......................................................................................... 32

Table 2.1 Participants baseline characteristics table. Mean (SD) and Count (%) for continuous and categorical variables, respectively................................................................. 45

Table 2.2 Results from two regression (baseline) and two mixed models (progression) on WMH and WM health, Genu-FA ........................................................................................................... 48

Table 2.3 Results from two standardized regression models on WMH with and without controlling for WM health, Genu-FA .......................................................................................................................... 51

Table 3.1 Characteristics of the MC and FV data sets. ........................................................................ 75

Table 3.2 Directional stability. ................................................................................................................ 76

Table 3.3 Metrics from clinical evidence ............................................................................................. 77

Table 3.4 External Validation. .................................................................................................................. 78

Table 4.1 Characteristics of participants. Mean (SD) and Count (%) for continuous and categorical variables ................................................................................................................... 91
List of Figures

Figure 1.1 The interpretation of edges in CSD........................................................................................................... 18
Figure 1.2 The “gold standard” graph.......................................................................................................................... 21
Figure 1.3 Discovered causal structure without background knowledge & their Statistics.......................... 24
Figure 1.4 Discovered causal structure with background knowledge & their Statistics.............................. 26
Figure 1.5 Discovered causal structure with longitudinal data & their Statistics........................................... 28
Figure 2.1 Graphical abstract .............................................................................................................................................. 38
Figure 2.2 Association between baseline WM hyperintensity with (A) age, (B) sex, and (C) hypertension.......................................................................................................................................................... 47
Figure 2.3 Association between baseline WM health measured by Genu-FA with (A) age, (B) sex, and (C) hypertension.......................................................................................................................................................... 49
Figure 2.4 A Relationship between baseline WMH and baseline Genu-FA in the pre-defined age groups. B shows the estimated local causal structure among age, sex, hypertension, WM health, and WMH, generated by the CSD algorithm. .......................................................................................................................................................... 50
Figure 2.5 The causal structure graph discovered using the CSD algorithm. The estimated standardized coefficients were labeled next to each edge........................................................................................................ 52
Figure 2.6 Important Relationships between midlife modifiers and WMH/WM health discovered by the CSD algorithm. .......................................................................................................................................................... 58
Figure 3.1 Study design and evaluations ..................................................................................................................... 67
Figure 4.1 The structure of the latent and observed variables.................................................................................... 88
Figure 4.2 Root Mean Square Error (RMSE) of the estimated latent effects.......................................................... 94
Figure 4.3 Root Mean Square Error (RMSE) and insignificance across various sample sizes........ 96

Figure 4.4 The EHR study. A) the assumed causal structure of the EHR latent and observed
variables. B) the RMSE of the proposed and SEM based method............................................ 98
Introduction

With the development of machine learning, there has been a great interest towards applying these advanced modeling techniques for addressing health or biomedical challenges. Machine learning models have demonstrated considerable success across various healthcare domains, with notable applications in translational bioinformatics[1 2] and image-based diagnostics[3-5]. Despite the promising results, the practical integration of AI technics into clinical practice has remained challenging[6]. Machine learning algorithms are known to have various issues when adapted to the real-world applications. These challenges include interpretability of the results from black-box models, prediction bias, fairness, and robustness. Sophisticated methods, such as deep learning, typically are less likely to underfit but suffer more from the lack of robustness and interpretability.

Regardless of their unique architectures and specific objectives, all machine learning algorithms are fundamentally built upon assumptions. Consider the case of linear regression which is commonly utilized for effect estimation. These models rely on four fundamental assumptions: linearity, homoscedasticity, independence, and normal distribution of the residuals. When any of these underlying assumptions is violated, the conclusion drawn from the model can be misleading. One of the most prevalent assumptions that is often violated in health informatics or biomedical informatics is ignoring potential confounders. **Confounder** (denoted as \( C \)) is defined as the variable that affects the outcome \( Y \) of interest and another predictor \( X \) (variables) and \( C \) is not on the causal path between \( X \) and \( Y \). Ignoring such confounder \( C \), estimating the treatment effect of \( X \) on the outcome \( Y \), can lead to a considerable bias in the treatment effect estimation.
In addition to the confounder issue, other significant considerations often remain unaddressed. Consider a scenario where all variables, including potential confounders, are observed, and we correctly deal with the confounder issue as mentioned before. Even in this context, if a variable $D$ is dependent on the outcome $Y$, the estimated treatment effect of $X$ on $Y$ can still be biased. When fitting a causal inference model with all variables as predictors, including $D$, it could yield a superior prediction accuracy compared to a modeling that excluding $D$. However, this inclusion can lead to a significant bias in the estimated treatment of $X$ on $Y$. 

The examples discussed in the previous paragraph highlight the critical importance of understanding the data generation process, particularly the causal relationships among variables, and the necessity of approaching problems from a causal perspective[7 8]. In the rest of this section, I will first introduce the concept of causality. Then, I will discuss the two important tasks in the field of causality: causal inference (CI) and causal structure discovery (CSD). Finally, I will expand on causal structure discovery, an area that has not been as extensively explored as causal inference. In this section, I will introduce key methods within this domain and pinpoint the knowledge gaps that my thesis seeks to fill.

Judea Pearl, the 2011 Turing award winner, has made significant contributions to causal analysis with a framework based on mathematics and probability theory[9]. In brief, it defines that a causal effect from $X$ to $Y$ exists if manipulating $X$ directly changes $Y$ accordingly, but intervening on the value of $Y$ does not change the value of $X$. This causality framework, also known as do-calculus, enables a rigorous analysis within the realm of causality. Building upon these foundational concepts, Pearl and his colleagues further proposed the idea of “Ladder of
Causation”[10] or “Pearl Causal Hierarchy” (PCH)[11]. These concepts systematically distinguish between conditional probability, causality, and counterfactual effects[12], providing a clear framework for understanding of causal relationships.

The PCH consists of three layers of causal hierarchy: association, intervention, and counterfactuals. (1) **Association.** As defined by Pearl, *associational* relationships are directly observable from plain data through conditional independence. This layer is the foundation for most modern machine learning methods, which are rooted in probability theory and focus on estimating the conditional relationships between variables, despite their varied architectures and objective functions. (2) **Intervention.** The second layer in the Pearl Causal Hierarchy concerns interventional relationships, which refers to causality. This effect is defined by the action of manipulation or “doing”, observing the changes in $Y$ in response to changes in $X$. It is important to note that, as proved by[11], corollary 1, one cannot answer higher layer questions only using the knowledge from a lower layer. In other words, merely given the observation datasets, without knowing the underlying causal structure, any inference on layer 2, *interventional* is invalid. However, it is possible to infer the association from either observational data (layer 1), as well as layer 2 *interventional* data. (3) **Counterfactuals.** The third and final layer of PCH is *counterfactual* relationships. As implied by its name, this layer focuses on estimating effects that go counter to the facts. It involves speculating how the outcome variable $Y$ would have responded if the predictor $X$ had assumed a different value than what was actually observed. Counterfactual effect holds significant practical relevance, for instance, in scenarios such as medical treatment evaluations. An example could be assessing whether a patient who survived after taking medication $A$ would have also survived had they taken medication $B$ instead. This
concept is integral to precision medicine or individualized treatment effect estimation, where the outcomes of different treatment options are compared to find the most beneficial one for each patient. Notably, in reality, only one outcome for each patient is observable, so that estimating counterfactual effect is extremely challenging[13].

In the past few decades, modern machine learning has been focusing on accurately learning associations. Starting from simple linear regression to deep neural networks, algorithms have become more capable of learning the non-linear relationships under various challenging scenarios, such as high-dimensionality, unstructured data (language models), small sample size, unsupervised learning context. Despite these advancements, the relationships these algorithms learn remain confined to the first layer of the PCH. Consequently, interpreting the result with causal context is invalid and misleading. In some applications, the causal interpretation may not be necessary. High predictive accuracy alone can be sufficient to support the transition from research to real-world applications, as evidenced in fields like image classification and autonomous driving systems. However, in the domain of health informatics, convincing healthcare professionals and patients to trust AI systems requires moving beyond mere associative relationships to making substantiated causal conclusions.

Researchers have increasingly focused on bridging the gap between associative knowledge and causality. A significant area is causal inference. Common methods include direct adjustment, inverse probability weighting[14], and propensity matching[15]. Unlike randomized clinical trials (RCTs), where confounding variables are adjusted during the random assignment of participants to control and treatment group, causal inference methods adjust for the confounders
by learning the confounding relationships from the observational data. Causal inference methods are increasingly being recognized as alternatives to RCTs for treatment effect estimation.[15-18] However, one important assumption that causal inference methods make is that the given causal structure is (partially) known, i.e., all potential confounders of the treatment assignment and the outcome are included in the model.

Another direction toward causality is CSD[19 20]. Unlike causal inference which focuses on quantifying specific effect sizes, CSD methods employ statistical techniques to infer the parameters defining causal relationships among variables. Additionally, there are specialized causal methods that build upon do-calculus[21], as well as methods that use observational data and experimental data[22] to learn the causal relationships. One can also think of CSD methods as a preliminary step of causal inference, since all causal inference methods rely on an underlying assumption of a pre-existing causal structure. A causal structure refers to a set of causal relationships among a set of variables. This is distinct from Bayesian network[23], relationships might not necessarily be causal. To distinguish the two terms, causal structure is also often called Causal Bayesian Network (CBN).

As proven in PCH, it is not possible to extract layer 2, interventional knowledge from layer 1 data (observational data). To address this challenge, CSD methods impose a set of fundamental assumptions. Common assumptions shared by most algorithms include the Markov condition, the causal faithfulness condition, and the causal sufficiency condition[19]. Some CSD methods relax some of the assumptions, which potentially enhancing flexibility or applicability, while often results in increased complexity in both computational processes and interpretational clarity.
Generally, CSD methods are classified into two main categories: constraint-based and score-based methods. The PC algorithm[19] is one of the most representative constraint-based methods. The discovery or search procedure can be separated into two phases. In the first phase, it starts with a fully connected undirected graph where each variable is presumed to potentially cause or be caused by every other variable. The algorithm then proceeds to eliminate edges based on (conditional) independence tests. For example, if \( X \) and \( Y \) are independent judged by a pre-defined threshold, PC algorithm removes the edge that connects \( X \) and \( Y \). The independence rules out the possibility of a direct causal relationship between them. The search stops when no further edges can be removed. The output of the first phase is a graph with undirected edges (it is called the skeleton). The second phase of PC orients the edges based on the idea of a collider: when variables \( X, Y, \) and \( Z \) are connected with the following pattern, \( X-Y-Z \), and \( X, Z \) are not directly connected, PC orients the relationship as \( X \rightarrow Y \leftarrow Z \) if conditional on \( Y \), \( X \) and \( Z \) are conditionally dependent. \( Y \) is the collider, \( X, Y, \) and \( Z \) form a “V” shape structure.

One important assumption that PC algorithm (and many other CSD algorithms) makes is that there is no unobserved confounder. To recall, a confounder \( C \) is defined as the variable that affect both treatment variable \( X \) and the outcome variable \( Y \). However, it is not always the case. Another constraint-based CSD method, Fast Causal Inference (FCI)[24] algorithm relaxes the assumption of no unobserved confounder. The structure searching procedure of FCI is similar to PC, but the orientating phase and the resulting graph of FCI is different. To relax the no unobserved confounder assumption, FCI introduces a new type of orientation structure, the “Y” structure. Given four variables \( W_1, W_2, X, \) and \( Y \), a “Y” structure is formed as \( W_1 \rightarrow X \leftarrow W_2, \)
and $X \rightarrow Y$. Within this “Y” structure, the oriented edge between $X$ and $Y$ is guaranteed to be non-confounded. For details, readers are referred to [19].

The greedy equivalence search (GES)[25 26] represents the other type of CSD algorithm, the score-based algorithms. Different from constrained-based algorithm, like FCI and PC, GES depends on a predefined score (e.g. Bayes Information Score[27]) to estimate the causal structure. Recent CSD methods extend the previous methods by relaxing many of the assumptions: handling non-linear casual relationships[28], cyclic causal relationships[29], individualized causal structures[30 31]. Additionally, integration of deep learning into CSD[32-36] has opened new methodological directions including the use of generating adversarial networks[37-39] and reinforcement learning[40-42].

This thesis aims to make contributions to the field of health informatics, particularly in the realm of causal analysis using biomedical datasets. It focuses on two primary objectives:

(1) **Exploring and Providing Guidance on Existing CSD Methods**: This involves an exploration of current CSD methodologies, aiming to provide comprehensive guidance on how these can be effectively utilized to extract biomedical causal knowledge. There is a notable gap in the literature regarding practical, example-driven guidance for this task, which this thesis seeks to address.

(2) **Proposing Improvements and Developing New Methodologies**: The thesis proposes innovative CSD/CI methodologies that are specifically designed to overcome the practical challenges prevalent in biomedical datasets.
The thesis is structured into four chapters, each delving into different aspects of causal analysis:

**Chapter I** of this thesis validates and compares existing computation CSD algorithms on discovering the causal structure among Alzheimer's disease (AD) related biomarkers. To enable the evaluation of CSD algorithms, a gold standard graph among biomarkers related to AD was constructed. We compared the performance of the three CSD algorithms PC, Fast causal inference (FCI), and Fast Greedy Equivalence Search (FGES), which is a fast version of GES, on discovering the gold standard graph. Several tools to improve the performance of CSD methods were also tested.

**Chapter II** applies CSD methods to discover causal relationships related to Alzheimer’s disease, and then uses causal inference method to estimate the effect sizes of each biomarker. This chapter pays attention to white matter hyperintensities (WMH), which are hyperintense patches on T2-weighted or fluid attenuated inversion recovery images. WMH are important to measure because they are the common manifestations of cerebrovascular disease and have significant impact on motor and cognitive function. To truly study the protective factors on WMH, we need to be able to study early changes to the brain prior the formation of WMH and the complex association between risk factors and protective factors and the influence on early brain changes. The goal of this chapter is to leverage both multi-modal imaging and clinical information along with CSD/CI methods to map the evolution of WMH.

**Chapter III** proposes a novel CSD algorithm specifically designed for discovering causal relationships from EHR datasets. Motivated by the findings from Chapter I and Chapter II, which reveal that large sample sizes and the longitudinal data significantly improve the accuracy and
recall of CSD methods in both the causal edge discovery and orientation. While EHR data offer an extensive and longitudinal collection of diverse patient data including demographics, disease histories, prescriptions, and laboratory results, making EHR an ideal source for CSD task. However, EHR datasets also present challenges in knowledge discovery. First, EHR data as it exists in the system does not follow any study design. Billing codes in particular are recorded for reimbursement purposes and do not distinguish between new incidences and pre-existing conditions. Secondly, time stamps in EHR can be unreliable. The time stamp of a diagnosis often does not coincide with the onset time of the disease, but rather reflects the documentation time. Our objective is to develop methodology that tackle these EHR-related challenges. Specifically, (1) we propose a data transformation procedure that distinguishes new incidences from pre-existing conditions, allowing the subsequent CSD algorithm to make the appropriate study design considerations. (2) We also develop a CSD algorithm that can infer the direction of causal relationships more robustly using longitudinal information and takes the above study design considerations into account.

Chapter IV introduces a novel causal inference method named Missing-Mechanism-Adjusted Structural Equation Models SEM (MMA-SEM) that aims to estimate the effects among latent variables in the presence of missing data. As missing data is one of the most challenging problems that exist in biomedical data, such as EHR. This method operates under the assumption that the structure among variables and the causes of missing values are known. We evaluate the proposed algorithm on synthetic datasets and a real EHR dataset. To further validate our approach, we also conducted experiments under conditions where this assumption was violated. The MMA-SEM algorithm stands out for its sophisticated integration of the missing data
mechanism. It adeptly leverages both the missing and observed data to make inferences about latent variables. Employing the Alternating Direction Method of Multipliers, the algorithm iteratively refines the estimates of missing data, latent variable values, and coefficients. This refinement is achieved by drawing on insights from both observed data and inferred latent variables, showcasing the algorithm's capability in handling complex data scenarios inherent in EHR systems.

Ultimately, the research presented in this thesis offers practical examples of applying CSD and causal inference methods to address biomedical problems. It also proposes two novel methods to navigate prevalent data challenges, which are crucial for effectively utilizing EHR data and extracting meaningful causal relationships.
Chapter I: Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer’s Pathophysiology

Adapted with permission from X Shen, S Ma, P Vemuri, G Simon. Scientific Reports 2020;10(1):2975 doi: 10.1038/s41598-020-59669-x.

Copyright 2020 Scientific Reports.
Overview

Causal Structure Discovery is the problem of identifying causal relationships from large quantities of data through computational methods. With the limited ability of traditional association-based computational methods to discover causal relationships, CSD methodologies are gaining popularity. The goal of the study was to systematically examine whether (i) CSD methods can discover the known causal relationships from observational clinical data and (ii) to offer guidance to accurately discover known causal relationships. We used Alzheimer’s disease (AD), a complex progressive disease, as a model because the well-established evidence provides a “gold-standard” causal graph for evaluation. We evaluated two CSD methods, Fast Causal Inference and Fast Greedy Equivalence Search in their ability to discover this structure from data collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI). We used structural equation models (which is not designed for CSD) as control. We applied these methods under three scenarios defined by increasing amounts of background knowledge provided to the methods. The methods were evaluated by comparing the resulting causal relationships with the “gold standard” graph that was constructed from literature. Dedicated CSD methods managed to discover graphs that nearly coincided with the gold standard. For best results, CSD algorithms should be used with longitudinal data providing as much prior knowledge as possible.

Introduction

Big data analytics, machine learning, and deep learning have garnered significant interest in the health science fields[43 44]. Due to their excellent predictive accuracy, they are increasingly employed for disease diagnosis and risk prediction[3]. However, in many biomedical applications, achieving high prediction accuracy in and by itself is not the primary goal;
discovering the risk factor or mechanism that can be altered is often the primary research question.

Today’s machine learning applications are largely based on associations. Even though a risk factor may be associated with the disease, it does not necessarily mean that it can alter the disease process. In early 2018, a Phase 3 trial called “TOMMORROW” tested the effect of a diabetes drug on reducing Alzheimer’s disease dementia risk [45]. The study measured amyloid deposition, which is an early sign of Alzheimer’s disease and is also associated with diabetes. However, since diabetes is not causal to amyloidosis, the study failed in the interim analysis [46 47]. For a successful intervention, the risk factor we intervene on should have a causal (rather than merely associative) relationship with the disease outcome.

Clinical research is predominantly focused on causal relationships. Hypothesis-driven clinical research, for example, often assumes a causal structure, a set of causal relationships among biomarkers and outcomes, and researchers estimate the effect size of these relationships (e.g. causal inference). In such research, drawing a causal conclusion is valid, because prior knowledge ascertains that the relationships are indeed causal. However, when there is no knowledge of the causality, the causal structure itself needs to be discovered from data through a process known as causal structure discovery. A commonly used but incorrect practice is to assume a partial causal structure and adjust it based on output statistics of the fitted model using methods such as structural equation models (SEM).
In this work, using AD biomarkers as the predictors and cognition as the outcome, we set out to determine an optimal way to discover causal relationships. We used AD as a model for this problem because the AD biomarker cascade is well understood[48] and the causal relationships between the primary predictors has also been well characterized such that a “gold standard” graph can be constructed. Further, the public data set of Alzheimer’s disease neuroimaging initiative (ADNI) has extensive longitudinal data available that is conducive for the systematic comparisons planned in this manuscript. Here, we focus on comparing the results from dedicated causal discovery algorithms and a searching algorithm based on SEM, with our “gold standard” graph. We also investigated the reason behind common mistakes and explored methods to prevent them. These experiments allowed us to provide guidelines for discovering causal structure using observational data.

**Background**

*Causal structure Discovery Algorithms*

Informally, causation is defined as a relationship between two variables X and Y such that changes in X lead to changes in Y[49]. The key difference between association and causation lies in the potential of confounding. Suppose that no direct causal relationship exists between X and Y but rather a third variable Z causes both X and Y. In this case, even though X and Y are strongly associated, altering X will not lead to changes in Y. Z is called a confounder. More formally, causation is a direct effect between A and B that remains after adjusting for confounding. Confounding can be observed or unobserved (latent).

*Causal structure* is the set of causal relationships among a set of variables, and
causal structure discovery is the problem of learning the causal structure from observational data. Dedicated causal structure discovery algorithms exist and can be separated into two subtypes, constraint-based and score-based. The constraint-based algorithms construct the causal structure based on conditional independence constraints, while the score-based algorithms generate a number of candidate causal graphs, assign a score to each, and select a final graph based on the scores. In this study, we selected one prominent algorithm from each type: Fast Causal Inference Algorithm (FCI), which is a constraint-based algorithm, and Fast Greedy Equivalence Search (FGES), which is a score-based algorithm. For brevity, we give a high-level description for FGES and FCI. For more detailed descriptions, we refer the reader to the references\[26 50 51\]. Both of the two methods can adjust for observed confounding and one of the algorithms, FCI, has some ability to discover latent confounding.

**Fast Causal Inference (FCI)**

The central concept behind constraint-based causal discovery algorithm is the idea that different causal structures imply different independence relationships. For example, the causal relationship A->B->C, implies that variable A is independent of C given B. On the other hand, when A->C<-B, A and B are independent (unconditionally), but become dependent conditional on C. The latter structure is called the “V” structure (also known as collider) which has a unique independence relationship compared with other causal relationships. In fact, it is one of the “primitives” that constraint-based algorithm, like FCI, looks for.

A feature specific to FCI even among constraint-based methods is its ability to discover latent (unobserved) confounders. This is enabled by another primitive, the “Y” structure. Four
variables define a “Y” structure when they have the following causal relationships: W1 -> X <\-\ W2 and X -> Y. Within the “Y” structure, both W1 and W2 are independent of Y conditional on X. This conditional independence helps rule out the possibility of an unmeasured confounder between X and Y. In other words, when FCI finds a “Y” structure in the graph, the causal relationship from X to Y is guaranteed to be unconfounded; otherwise, FCI assumes that possibly unobserved confounders exist[52].

**Fast Causal Inference (FCI) algorithm:** FCI constructs a causal graph starting with a fully connected undirected graph, and removes edges that connect conditionally independent variables. In the second phase, it orients edges by identifying the “V” and “Y” structures, and tries to orient the remaining edges based on a set of rules which have been explained in detail elsewhere[19 50].

**Fast Greedy Equivalence Search (FGES) algorithm:** The Greedy Equivalence Search (GES) algorithm also has two phases. In the first phase, it starts with a graph containing no edges (corresponding to all variables being independent of each other) and greedily adds edges (dependencies) one at a time in the orientation that minimizes the Bayes Information Score[27] (BIC), which is likelihood penalized for complexity to reduce overfitting. GES then removes edges one at a time as long as it decreases the BIC. The FGES[51] algorithm used in this work is simply a “fast” (parallelized) version of GES[25 26]. Similarly to FCI, FGES also relies on the “V” structures to orient edges. The implied likelihood of the “V” structure is unique while the likelihoods of A->B->C, C->B->A and A<-B->C are the same. Thus, FGES will select the “V” structures when it implies a higher likelihood than other structures.
**Structural Equation Modeling (SEM):** Structural Equation Modeling (SEM) is a family of statistical models, which, given the underlying causal structure, can estimate the effect size (and other statistics as well) of each relationship[53]. SEM can also suggest modifications to the given causal structure to improve model fit statistics.

While SEM was not designed to discover the causal structure, it is not uncommon to use SEM’s suggested modifications to “refine” the graph structure. This feature can be exploited to iteratively build a causal graph, in each iteration, adding one edge as per the suggestion by SEM. We implemented this (incorrect) searching method under two scenarios: (1) starting from the empty graph (Causal discovery); and (2) starting from a graph obtained by deleting 1 or 2 edges from the “gold standard” graph. Note that, within the scope of this paper, we use the term “SEM” to represent the algorithm that uses SEM to conduct edge searching, not to estimate the effect size.

**Key differences between the Algorithms**

Both SEM and FGES are stepwise algorithm which modify structure by adding or deleting edges. The biggest advantage of FGES is that it extends the search space by transforming the current structure to other “equivalent” structures. For example, given the edge A -> B is in the A, B, and C graph. SEM will try adding one directed edge between C and A or C and B, yielding four possibilities. However, FGES considers more possibilities as it can also reverse the existing edge A->B to A <- B, yielding four additional possible structures.
Other than the searching strategies (Constraint versus Score based), FCI also differs from the other two algorithms in its assumption about causation: both SEM and FGES operate under the assumption of no unmeasured confounders. In other words, all the confounding variables are measured in the dataset. FCI, however, relaxes this assumption, and reports an unconfounded relationship only when it encounters a “Y” structure[54].

FCI and FGES algorithms are implemented in the Tetrad software package (Version 6.5.4). Figure 1.1 shows the interpretations of different edge types in the output graph. For SEM, we used the R package ‘lavaan’[55].

<table>
<thead>
<tr>
<th>Present Relationships</th>
<th>Absent Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 A ——— B</td>
<td>A is the cause of B</td>
</tr>
<tr>
<td></td>
<td>B is not cause of A</td>
</tr>
<tr>
<td>2 A ——— B</td>
<td>A is the cause of B</td>
</tr>
<tr>
<td></td>
<td>or B is the cause of A</td>
</tr>
<tr>
<td>3 A ——— B</td>
<td>A is not a cause of B</td>
</tr>
<tr>
<td></td>
<td>and B is not a cause of A</td>
</tr>
<tr>
<td>4 A o—— B</td>
<td>Either A is a cause of B</td>
</tr>
<tr>
<td></td>
<td>or there is an unmeasured confounder of A and B</td>
</tr>
<tr>
<td></td>
<td>B is not a cause of A</td>
</tr>
<tr>
<td>5 A o—— B</td>
<td>Exactly one of the following holds:</td>
</tr>
<tr>
<td></td>
<td>1. A is a cause of B</td>
</tr>
<tr>
<td></td>
<td>2. B is a cause of A</td>
</tr>
<tr>
<td></td>
<td>3. There is an unmeasured confounder of A and B</td>
</tr>
<tr>
<td></td>
<td>4. Both 1 and 3</td>
</tr>
<tr>
<td></td>
<td>5. Both 2 and 3</td>
</tr>
</tbody>
</table>

The output graph of SEM contains edge type 1; FGES contains edge types 1,2; FCI contains edge types 1,3,4,5.

**Figure 1.1** The interpretation of edges in CSD

**Method**
Data

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD)[56]. There are three phases of the ADNI study where the last one, ADNI3, is still ongoing. All study participants provided written informed consent, and study protocols were approved by each local site’s institutional review board. All methods were carried out in accordance with the relevant guidelines and regulations. For up-to-date information, see www.adni-info.org. IRB Review was not required since the ADNI data is de-identified and publicly available for download. We focused our study on the first two: ADNI 1 and ADNI 2/GO. The variables extracted from the data are fludeoxyglucose PET (FDG), amyloid beta (ABETA), phosphorylated tau (PTAU), apolipoprotein E (APOE) ε4 allele; demographic information: age, sex, education (EDU); and diagnosis on AD (DX). Table 1.1 presents summary statistics of the data set. After removing records with missing values, there are 1008 participants remaining with at least one complete record, and 266 with a regular two-year follow-up visit.

Table 1.1 Characteristics for Continuous and Categorical Variables.

<table>
<thead>
<tr>
<th>Demographic variables</th>
<th>Label</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>AGE</td>
<td>74.09 (7.46)</td>
</tr>
<tr>
<td>SEX</td>
<td>SEX</td>
<td>0.55 (0.50)</td>
</tr>
<tr>
<td>Education Level</td>
<td>EDU</td>
<td>16.15 (2.71)</td>
</tr>
</tbody>
</table>

Biomarkers
The “Gold standard” graph

The AD biomarker cascade has been evaluated widely. The deposition of ABETA in the brain is an early event in the disease process and is captured through the decrease in CSF ABETA. The only consistently shown risk factors for ABETA are age and the number of APOE4 alleles[47 57 58]. ABETA causes downstream neurofibrillary tangle formation and subsequently neurodegeneration, both of which are captured by metabolic dysfunction via FDG-PET[59] and PTAU increase measured on CSF[60]. The two markers FDG-PET and CSF PTAU are the strongest predictors of cognitive dysfunction or diagnosis[61 62] (in comparison to ABETA).

Education, a surrogate of cognitive resilience, influences an individual’s cognitive status[63]. All of these are well established relationships in the literature. There are weaker causal associations such as sex influencing some of these associations which we did not regard to evaluate the algorithms because the impact of these associations is much smaller in comparison to the main effects considered in the “gold standard” graph. The relationships described above are shown in figure 1.2.
Background Knowledge and cross-sectional vs Longitudinal data

To constrain the relationships that the algorithms can discover, background knowledge can be provided in the form of must-have or must-not-have (prohibited) edges. In this paper, we defined three degrees of background knowledge as: (Level 1) *No knowledge*: the discovered structure purely reflects the data; no edges are prohibited. (Level 2) *Trivial background knowledge*: (a) edges among demographic variables are prohibited (although association between them can remain) (b) edges from biomarkers or diagnosis to demographic variables are prohibited. (Level 3) *Longitudinal*: in addition to the edges prohibited in Level 2, edges pointing from a later time point to an earlier time point are also prohibited.

Study design
Causal discovery study

We extracted two data sets from ANDI for this part of the study: one was a single cross-sectional, and data was collected at the baseline visit made by each participant. The second one is longitudinal, where we included data from two cross-sections: the baseline visit, and the visit made at the 24 months. Records with missing data were removed from further study.

To generate robust results, both the cross-sectional and the longitudinal data were bootstrapped 100 times at the participant’s level. Then, the three algorithms, SEM, FCI and FGES were tested on all bootstrap samples for evaluation, incorporating the three different degrees of knowledges that were described in the previous section.

SEM-recovery study

Since most researchers would start with a hypothesized graph and only use SEM to add edges, we also tested SEM under this assumed use case: we initialized (hypothesized) graphs by deleting each single edge and each pair of edges from the “gold standard” graph, and then tested whether SEM can recover the deleted edges after no more than five iterations of edge adding. We chose five in this study because more than five times of edge adding will result in a graph with low recall.

Evaluation Metrics

To assess the performance of methods, we defined following evaluation metrics. An edge is correct, if and only if the same edge exists in the “gold standard” graph and the orientation of the edge coincides with the orientation in the “gold standard” graph; an edge is semi-correct, if
and only if the same edge exists in the “gold standard” graph and its orientation does not contradict with the true orientation of the edge in the “gold standard” graph; And finally, an edge is incorrect if the edge does not exist in the “gold standard” graph or if it exists but its orientation is the opposite of the true orientation.

We will present the following metrics:

1. Number of correct, semi-correct, incorrect edges
2. Precision: the proportion of correct or semi-correct edges over all edges reported by the algorithm
3. Recall: the proportion of edges in the “gold standard” graph that are correctly or semi-correctly reported
4. Occurrence rate: the percentage of the adjacency shows in the result of the 100 bootstrap runs

Results

Causal discovery study

The discovered causal structures generated by SEM, FCI, and FGES algorithms across three degrees of prior “knowledge” are shown in this section. The behind-the-scenes mechanism of typical mistakes will be examined further in the discussion section.

Experiment 1: Without background knowledge
In Figure 1.3, we present the edges with at least 80% occurrence rate in the 100 bootstraps samples (The number located near each edge). The edges that were not in the gold standard graph are colored in red. The numbers on the right of each graph are the precision, recall, and the number of correct, semi-correct, and incorrect edges averaged over the 100 bootstrap samples. To ease direct comparison, the variables are laid out almost identically: the same variable occupies the same relative location in all three graphs.
SEM was only able to retrieve two correct edges from the “gold standard” graph (with average precisions 0.24 and recall 0.30), while FCI and FGES found 4 out of 8 edges correctly or semi-correctly (precision 0.24 and 0.44, recall 0.46 and 0.6 correspondingly). Both FCI and FGES successfully recovered the causal relationship between genetic variable APOE41 with ABETA, ABETA with FDG, and FDG, PTAU with DX. However, the algorithms failed to determine the directionalities of some of the relationships. We also observed that all three algorithms reported edges from biomarkers to demographic variables which are certainly errors (e.g. ABETA causes APOE42 in FCI’s graph). It is important to note that some well-established relationships such as age and amyloid as well as education and diagnosis were not discovered in any of the graphs which did not have background knowledge.

*Experiment 2: Addition of trivial background knowledge*
Figure 1.4 presents the causal structures discovered by the three algorithms incorporating trivial background knowledge: demographic variables cannot be caused by other demographic variables nor by biomarkers (e.g. participant’s age is not affected by education or ABETA level). The structure of Figure 1.4 is analogous to Figure 1.3.

While all the methods made several mistakes, there were significant improvements when trivial background information was added. Some of the incorrect causations found by SEM are actually indirect causal relationships in the “gold standard”. For example, the effect from APOE42 to DX
is an indirect effect that flow through ABETA in the ‘gold standard’ graph. All three algorithms discovered the edge ABETA -> DX. Though it is not a direct casual effect in our “gold standard” graph, ABETA is a common cause of FDG reduction as well as PTAU increase and both FDG and PTAU lead to DX. Therefore, the effect of ABETA and DX could be anticipated. Both FCI and FGES reported a falsely-directed edge between PTAU and DX. We will see this error is corrected by using longitudinal data. The FCI algorithm also reported unmeasured confounders between PTAU and DX with FDG, which are interesting hypotheses that need further studies. Among the three algorithms, SEM achieved the lowest performance (Precision 0.31, recall 0.39) while FCI and FGES achieved higher and substantial higher performance (FCI: 0.42 precision and 0.55 recall; FGES 0.71 precision and 0.68 recall).

Experiment 3: Addition of longitudinal data and trivial background knowledge
Figure 1.5 presents the most frequent edges discovered by the three algorithms and their performance metrics. The layout of the graphs is different from the previous Figures because they are built on the longitudinal data set. All nodes associated with biomarkers or diagnosis hence appear twice: once with their baseline value (denoted by ‘.0’ suffix) and once at 24 months (denoted by the ‘.24’ suffix). Most of the statistics further improved relative to previous results and FGES recovered a graph with only one incorrect edge.
The performance of the SEM algorithm trailed behind the other two dedicated causal discovery algorithms. In some of the bootstrap runs, SEM missed the direct effects between the longitudinal measurements of the same biomarker (e.g. the estimated probability of SEM discovered the edge ABETA.0 -> ABETA.24 is 0.47 where FCI and FGES always include this edge).

The FCI algorithm further identified that PTAU at initial visit has an effect on the diagnosis (AD) at 24 months. In other words, PTAU may have a lagged effect on AD diagnosis which is a highly plausible hypothesis. We also observed that AGE and EDUCATION lead to different FDG and diagnosis at the first visit, but not directly to the assessment at 24 months (after adjusting the assessment at baseline visit).

The FGES algorithm incorporated the longitudinal data and successfully discovered the FDG to DX edge. It also removed the incorrect edges from DX to PTAU. Furthermore, with longitudinal data, the previously undirected edges identified by FGES got directed without compromising the overall precision and recall.

**SEM-recovery study**

Table 1.2 shows the statistics when we tested SEM’s ability to recover deleted edges from the “gold standard” graph. In each run, we deleted a single edge or a pair of edges. The “fully recovery rate” represents the percentage of runs in which SEM managed to fully recover the deleted edge(s). The “precision” and “recall” columns are defined the same way as in the previous experiments. As we can see from Table 1.2, when we removed only a single edge, the
recovery rate is very low (12.5%). When we removed two edges from the “gold standard” graph, SEM was unable to recover the true graph.

<table>
<thead>
<tr>
<th>Number of edges removed</th>
<th>Fully recover rate</th>
<th>Precision. Mean</th>
<th>Recall. Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.125</td>
<td>0.67</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.70</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 1.2 Recovery rate of edges

Discussion

In this study, we compared three different methodologies to recreate the known ground-truth causal structure based on an observational dataset using three different degrees of “knowledge”. We used Alzheimer’s disease data from ADNI which is a well characterized openly accessible data set. Since the relationships among biomarkers and diagnosis in Alzheimer’s disease are well understood, we began with a “gold-standard” causal structure based on the existing literature. Then, we applied three algorithms to discover this causal structure from data. This work highlights the common errors made by the different algorithms and offers us with ideas and suggestions to avoid these errors. In the end, a detailed guideline on how causal discovery algorithm can be applied to discover high-quality causal relationships was provided.

Each of the three algorithms used in this work represents a class of algorithms with its specific characteristics. Two of the algorithms, FCI and FGES, are dedicated causal discovery algorithms, while the third one, SEM, is primarily designed as a confirmatory tool. The dedicated causal discovery algorithms outperformed SEM across all three degrees of background knowledge. This is not surprising, because SEM is not specifically designed to discover causal structure; statistics reported by SEM only indicate possible adjustments to the a priori user-
defined causal structure. What is surprising is the extent to which FGES outperformed SEM, since both FGES and SEM optimize the same criterion, which is BIC. The key difference between FGES and SEM is the scale of the underlying search space: FGES considers a broader array of graphs, all graphs that have the same dependence structure (same set of conditional independence relationships among the variables). From the SEM-recovery experiment, we also observed that the SEM’s suggestions for adding edges are generally not reliable. These edges may maximize BIC in SEM’s limited search space, but these are not the overall optimal edges: FGES, with its larger search space, managed to (almost perfectly) recover the “gold standard” graph.

With FCI and FGES having similar search spaces, the main differences between them lies in their search algorithm. The performance of the score-based algorithm FGES was higher and was more stable than the constraint-based algorithm FCI in our study. The decision making of FCI was affected by the incorrect independence tests introduced by selection bias or data artifacts. These mistakes propagated to other parts of the graph through generating incorrect “V” or “Y” structures and eventually caused damage to large portions of the graph. In contrast, score-based algorithms consider the likelihood of the global structure while making local decisions; so, these errors remain localized. This explains why the discovered structure of FGES before or after adding trivial knowledge are more consistent. FCI has the advantage of being able to relax the typical assumption of no unmeasured confounders. This relaxation can be useful when either identifying unmeasured confounders or finding unconfounded causal relationships is important. In our study, FCI found that the relationship from ABETA to FDG is un-confounded and that unmeasured confounders may exist between FDG, DX (Figure 1.4-FCI)
We further investigated the mistakes that FCI and FGES made. We grouped these mistakes into three categories and described their causes and work-arounds in Table 1.3.

Table 1.3 Typical problems and solutions

<table>
<thead>
<tr>
<th>Error</th>
<th>Location</th>
<th>Reason for error</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 EDU and SEX</td>
<td>Figure 1.3-FCI</td>
<td>Selection bias</td>
<td>Add trivial knowledge</td>
</tr>
<tr>
<td>2 APOE4, PTAU and FDG</td>
<td>Figure 1.4-FCI</td>
<td>Selection bias or Artifacts and No background knowledge</td>
<td>Longitudinal data</td>
</tr>
<tr>
<td>3 PTAU → DX</td>
<td>Figure 1.5-FGES</td>
<td>Small sample size</td>
<td>Increase sample</td>
</tr>
</tbody>
</table>

1. The first kind of error happens when artifacts in the data induce incorrect edges. For example, FCI reported an edge between EDU and SEX (Figure 1.3-FCI), because in our sample, the average education level of male participants is higher. Avoiding such incorrect edges is important because they can potentially create incorrect “V” or “Y” structures that jeopardizes the remaining causal discovery steps. Adding trivial background knowledge can resolve this problem by preventing the algorithms from treating association as causation.

2. When universally accepted background knowledge is not available, compensating for data artifacts is more difficult and can have distant downstream effects. For example, the APOE42-PTAU-FDG structure was inferred as a “V” structure (APOE42→PTAU←FDG) in some of the bootstrap runs. This error was a result of a single incorrectly inferred independence between APOE42 and FDG from the sample data. This error propagated through three associations between (1) APOE42 and PTAU, (2) PTAU and FDG, and (3) APOE42 and FDG conditional on PTAU, which led to a “V” structure among the three. We
cannot prevent this edge using background knowledge unless we know the conditional independence relations between APOE42 and FDG beforehand. The use of longitudinal data helped correct these mistakes; as we can see in Figure 1.5-FCI, the substructure was corrected when longitudinal data was used.

3. While longitudinal data provides a solution to a number of the problems, the requirement of repeated observations can constrain the sample size and introduce some errors of its own. For example, FCI discovered a possible lagged relationship between PTAU at time 0 and DX at month 24 (Figure 1.5-FCI), however, the same relationship is not observed in the results from FGES (Figure 1.5-FGES) -- possibly due to the small sample size.

In the longitudinal study, local structure across time points are not guaranteed to be the same. For example, in the graph discovered by FGES, ABETA.0 causes PTAU.0 but ABETA.24 does not cause PTAU.24. The reason is that ABETA.0 is a common parent of ABETA.24 and PTAU.0. PTAU.24 is conditionally independent of ABETA.24 given either ABETA.0 or PTAU.0. This conditional independence relationship implies that in the presence of ABETA.0 or PTAU.0, ABETA.24 is not needed to explain the variation in PTAU.24.

Even though the final graph learnt from observational data matched the “gold standard” graph closely, our conclusions still depend on the correctness of the “gold standard”. In general, we have high confidence in the “gold standard” graph as the biological mechanism behind the AD biomarker cascade is well understood and FGES managed to discover the “gold standard” almost perfectly. However, it is highly possible that FDG and PTAU only explain part of the effect from
ABETA on diagnosis of AD; a direct causation from ABETA on DX could exist. Although we recommend longitudinal data, collecting data longitudinally is often costly which typically results in a smaller sample size. Small sample size lowers the statistical power in causal discovery algorithm which is a trade-off. We tried reducing the sample size by 50% and 75% and conducted the same analysis. When we reduced the sample size by 50%, the total numbers of discovered edges across 100 bootstrap iterations reduced. However, edges that were consistently discovered on the full sample were consistently discovered on the reduced sample as well. We achieved similar precision and recall. When we further reduced the sample size (by a total of 75%), the total number of edges further reduced. While the most frequently discovered edges continued to get discovered, the number of “noise edges”, edges that were discovered only in a few bootstrap iterations, increased.

In conclusion, dedicated causal discovery algorithms outperformed SEM in discovering the causal structure. In real-world data analysis, data quality impacted the correctness of the discovered structure. Incorporating prior knowledge and using longitudinal data can improve the discovered result by preventing algorithms from make some potential mistakes.

Acknowledgement

This study is supported by the National Institutes of Health (NIH) grant #AG056366. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NIH. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded
by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
Chapter II: Causal structure discovery identifies risk factors and early brain markers related to evolution of white matter hyperintensities

Adapted with permission from X Shen, S Raghavan, SA Przybelski, TG Lesnick, S Ma, RI Reid, et al. NeuroImage: Clinical 35, 103077

Copyright 2022 NeuroImage: Clinical
Overview

Our goal was to understand the complex relationship between age, sex, midlife risk factors, and early white matter changes measured by diffusion tensor imaging (DTI) and their role in the evolution of longitudinal white matter hyperintensities (WMH). We identified 1564 participants (1396 cognitively unimpaired, 151 mild cognitive impairment and 17 dementia participants) with age ranges of 30-90 years from the population-based sample of Mayo Clinic Study of Aging. We used computational causal structure discovery and regression analyses to evaluate the predictors of WMH and DTI, and to ascertain the mediating effect of DTI on WMH. We further derived causal graphs to understand the complex interrelationships between midlife protective factors, vascular risk factors, diffusion changes, and WMH. Older age, female sex, and hypertension were associated with higher baseline and progression of WMH as well as DTI measures ($P \leq 0.003$). The effects of hypertension and sex on WMH were partially mediated by microstructural changes measured on DTI. Higher midlife physical activity was predictive of lower WMH through a direct impact on better white matter tract integrity as well as an indirect effect through reducing the risk of hypertension by lowering BMI. This study identified key risks factors, early brain changes, and pathways that may lead to the evolution of WMH.
INTRODUCTION

Leukoaraiosis or white matter hyperintensities (WMH) are the hyperintense patches on T2-weighted or fluid attenuated inversion recovery (FLAIR) images. WMH are important because they are a common manifestation of cerebrovascular disease (CVD) and have a significant impact on motor and cognitive function[64 65]. They are also a predictor of increased risk of stroke and dementia[66]. Older age and hypertension are the two well known risk factors associated with the development of WMH[67-70]. Recent studies show that females have higher WMH load suggesting that sex differences play a role in the evolution of WMH[67 71]. Even with vast literature on risk factors of WMH, a clear mechanistic understanding of the factors related to the evolution of WMH is still lacking.

**Figure 2.1** Graphical abstract
Emerging evidence shows significant microstructural changes on diffusion tensor imaging (DTI) even before the appearance of WMH[72 73]. Given the heterogeneity in the formation of WMH, our first hypothesis was that measuring early diffusion changes will aid in predicting future formation of WMH beyond the classical risk factor - hypertension. Further, evaluating midlife risk factors in the context of early diffusion changes are likely to lead to a better understanding of the mechanisms. Therefore, our second hypothesis was that midlife risk factors have an impact on WMH through early diffusion changes and measuring these early diffusion changes will allow us to identify the pathways for prevention of WMH.

Given these gaps in knowledge, our goal was to investigate the complex relationships between age, sex, midlife risk/protective factors, diffusion changes, and WMH. We evaluated this question using a large dataset (n=1564) of longitudinal multi-modal imaging data (WMH via FLAIR MRI and fractional anisotropy or FA from DTI MRI), clinical information (related to hypertension, dyslipidemia, and diabetes), and midlife risk factors (physical activity, cognitive activity, and smoking) from the population-based sample of Mayo Clinic Study of Aging (MCSA). In addition to traditional statistical regression models, we leveraged causal structure discovery (CSD) models to study the interactions and pathways that lead to worsening WMH. The CSD methods have many strengths. As compared to traditional regression models for a specific outcome, they can model the complex (direct and indirect) relationships among all variables at the same time. Given a set of assumptions[9], such as no unobserved confounders, the Causal Faithfulness condition, and the Causal Markov condition, the relationships in the causal graph admit a ‘causal’ interpretation. As compared to traditional structural equation modeling and path analysis, the CSD methods are data driven and can incorporate existing
knowledge regarding the domain of interest.[9] These discovered relationships help investigate causal interpretations whereby manipulating one variable can alter other variables. These models stand in contrast to models that investigate associative relationships and have demonstrated great success in many domains such as causal relationships in anxiety disorder, image recognition, or climate prediction[74-76].

MATERIALS AND METHODS

Selection of Participants

Participants were selected from MCSA, an epidemiological sample of residents living in Olmsted County, Minnesota. Olmsted county population was enumerated in the Rochester Epidemiology Project (REP) medical records-linkage system[77 78]. The details of the study design were published elsewhere [79 80]. The inclusion criteria were cognitively unimpaired elderly with an age range of 30-90 years with usable 3T FLAIR MRI, DTI, and cardiovascular and metabolic risk factor information.

Standard protocol approvals, registrations, and patient consents: The study was approved by the Mayo Clinic and Olmsted Medical Center institutional review boards and written informed consent was obtained from all participants.

Imaging

Assessment of WMH on FLAIR scans

All MRI images were acquired from 3T MRI systems (GE Healthcare). Both 3D MPRAGE and 2D FLAIR image were used to calculate WMH volume. The acquisition and analysis of the FLAIR images were described previously[81]. In brief, the possible WMH voxels on FLAIR
images were identified initially through clustering via connected components\[81\]. We then
masked the FLAIR images using white matter (WM) masks derived from 3D MPRAGE
segmentation to exclude the false-positive WMH voxels. These masks were edited by a trained
analysts to remove non-WMH voxels.

**DTI**

The DTI acquisition protocol used a single-shot echo-planar imaging sequence with an isotropic
resolution of 2.7 mm, 5 non diffusion-weighted images, and 41 b=1000 s/mm\(^2\) diffusion-
encoding gradient directions spread over the whole sphere. The data were preprocessed by skull
stripping\[82\], denoising\[83\], correcting for head motion and eddy current distortion\[84\], Gibbs
ringing\[85\], and then debiasing\[86\]. Diffusion tensors were then fit using a nonlinear least
squares fitting algorithm implemented in dipy,\[87\] and then FA maps were generated. ANTs
(Advanced normalization tools symmetric normalization)\[88\] was used to non linearly register
each participant’s FA image to the in-house- version of John’s Hopkins University “Eve” WM
atlas\[89\], and then regional FA measures were computed. FA measures were computed for 12
WM tracts including genu of the corpus callosum, splenium of the corpus callosum, body of the
corpus callosum, parahippocampal cingulum, cingulum, inferior temporal WM, superior
longitudinal fasciculus, corticospinal tract, anterior limb of the internal capsule, posterior limb of
the internal capsule, inferior fronto-occipital fasciculus, and uncinate fasciculus.

Quality assessment was performed by the visual inspection of each participant’s DTI image
acquisition (both raw images and DTI based FA/MD images) and registration of the JHU atlas to
the images. In those who had potentially usable data, we found that a small percentage (0.28%)
of corresponding DTI scans failed quality control based on visual inspection of the DTI image
acquisition and processing, and those scans were not used. Separately from QC, the FA and MD values were characterized using the median of voxel values within each JHU region of interest. The median was used to reduce the partial volume contamination from the edge voxels of each region. We also excluded the voxels with MD > 2 × 10^{-3} or < 7 × 10^{-5} mm²/s as they were mostly cerebrospinal fluid or air, respectively. Moreover, we excluded the smallest JHU regions (cuneus WM, fusiform WM, lingual WM, and precuneus WM, all typically < 7 diffusion voxels) as they were too small for reliable registration onto the corresponding subject structure, and liable to be dominated by edge voxels.

**Assessment of cardiovascular and metabolic risk factors**

The medical history of the participants was obtained from a combination of in person clinical visit or the REP medical record linkage system. We obtained the body mass index (BMI, kg/m²), metabolic syndromes such as low-density lipoprotein (LDL), high density lipoprotein (HDL), triglycerides, systolic and diastolic blood pressure (SBP≥140 mmHg and DBP≥90 mmHg), and total cholesterol from REP. We also utilized nurse abstracted data from MCSA on the history of vascular risk factors including type 2 diabetes mellitus, hypertension, and dyslipidemia[90]. In addition, we utilized smoking status ascertained at the time of the clinical visit (never, current, and former). We estimated midlife physical and cognitive activity summary scores from questionnaires that were published previously[91] which summarize engagement in several physical and cognitive activities.

**Statistical analyses**

42
All statistical analyses were conducted using R (v.4.0.0) and we used R package rcausal (v.1.1.1) for causal discovery. Characteristics of the participants were summarized as mean (standard deviation) for the continuous variables, and count (%) for the categorical variables stratified by age groups. The differences between age groups were compared using ANOVA and $\chi^2$ test for continuous and categorical variables, respectively. We computed total intracranial volume using an in-house modified SPM software implementation and used WMH as a percentage of total intracranial volume (dividing by the total intracranial volume). Dividing by total intracranial volume accounts for differences in head size and is common practice in the field[92]. WMH was log-transformed for normality.

*Risk factors associated with baseline WMH and its progression.* To analyze the risk factors of WMH and WM health at baseline, we used stepwise linear regression to select predictors from cardiometabolic, clinical, and demographics risk factors. (We excluded HbA1c because of missingness.) Samples with missing data elements were removed from the study. Longitudinal WMH scans were measured cross sectionally and used as the measurement for WMH progression. We showed the scatterplot of WMH and age, as well as the boxplots of WMH for (i) patients with and without hypertension and (ii) male and female patients across the five age groups. An unpaired t-test was used to test the WMH differences between the presence and absence of hypertension (and male and female sex) within each age group. To quantify the effect of age, sex, and hypertension on baseline WMH, an ordinary least squares multiple regression model was constructed regressing the WMH on age, sex, and hypertension status. For the longitudinal analysis, we modeled the WMH progression using linear mixed effect models (R
package lmerTest, lme4) while controlling the same set of variables as in the baseline analysis.
Participants with only the baseline visit were removed from this part of the analysis.

Risk factors associated with baseline WM health measured by DTI and its progression. We conducted the same set of analyses for WM health as we did for WMH.

Mediation effect of WM health on WMH. We first showed the scatterplot of WM health and WMH. We then applied mediation analysis to study the mediation effect of WM health on WMH. The steps of the mediation analysis can be found here[93]. We fitted two regression models to log-transformed WMH while controlling for the direct causes of WMH and WM. One model included WM health, and one model did not. All variables were \( z \)-transformed, so that the standardized effect sizes can be compared across models.

Effect of midlife cardiovascular risk factors on WMH and WM health (pathway to potentially preventing progression of WMH). To study the effect of intervention on WMH, we added three midlife modifiers (midlife physical activity, midlife cognitive activity, and up-to-midlife smoking status) to the models. Statistically significant (\( P < 0.05 \)) midlife modifiers were kept in the model. To further explore the causal relationships between WMH and its potential midlife modifiers, we applied a CSD algorithm[19 20], Fast Greedy Equivalence Search[94] (FGES) with Fisher-z score and alpha 0.01, to generate the causal graph underlying the relationships. We ran 100 bootstrap iterations on the cohort and extracted relationships that appeared in more than half of the iterations.
RESULTS

We included 1564 participants, 1396 cognitively normal, 151 with mild cognitive impairment and 17 with dementia from the population-based sample of Mayo Clinic Study of Aging (MCSA). The demographics, clinical, late-life vascular risk factors, midlife risk factors, and imaging biomarkers of the participants by decade (30-49, 50-59, 60-69, 70-79, and 80-89) are shown in Table 2.1. 54%, 51%, 50%, 52%, and 58% of participants were male in the age range of 30-49, 50-59, 60-69, 70-79, 80-89, respectively. There were 10%, 31%, 53%, 67%, and 82% hypertensive participants across the age groups as expected. All the late-life chronic condition measurements and midlife risk factors were significantly different between the age groups ( \( P < 0.007 \) and \( P < 0.001 \) respectively). The percentage of participants diagnosed with diabetes, hypertension, and dyslipidemia increased as the baseline age increased; in the age group 80-89, there were 19%, 82%, and 86% of participants with these diagnoses, respectively. Six hundred and thirty-one (40.3%) participants had more than one scan and the rest of them only had baseline imaging data available. To make use of all the available data, we conducted analyses on both the cross-sectional and longitudinal data.

Table 2.1 Participants baseline characteristics table. Mean (SD) and Count (%) for continuous and categorical variables, respectively

<table>
<thead>
<tr>
<th>Demographics &amp; APOE</th>
<th>30-49, n=59</th>
<th>50-59, n=262</th>
<th>60-69, n=540</th>
<th>70-79, n=407</th>
<th>80-89, n=296</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>40.02(5.24)</td>
<td>54.64(2.46)</td>
<td>64.59(2.84)</td>
<td>74.51(2.91)</td>
<td>83.52(2.94)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>32(54%)</td>
<td>133(51%)</td>
<td>272(50%)</td>
<td>212(52%)</td>
<td>171(58%)</td>
<td>0.32</td>
</tr>
<tr>
<td>APOE e4</td>
<td>14(24%)</td>
<td>79(30%)</td>
<td>154(29%)</td>
<td>124(30%)</td>
<td>78(26%)</td>
<td>0.661</td>
</tr>
<tr>
<td>BMI</td>
<td>28.31(5.87)</td>
<td>29.21(5.4)</td>
<td>29.49(5.43)</td>
<td>28.21(4.47)</td>
<td>26.98(4.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Laboratory results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDL</td>
<td>102.42(31.32)</td>
<td>109.13(30.92)</td>
<td>103.76(29.77)</td>
<td>97.3(29.03)</td>
<td>90.89(27.32)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HDL</td>
<td>55.08(18.03)</td>
<td>55.64(17.85)</td>
<td>55.49(17.05)</td>
<td>55.76(17.36)</td>
<td>55.09(16.51)</td>
<td>0.991</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>119.56(70.17)</td>
<td>128.87(65.16)</td>
<td>130.76(62.01)</td>
<td>125.95(59.16)</td>
<td>123.17(53.95)</td>
<td>0.425</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>181.42(40.46)</td>
<td>190.57(33.87)</td>
<td>185.42(35.61)</td>
<td>178.26(36.32)</td>
<td>170.6(33.89)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HbA1c</td>
<td>5.87(1.32)</td>
<td>6.04(1.12)</td>
<td>5.93(1)</td>
<td>5.95(0.79)</td>
<td>5.97(0.76)</td>
<td>0.912</td>
</tr>
<tr>
<td>SBP</td>
<td>126.19(15.02)</td>
<td>131.48(16)</td>
<td>139.13(18.45)</td>
<td>141.15(18.5)</td>
<td>142.49(20.13)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DBP</td>
<td>78.42(10.96)</td>
<td>79.16(9.34)</td>
<td>78.2(9.98)</td>
<td>74.14(10.05)</td>
<td>71.12(10.86)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Current Clinical Diagnosis

| Cognitively Impaired | 2(3%) | 7(3%) | 29(5%) | 62(15%) | 68(23%) | <0.001 |
| Diabetes             | 3(5%)  | 31(12%) | 73(14%) | 74(18%) | 56(19%) | 0.007  |
| Hypertension         | 6(10%) | 82(31%) | 285(53%) | 273(67%) | 244(82%) | <0.001 |
| Dyslipidemia         | 20(34%) | 169(65%) | 424(79%) | 332(82%) | 254(86%) | <0.001 |

Midlife risk factors

| Midlife physical activity | 8.27(3.55) | 7.39(3.66) | 9.1(4.43) | 9.11(4.27) | 9.15(4.66) | <0.001 |
| Midlife cognitive activity | 17.07(6.59) | 20.29(8.27) | 18.62(9.22) | 20.94(8.61) | 20.87(9.49) | <0.001 |
| Smoking status         | 21(36%) | 92(35%) | 246(46%) | 205(50%) | 119(40%) | <0.001 |

MRI & DTI

| Abnormal WMH | 0(0%) | 0(0%) | 24(4%) | 65(16%) | 110(37%) | <0.001 |
| Baseline WMH  | 0.18(0.09) | 0.31(0.22) | 0.56(0.61) | 1.02(0.79) | 1.74(1.35) | <0.001 |
| Genu of the corpus callosum FA | 0.64(0.04) | 0.63(0.04) | 0.61(0.04) | 0.59(0.05) | 0.56(0.05) | <0.001 |
| Splenium of the corpus callosum FA | 0.69(0.03) | 0.69(0.03) | 0.69(0.04) | 0.68(0.04) | 0.66(0.05) | <0.001 |
| Body of corpus callosum FA | 0.62(0.03) | 0.61(0.04) | 0.6(0.04) | 0.59(0.04) | 0.56(0.05) | <0.001 |
| Anterior limb of the internal capsule FA | 0.59(0.03) | 0.59(0.03) | 0.59(0.03) | 0.57(0.04) | 0.56(0.04) | <0.001 |

Scan statistics

| Scan Interval, years | 1.25(0.44) | 1.31(0.45) | 1.73(0.77) | 1.78(0.87) |
| Participants with 2 Scans | 0(0%) | 70(27%) | 200(37%) | 129(32%) | 80(27%) |
| Participants with 3 Scans | 0(0%) | 11(4%) | 43(8%) | 62(15%) | 36(12%) |

Abbreviations: APOE e4: Apolipoprotein E epsilon 4; BMI: body mass index; HDL: high density lipoprotein; LDL: low density lipoprotein; SBP: systolic blood pressure; DBP: diastolic blood pressure; HbA1c: glycated hemoglobin.
Risk factors associated with baseline WMH and its progression

**Figure 2.2** Association between baseline WM hyperintensity with (A) age, (B) sex, and (C) hypertension

The overall distribution of log-transformed WMH by age is shown in the left panel, Figure 2.2A Starting from age 50, the WMH increases linearly as age increases. The right panel (Figure 2.2B and 2.2C) shows the difference in WMH between males and females, and between participants with and without hypertension (normotensive). Student t-tests were conducted to test the within-group differences as shown at the top of each boxplot. WMH significantly differed by sex in late-life ($P \leq 0.005$ in 60-70, 70-80, and 80-90 years of age). There were significant differences in WMH between hypertensive and normotensive participants for ages 60-70 and 70-80 ($P < 0.01$). In WMH models with all vascular risk factors including laboratory results shown in Table 2.1 as predictors, only presence of hypertension, increasing age, and female sex were significantly
associated with both baseline and progression of WMH. The final models are shown in the regression models #1 & #2 ($P<0.001$, Table 2.2). Neither diabetes nor hyperlipidemia showed a significant association with baseline WMH or WMH progression after adjusting for age, sex, and hypertension.

**Table 2.2** Results from two regression (baseline) and two mixed models (progression) on WMH and WM health, Genu-FA

<table>
<thead>
<tr>
<th>Model #</th>
<th>Outcome</th>
<th>Predictors</th>
<th>Estimated effect size</th>
<th>SE</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td># 1</td>
<td>Baseline WMH</td>
<td>Age</td>
<td>0.053</td>
<td>0.002</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>-0.183</td>
<td>0.033</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>0.172</td>
<td>0.037</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td># 2</td>
<td>WMH progression</td>
<td>Age</td>
<td>0.060</td>
<td>0.003</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>-0.207</td>
<td>0.049</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>0.194</td>
<td>0.053</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td># 3</td>
<td>Baseline Genu-FA</td>
<td>Age</td>
<td>-0.002</td>
<td>0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>0.012</td>
<td>0.002</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>-0.010</td>
<td>0.002</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td># 4</td>
<td>Genu-FA progression</td>
<td>Age</td>
<td>-0.002</td>
<td>0.0002</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>0.013</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>-0.011</td>
<td>0.0036</td>
<td>0.003</td>
</tr>
</tbody>
</table>

WMH measurements are log-transformed. Higher WMH and lower Genu-FA represent poor brain health. HTN: Hypertension. Other potential variables were dropped out from the stepwise regression. 1564 participants are included in baseline analysis where 1396 are cognitively normal; 740 participants are included in the progression analysis, where 632 are cognitively normal.

**Risk factors associated with baseline diffusion changes measured by DTI and its progression**
We originally considered 12 WM tracts to measure WM health and to evaluate their usefulness in predicting WMH. Because FA of all the WM tracts are highly correlated, we considered FA of the genu of the corpus callosum (Genu-FA) which consists of thin fibers in the anterior part of the corpus callosum and has greater vulnerability to aging and neurovascular damage[95]. Previous data supports the use of Genu-FA as an important marker of cerebrovascular disease[96 97]. We also considered three other tracts that were also predictive of WMH for comparison - FA of the splenium of the corpus callosum (Splenium-FA), FA of the body of the corpus callosum (Body-FA), and FA of the anterior limb of the internal capsule (ALIC-FA) as presented in Supplement material, section sensitivity analysis.
In a model with Genu-FA as an outcome and vascular risk factors as predictors, only age, sex, and hypertension were associated with Genu-FA. As shown in Figure 2.3A, Genu-FA decreases as age increases. Genu-FA significantly differed by sex in late-life with most significant findings in the 60-69 ($P<0.001$), and 80-89 ($P=0.007$) age groups (Figure 2.3B). There were significant differences in Genu-FA between hypertensive and normotensive participants for all ages across late-life ($P\leq0.006$) (Figure 2C). Predictions of baseline and progressive WM integrity through Genu-FA are shown in Models #3 & #4 in Table 2. Older age, female sex, and hypertension were significantly associated with lower baseline and longitudinal Genu-FA in both models ($P<0.001$, $P\leq0.003$ respectively).

**Genu-FA mediates the effect of hypertension and male sex on WMH**

![Figure 2.4](image)

**Figure 2.4 A** Relationship between baseline WMH and baseline Genu-FA in the pre-defined age groups. **B** shows the estimated local causal structure among age, sex, hypertension, WM health, and WMH, generated by the CSD algorithm.

When we looked at the relationship between WMH and Genu-FA as illustrated in the scatterplot of Figure 2.4A, we found high correlation between the two variables that increased with age. To
further investigate the relationship between Genu-FA and WMH, we built two standardized regression models of WMH, controlling for age, sex, hypertension. In model #5, we included Genu-FA but did not include in model #6 (Table 2.3). Our goal with this comparison was to understand the differences in standardized effect sizes between the two models. In model #5 from Table 2.3, Genu-FA was significantly associated with WMH ($\beta=-0.29, P<0.0001$).

Compared to model #6, where Genu-FA was not included, the effect size from male sex and hypertension both dropped by approximately 40% ($\beta=-0.1, 0.09$ to $-0.06, 0.06$). This mediation analysis demonstrated that the effect of sex and hypertension was mediated by Genu-FA. The above conclusion was also confirmed in longitudinal analysis (Supplementary Table S1).

Figure 2.4B shows the causal structure among the risk factors and WMH discovered by the CSD algorithm. A directed edge between two variables indicates a direct causal relationship. The relationships in the discovered graph are consistent with our previous finding as well as the mediation analysis.

**Table 2.3** Results from two standardized regression models on WMH with and without controlling for WM health, Genu-FA

<table>
<thead>
<tr>
<th>Model #</th>
<th>Outcome</th>
<th>Predictors</th>
<th>Standardized effect size</th>
<th>SE</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td># 5</td>
<td>Baseline WMH</td>
<td>Genu-FA</td>
<td>-0.29</td>
<td>0.020</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age</td>
<td>0.52</td>
<td>0.020</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>-0.06</td>
<td>0.017</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>0.06</td>
<td>0.018</td>
<td>0.0004</td>
</tr>
<tr>
<td># 6</td>
<td>Baseline WMH</td>
<td>Age</td>
<td>0.65</td>
<td>0.019</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>-0.10</td>
<td>0.018</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HTN</td>
<td>0.09</td>
<td>0.019</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

WMH measurements are log transformed before the standardized regression models.
Identifying factors related to the evolution of white matter hyperintensities

**Figure 2.5** The causal structure graph discovered using the CSD algorithm. The estimated standardized coefficients were labeled next to each edge.

Based on previous findings, we studied the relationship between midlife risk factors and WMH. We included three midlife indicators, midlife cognitive activity, midlife physical activity, and smoking status. Table S2 shows that smoking status was strongly correlated with WMH ($\beta=0.05$, \(\sigma=0.05\)).
and midlife physical activity was significantly associated with WM health (Genu-FA) \((\beta=0.06, P<0.008)\) after adjusting for other predictors.

To study the effects of risk and protective factors on WMH, causal relationships, not merely associations, are needed. Therefore, we applied the CSD algorithm to all the sets of predictor variables. On the basis of the Fast Greedy Equivalence Search [94] (FGES), the causal structure discovered by the CSD algorithm is shown in Figure 2.5 The complex relationships were described in one graph where an edge implies a direct causal effect. We then fitted multiple regression models to estimate the effect sizes for each edge.

While older age had an impact on all variables, male sex was associated with higher frequency of smokers, less cognitive activity, higher Genu-FA, and lower WMH. Physical and cognitive activity were highly correlated; there might exist other unobserved variables that are common between the two variables. Smoking (ever/never, measured up to midlife) was associated with greater later life WMH load, as has been observed in many other studies. The associations between hypertension and WM health/WMH that have been described in previous sections are also shown as a direct relationship in the graph. A key finding from the CSD graph was that greater physical activity in midlife was associated with healthier WM outcomes (mainly higher Genu-FA). The effect of greater physical activity had two pathways to better WM outcomes – 1) higher midlife physical activity was associated with higher Genu-FA (directly) and with lower WMH (through its effect on Genu-FA); 2) lower midlife physical activity was associated with higher BMI and higher BMI was associated with higher frequency of hypertension at the time of the scan which in turn contributed to lower WM health (lower Genu-FA and higher WMH).
The sex specific observations were interesting. Male sex was associated with higher frequency of smokers and had an impact on WM health through 1) direct impact on WMH and 2) through higher BMI contributing to greater frequency of hypertension. However, males in general had better WM outcomes as also confirmed by regression models suggesting that the net impact of male sex on WM outcomes was positive. While most of the relationships in the graph detected were plausible, there was one suggested relationship between higher cognitive activity and higher BMI that was not intuitive and should be interpreted with caution. Additional longitudinal data in future studies will be able to discern the relationship between physical activity, cognitive activity, and BMI which are highly inter-related and complex.

DISCUSSION
In this large dataset of multi-modal imaging and clinical information, we investigated the relationships among WM health, WMH, and their risk factors. The major conclusion of the study was to identify pathways, particularly measuring early brain changes, to prevent WMH. Specifically, our main findings were: i) the effect of older age, hypertension, and female sex on greater WMH were mediated by diffusion changes seen on DTI; ii) midlife physical activity can aid in maintaining better microstructural health (and lower WMH) directly as well as indirectly through reduction of BMI which is associated with lower frequency of hypertension; iii) measuring Genu-FA can aid in prediction of WMH trajectories in the population; iv) further, the causal modeling of complex relationships identified key pathways related to the evolution of WMH.

Diffusion changes predictive of WMH trajectories
It has been widely shown that hypertension impacts WMH [98-100], and there is also emerging evidence that vascular risk factors are associated with WM damage [95 101 102]. Our results from models 1 and 3 (Table 2.2) provide further evidence for the association between hypertension and WMH and its relationship with frontal WM (lower Genu-FA). Consistent with our findings, previous studies have reported associations of hypertension with lower FA in the genu of corpus callosum[103 104]. However, the actual mechanism underlying the association between hypertension and WM health is more complex and not clearly understood. Hypertension may cause vascular impairment leading to vascular remodeling and reduced vascular reserve, which may cause arteriosclerosis, microatheroma, and microaneurysms. These processes result in reductions of blood flow, which can in turn cause myelin damage and gliosis[105 106] and the pathologies often observed as WMH on MRI [107 108]. Our data also extends the past research from our group and others, showing that hypertension accelerated the progression of WMH[98 109 110]. In additional analyses, we observed that the impact of vascular health on WMH were mediated by WM health. Although hypertension was useful for prediction of future progression of WMH, WM microstructural integrity provides more information which is likely applicable earlier than WMH. Our findings are consistent with prior evidence showed that information captured by lower microstructural integrity was an independent predictor of conversion of normal appearing WM to WMH[72 111-113]. Furthermore, a previous study demonstrated distinct WM microstructural patterns with increasing WMH load even before the formation of lesions and pleiotropic effects, suggesting WM changes as the early measure of WMH[114]. Previous longitudinal studies also have shown that microstructural WM alterations measured by FA and MD are correlated with future WMH[65 72]. More recently, we also found a stronger association between Genu-FA and WMH that lend support of the frontal WM damage due to
systemic vascular health prior to WMH[115]. However, there was no established link for a mediating effect of Genu-FA on WMH that makes these study findings more unique. This new finding might indicate the importance of targeting early changes in WM as a predictor of future changes of WMH in middle-aged and older adults. These findings also support the usefulness of DTI in clinical prevention trials targeting modifiable vascular risk factors and the reduction of WMH.

**Sex differences in WM health**

The current study also showed a sex-specific association with microstructural integrity of WM and found a greater vulnerability in females with greater baseline WMH burden and progressive WMH. These findings extend our past research[71 98] and several others [116-118], suggesting that sex is an important factor contributing to WMH. The possible explanations include genetic and hormonal factors, sexual dimorphism in WM microstructure, and their influence on CVD. A higher genetic heritability of WMH has been shown in women compared to men[119]. The major hormonal changes in women are associated with pregnancy and menopause. Studies have suggested the link between hypertensive pregnancy disorders such as preeclampsia and WMH[120 121]. It is also possible that reduction in estrogen after menopause may have an influence and suggested that hormone replacement therapy might protect the WM integrity[121-123]. However, a recent study in women of the Kronos Early Estrogen Prevention Study failed to reveal any significant association of menopausal hormone treatments with WMH[124]. Although studies often considered a possible association of bilateral oophorectomy with WMH formation and progression[120], no significant evidence is available yet[125]. Evidence also suggested that
a higher frequency of arterial stiffness in women might lead to increased WMH burden through cerebrovascular remodeling[126].

Interestingly, the male effect on WMH was mediated through better WM health, suggesting that organization of WM fiber tracts are relatively preserved in men. The mediating effect of higher Genu-FA contributing to lower WMH in men has not been reported previously, to our knowledge. However, prior studies of the corpus callosum revealed inconsistent findings in association with sex in which some reported higher FA in men[127 128], others found higher FA in women[129 130]. Animal studies have shown that there is a greater proportion (and area) of myelinated fibers in males than females in the corpus callosum[131], which could explain greater white matter reserve in men to ongoing damage. Another possible mechanism may be the different protein composition in myelin sheaths and the greater turnover of oligodendrocytes observed in females[132]. Further research is warranted to investigate the sexual dimorphism of WM health and WMH.

Factors driving WM changes
**Figure 2.6** Important Relationships between midlife modifiers and WMH/WM health discovered by the CSD algorithm.

Blue and black arrows represent negative and positive (respective) causal relationships from the source to the target and the dashed arrow represents a statistically significant relationship with standardized effect size less than 0.2, while solid arrows represent standardized effect sizes at least 0.2.

The local relationships between midlife modifiers and WM measurements extracted from Figure 2.5 are shown in Figure 2.6. Consistent with previous findings [133 134], we found smoking is associated with WMH which were confirmed by both the regression and CSD analyses. These associations between cigarette exposure and WMH were also well established in two large samples from the UK Biobank [101 135]. Our findings further provide evidence for an association between smoking effects and poorer WM integrity. As expected, there was an association between smoking and lower Genu-FA, these anterior fibers are most susceptible to
normal and neurovascular aging[95 101]. In addition, midlife physical activity showed an indirect association with WMH through the path of WM health. There are also reports on the short and long term effects of physical activity on brain health. Evidence suggests that higher levels of physical activity attenuate the amount of WMH in middle aged and older adults[136 137]. A recent study from Atherosclerosis Risk in Communities (ARIC) suggests that greater levels of midlife and late-life physical activity may reduce cerebrovascular lesions in late-life[138] and hence aid in preventing cognitive decline. Although, there is very little evidence on the effect of midlife physical activity on midlife WM integrity, a more recent study suggests that midlife aerobic exercise may prevent or slow down the detrimental age-related WM fiber integrity degradation[139]. They also demonstrated greater FA in the genu, superior longitudinal fasciculus, and uncinate fasciculus in middle aged aerobically trained adults compared to middle aged sedentary.

It is also well known that hypertension affects WMH[98 110 140 141], however, the indirect pathway through BMI was not established before. Interestingly, BMI association with WMH has been reported[142 143], with higher volumes in women, especially in the deep WMH[116]. In the present work, using a causal model, without any prior information, the model identified an indirect relationship with WMH. Further investigations are needed to confirm this hypothesized mechanism and better establish the interrelationships. Altogether, our data revealed the neuroprotective mechanism by which better midlife physical and cognitive activity and not smoking may protect against cerebrovascular sequelae. Further, we provided evidence that DTI based measures may be appropriate as surrogate measures of WMH for clinical trials targeting vascular risk factors.
While studying the midlife modifiers of WMH, we implemented causal structure discovery to generate the causal graph among risk factors and WMH. The discovered graph was then compared with the results from traditional statistical analyses. The workflow that embedded CSD method provides benefits in the following ways: (1) Under a set of assumptions, the relationships discovered from CSD methods have a causal interpretation. For example, a causal relationship from X to Y has the interpretation that changes in X lead to changes in Y. We also note that these causal conclusions are potential hypotheses that still need to be validated with further experiments; (2) When the goal of the study is not merely predicting the outcome, knowing the structure among risk factors helps understand the underlying biological mechanism with reasonably high accuracy [144]; (3) In our current workflow, the CSD model was mainly used for validating the conclusion from well-established traditional methods. However, it can also be used for hypothesis generation. We can design the regression analyses based on the causal graph obtained through CSD methods. The workflow of how the CSD model was used may change the way of investigating the relationship between risk factors and targets of interests. The goal of the study is not to study all variables that influence WM outcomes. In our study, the framework is focused on finding early brain measures and early (risk and protective) factors that influence WMH trajectories. We tested this framework by modeling brain measures as outcomes (i.e. WMH and DTI) using regression models and also modeled the complex relationships using CSD with all factors in a single model. Our goal was to generate test hypotheses but not design treatment trials. Given the large number of variables, causal analysis can be leveraged to identify critical variables for further rigorous investigation. The method allows identification of key variables that can be potentially targeted for treatment trials.
**Strengths and Limitations**

The study has several strengths. First, the availability of large data set with diffusion, FLAIR, and T1 weighted MRI across the adult lifespan enabled us to study their complex causal relationship. Second, the information on the vascular risk factors based on their medical records through REP was crucial, although none of laboratory values showed significance. The third strength was the newly implemented workflow, which by combining the CSD methods with regression analysis, may improve the design of prevention trials.

The study also has some limitations. First, although we considered the longitudinal aspect, only 10 percent of participants had 3 or more observations, and more than half of participants have one available visit. To make use of all available data, we conducted pairs of analyses, one regression analysis for baseline observations, and one mixed effect analysis for longitudinal data. Ideally, longitudinal data can provide more knowledge when studying the mixed effect, and we have shown before that the temporal information naturally provided by the longitudinal data also helps the CSD method to achieve higher accuracy[144]. Second, we only considered the global measure of WMH and did not consider regional variability. Third, relatively shorter longitudinal observation intervals. Fourth, though we did extensive bootstrapping to confirm our study findings, we acknowledge that the uniqueness of this dataset with extensive clinical and imaging data limits our ability to replicate this in an independent dataset. Fourth, a possible limitation may be that many laboratory values were under control by medications, so they are likely to become less predictive compared with the disease diagnosis code. Fifth, the smaller sample size in the cognitively impaired participants reflects the population-based sample. In the future, we would be able to
extend the study to a large sample to investigate the complex interplay between regional measures of WM health and WMH.

**Conclusions**

The present study demonstrated a significant age, sex, and hypertension association with WMH. The relationship between vascular factors and WMH can be better explained by early changes in WM microstructural integrity. The midlife modifiers emerged as important components of WM health. Hence midlife may be the relevant window for prevention of late-life WMHs and measuring microstructural integrity using DTI can aid in better design and monitoring of prevention trials.
Chapter III: A novel method for Causal Structure Discovery from EHR data: A demonstration on type-2 diabetes mellitus

Adapted with permission from X Shen, S Ma, P Vemuri, MR Castro, PJ Caraballo, GJ Simon
Scientific reports 11 (1), 21025

Copyright 2021 Scientific Reports.
Overview

Modern AI-based clinical decision support models owe their success in part to the very large number of predictors they use. Safe and robust decision support, especially for intervention planning, requires causal, not associative, relationships. Traditional methods of causal discovery, clinical trials and extracting biochemical pathways, are resource intensive and may not scale up to the number and complexity of relationships sufficient for precision treatment planning. Computational causal structure discovery (CSD) from electronic health records (EHR) data can represent a solution, however, current CSD methods fall short on EHR data. This paper presents a CSD method tailored to the EHR data. The application of the proposed methodology was demonstrated on type-2 diabetes mellitus. A large EHR dataset from Mayo Clinic was used as development cohort, and another large dataset from an independent health system, M Health Fairview, as external validation cohort. The proposed method achieved very high recall (.95) and substantially higher precision than the general-purpose methods (.84 versus .29, and .55). The causal relationships extracted from the development and external validation cohorts had a high (81%) overlap. Due to the adaptations to EHR data, the proposed method is more suitable for use in clinical decision support than the general-purpose methods.

Introduction

Diagnostic tools based on artificial intelligence (AI) have recently demonstrated human-like performance[3 5 43 145], owing their high performance to their ability to synthesize information from many features. Consistent with this observation, national initiatives such as the Precision Medicine Initiative[146] and the Learning Health Systems[147] encourage the inclusion of a wide-range of information about the patient into the decision making process. Increasingly, clinical
decision support systems start to include treatment planning and selection tools\[148\]. Such tools require causal knowledge, not merely the associations (correlations). Intervening on correlates rather than causal factors of the disease leads to lack of efficacy, under- or overtreatment, and in worst case, to iatrogenic harm\[149\].

The gold standard for discovering causal relationships is conducting a randomized clinical trial or elucidating the underlying biochemical pathways. In many cases, clinical trials are impractical, unethical, if not outright impossible. Computational causal structure discovery (CSD) methods to discover causal relationships have demonstrated great success in many domains\[74-76\] and their application to EHR data could offer a solution for causal discovery from observational real world medical data. However, to unlock their full potential, these general-purpose algorithms need to be adapted to address study design and data quality challenges specific to the EHR data.

We propose an algorithm with three adaptations. First, we incorporate study design considerations. EHR data as it exists in the system does not follow any study design. Billing codes in particular are recorded for reimbursement purposes and do not distinguish between new incidences and pre-existing conditions. Understanding this difference is critical for study design. Second, \textit{time stamps can be unreliable}. The time stamp of a diagnosis often does \textbf{not} coincide with the onset time of the disease, but rather reflects the documentation time. In some cases, the temporal ordering of diseases may be reversed. Partly for this reason, general purpose CSD algorithms applied to the EHR data occasionally report “causal” relationships that are in the opposite direction of the natural disease progression. Third, \textit{general-purpose CSD methods sometimes fail to orient edges}. Even when a clear causal direction exists and is not masked by data artifacts, CSD algorithms can have
difficulty distinguishing the cause from the effect due to statistical equivalence[9]. Leveraging the longitudinal nature of EHR data and incorporating time information as part of the causal discovery process can enhance the identification of edge orientation.

In this paper, (1) we propose a data transformation procedure that distinguishes new incidences from pre-existing conditions, which allows us to determine the temporal order of the disease-related events despite the inaccurate (or rather noisy) timestamps in the EHR data. (2) We then present modifications to an existing CSD method, (Fast) Greedy Equivalence Search (GES)[51 150], to infer the direction of causal relationships more robustly using longitudinal information and takes the above study design considerations into account.

We demonstrate this methodology through the clinical example of type-2 diabetes mellitus (T2D), its risk factors and complications. T2D is an exceptionally well-studied disease with numerous clinical trials having produced a vast knowledge base, making the evaluation of the methodology possible. The goal of this work is not to uncover new causal relationships in diabetes, but to present a novel methodology for discovering causal relationships from EHR data that are sufficiently robust to support model development for clinical decision support tools. While we use T2D as our use case, we expect our methods to generalize to other diseases, typically chronic diseases, that exhibit similar characteristics and suffer from the same EHR shortcomings.

**Methods**

**Study Source and Population**
This retrospective cohort study utilized EHR data sets from two independent health systems, Mayo Clinic (MC) in Rochester, Minnesota and M Health Fairview (FV) in Minneapolis, Minnesota. Two 2-year time windows 2003-2004 and 2006-2007 for MC; and 2008-2009, and 2011-2012 for FV were defined. Dates for the time windows differed between MC and FV due to data availability. We extracted diagnoses, prescriptions, laboratory results, and vital signs from the two EHR data sets with the same inclusion and exclusion criteria: patients must have at least two blood pressure measurements, one before the first time window and one after the second time window; aged 18+ at the end of the first time window; and sex and age must be known. Figure 3.1A shows an overview of the study design of MC EHR (the study design for FV is similar). We used the MC EHR as the development cohort.

Figure 3.1 Study design and evaluations
A. Overview of the study design for Mayo Clinic (MC) EHR. B. The workflow of the internal evaluation. Three methods FGES+raw, FGES+transf, and the proposed algorithm were compared using stability, precision, and recall. Orange color highlights the proposed method (Method 3). C. The workflow of external comparison. The proposed method was applied to two datasets, MC and M Health Fairview (FV), and the resulting graphs were compared.

Variables

Diagnosis codes are aggregated into the disease categories of obesity, hyperlipidemia, pre-diabetes, type 2 diabetes mellitus, coronary artery disease, myocardial infarction, heart failure, chronic renal failure, cerebrovascular disease, and stroke based on ICD-9 and codes following our previous work[151]. Medications indicated for the above conditions were rolled up into NDF-RT therapeutic subclasses. Relevant laboratory results and vital signs were categorized based on cutoffs from the American Diabetes Association guidelines[152].

Causal Structure Discovery

A relationship between two events is causal if manipulating the earlier event causes the other (later) event to change. For example, prescribing a medication reduces the probability of downstream events (complications). Causation differs from association. For example, blood sugar is associated with risk of stroke: diabetic patients with higher blood sugar have a higher risk of stroke; however, this relationship is likely not causal in diabetic patients since attempts to reduce the risk of stroke by reducing blood sugar consistently failed in clinical trials[153 154]. If two events share a common cause (a confounder) and are not otherwise causally related, then manipulating one event
will not affect the other variable as long as the common cause remains unchanged. The confounder can be observed or latent. The term causal structure refers to the set of all existing causal relationships among all events and can be visualized as a graph. The causal graph consists of nodes, which corresponds to events, and the nodes are connected by edges that denote causal relationships. General-purpose CSD methods are designed to work with observational data to derive a causal structure that are consistent with the joint probability of the data.

Several general-purpose CSD algorithms have been proposed and the interested reader is referred to the Supplements II where we present an overview of the major methods. In this work, we focus on (Fast) Greedy Equivalence Search (FGES) as the comparison method, because we previously found it to outperform other CSD methods[144]. Briefly, FGES finds the optimal causal graph by a greedy search guided by a goodness-of-fit score (e.g. BIC or BDeu) over all possible graphs. Particularly, it starts with an empty graph, and iteratively adds individual edges that maximize the score given the current graph, until adding edges no longer improves the score. Then, FGES iteratively removes individual edges that maximizes the score, until edge removal ceases to improve the score. The output of FGES is a pattern, which can contain undirected edges, where the causal effect direction could not be determined due to statistical equivalency. FGES has good mathematical properties and been shown to be consistent under a set of assumptions[26 51].

Proposed Methods

The workflow of the proposed methods is described in Figure 3.1B, method 3 (colored in orange). We propose two methods, a data transformation and a causal search method. The former method transforms the longitudinal EHR data into disease-related events, so that we can determine the
temporal ordering of events (diseases) despite inaccuracies in the EHR data and extracts all pairs of diseases where a clear precedence ordering exists. The search method constructs the causal graph using the transformed data and the set of precedence pairs.

**Data Transformation Method**

A disease-related event is defined as a diagnosis, a prescription, an abnormal lab result, or abnormal vital sign. An event is incident if it occurs in the second time window but is not present in the first time window although the patient is observed in the first time window. Conversely, a disease event is pre-existing if the patient presented with it in or before the first time window. An event $A$ precedes another event $B$ if among patients who have both $A$ and $B$ in the second time window, $B$ is significantly more likely to be incident than $A$. Note that precedence implies neither causation nor association; however, if a causal effect exists, it must follow the precedence direction.

Formal mathematical definitions of these concepts can be found in the Supplement I. The output from this step is (i) an event-based data set consisting of the incident and pre-existing conditions for each patient in each of the two time windows, (ii) a set $\mathcal{C}$ of precedence relationships of all pairs $(v_i, v_j)$ of events for which event $v_i$ clearly precedes $v_j$.

**The proposed CSD Search Algorithm**

Given $\mathcal{C}$, we construct the causal graph $\mathcal{G}$ by iteratively adding edge $(v_i, v_j)$ from $\mathcal{C}$ that maximizes the goodness of fit of $\mathcal{G}$. The orientation of this edge must be consistent with the precedence relationship, namely from $v_i$ to $v_j$. The goodness of fit is defined by the BIC criteria.
Let $X^{(1)}, X^{(2)}$ denote the data sets collected in the two distinct time windows, where $X^{(2)}$ follows $X^{(1)}$. The likelihood of the $G$ is

$$
\mathcal{L}(G|X^{(1)}, X^{(2)}) = P(X^{(2)}, X^{(1)}|G) \\
= P(X^{(2)}|X^{(1)}, G) P(X^{(1)}|G) \\
= \prod_s \prod_v P(v_s^{(2)}|x_s^{(1)}, G) P(x_s^{(1)}|G) \\
= \prod_s \prod_v P(v_s^{(2)}|pa(v, G)_s^{(1)}) P(x_s^{(1)}|G), \tag{4}
$$

where $x_s^{(t)}$ is the observation vector for subject $s$ at the cross-section $t$; $v_s^{(t)}$ is the observation of variable (event) $v$ for subject $s$ at the cross-section $t$; and $pa(v, G)_s^{(1)}$ is the observation vector for the parents of $v$ in the causal structure $G$, at cross section 1 for subject $s$.

The algorithm estimates $P(v_s^{(2)}|pa(v, G)_s^{(1)})$ using logistic regression on the subjects that do not have $v$ at the first cross section and are under observations for both cross sections. For subjects who have $v$ at the first cross section, the probability of having $v$ at the second cross section is 1.

Since $G$ represents the transition graph, the term $P(x_s^{(1)}|G)$ is a constant.

Finally, the BIC score is

$$
\text{BIC}(G) = -2\ln \mathcal{L}(G|X^{(1)}, X^{(2)}) + \ln(n)|G|, \tag{5}
$$
where $n$ is the number of observations that are common in the two cross sections, and $|\mathcal{G}|$ is the number of edges in the causal structure $\mathcal{G}$.

**Algorithm 3.1** The proposed causal search algorithm.

Mathematically, Algorithm 3.1 describes the proposed algorithm for constructing the causal graph $\mathcal{G}$. $\mathcal{G}$ is a directed acyclic graph (DAG), with nodes representing variables and edges representing causal effects between a pre-existing and an incident variable.

**Evaluation**

**Clinical evidence**

The standard way to evaluate CSD methods is to compare the resulting graph to a gold standard graph. However, such a gold standard graph does not exist and possibly many relationships are unknown. However, there exists (i) Associative Evidence: a large body of observational studies
documenting risk factors and outcomes for diabetes. Results from these studies have already been distilled into summaries.[21] (ii) Clinical trials can support both the existence (positive) and also the lack (negative) of hypothesized causal relationships. We compiled a list of causal relationships from clinical trials considering 175 clinical trials with a primary endpoint of any of the conditions we studied, including composite end points. We excluded trials with inclusion criteria that are too strict (trial results would not generalize to our population) and the interventions that are out of the scope of our study. 14 trials remained yielding 19 positive and 18 negative causal relationships. These trials and the evidence they produced are listed in Supplement III, Table S1. These relationships are used as causal evidence to compute recall.

**Internal Evaluation**

We evaluated the method and the resulting graphs from the following four perspectives.

*Stability.* We run 1000 bootstrap replicas on the development cohort. An edge has ambiguous orientation if it is present in at least half of the 1000 graphs (edge is not noise) and both orientations appear in at least 30% of the graphs that contain this edge (it does not have a dominant direction). We report the percentage of ambiguous edges.

*Precision.* Based on the causal graph derived from the training cohort, an edge is incorrect if there is no associative evidence of a relationship between the two events; or if causal evidence specifically indicates the lack of a causal relationship. We define precision as one minus the proportion of incorrect edges among the discovered edges.
Causal recall. Causal recall is computed on a single graph discovered from the training cohort, quantifying the percentage of the known causal relationships discovered. A known causal relationship from $A$ to $B$ is discovered if there is a node in the graph that maps to $A$, another node that maps to $B$ and (a) a direct causal relationship $A \rightarrow B$ in the graph exists or (b) a causal path $A \rightarrow X \rightarrow B$ exists and no causal evidence states that in patients with $X$, $A$ does not cause $B$. For example, if the evidence states that blood pressure (without specifying whether it is systolic or diastolic) increases the risk of stroke, then the path $\text{sbp} \rightarrow \text{cevd} \rightarrow \text{stroke}$ would satisfy this relationship.

Associative recall. Associative recall is also computed on a single graph discovered from the training cohort and it quantifies the percentage of known associative relationships that can be explained by the discovered causal graph. An associative relationship between $A$ and $B$ is explained by the graph if there is a node in the graph that maps to $A$, another node that maps to $B$, and a path between $A$ and $B$ exists in the graph.

External Validation

We performed 1000 bootstrap replications on both data sets independently using the proposed method. On each data set, all edges from the 1000 graphs were pooled, resulting in two sets of pooled edges. We compared these two sets and pointed out the edges that were discordant between the MC and FV data.

Method Comparison
Figure 3.1B depicts an overview of the method comparison. Three methods are compared, (i) **FGES+raw**: FGES is applied directly to the raw data; (ii) **FGES+transf**: data is transformed using the proposed transformation method and FGES is applied to the transformed data; and (iii) **Proposed**: the proposed search algorithm is applied to the transformed data. Comparing FGES+raw and FGES+transf isolates the effect of the proposed transformation method, and comparing FGES+transf and Proposed highlights the effect of the proposed search algorithm.

**Results**

**Baseline characteristics**

Table 3.1 presents descriptive statistics for the MC and FV data sets at the end of the first time window and incidence rates for the diseases in the second window. Differences between datasets are tested through the t-test (for age) and the chi-square test (all other variables).

**Table 3.1 Characteristics of the MC and FV data sets.**

For age, mean (sd) is indicated; for the disease-related events, percentage (%) of positive is indicated. New events rate at the second time windows is reported.

<table>
<thead>
<tr>
<th></th>
<th>MC (N = 57332)</th>
<th>FV (N = 7946)</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>events in window 1</td>
<td>new events in window 2</td>
<td>events in window 1</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>48.1 (18.2)</td>
<td>50.4 (14.6)</td>
<td>0.000</td>
</tr>
<tr>
<td>male</td>
<td>0.43</td>
<td>0.34</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Vitals &amp; labs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI ≥ 25 &amp; ≤ 30</td>
<td>27.1</td>
<td>2.9</td>
<td>27.5</td>
</tr>
<tr>
<td>BMI ≥ 30</td>
<td>32.6</td>
<td>3.6</td>
<td>43.1</td>
</tr>
<tr>
<td>SBP ≥ 140</td>
<td>10.3</td>
<td>3.4</td>
<td>4.5</td>
</tr>
<tr>
<td>DBP ≥ 90</td>
<td>2.3</td>
<td>1.0</td>
<td>1.6</td>
</tr>
<tr>
<td>LDL ≥ 130</td>
<td>18.4</td>
<td>3.6</td>
<td>15.4</td>
</tr>
<tr>
<td></td>
<td>Number of distinct edges</td>
<td>Ambiguously oriented</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------</td>
<td>----------------------</td>
<td></td>
</tr>
<tr>
<td>FGES+raw</td>
<td>125</td>
<td>45%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.2** Directional stability.

The table shows the number of distinct edges that appeared in half of the 1000 bootstrap replications, and the percentage of ambiguously oriented edges.
Directional stability

The proposed data transformation reduced the percentage of ambiguously oriented edges from 45% to 24%, and finally, the proposed search method eliminated ambiguously oriented edges (Table 3.2).

Table 3.3 Metrics from clinical evidence

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Associative Recall</th>
<th>Causal Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. FGES+raw</td>
<td>0.294</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2. FGES+transf</td>
<td>0.549</td>
<td>0.985</td>
<td>1.000</td>
</tr>
<tr>
<td>3. Proposed</td>
<td>0.838</td>
<td>1.000</td>
<td>0.947</td>
</tr>
</tbody>
</table>

Correctness and completeness

Table 3.3 shows the precision, associative recall and causal recall of the graphs discovered by the three methods. All three methods achieved almost perfect recall; FGES+raw achieved the lowest precision of 0.294: less than third of the events reported as causally related are even associated. By using the proposed transformation, the precision increased to 0.55, but almost half of the reported causal relationships are still incorrect. Finally, the proposed method achieved a precision of 0.838. We present the causal graph discovered by the proposed methods in the Supplement IV, Figure S1.
We compared the graphs discovered from the MC and FV data sets. There are 74 distinct edges that were discovered from at least one of the data sets. Sixty (81%) edges coincided across the two datasets, while 14 (19%) differed. Table 3.4 shows the discordant edges, the percentage of bootstrap iterations in which the edge was present and the main reason for the discordance.

**Table 3.4 External Validation.**

The table describes the edges that were discordant between the Mayo Clinic (MC) and M Health Fairview (FV) data sets. It shows the percentage of the bootstrap iterations in which the edge was discovered at MC and FV and a brief reason for the discrepancy.

<table>
<thead>
<tr>
<th>Edge</th>
<th>Discovery %</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC</td>
<td>FV</td>
</tr>
<tr>
<td>hdl→trigl</td>
<td>0</td>
<td>91.7</td>
</tr>
<tr>
<td>htn.dx→crf</td>
<td>88.5</td>
<td>0.1</td>
</tr>
<tr>
<td>trigl→dm.dx</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>trigl→hl.tx</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>ldl→hl.dx</td>
<td>72.1</td>
<td>0</td>
</tr>
<tr>
<td>fasting.125→dm.dx</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>trigl→fasting.125</td>
<td>99.5</td>
<td>0.2</td>
</tr>
<tr>
<td>dbp→htn.tx</td>
<td>91.5</td>
<td>0</td>
</tr>
<tr>
<td>sbp→hl.tx</td>
<td>99.3</td>
<td>1.7</td>
</tr>
<tr>
<td>sbp→htn.tx</td>
<td>100</td>
<td>29.1</td>
</tr>
<tr>
<td>trigl→htn.tx</td>
<td>83.7</td>
<td>0</td>
</tr>
<tr>
<td>chf→mi</td>
<td>0</td>
<td>67.6</td>
</tr>
<tr>
<td>ldl.dx→trigl</td>
<td>0</td>
<td>87.6</td>
</tr>
<tr>
<td>hl.tx→cad</td>
<td>0</td>
<td>74.3</td>
</tr>
</tbody>
</table>

There are three broad reasons for differences in edges. The main reason, affecting half of the edges was that of policy differences. These include preferred lab results (A1C vs FPG) and decisions regarding therapeutic interventions. The second reason, affecting four edges, is a lack of clear
precedence in the relationships among the events. For example, the abnormal TG→HL treatment edge was not discovered at FV because the first abnormal TG precedes or follows the HL treatment in statistically equal proportions. The final reason, affecting the remaining three edges, is differential degree of confounding between the two sites. For example, SBP is a confounder of CHF and MI. When the algorithm fails to detect the SBP→MI edge, the effect of SBP on MI was shown through CHF (which depends on SBP more than MI). For the HL diagnosis→TG edge, the common cause is BMI, and for the HL treatment→CAD edge, it is LDL. The reason for differential confounding was likely a combination of population and institutional differences as well as data artifacts.

**Discussion**

We proposed a new data transformation method and a new search algorithm specifically designed for EHR data, and showed how the resulting graph achieved close to 90% precision (90% of the edges were correct), almost 100% recall (the graph could explain all known associations and almost all known causal relationships), and the graph was remarkably stable in face of data perturbation (no edge disappeared or changed direction). Due to its built-in facility to handle inaccuracies in the EHR data and study design considerations, our method outperformed general purpose methods by a large margin.

While the two graphs from the two independent health systems are reassuringly similar, small differences exist. None of these differences implies an incorrect physiological or pathophysiological effect. Among the 14 edges that differed, seven captured differences between the population and the institutions, such as institution-specific triggers for prescriptions and the
use of different laboratory tests for the same purpose (fasting plasma glucose versus A1c). Depending on the goal of the modeling, it may be desirable to include such differences. We believe that the discovered causal graphs offer adequate information about causal (including confounding) factors to support the development of clinical decision support models and can also support clinical research efforts.

*Generalizability beyond diabetes.* The proposed method was demonstrated on type 2 diabetes, but it can generalize to other applications as long as the target application benefits from some of the improvements: reducing the impact of inaccuracies in the EHR data, accounting for the temporal ordering of events and distinguishing pre-existing and incident conditions.

*Future work:* The algorithm requires longitudinal data with at least two time-points. Different disease and their symptoms might manifest at different rates, incorporating this knowledge into the discovery process may enhance the performance of the algorithms. Further, variable semantics (such as SBP and DBP being measures related to hypertension) is an essential component of the proposed algorithm, but it is not always available in a computable form. Further, both datasets in this study are from the Midwest with a predominantly white patient population. The generalizability of the discovered causal relations can be further tested by examining a broader patient population.

**Conclusions**

We have demonstrated that the graph produced by the proposed transformation and search algorithm is more stable across bootstrap iterations and as complete as other methods yet it
contained substantially fewer errors (had higher precision) than graphs produced by general-purpose methods. The resulting graph was successfully validated using longitudinal EHR data from an independent health system. We conclude that the proposed method is more suitable for use in clinical studies using EHR data.

**Funding acknowledgement**

This work is supported by the National Institutes of Health (NIH) grants AG056366, TR002494, and LM011972. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NIH.
Chapter IV: Missing-Mechanism-Adjusted SEM (MMA-SEM): Enhancing Causal Inference in EHR
Overview

Causal inference aims to estimate causal effects and draw causal conclusions from observational data. Electronic Health Records (EHR) offer a valuable source for this purpose. However, the challenges arise when there are missingness in the observed dataset. We demonstrate that even with a correctly specified data generation structure, missing data can still lead to biased effect size estimates. To address this issue, we proposed a new causal inference method that leverages both the missing information and available observational data to estimate the causal effects, called Missing-Mechanism-Adjusted structural equation modeling (MMA-SEM). In this paper, we compared the performance of the MMA-SEM with the state-of-art method, Structural Equation Modeling. We evaluated these methods for their accuracy in estimating the latent causal effect sizes in the presence of various missingness rates, sample sizes and assumption scenarios. The proposed causal inference method achieved a higher insignificant rate and reduced mean square error in estimating latent treatment effects on the synthetic data. In EHR datasets, MMA-SEM also outperforms other methods when the prevalence of missing data exceeds 6%. In summary, our proposed method, which integrates knowledge of missing data mechanisms, effectively reduce estimation error in EHR datasets. It presents a significant improvement over traditional approach.

Keywords—causal inference, EHR, latent effects, SEM

Introduction

Causal inference has emerged as a widely recognized alternative to randomized clinical trials for estimating causal effects, particularly in instances where randomized clinical trials are impractical due to ethical considerations or high cost[155]. While inverse probability weighting[156 157] and
propensity score matching[15] are popular causal inference approaches, this paper focuses primarily on structural equation modeling (SEM)[158-160]. SEM stands out as it can examine relationships among all observed and latent variables, offering a more comprehensive approach than propensity matching or weighting methods, which primarily focus on estimating (typically) a single or a few treatment effect(s). The accuracy and generalizability of these models fundamentally rely on the quantity and quality of the data.

The rise of electronic health record (EHR) systems has made vast amounts of data accessible. However, as the EHR system is primarily designed for reimbursement purposes, leveraging it for inference and biomedical knowledge discovery remains challenging[161-163] for two main reasons. (1) There is an interpretation gap between the variables collected in the EHR and disease concepts our biomedical knowledge is based on. Consider, for instance, the task of analyzing the impact of hypertension (HTN) disease on stroke. Relevant biomarkers related to hypertension include systolic and diastolic blood pressure (SBP, DBP), HTN-related diagnosis codes, and prescriptions of anti-HTN medication. However, none of these observable measurements alone truly represent the severity of the disease. Severity is unmeasured (latent) in the EHR, thus we focus on latent-variable SEMs, where for each disease, information from the corresponding observed variables is summarized into a latent variable. (2) Missing data[164-168] in the EHR extends beyond instances where data omissions are random events, such as inadvertently forgetting to document some data elements [169]. Missingness in EHR often relates to underlying condition. For example, when a physician presumes that the patient does not suffer from a disease, the corresponding lab tests will not be ordered, resulting in missing data. Such missingness depends on the underlying unobserved lab result, which is more likely to be normal.
Method to address missing data include either discarding the missing records or imputing the missing values. Most imputation methods are designed for a specific missingness mechanism, and they work reasonably well when their underlying assumptions are met[170 171]. However, as we demonstrated in this paper, missingness can affect the distribution of the variable to a degree where the causal structure of the problem gets distorted. Even when the true data generating structure is specified correctly, imputing missing values under the wrong missing data mechanism can alter the association structure in the data, possibly deleting correct edges or opening new “causal” pathways. Failure to correctly account for the missingness can lead to biased results in causal inference.

In this study, we introduce a novel causal inference method named Missing-Mechanism-Adjusted SEM (MMA-SEM) that aims to estimate the effects among latent variables in the presence of missing data. This method operates under the assumption that the structure among variables and the causes of missing values (which we refer to as the causal structure of missingness) are known. While this is often the case in healthcare, to further validate our approach, we also conducted experiments under conditions where the causal structure of missingness was intentionally mis-specified, simplified to the assumption that missingness only depends on the underlying disease severity. Innovations in this work is that the proposed algorithm incorporates the causal structure of missingness, leveraging both missing and observed data to iteratively refines missing data, latent variable values, and coefficient estimates. We evaluate the proposed algorithm on synthetic datasets and a real EHR dataset.
**Material & Method**

**Structural Equation Modeling**

SEM is a causal inference method that integrates factor analysis and multiple regression to analyze the structural and causal relationships between observed and latent variables. SEM allows for the specification of the relationships between observed and latent variables through a series of structural equations, represented by a directed graph, as shown in Figure 4.1. More formally, let $Z$ be an $N \times p$ matrix of $p$ latent variables. These variables are causally related, and their relationships are a priori known and encoded in a weighted adjacency matrix $B$, where weights $B_{ij}$ denote the causal effect of $Z_i$ on $Z_j$. When no edge exists from $Z_i$ to $Z_j$, $B_{ij} = 0$. Since the causal relationships among the latent variables are known, the zero elements in $B$ are known, but the non-zero weights need to be estimated. The main objective of the SEM problem is to estimate these weights.

Similarly, let $X$ be an $N \times q$ matrix of $q$ observed variables. The causal relationships between $Z$ and $X$ are a priori known or can be assumed. Let $A$ be a $p \times q$ weighted adjacency matrix with $A_{ij} = 0$ iff no edge exists from $Z_i$ to $X_j$ such that,

$$X = ZA + \xi,$$

where $\xi$ are errors. Since the causal relationships are known, the zero elements in $A$ are known, but the non-zero weights in $A$ need to be estimated. In the example graph in Figure 4.1, $z_i$ represents latent variables, $x_{i,j}$ represents the observed variables, and directed edges, e.g., $z_1 \rightarrow x_{1,1}$, indicates a non-zero coefficient of latent variable $z_1$ in the structural equation for $x_{1,1}$, such that $x_{ij} = a_{ij}z_i + \epsilon_{ij}$.
Formal Problem Definition

In this paper, we focus on the estimation of causal effects among latent variables, a key distinction from Confirmatory Factor Analysis (CFA), which does not consider these latent effects. These causal effects are denoted as $\beta_k$ in Figure 4.1, the direct edges between $z_i$. $z_i = \sum_k \beta_k z_k + \zeta_i$. Given the structural relationships among latent and observed variables, our objective is to estimate these causal effect sizes among latent variables, while acknowledging the potential of missingness in the observed variables.

Proposed MMA-SEM

Expanding upon the foundational framework of SEM, the proposed MMA-SEM iteratively learns relationships among latent variables $Z$, observed variables $X$ (with some observations missing), and indicators of data missingness $Y$, while concurrently determining the coefficients that describe the relationship among them. As described in Figure 4.1, the proposed algorithm aims at optimizing these variables and coefficients based on observed data. To achieve this, MMA-SEM employs the alternating block gradient descent method.
Figure 4.1 The structure of the latent and observed variables.
The Z denotes the unobserved latent variables, and the arrows represent the directed causal relationships. The observed variables are denoted as X in the figure. Noted that the figure only shows 4 latent variables while the simulation study included 20 latent variables.

Causal structure of missingness. We introduce Y, an $N \times r$ matrix of $r$ observed binary missing indicator variables. They indicate the missingness in $X$. Let $C$ be a $p \times r$ weighted adjacency matrix linking $Z$ to $Y$ be a set of causal edges from latent variables to missingness indicators. Let $c_0$ be a vector of intercepts such that,

$$\logit(Y) = 1_N c_0 + ZC,$$

where the logit function is applied elementwise. The zero elements in $C$ are a priori known, the non-zero elements need to be estimated. Thus, the missingness of $X$ depends on $Z$ and $Y$. 
We solve the problem by minimizing the negative log likelihood

$$\min_{\mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}, c_0} \frac{1}{2} \| \mathbf{Z} - \mathbf{ZB} \|^2 + \frac{1}{2} \| \mathbf{X} - \mathbf{ZA} \|^2 + (-1) \left[ Y_{ij} U_{ij} - \log(1 + e^{u_{ij}}) \right].$$

where $\mathbf{U} = \mathbf{1}_N c_0 + \mathbf{ZC}$.

We solve this optimization problem using ADMM. In order to separate the objective into easily solvable parts, we create $\mathbf{V}$, an $N \times p$ matrix, as a “twin” of $\mathbf{Z}$, and, similarly, we create $\mathbf{W} = \mathbf{1}_N c_0 + \mathbf{VC}$, as a “twin” of $\mathbf{U}$, $\mathbf{W}$ differs from $\mathbf{U}$ in that it uses $\mathbf{V}$ instead of $\mathbf{Z}$. Under the constraint that $\mathbf{V} = \mathbf{Z}$, the objective becomes

$$\min_{\mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{C}, c_0} \frac{1}{2} \| \mathbf{Z} - \mathbf{ZB} \|^2 + \frac{1}{2} \| \mathbf{X} - \mathbf{ZA} \|^2 + (-1) \left[ Y_{ij} W_{ij} - \log(1 + e^{w_{ij}}) \right]$$

s.t. $\mathbf{V} = \mathbf{Z}$.

The augmented Lagrangian is

$$\mathcal{L} = \frac{1}{2} \| \mathbf{Z} - \mathbf{ZB} \|^2 + \frac{1}{2} \| \mathbf{X} - \mathbf{ZA} \|^2 + (-1) \left[ Y_{ij} W_{ij} - \log(1 + e^{w_{ij}}) \right] + \frac{\rho}{2} \| \mathbf{Z} - \mathbf{V} \|^2 + \mathbf{1}_N^T (\Lambda \circ (\mathbf{Z} - \mathbf{V})) \mathbf{1}_p$$

where $\circ$ denotes element-wise multiplication, $\Lambda$ is an $N \times p$ matrix of Lagrange multipliers, $\Lambda \succ 0$, and $\rho$ is a scaler. The objective is to minimize $\mathcal{L}$ in $\mathbf{Z}, \mathbf{V}, \mathbf{A}, \mathbf{B}, \mathbf{C}$ and $c_0$, under the constraint that $V_j$ is standard normal.

Details of the ADMM update equations and the initialization of the problem are included in the Supplement.

**Experiment Design**

We designed two experiments, one based on synthetic data with known true parameter values and another experiment using real-world EHR data. The synthetic data experiment assessed the
performance of MMA-SEM under varying missingness percentages, different sample sizes, and its resilience under assumption violations. In the EHR dataset analysis, we evaluated the method's performance in real-world data scenarios, focusing on preserving the correct causal structure and accurately estimating effect sizes among latent variables. All analyses were implemented in R (version v4.0.0; R package: mice, v.3.14.0 for multiple imputation[172], lavvan v.0.6-9 for SEM implementation[173])

**Synthetic Data Experiment**

To evaluate the performance of the proposed method and compare it with SEM-based methods, we conducted a simulation study. The data were simulated through the following steps. (1) We simulated twenty latent variables and the relationships among them randomly, denoted as Z and \( \beta \) in Figure 4.1. (2) We constructed three observed variables for each latent variable, denoted as X in the Figure 4.1. The number three was chosen in accordance with the guidance for factor analysis. We also ensured the correlation among observed and latent variables were similar in scale to that of the real EHR data. (3) Finally, we masked out observations to create missingness. Detailed simulation parameters can be found in the supplemental materials.

With the synthetic data, we conducted two studies. In study 1, we systematically varied the degree of missingness and introduced a violation of the algorithm's underlying assumptions by assuming an unknown missing data generation mechanism. Specifically, MMA-SEM assumes the correct missingness generating mechanism, where ancestor (topologically) latent variables determine the missing probability of observed variables. However, in this part, the corresponding latent variables that determines the missingness probability were assumed unknown. Instead, the algorithm assumed the missingness was generated by the parent latent variables. For example,
MMA-SEM assumes the missingness of $x_{4,3}$ is determined by $z_1$, but in reality, it is $z_4$. In study 2, we systematically varied the degree of missingness and the sample sizes to access the robustness of MMA-SEM under these conditions.

**EHR Experiment**

This retrospective cohort study utilized the EHR dataset from Mayo Clinic with 37,488 patients. Two time-windows, 2000-2005 (baseline) and 2005-2010 (follow-up), were defined. We extracted diagnoses, prescriptions, laboratory results, and vitals from adult (age 18+ on Jan 1st, 2005) primary care patients (defined as having two blood pressure (BP) measurements one before Jan 1st, 2000, one after Jan 1st, 2010). Since we wish to control missingness, patients with missing data were also removed from the study. After filtering by the exclusion criteria, 18,318 patients remained. Detailed characteristics of the patients can be found in Table 4.1.

**Table 4.1** Characteristics of participants. Mean (SD) and Count (%) for continuous and categorical variables

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No Diabetes (N=36506)</th>
<th>Diabetes (N=982)</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>47(16.9)</td>
<td>57.8(15.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>BMI</td>
<td>28.6(7.6)</td>
<td>33.2(10.5)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SBP</td>
<td>124.5(12.8)</td>
<td>132.9(12.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DBP</td>
<td>74.6(7.3)</td>
<td>71.8(7.5)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LDL</td>
<td>115.6(28)</td>
<td>101.3(25.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HDL</td>
<td>54.4(15)</td>
<td>48.2(14.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>139.7(75.2)</td>
<td>175.2(94.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>98.8(18.5)</td>
<td>167.6(49.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex</td>
<td>22160(60.7%)</td>
<td>508(51.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Obesity</td>
<td>1613(4.4%)</td>
<td>142(14.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HTN</td>
<td>13485(36.9%)</td>
<td>808(82.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HTN medication</td>
<td>9695(26.6%)</td>
<td>750(76.4%)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
We extracted variables related to obesity, hypertension (HTN), and hyperlipidemia (HL) from the baseline time window; and collected variables related to diabetes mellitus (DM) from the follow-up time window. Variables extracted from each disease group are:

- **Obesity group**: body mass index (BMI), obesity diagnosis
- **HTN group**: Systolic blood pressure (SBP), diastolic blood pressure (DBP), HTN diagnosis, HTN medications
- **HL group**: high-density lipoprotein (HDL), low-density lipoprotein (LDL), Triglyceride (Trig), HL diagnosis, HL medication
- **DM group**: fasting glucose (fasting), DM diagnosis, DM medication.

Each disease group was represented by a latent variable approximating the severity of the disease for individual patients. To initialize these latent variables by fitting a logistic model for each disease and using the predicted probability of disease as the latent disease severity measure. In each logistic model, age, gender, and the variables within the corresponding disease group were used as covariates, while the diagnosis was used as the outcome. Additionally, we extracted diagnosis data for two diseases that potentially related with the above chronic conditions, coronary artery disease (CAD) and congestive heart failure (CHF).
Simulation process. (1) In the fully simulated experiments, the gold standard graph and true effect sizes can be directly obtained from the simulation procedure. In the EHR experiment, we learned the graph from the complete data using the Greedy Equivalence Search[26 94], a causal structure discovery algorithm. We consider this as our “gold standard” causal structure. (2) We used the complete data and the “gold standard” causal structure from Step 1 to infer the effect sizes among the latent disease variables, which we considered as the “ground truth”. (3) We simulated missingness in the data. The probability of deleting an observation (making an observation missing for a patient) was determined based on their disease risk, where higher disease risk led to a lower probability of missingness in the corresponding lab, vitals, diagnosis, and medication of the disease. (4) We ran MMA-SEM on the resulting data (with non-randomly missing values), estimated the causal effect sizes among the latent disease variables and compared them to the “ground truth” effect sizes. We repeated steps (3) and (4) for different missing variables rates.

Evaluation method and metrics

To evaluate the performance of SEM-based methods and the proposed MMA-SEM method, we focus on two main aspects: (1) whether the causal structure was preserved (i.e., causal edges were not lost); and (2) the accuracy of the estimated effect size among latent variables. Each experiment was repeated 100 times to obtain empirical estimation of the standard deviation. Our evaluation focused on testing:

(1) Preserve the correct causal structure, i.e., not assign a statistically zero effect to a true causal edge.
(2) Accurately compute the effect sizes of the latent edges among the latent variables. We calculated root-mean-square error (RMSE) of the estimated effect size and its empirical confidence interval (+/- standard deviation).

**Results**

a. *Synthetic Data Experiment*

**Study 1: missing rate & assumption violation**

![Figure 4.2 Root Mean Square Error (RMSE) of the estimated latent effects.](image)

The plot shows the mean RMSE for the five tested methods across various missing percentage levels ranging from 10% to 33%. The proposed method, MMA-SEM, is compared to three SEM-based methods. The “+” represents the combination of SEM with a specific missing imputation.
method. Additionally, “MMA-SEM+wrg” represents the test where MMA-SEM incorrectly assumed the missing probability is determined by the parent latent variables.

Figure 4.2 shows the RMSE for the estimated effect sizes among latent variables. The horizontal axis represents the percentage of missing for observed variables, which ranges from 10% to 33%. Each line of points represents an effect estimation method, and the bar surrounding each point represents the empirical confidence interval for the RMSE (+/- one standardized deviation). We tested three missing data imputation methods, multiple imputation by chained equations (MICE)[174], mean imputation, and zero imputation. Subsequently, SEM was applied to each imputed dataset. These methods were labeled as ‘SEM+mice’, ‘SEM+mean’, and ‘SEM+zero’, respectively. We also evaluated list-wise deletion as a strategy to work around missingness, however, the results were orders of magnitude worse than those obtained from other methods, so we have relegated these results to the Supplement. As the percentage of missingness increases, the RMSE rises for both the proposed MMA-SEM and SEM-based methods. The proposed method (labelled with “MMA-SEM”) largely outperformed the SEM-based methods across all three missing imputation methods.

To evaluate the resilience of the proposed method when the missingness mechanism was not specified correctly, we included a test where the algorithm incorrectly assumes the missingness generation process. It is labeled as “MMA-SEM+wrg”. From the results, it can be concluded that the proposed method achieved lower RMSE when the missingness is lower than 25%. However, when the missingness percentage exceeds 30%, the mis-specified missing mechanism
jeopardized the performance of the proposed method, causing it to underperform in comparison to the MICE-based methods. Despite this, the “MMA-SEM+wrg” still demonstrates a certain degree of stability in the presence of violated assumption.

**Study 2: sample size**

![Graph](image)

**Figure 4.3** Root Mean Square Error (RMSE) and insignificance across various sample sizes.

A) The mean RMSE for the four tested methods as the sample size varies from 200 to 15000. B) The average percentage of latent effects that were found to be insignificant for each method at different sample sizes.

In this analysis, we conducted a comparison between the proposed method and the SEM-based methods across various sample sizes, ranging from 200 to 15,000. We simulated data with 20 latent variables and 3 observed variables for each latent variable. The missingness rate was fixed
to 25%. Figure 4.3A shows the RMSE for both the proposed method and the SEM-based methods corresponding to a specified sample sizes. Figure 4.3B showcases the percentage of latent edge effects that are insignificant, defined by an estimated p-value exceeding 0.05. An insignificant p-value implies that the edge in question does not exist, and its effect is statistically indistinguishable from zero.

Across all sample sizes, MMA-SEM consistently outperformed the SEM methods, with lower RMSE and narrower confidence intervals. For sample sizes below 2000, the SEM methods tended to produce edge estimations that were either extremely large or small, resulting in a large RMSE and broader confidence interval. Some confidence intervals were omitted from the graph as they exceeded the graphical scale.

In Figure 4.3B, it can be observed that as the sample size increases above 500, fewer than 1% of the existing edges remained insignificant for ‘SEM+mice’ and MMA+SEM. Almost all edges become significant as sample size increases to 5000. Among the three SEM-based imputation methods, MICE produced stable effects estimation while the other two SEM-based imputation methods lacked such stability.

b. EHR Experiment
Figure 4.4 The EHR study. A) the assumed causal structure of the EHR latent and observed variables. B) the RMSE of the proposed and SEM based method

Figure 4.4 shows the graph that describes the relationships among latent and observed EHR variables. The structure among latent variables was learned using the Greedy Equivalence Search algorithm, a causal structure discovery algorithm. Since the observed variables were collected from real EHR, SEM with complete data was used to infer the true latent effect sizes. This graph and the estimated effect sizes from it were used as the “ground truth”. In other words, the estimated effect sizes served as a benchmark for comparing the performance of MMA-SEM and SEM.

The graph on the right is the result graph as we shown in the synthetic data experiment, where x-axis represents the missing percentage, y-axis is the RMSE. As observed from the graph, when
the percentage of missing was smaller than 8%, the proposed method and the SEM method performed similarly in terms of RMSE, with the proposed method exhibiting greater stability (i.e., a smaller standard deviation). However, for larger percentages of missing (greater than 9%), the proposed method clearly outperformed the SEM method. These findings suggest that the proposed method outperforms SEM approach, particularly when missing data is more prevalent.

Discussion

In this study, we focused on estimating effect sizes among latent variables in the presence of non-randomly missing observations. Contrary to existing methods, we focus on effect sizes among latent variables, as opposed to effect sizes among observed variables or effect sizes between observed and latent variables, because in health care, disease concepts that appear in our clinical knowledge do not correspond well to observed variables in the data set. For example, when clinical knowledge states that high blood pressure increases the risk of stroke, the disease concept is high blood pressure, which is represented in the EHR by multiple observed variables: vital signs (systolic and diastolic blood pressure) as well as diagnosis codes and even prescriptions often signal the increased severity of the condition. The role of the latent variables in this study is to represent these disease concepts; and the effect sizes among diseases are represented by the effect sizes among the corresponding latent variables.

We particularly explored the challenges associated with missing data in EHR datasets. In healthcare, missingness is often non-random, signaling that the unobserved value of a variable can be normal and is missing because there was no need to measure it; or conversely, the condition is extremely severe and there is no time to measure it. In case of such non-random
missingness, traditional methodologies often fall short. Their inability to handle such missingness correctly, can lead to biased estimations, sometimes to the extent, where despite correctly specifying the causal structure of the problem, SEM reports the data to be inconsistent with the causal structure (and estimating some effect sizes as indistinguishable from 0).

Fortunately, in healthcare, the causal mechanism that caused the missingness is often known or at least can be assumed. We introduced the Missing-Mechanism-Adjusted SEM (MMA-SEM), a novel approach specifically designed to mitigate biases in causal effect size estimates. By seamlessly incorporating the causal mechanism of missingness, MMA-SEM iteratively refines the model fit, ensuring more reliable and accurate estimations of causal effects, even when the data is incomplete. Through comprehensive experiments on both simulated datasets and real-world EHR data, we demonstrated (i) MMA-SEM successfully preserved the correct causal structure in the sense that it never estimated the causal effect size of an existing causal edge as 0; (ii) the estimated effect sizes, especially in scenarios with high missingness, were always less biased than competing methods; and (iii) even when the casual structure of missingness was specified incorrectly, MMA-SEM never had statistically significantly larger bias than competing methods.

In contrast, we found that when SEM was used with standard missing-at-random imputation, the causal structure specified in the model was distorted to a level where SEM could not effectively fit the data, despite having a large dataset of 10,000 observations. The proposed method, which handles missingness properly, resulted in significantly lower errors (RMSE) in estimating the effect sizes and also lower rates of lost causal edges (causal edges with 0 effect size).

Furthermore, even when the assumed missingness generation mechanism was incorrect, i.e., the
missingness is not introduced by the assumed latent variables, the proposed method was still able to estimate the causal effect, and this was achieved without a significant decrease in accuracy. By mimicking real-world EHR data collection and analysis processes, and introducing missingness manually, results from our EHR experiment echoed the findings from the synthetic data experiment. These findings demonstrate the pitfalls of inadequately addressed missingness in causal effect size estimations.

The proposed method is innovative and contribute to the ability to estimate causal effects among hidden variables. The proposed method is capable of incorporating the causal mechanism of missingness and leveraging both the missing information and available observational data to learn the underlying latent variables. In MMA-SEM, the latent variables are estimated by their causal observed children as well as their latent causal parents and latent causal children. This provides a unique way to estimate the value of the latent variable when the observed children have missing values. This integration prevents distortion of the causal structure due to reliance on incorrect imputation methods. Notably, the RMSE of the results improves with more information provided about the missingness mechanism.

**Limitations.** This study has some limitations. First, the proposed method assumes that the underlying missing mechanism is known, which is often but not always the case in real-world data. We have demonstrated, however, that the simplistic assumption that the missingness of the observed variables only depends on its direct latent parent, still improves performance over not assuming any particular missingness mechanism. Secondly, the missingness in EHR could also be due to a combination of missing at random, missing completely at random and missing not at
random. The proposed method offers no advantage (but is also not disadvantageous) in these scenarios. Applying our method for missing-at-random or missing-completely-at-random data is neither advantageous nor disadvantageous; all methods can solve these relatively simple problems. Third, the method is based on the assumption that the missingness of the observed variable depends on latent variable. While this is a very reasonable assumption for the reasons we explained earlier, in some cases, the latent variable may not be uniquely identifiable (e.g., there is only one observed variable).

**Conclusion.** The study introduced the Missing-Mechanism-Adjusted SEM (MMA-SEM) for estimating causal effect sizes in the presence of missing data, particularly in the context of modeling of EHR data with latent disease concepts. Our extensive testing on synthetic and real-world datasets demonstrated MMA-SEM’s ability to preserve the correct causal structure of the problem and produce effect size estimate that are less biased than those produced by competing methods.
Discussion and Conclusion

Unlike associations, causality enables the modeling of manipulatable relationships, a capability that is highly valuable in the fields of health informatics and biomedical research. Identifying biomarkers that have a causal effect on the target disease can open the door for the development of specific drugs. Additionally, understanding the causal structure when estimating the effect of a treatment effect can enhance accuracy and help avoid common pitfalls. Because of these advancements, methods in CSD and CI have evolved over recent decades. Nevertheless, the field continues to face several challenges. The four chapters of my research were designed to explore the possibility and solve these challenges step by step:

i. One of the primary challenges in the field of CSD is the validation and interpretation of the results with real-world data. While there is an abundance of new CSD algorithms, most validations have been limited to simulated datasets where the underlying structure is predefined and known. This approach fails to address how these algorithms perform with the complexities of real-world data. Moreover, there is a noticeable gap in research regarding the applicability of CSD methods. Researchers from other fields of study often prefer traditional statistical analysis tools over CSD tools. This preference likely arises from the challenges in implementing CSD methodologies (software tools or packages) and interpreting the complex results they yield. The key to overcoming this challenge is to demonstrate that CSD methods can produce results that are not only reliable but also convincing. The first chapter explores the merits of existing CSD methods on discovering the causal relationships in real-world data in the biomedical field. We show that the dedicated CSD method, FGES, can achieve a precision and recall of 0.5, which is reasonable considering that a single CSD task can be viewed as a combination of multiple multi-
class classification problems. We also found that with the help of background knowledge (true edges) and longitudinal data (information on time), the precision and recall could be further improved to nearly 0.8.

ii. Following the validation of CSD methods in Chapter I, Chapter II solved a real-world biomedical problem with causal analysis pipeline. The goal of the study focused on identifying early brain measurements and early risk or protective factors that influence WMH trajectories. In the first step of our analysis, we implemented CSD to generate a causal graph among a set of risk factors and WMH. Based on the discovered graph, we then employed a CI method to calculate the exact causal effect sizes from midlife modifiers of WMH. This causal analysis workflow that incorporates CSD provides benefits in the following ways: (1) **Causal interpretation of relationships.** Under some assumptions, the relationships discovered from CSD methods have a causal interpretation, not merely association. (2) **Understanding of underlying biological mechanisms.** When the goal of the study is not merely predicting the outcome, but also extracting the structure among risk factors, helps understand the underlying biological mechanism which is also helpful in CI estimation process. (3) **Hypothesis generation and validation.** In our current workflow, the CSD model was also used for validating the conclusion from well-established traditional methods. Nonetheless, it can also be used for hypothesis generation.

The present study demonstrated a significant age, sex, and hypertension association with WMH. The relationship between vascular factors and WMH can be better explained by early changes in WM microstructural integrity. The midlife modifiers emerged as important components of WM
health. Hence midlife may be the relevant window for prevention of late-life WMHs and measuring microstructural integrity using DTI can aid in better design and monitoring of prevention trials.

iii. Many CSD methods such as PC, FCI and GES often fail to produce a fully directed acyclic graph, where every edge is directed and its direction indicates the causal effect direction. This limitation poses a significant difficulty in practical applications, as interpreting unoriented edges within the graph is challenging. Chapter I & II have demonstrated that two important factors, longitudinal data (temporal information) and sample size significantly affect the performance of CSD analysis in extracting causal knowledge and orientating edges. EHR data naturally meet the needs for extracting causal knowledge, however, it also posed its own challenges. We proposed an algorithm with three adaptations. First, we incorporated study design considerations. EHR data as it exists in the system does not follow any study design. Second, transformation of the time stamps in EHR. General purpose CSD algorithms applied to the EHR data occasionally report “causal” relationships that are in the opposite direction of the natural disease progression. Third, leveraging the longitudinal nature of EHR data and incorporating time information as part of the causal discovery process can enhance the identification of edge orientation. We have successfully demonstrated that the graph produced by the proposed transformation and search algorithm is more stable across bootstrap iterations and as complete as other methods yet it contained substantially fewer errors (had higher precision) than graphs produced by general-purpose methods. The resulting graph was successfully validated using longitudinal EHR data from an independent health system. Based on these
findings, we concluded that the proposed method is more suitable for use in clinical studies using EHR data.

iv. Real-world data are often subject to errors, which, while individually insignificant, can have far-reaching effects in CSD and CI methods. These errors can propagate throughout the causal graph during the process of structure updating or effect estimation, leading to compounded inaccuracies. In Chapter IV, our study focused on estimating effect sizes among latent variables in the presence of non-randomly missing observations. Contrary to existing methods, we focus on effect sizes among latent variables, which typically estimate effect sizes among observed variables or effect sizes between observed and latent variables. Our study is particularly relevant in healthcare field that disease concepts that appear in our clinical knowledge do not correspond well to observed variables in the data set.

The study introduced Missing-Mechanism-Adjusted SEM as a new method for estimating causal effect sizes in the presence of missing data, particularly in the context of modeling of EHR data with latent disease concepts. Our comprehensive evaluation using both synthetic and real-world datasets have demonstrated MMA-SEM's ability to preserve the correct causal structure of the problem and produce effect size estimate that are less biased than those produced by competing methods.

Given the completion of the four chapters, my research has the following contribution to science, heath informatics, and medicine:
i. Chapter I explored the possibility of applying CSD algorithms to health/biomedical informatics problems. We have shown that even though CSD methods were not designed for health science specifically, they can effectively be employed to solve problems within health informatics, particularly in discovering causal relationships among variables of interest. In the realms of health informatics and medicine where causal relationships are crucial to understand the underlying biological mechanisms or drug effects, the findings from Chapter I are significant. They indicate that, with certain adaptations (such as background knowledge and longitudinal data), CSD methods are capable of successfully discovering the underlying causal relationships. It is also important to note that in situations involving extensive variables, such as genomic data, computational CSD becomes not just advantageous but often the only viable approach for causal discovery. This is because traditional, non-computational methods would be impractical or impossible to analysis such great number of variables. This Chapter not only served as a benchmark for application of CSD but also underscores their significant potential to contribute to advancements in health informatics and medical research.

ii. The contributions of Chapter II are twofold: the biomedical aspects of White Matter Hyperintensities and the development of a causal analysis pipeline. On the biomedical aspect, this chapter revealed that midlife lifestyle factors, specifically smoking and physical activity, have a causal effect on future WMH trajectories. WMH are a common manifestation of cerebrovascular disease and significantly impact motor and cognitive functions. This biomedical discovery is crucial. It holds merit for its potential to improve the design and monitoring of prevention trials, offering new path for managing these health risks.
Secondly, the causal analysis pipeline implemented in this study exemplifies the integration of Causal Structure Discovery as a guiding framework in traditional Causal Inference. This approach validates the common assumptions made during the CI process using CSD, providing a robust methodological model. This pipeline significantly contributes to the field of health informatics and broadly impacts health science research. Its innovative approach to combining CSD and CI methodologies not only enhances the reliability of causal conclusions but also sets a precedent for future research in health informatics and related disciplines.

iii. Chapter III introduces a novel method specifically designed for uncovering causal relationships from Electronic Health Records (EHR) data. Given that EHR use is expanding and is rapidly becoming a valuable health and biomedical data source, its potential for causal discovery is immense yet largely unused due to its limitations. The method proposed in this chapter bridges this gap, enabling the effective extraction of causal knowledge from EHR data. This advancement not only makes better use of the rich information contained within EHRs but also marks a significant step forward in the application of causal discovery techniques in the realm of health informatics.

iv. Chapter IV proposed a novel Causal Inference method, designed to address the challenge of non-randomly missing observations. The non-randomly missing problem is common in biomedical datasets such as Electronic Health Records. In EHRs data, it is often observed that the missingness of a lab or vitals measurement is dependent on the unobserved conditions of a disease. The proposed method incorporates the non-random missingness, along with causal inference techniques to estimate effect sizes.
Similar to Chapter III, this innovative methodology effectively narrows the gap between theoretical causal inference and its practical application in EHR data. This significant development addresses the prevalent issue of missing data in EHRs, thereby greatly facilitating the extraction of causal knowledge. This breakthrough not only contributes substantially to the field of health informatics but also enhances the overall utility and effectiveness of EHR data in clinical research and practice.
Reference

16. Charlot M, Grove EL, Hansen PR, et al. Proton pump inhibitor use and risk of adverse cardiovascular events in aspirin treated patients with first time myocardial infarction:


45. Rogers MB. There's No Tomorrow for TOMMORROW. ALZFORUM 2018.


79. Roberts RO, Geda YE, Knopman DS, et al. The Mayo Clinic Study of Aging: design and


92. Wardlaw JM, Chappell FM, Valdés Hernández MDC, et al. White matter hyperintensity


104. de Groot M, Cremers LG, Ikram MA, et al. White Matter Degeneration with Aging:


140. Wartolowska KA, Webb AJS. Midlife blood pressure is associated with the severity of white matter hyperintensities: analysis of the UK Biobank cohort study. Eur Heart J 2021;42(7):750-57 doi: 10.1093/eurheartj/ehaa756 [published Online First:


170. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology 2014;14:118-18 doi:


