Causality Methods for Dementia Risk Inference

Suryansh Manocha

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2024

Abstract

Understanding the causal factors contributing to dementia onset is crucial for developing preventive interventions. This study investigates the utility of causal structure discovery (CSD) methods in identifying such causal factors from longitudinal data. We evaluate the performance of various temporal and atemporal CSD models in recovering the true causal graph, guided by a ground truth causal diagram curated with domain expertise. Our findings indicate that recent constraint-based CSD algorithms like CIM outperform established baselines in uncovering the underlying causal structure, with temporal models exhibiting superior accuracy by leveraging longitudinal information. Quantitative analyses reveal age and life satisfaction as the primary direct causal influences on dementia onset, followed by BMI and smoking. We also identify significant indirect causes like executive function and social engagement. To integrate causal discovery with predictive modeling, we propose a novel framework combining interpretable machine learning techniques for feature selection and CSD models for causal inference, augmented with general prior knowledge. Evaluations demonstrate this approach's potential to derive robust, actionable insights closely aligned with the ground truth causal model. While inherent limitations exist, our work highlights the prospects of combining interpretable AI and causal reasoning methods to advance dementia research and prevention efforts.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Suryansh Manocha*)

Acknowledgements

I wish to extend my sincere appreciation to those whose contributions have been integral to the successful completion of this research endeavour. Firstly, I am deeply grateful to my supervisor, Dr. Sohan Seth, for his invaluable expertise in this field and his consistent provision of feedback and guidance throughout both this project and my previous year's endeavour. His mentorship has played a significant role in forming a cohesive body of work spanning across consecutive years.

Dr. Seth's expansive network of academics has helped facilitate additional insights, resources, and opportunities for collaboration, enriching the depth and breadth of my work. In particular, I would like to thank Dr. Lucy Stirland for her assistance in creating the ground truth causal graphs, and her understanding of the causal relationships between covariates leading to dementia.

Additionally, I extend my heartfelt thanks to the members of the Data Science Unit within the Informatics Research Group for graciously welcoming me and offering invaluable feedback. I am particularly appreciative of the opportunities afforded to me to present my research findings to the unit on multiple occasions throughout the year. Notably, I am thankful for the privilege of presenting in the Data Science for Mental Health and Wellbeing (DS4MHW) workshop, a significant platform for interdisciplinary scholarly exchange (accessible at: `https://web.inf.ed.ac.uk/data-science-unit/engagement/workshop`).

Table of Contents

Chapter 1

Introduction

1.1 Motivation

With over 10 million new cases of dementia reported annually worldwide [1], and up to 40% of these cases potentially preventable or delayed by addressing health risk factors [2], understanding the etiology of dementia is paramount. Despite a relatively high diagnosis rate, exemplified by the 62% diagnosis rate in England in 2022 [7], the focus should shift towards comprehending the underlying causes of dementia rather than solely diagnosing the condition [105]. Many research endeavours leveraging historical datasets, such as longitudinal data [29], predominantly concentrate on constructing predictive models [68, 17, 84], often utilising dementia symptoms as primary predictive variables [19]. While these associative models serve their purpose in diagnosis, they fall short in unravelling the intricate causal relationships and contributing factors leading to dementia onset [60].

Bennett et al.'s study [20] compellingly illustrates the intricate challenge of establishing causality between depression and later-life dementia, with findings suggesting a bidirectional relationship that complicates the determination of whether depression predisposes individuals to dementia or vice versa. This inherent ambiguity, while emphasising the fundamental limitations of associative models in definitively establishing the directionality of causal relationships, also underscores the opportunity for exploring alternative methodologies like causal discovery models [37]. These models offer a systematic approach to infer the directionality of causation from observational data by considering various assumptions and model structures. Additionally, longitudinal studies provide a unique opportunity to delve into temporal order, wherein events preceding others are construed as causes, and subsequent ones as effects, thereby facilitating the exploration of causal pathways; [111] claims that Causal Structure Discovery (CSD) should be used with longitudinal data to achieve the best results. However, it is crucial to exercise caution, as highlighted by Lagnado and Sloman [72], against an excessive reliance on temporal order, as it can potentially lead to erroneous conclusions when temporal cues are misleading.

The present study aims to address several key research questions pertaining to the onset and prevention of dementia. Firstly, it seeks to identify the primary causal factors

contributing to dementia onset, focusing specifically on modifiable or intervenable covariates preceding the manifestation of the condition. Secondly, the study delves into the efficacy of CSD in discovering the true causal graph within the dataset, and whether this poses a viable way to discover the underlying causal structure if the ground truth graph was not known. Lastly, the study examines the potential synergy between interpretable machine learning (IML) techniques and causal inference in generating deeper insights compared to their individual applications. It scrutinises the practicality of employing these approaches without domain-specific knowledge and explores whether causal discovery models, coupled with basic general knowledge, can offer comparable insights.

1.1.1 Previous Work (MInf1)

In our prior investigation detailed in MInf1 [84] we sought to develop predictive models utilising survival analysis techniques, specifically Temporal Random Survival Forest [35], whilst accounting for competing risks [98]. Subsequently, we employed these predictive models for inferential analysis to find the covariates most strongly associated with predicting dementia onset and non-onset. To achieve this, sophisticated interpretability techniques, including LIME [103] and counterfactual explanations [73], were utilised to provide insights into the model's decision-making process, with the objective of inferring the associations revealed by the models.

In our previous study in MInf1, we explored not only global explanations [80], but also local explanations - selecting two specific participants from the study whose survival probabilities were vastly different; we revisit these individuals in the current study to expand upon our findings. Our previous research in MInf1 indicated that, while the choice of interpretation technique significantly influenced the explanation, unsurprisingly, cognitive and social factors emerged as top predictors of dementia. However, upon further analysis in this study, we discovered that many of these predictors were symptoms of dementia, rather than causes, which, as we discussed above, may be useful for diagnosis, but is not useful for interventional analysis, whereby a person can be informed of counterfactual conditions with variables amenable to intervention.

Graphical causal models serve as powerful exploratory tools, mapping various causal pathways to a specific outcome and illustrating interrelationships among variables [108]. In the context of dementia research, the application of a causal graph may unveil, for instance, that diminished social interactions precede depression, subsequently contributing to dementia - a subtlety not discernible through associative interpretability techniques alone. Such techniques might only indicate a strong correlation between depression and dementia, failing to account for the confounding variable of social interaction; this notion is supported by research [70], which suggests that inadequate social interaction and incident dementia are comparable risk factors to depression. Moreover, it is well-established that a lack of social interaction can lead to depression [92, 91], further reinforcing the presence of the existence of a confounding variable. The potential of modelling these intricate relationship dynamics using domain knowledge to construct a ground truth graph is considerable; not only can it account for confounding variables, but this graph can then inform the training of our graphical causal model on

our dataset under these modelling assumptions, allowing us to discern which variables might contribute most significantly to the dementia outcome.

1.2 Literature Review

Randomised controlled trials (RCTs) are widely acknowledged as the gold standard in medical research for establishing causal relationships between interventions and outcomes [47]. However, their implementation is often hindered by exorbitant costs and logistical constraints, prompting researchers to turn to observational studies (like ours) as a practical alternative [39]. Utilising observational data, survival machine learning models offer a means to predict the likelihood of individuals developing certain health conditions, as demonstrated in MInf1 [84]; despite their predictive power, the increasing complexity of ML models has rendered them opaque to interpretation [69]. This opacity presents challenges for domain experts seeking to comprehend the underlying mechanisms driving model predictions [55]; while IML/XAI explanations serve as initial steps towards understanding model outputs, they often fall short in providing comprehensive insights. To achieve a level of ”explainable medicine”, it becomes imperative to delve deeper into causality, emphasising the need to understand not only what factors contribute to health outcomes but also why they do so [56].

In biomedical applications, the primary research goal often extends beyond achieving high prediction accuracy to encompass the identification of influential risk factors or underlying mechanisms that can be altered [111]. Machine learning methodologies predominantly rely on discerning associations within data [111]; consequently, interpretations derived through IML techniques primarily unveil statistical relationships inherent in the observed data, characterized by conditional independence [89]. In accordance with Pearl's three-layer causal hierarchy [96], IML predominantly operates at Level 1, focusing on associations $P(y|x)$, characterised by activities akin to ”seeing,” and inquiries probing how changes in one variable influence beliefs about another - within our context, this would be the question of “What does a high BMI tell me about dementia likelihood?”. However, in biomedical research, a demand persists to go beyond the limitations of IML methodologies [55]; by modelling the causal connections, for example, in our research, we model the ground truth graph with the help of a domain expert. This enables exploration at Level 3 of Pearl's hierarchy, which deals with counterfactuals $P(y_x \mid x', y')$ wherein the primary activities involve retrospection, and inquiries revolve around establishing causality - in our context this would be the question of “If the patient had lower BMI, would they have lower risk of dementia?”.

The study [67] asserts that IML techniques can be formalised as a statistical process, progressing from correlation to causation. This formalisation enhances human understanding of IML explanations, enabling individuals to then articulate causal propositions based on such interpretations. In the context of our MInf1 endeavours [84], our objective was precisely this — to scrutinise the associations identified by models using IML techniques. Our research findings presented associations, leaving it to humans to extrapolate potential causal associations from these correlation-based insights. [67] formalises this notion, with IML producing explanations, and the human interpreter probing the questions to the IML model and making causal interpretations based on

their human interpretations. However, it is crucial to acknowledge that inherent biases permeate each stage of this process, ranging from sampling bias to the capabilities of the methodology employed, to human bias. The propagation of biases [10] warrants meticulous consideration, necessitating an awareness of the assumptions and limitations inherent in each stage of the methodology adopted. Nevertheless, this framework assumes that the human interpreter possesses domain expertise, such that they are able to interpret the explanations produced by the IML model and infer causal relations from them, which is unlikely to be the case, thereby heightening the risk of human bias. To address this challenge, we propose an extension of this framework, incorporating CSDs to infer causal relationships, while still retaining a degree of human oversight. This entails allowing the human interpreter to impose constraints on the model, such as specifying that no other variable can cause age. This augmentation is necessary as CSDs often encounter difficulty in discerning causal directionality [101]; through this hybrid methodology, we endeavour to reduce the impact of human bias while leveraging CSDs to enhance causal inference capabilities.

Rawal et al. [99] proposes a novel hybrid approach that combines causal discovery and explainable AI (XAI) techniques to create more robust and trustworthy AI/ML systems, especially when working with observational data lacking ground truth causal information. The key novelty lies in using causal discovery methods to identify causally relevant features from the data, and then comparing/validating these against the feature importance explanations obtained from correlation-based XAI techniques like SHAP. By accounting for both causal relationships between features (via causal discovery) as well as feature relevance based on correlations (via XAI), this ”causal explainable” approach aims to provide more robust explanations that go beyond mere correlations to also capture causal effects. It argues that combining causal discovery and XAI provides an additional layer of robustness that cannot be achieved by either method alone. However, their methodology exhibits certain limitations, which we aim to address in our research. Firstly, the authors only use one IML model (SHAP), which limits the outcomes of the study, drawing from our findings in MInf1 [84], we recognize that the choice of interpretation technique significantly influences the insights extracted from the data; which is why we use a collection of standardised IML approaches. Furthermore, the exclusive use of a singular model for causal discovery (PC) without explaining its selection rationale or the alignment of its assumptions with the dataset presents a notable gap. Furthermore, their methodology could be strengthened by incorporating constraints, as seen in Figure 4 [99], where bidirectional links between variables present logical inconsistencies. Even without domain expertise, such constraints could have been incorporated into the model, mirroring our approach in the proposed pipeline. Additionally, employing a qualitative comparison between XAI outcomes and CSD outcomes may introduce cognitive biases, as highlighted by Blanco et al. [23], potentially leading to the erroneous attribution of causation to recent or convenient confounders correlated with the outcome.

With regards to this specific dataset, to our knowledge, only one study exists [13] that also focuses on causality, which examines if engagement in a variety of different lifestyle activities can slow the rate of cognitive decline as older adults age; and finds that two factors, moderate-intensity physical activity and learning activities resulted in

significant positive impact on cognitive function. The study focuses on causal inference in observational data, employing a methodology centred around matching treatment and control groups over time. The approach uses a matching algorithm to pair samples, applies statistical tests to assess the impact of candidate variables on cognitive health, and addresses the issue of multiple comparisons through adjustment methods. Notably, the study assumes that temporal associations imply causal relationships and emphasises estimating treatment effects. This is in contrast to the structural causal models that we use, which aim to uncover the underlying causal structure in observational data by identifying direct causal relationships among variables, typically represented as a directed acyclic graph (DAG). This way, the underlying assumptions that are made by our causal model are made more explicit.

Similarly, Suen et al. (2022) [122] investigates the creation of causal graphs across different time points using the ELSA-Brasil dataset, particularly examining changes in the causal structures amidst varying mental symptoms pre-COVID, during-COVID, and post-COVID. Their endeavour to compare causal graphs across time periods, while insightful, potentially introduces bias due to unobserved or uncontrolled confounding variables [130]. Moreover, the inherent stochasticity of causal discovery methods, as highlighted in their study, poses challenges in drawing definitive conclusions regarding temporal changes in causal relationships. In contrast, our study addresses these concerns by leveraging temporal causal models, which account for sequential event orders, dynamic relationships, and changes over time [50]. This approach facilitates the consideration of time-varying confounders, captures lagged effects, and maintains statistical power through the utilisation of the entire sample size. By adopting a unified temporal framework, we ensure consistency in modelling assumptions, mitigate data fragmentation.

In [82], the authors advocate for the fusion of domain knowledge with CSD algorithms, suggesting improved outcomes. However, this raises a practical query: if one already possesses domain expertise, why resort to a CSD algorithm? In our research, we address this concern meticulously in our proposed framework (Section 4.4) - we provide our models with rudimentary domain knowledge, to allow for scenarios where expert domain knowledge is not available. For instance, we impart to our models the fundamental understanding that no variable can cause 'Age,' thereby establishing 'Age' as a source node in the graph (Section 4.2.1).

1.3 Research Questions

We aim to answer the following research questions as part of this study:

Question 1: To what extent can causal structure discovery (CSD) methods identify the true causal graph in the dataset? Considering our exposure to longitudinal data, do temporal CSD models outperform their generic counterparts?
Recent advancements in causal discovery [142, 43] propose novel methodologies tailored for different dataset types, such as longitudinal data [119].

Question 2: What are the primary causal factors contributing to dementia onset, specifically targeting covariates that can be modified or intervened upon before

the onset of dementia?
Aiming to identify actionable insights empowering individuals to take preventive measures. Unlike predictive models, the focus here lies not on the outcomes of dementia, but solely on the causes, thereby enabling interventions.

Question 3: Can the combined utilisation of interpretable machine learning techniques and causal inference offer superior insights compared to using either approach individually?
Can we develop a predictive model that not only employs interpretable machine learning techniques but also provides causal explanations for its decisions?

1.4 Objectives & Contributions

Given the ongoing evolution of causality research and its expanding body of literature [86], we aim to leverage the latest advancements in this field to address the questions posed above. To make explicit the contributions of this project *(also reflected in figure 4.1)*:

- Perform causal discovery on the ELSA dataset.
  This brings novelty, as there are no other papers that look at causal discovery for dementia on the ELSA dataset.
- Comparison of temporal causal models to their atemporal counterparts.
  This brings novelty, since there are not only no papers on ELSA utilising temporal causal discovery models, but to our awareness, there are also no papers on dementia in general employing temporal causal discovery models.
- Utilise recent advancements in causal discovery algorithms - previous studies typically rely on established algorithms like PC.
  This introduces novelty, as existing literature predominantly relies on older algorithms like PC, whereas our study explores newer algorithms, such as CIM introduced in 2023 [119].
- Feature importance utilising both the CSD model and the ground truth model.
  This brings novelty, as we collaborated with a domain expert to construct a ground truth graph specific to this dataset. No such prior ground truth exists, nor causal inference evaluation on this dataset. Moreover, we utilise recent work in feature importance derived from causal explanations, such as the Shapely approach by [63] from 2020.
- Comparison of causal inference to inference based on IML techniques.
  For example, which features are the most relevant from a causal perspective, as compared to a purely IML perspective?
- Proposal of a new framework expanding beyond traditional IML techniques, integrating considerations of causality.
  To our knowledge, this is among the first research to integrate both IML and causal discovery techniques in unison.

Chapter 2

Background

2.1 Causal Inference

Causal inference is the process of deriving cause-and-effect relationships from data, crucial across various disciplines such as medicine, social sciences, and machine learning. Unlike traditional statistical analysis, which primarily focuses on describing associations between variables, causal inference seeks to uncover the underlying mechanisms generating observed data and predict the consequences of interventions or system changes [95].

Consider the causal model depicted in Figure 2.1. How does manipulating the value of feature A impact feature B? We denote this causal relationship as $P(B|\text{do}(A))$, distinguishing it from the observed statistical relationship $P(B|A)$, which is confounded by the influence of C. The operator $\text{do}(A)$ signifies an intervention to alter the value of A, leading to a new causal model where A is disconnected from its causal parents [95]. Notably, in a randomised experiment, we do not need to enforce the do operator, as $P(B|\text{do}(A)$ is equivalent to $P(B|A)$ [94], however, in the context of our observational data that we investigate, this needs to be enforced.

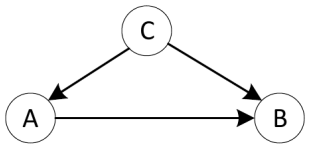


Figure 2.1: Causal graph showing a confounding variable C on the effect of A on B

One of the main goals of causal inference is to estimate the effect of an intervention or treatment on an outcome variable. The causal effect of a treatment T on an outcome Y is typically defined as the difference between the potential outcomes [106] under different treatment values:

$$\text{Causal Effect} = Y(T=1) - Y(T=0)$$

Since we cannot observe individual-level causal effects, we often focus on estimating

average causal effects over a population, such as the Average Treatment Effect (ATE) , Average Treatment Effect on the Treated (ATT), and Average Treatment Effect on the Untreated (ATU). [54]

2.1.1 D-separation

D-separation is a concept used to determine the absence of causal effects between variables in a directed acyclic graph (DAG) [46]. It provides rules for assessing if a path between two variables is "blocked," indicating no causal influence is transmitted through that path [44]. Mathematically, a set of variables S d-separates variables X_1 and X_3 if S blocks all paths connecting them [46].

There are three scenarios for blocking paths [38]:

1. **Chain:** If S includes a variable X_2 on a direct path between X_1 and X_3, then X_2 blocks the path.
2. **Common Effect:** X_1 and X_3 can be d-separated by S if S includes the "collider" variable (with two incoming arrows) X_2, but excludes any of X_2's descendants. This is because conditioning on the collider itself introduces a spurious association between X_1 and X_3.
3. **Common Blocking:** If X_1 and X_3 are connected by a common cause, and S includes that common cause (X_3), then the path is blocked ($X_1 \leftarrow X_3 \rightarrow X_2$). Conditioning on the common cause X_3 eliminates the indirect effect it has on both X_1 and X_3.

D-separation enables the assessment of conditional independence between two variables given a set of observed variables, implying no causal effect between them. Conversely, if they are not d-separated, there might be a causal influence that needs to be considered when inferring relationships from data.

2.1.2 Markov Equivalence Class (MEC)

Markov Equivalence Class is a concept in causal inference that groups graphical models based on the conditional independence relationships they represent among observed variables. Two models belong to the same MEC, denoted mathematically as MEC(G), if they encode the same set of conditional independence statements using the concept of d-separation, regardless of the specific graphical structure. For instance, consider observed variables X_1, X_2, and X_3. If both graphs imply the statement $X_1 \perp\!\!\!\perp X_3 | X_2$ (meaning X_1 and X_3 are conditionally independent given X_2), they would be classified in the same MEC. The importance of MEC lies in its ability to identify the minimal set of conditions required to infer causal relationships from observational data, without relying on specific details of the underlying graphical model.

2.1.3 Kullback-Leibler Divergence

Kullback-Leibler (KL) divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. Given two probability dis-

tributions $P(x)$ and $Q(x)$, KL divergence quantifies the amount of information lost when $Q(x)$ is used to approximate $P(x)$. In essence, KL divergence measures the average difference between the log probabilities of the two distributions weighted by the probabilities in $P(x)$. Formally, KL divergence is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{2.1}$$

2.2 Causal Structure Discovery

Causal structure discovery models rely on several key assumptions [116]:

- **Causal Markov Condition:** Each variable is independent of its non-descendants, given its parents in the DAG. That is, if variables are d-separated in the graph G, then they are independent [75].

$$P(x_1, \ldots, x_n) = \prod_{j=1}^{J} P(x_j | \mathrm{Pa}_j, \varepsilon_j)$$

  Each variable x_j is conditionally independent of its non-descendants given its parents Pa_j and unobserved (noise) variables ε_j:

- **Causal Sufficiency:** The assumption that the causal graph is faithful to the probability distribution, meaning that all observed independencies are due to the causal structure.
- **Causal Faithfulness:** The observed conditional independencies in the data are exactly those implied by the causal Markov condition.
- **Causal Independence-based:** This approach relies on the assumption that causal structure can be inferred from statistical independencies in the data.

2.2.1 Constraint-based Algorithms

Constraint-based causal discovery algorithms use statistical tests to identify conditional independence between variables. These relationships are leveraged to build a skeleton of the causal structure, followed by an orientation step based on specific rules. The primary objective is to reconstruct a Completed Partially Directed Acyclic Graph (CPDAG) that captures the most informative set of constraints consistent with the underlying causal structure, while ensuring adherence to fundamental assumptions such as the causal Markov property and faithfulness (and possibly causal sufficiency).

2.2.1.1 Peter-Clark (PC)

The PC algorithm [115] begins by constructing a complete undirected graph representing all potential pairwise relationships between variables. It then iteratively refines this graph based on conditional independence tests, ultimately resulting in a directed acyclic graph (DAG) representing the causal structure.

The key steps of the PC algorithm can be summarised as follows [45]:

1. Form a complete undirected graph [137].
2. Eliminate edges between variables that are unconditionally independent (under the causal faithfulness property).
3. For each pair of variables connected by an edge, and for each variable connected to either of them, remove the edge if the variables are conditionally independent given the third variable.
4. Continue checking conditional independence involving subsets of variables until no further edges can be removed.
5. Orient edges based on certain structural patterns, such as v-structures [36] and orientation propagation, to establish causal directions.

The algorithm's output, often represented as a partially directed acyclic graph (PDAG), captures the Markov equivalence class of possible causal structures consistent with the observed data [139]. Notably, the PC algorithm is consistent under certain assumptions, including the Causal Markov and Faithfulness assumptions, i.i.d. sampling, and absence of latent confounders [9].

2.2.1.2 Missing Value PC (MVPC)

The Test-wise Deletion PC algorithm (TD-PC) [120] deviates from the traditional list-wise deletion approach by selectively removing records with missing values that are pertinent to the current Conditional Independence (CI) tests. This strategy, highlighted by [120], presents a more data-efficient approach in causal inference. However, TD-PC faces challenges in scenarios involving Missing At Random (MAR) and Missing Not At Random (MNAR) data; in these cases, TD-PC may erroneously infer causal edges due to the absence of certain conditions, potentially leading to the introduction of extraneous edges in the inferred causal skeleton [139].

The MVPC algorithm [126] addresses these challenges by applying corrections only to the CI tests that are influenced by missing data, thereby improving both data and computational efficiency. This is particularly beneficial in scenarios where data are MNAR or MAR, where traditional causal discovery methods, such as the standard PC algorithm, are prone to errors [126]. The MVPC algorithm can be summarised as follows:

Algorithm 1 Missing-value PC Algorithm

Initialise: Create a full undirected graph G with nodes V.
Discovering Causal Structure: Utilise the deletion-based PC algorithm to prune edges from G using test-wise deleted data.
▷ Step 2: Identifying Direct Influences of Missingness Indicators
for each variable $V_i \in V$ with missing data **do**
    for each $j \neq i$ **do**
        Test for conditional independence between R_i and V_j. If independent, V_j is not a direct influencer of R_i.
    end for
end for
▷ Step 3: Locating Possible Redundant Connections
for each pair V_i and V_j where $i \neq j$ **do**
    if V_i and V_j are adjacent and share at least one common neighbouring variable or missingness indicator **then**
        Consider the edge between V_i and V_j as potentially redundant.
    end if
end for
Apply corrective measures to eliminate redundant edges in G.

2.2.1.3 Causal Inference over Mixtures (CIM)

The CIM algorithm [119] uses a mixture of DAGs, and is capable of handling cycles, non-stationarity, latent variables, and selection bias that may exist within the data. This model operates by integrating longitudinal data, comprising multiple observations over time, to infer causal connections. By examining temporal sequences, CIM discerns the directionality of causal effects, aiding in the orientation of arrows within the causal graph [119].

Central to CIM is the concept of a mixture of DAGs, which represents various potential causal structures within the dataset. Each sample may adhere to a different DAG, reflecting the variability in causal processes across different conditions or populations. This flexibility allows CIM to accommodate scenarios where causal relationships may not be consistent across all samples [131].

Algorithmically, CIM functions by summarising the myriad of causal relationships from the mixture of DAGs into a cohesive graph. This graph encapsulates the most consistent causal directions observed across the diverse DAGs. Moreover, CIM is adept at handling cyclic patterns, unlike traditional models that struggle with cycles; it interprets cyclic causal processes through multiple DAGs, each offering a distinct view of the cyclical progression at various time points [119].

$$p(\mathbf{X}, \mathbf{T}) = \prod_{i=1}^{p} p(T_i) \prod_{i=1}^{p} p(X_i | \mathrm{Pa}_{\mathbf{T}}(X_i))$$

This equation represents the joint probability distribution of a set of variables X and a set of conditions T. $\prod_{i=1}^{p}$: The product over all variables, indicating that the joint

distribution is the product of individual distributions. $p(T_i)$: The probability distribution of each condition T_i. $p(X_i|Pa_{\mathbf{T}}(X_i))$: The conditional probability distribution of each observed variable X_i, given its parents $Pa_{\mathbf{T}}(X_i)$ in the DAGs.

2.2.2 Score-based

Score-based algorithms in causal discovery aim to identify the optimal causal model by maximising a scoring function, denoted as $S(G,D)$, which evaluates the fit of a given graph G to the observed data D [83]. The objective is to find the graph G^* that maximises this score [139] :

$$G^* = \arg\max_{G \in \mathcal{G}} S(G,D)$$

Here, $\mathcal{G}$ represents the space of all possible graphs.

2.2.2.1 Greedy Equivalence Search (GES)

The GES algorithm [11] has forward and backward search phases, beginning with an empty graph, GES iteratively adds or removes edges to improve the score. It assumes that the true causal structure can be inferred from the data under the faithfulness assumption, which aligns observed conditional independencies with d-separation statements in the graph [139]. In our case, GES employs the Bayesian Information Criterion (BIC) [90] as its scoring function, defined as :

$$\text{BIC}(G,D) = \log P(D|G) - 2\log n \cdot |G|$$

Here, $P(D|G)$ represents the likelihood of the data given the graph, n is the sample size, and $|G|$ is the number of parameters in the model. BIC seeks to strike a balance between the likelihood that a model assigns to the observed data and the number of parameters in the model. Therefore, if two models, M and M', have equal likelihood but M has fewer parameters, the BIC for M would be higher than the BIC for M'.

2.2.2.2 NOTEARS

NOTEARS [140] is distinct from GES by framing the problem as a continuous optimisation task. It exploits the fact that acyclic graphs can be characterised by the trace of the matrix exponential. NOTEARS minimises a loss function while ensuring the graph remains acyclic, represented as:

$$\min_W L(W,D) \quad \text{subject to} \quad h(W) = 0$$

Here, W is a weighted adjacency matrix representing the graph, $L(W,D)$ is the loss function, and $h(W)$ ensures acyclicity. Typically, $h(W)$ is implemented as $\text{tr}(e^{W \circ W}) - d$ [139], where tr denotes the trace operator, $e^{W \circ W}$ is the matrix exponential of the Hadamard product [57] of W, and d is the number of variables. NOTEARS efficiently discovers a graph structure without resorting to combinatorial search [139].

2.2.3 Functional Causal Models

2.2.3.1 Linear Non-Gaussian Acyclic (LiNGAM)

The LiNGAM algorithm [113] operates under the assumption that the data is generated by a linear acyclic model with non-Gaussian distributions, which allows for the identification of the causal order among variables. The core equation representing the LiNGAM model is $X = BX + E$ [45], where X is a vector of observed variables, B is a matrix representing the causal coefficients between variables, and E is a vector of non-Gaussian independent error terms. The algorithm seeks to estimate the matrix B by exploiting the non-Gaussian nature of the data, which allows for the identification of the causal direction, a task not possible with Gaussian data due to its symmetry. LiNGAM's approach is unique because it combines Independent Component Analysis (ICA) [113] with causal discovery, providing a statistically sound method for inferring causality from purely observational data without the need for controlled experiments.

2.2.3.2 Longitdinal LiNGAM

In Longitudinal LiNGAM [65], the core equation is modified to:

$$X(t) = \sum_{\tau=0}^{l} B(t, t-\tau)X(t-\tau) + E(t)$$

where $X(t)$ represents the observed variables at time t, $B(t, t-\tau)$ are matrices that describe the causal effects from variables at time $t-\tau$ to t, and $E(t)$ is a matrix of non-Gaussian external influences at time t. This model allows for the causal coefficients to change over time, reflecting the temporal evolution of the causal structure.

2.3 Evaluation Metrics

2.3.1 Structural Hamming Distance (SHD)

The Structural Hamming Distance (SHD) [125] is a metric for evaluating the performance of causal discovery algorithms, which works by identifying discrepancies between the learnt and the true model, guiding improvements in the learning process. A lower SHD indicates that the learnt model is closer to the true model, and thus, it has more accurately captured the underlying causal structure.

Algorithm 2 Structural Hamming Distance (SHD)

function SHD(G_1, G_2)
    $E_1 \leftarrow$ Edges of G_1
    $E_2 \leftarrow$ Edges of G_2
    $E_{\text{diff}} \leftarrow E_1 \oplus E_2$ ▷ Symmetric difference of edge sets
    $s \leftarrow |E_{\text{diff}}|$
    return s
end function

2.3.2 Structural Intervention Distance (SID)

The Structural Intervention Distance (SID) [97], unlike the SHD, which counts the number of incorrect edges, the SID considers the causal inference statements that can be made from a graph, and is therefore well-suited for evaluating graphs that are used for computing interventions.

It counts the number of pairs of nodes (i, j), where $i \neq j$, for which the intervention distribution from i to j is falsely inferred by the learned graph with respect to the true graph [97]. This is done based on a graphical criterion that checks whether the parent sets in the learned graph are valid adjustment sets for interventions in the true graph. Therefore, a lower SID indicates that the learned graph is closer to the true graph, making it a useful metric for evaluating the performance of causal discovery algorithms.

2.3.3 PageRank

PageRank [52], originally devised for ranking web search results, represents a link analysis algorithm applicable to directed graphs. It attributes a weight, referred to as the 'PageRank score', to each node within the graph, reflecting the node's relative significance based on its connectivity. The fundamental principle asserts that a node garners importance if it is linked to by other significant nodes. Mathematically, PageRank is formally defined as follows:

$$PR(v) = (1-d) + \sum_{u \in In(v)} \frac{PR(u)}{Out(u)} \tag{2.2}$$

where d is a damping factor (between 0 and 1) that accounts for the possibility of randomly jumping to any node in the graph. $In(v)$ denotes the set of nodes that have directed edges pointing to node v (in-links). $Out(v)$ denotes the number of outgoing edges from node u.

2.4 MICE Imputation

Multiple Imputation by Chained Equations (MICE) [127] stands out among various imputation methods due to its ability to mitigate bias in subsequent analyses conducted on datasets [85]. It has been observed to enhance feature selection compared to basic imputation techniques or leaving data incomplete [85]. MICE operates by sequentially fitting regression models for each variable containing missing data, conditioning on the other variables present in the dataset. This method offers adaptability in accommodating diverse variable types and has demonstrated efficacy even in datasets comprising thousands of observations and variables [53, 121].

The steps involved in the algorithm can be broken into [15]:

1. **Initial Imputation**: Begin by performing a simple imputation method, such as mean imputation, to fill in missing values across the dataset.

2. **Reset Placeholder Imputations**: Reset the placeholder imputations for a specific variable, effectively reverting them back to missing values.
3. **Regression Imputation**: Use the observed values of the variable from step 2 and regress them against other variables in the dataset. This regression model treats the variable of interest as the dependent variable and utilises the other variables as independent predictors. Replace the missing values for the variable with predictions obtained from the regression model.
4. **Iterative Cycling**: Repeat steps 2-3 for each variable with missing data, constituting one iteration or cycle. This iterative process ensures that missing values are successively imputed using regression models reflecting observed data relationships.
5. **Multiple Iterations**: Iterate steps 2-3 for a specified number of cycles, with imputations being updated at each cycle. This allows for refinement of imputed values over successive iterations [15].

MICE relies on the Missing at Random (MAR) assumption, which implies that the missingness in the data can be explained by the observed variables and not by the missing values themselves [15]. However, it should be noted that reliance on the MAR assumption necessitates cautious interpretation of results to evaluate the robustness of findings [117], and that implementing MICE when the data is not MAR could result in biased findings [15].

2.5 SHapley Additive exPlanations (SHAP)

SHAP [81], rooted in game theory [51], provides a comprehensive framework for attributing feature importance in predictive models. It assesses the contribution of each feature to the prediction of a specific instance by computing Shapley values [51], which are grounded in equitable payoff distribution in cooperative games. The fundamental concept revolves around measuring how much the prediction changes when a feature is included versus excluded from the model, averaged over all possible feature combinations [81]. A simplified mathematical expression for Shapley value estimation is as follows:

$$\phi_i(f,x) = \sum_{S \subseteq F \setminus \{i\}} [f_x(S \cup \{i\}) - f_x(S)]$$

where: $\phi_i(f,x)$ represents the SHAP value for feature i within prediction model f and instance x. F denotes the set of all features. S is a subset of features excluding feature i. $f_x(S)$ is the prediction of the model f when only the features in set S are known. The term $f_x(S \cup i) - f_x(S)$ represents the marginal contribution of feature i when added to the set S.

Chapter 3

Dataset

This study utilises data from the English Longitudinal Study of Ageing (ELSA) [16], a nationally representative longitudinal study tracking the health and well-being of individuals aged 50 and above in England. ELSA, initiated in 2002 and conducted biennially (one wave), collects extensive data on health, economic status, social participation, and well-being, comprising thousands of variables per participant. Our analysis focuses on a 14-year follow-up period from 2002 to 2016. After pre-processing, our temporal dataset includes 15,570 entries, of which 359 have been reported as having had dementia; our atemporal dataset includes 9,066 entries, of which 359 have been reported as having had dementia. For each participant, we consider the 15 features listed in Section 3.1.

3.1 Feature Selection

The underlying dataset used in this research extends the work from MInf1 [84], where we selected a subset of 226 features from the thousands available to us [16] by conducting thorough literature review on causal relationships associated with dementia. We create two separate datasets, which we pre-process separately. The first is a temporal dataset, which makes use of the longitudinal aspect of the study, and the second is an atemporal dataset, which ignores temporal aspects by only considering one entry per participant.

For the atemporal dataset, the objective was to retain a single entry for each unique participant. To achieve this, we established the following selection criteria:

1. If dementia was reported in any entry for a participant, that entry is selected.
2. Otherwise, the entry with the highest number of non-NAN values was chosen.

This approach aimed to maximise the representation of the event of interest (dementia) while addressing data missingness concerns [107]. However, 226 variables is too many to feasibly model in a causal DAG [28], especially since we will create a ground truth DAG with a domain expert. To facilitate further feature selection, interpretable machine learning (IML) techniques were applied to the temporal survival machine learning

model trained in MInf1 to identify the covariates most closely associated with dementia. Expanding on the methodology proposed in MInf1 [84], where covariate occurrences were counted in each IML model, the approach now incorporates the examination of scores produced by three distinct IML models. We standardise the scores produced by each model using min-max standardisation [87], and show the results in Figure 3.1.

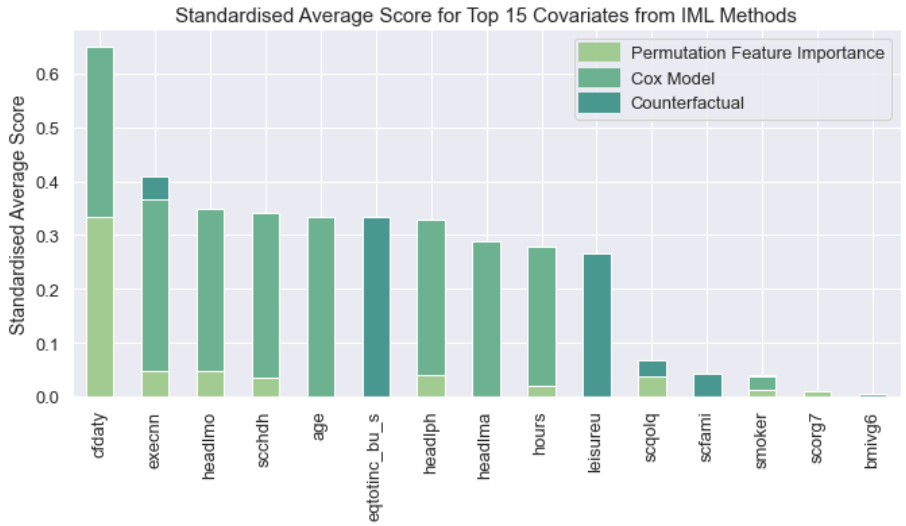


Figure 3.1: The 15 highest standardised average scores from covariates, as identified by three distinct interpretable machine learning techniques.

Figure 3.1 illustrates the 15 covariates with the highest standardised average scores across the three IML models. We use these identified 15 covariates as our features which we investigate throughout this study, a more detailed view of the data processing steps are shown in Figure 4.1. It is noteworthy that many of the top-ranked features pertain to cognition, likely symptomatic rather than causative factors of dementia, as indicated by the causal graph in Figure 4.2a. This underscores the imperative of investigating from a causal perspective - further discussion on this aspect is provided in Section 5.1, where we contrast the feature importance derived from a purely IML-driven approach against a causality-driven approach, to answer Question 3 (Section 1.3).

3.2 Handling Missing Data

The dataset utilised in our research exhibits an approximate 4% incidence of missing values. Direct application of existing causal discovery algorithms to incompletely observed data may lead to erroneous conclusions [135]. Consequently, the adoption of an imputation technique becomes imperative [78] to address these missing data entries.

In causal discovery datasets characterised by a small number of mostly discrete variables, such as ours, MICE [128] emerges as the preferred method for handling missing data, as indicated by the findings of the Witte et al. [138]. MICE outperforms test-wise deletion in such settings, demonstrating its effectiveness in preserving data integrity for accurate causal discovery, and MICE finds consistent application within the realm

of causal discovery [41, 58]. We use MICE to impute missing values in our temporal dataset, given its larger size and the proven efficacy of MICE on larger datasets [34]. Furthermore, since our atemporal dataset is derived from our temporal dataset (Section 3.1), employing MICE for imputation aligns with our data preprocessing strategy.

Figure 3.2 illustrates the impact of imputation on the distribution of select variables before and after the imputation process. The overall distribution remains consistent, albeit with certain features exhibiting a higher count post-imputation (e.g., 'execnn'), indicating a greater number of values that required imputation. Nonetheless, the fundamental distribution of values remains unchanged. It should also be noted that not all of our CSD methods used this imputed data, e.g. MVPC handles its own imputation.

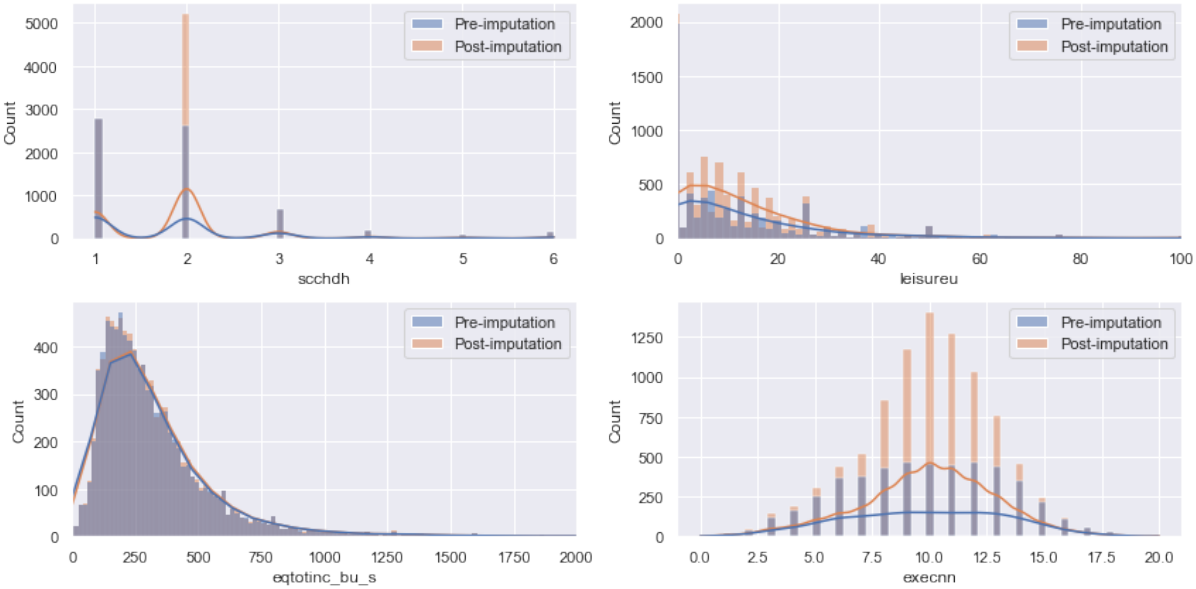


Figure 3.2: Figures showing the distributions of a subset of the variables chosen, before performing imputation, and after performing imputation; for the temporal dataset.

Figure 3.3 demonstrates that participants within the dementia group are evenly distributed among the rest of the study participants. This observation is essential, as an uneven distribution would cast doubt on the credibility of the data-generating process, potentially compromising our ability to make reliable causal assumptions [31]. Moreover, the figure indicates a substantial representation of participants in the dementia group, thereby enhancing the informativeness of our causal discovery models.

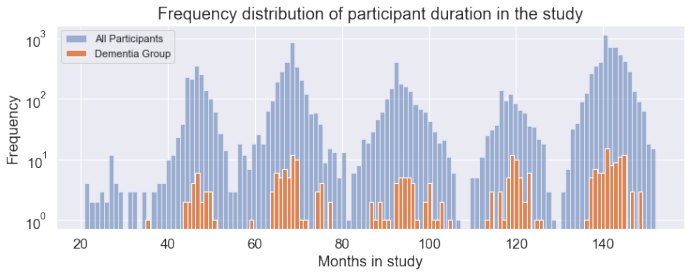
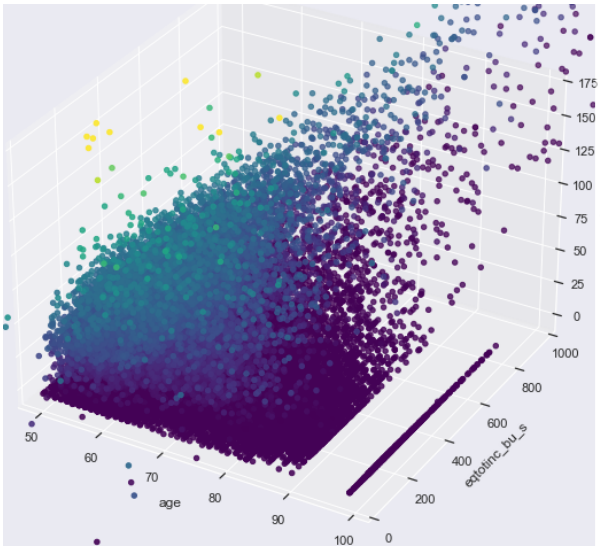


Figure 3.3: The duration that participants have stayed in the study.

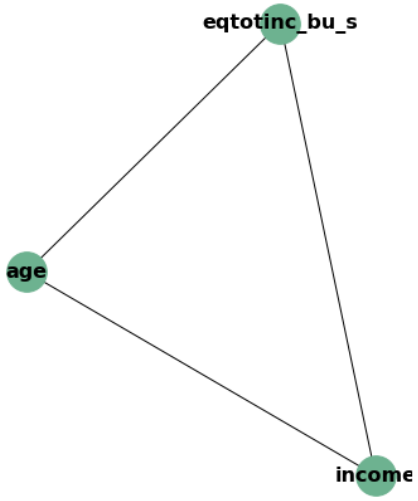
3.3 Causal Investigation

In this section, we aim to substantiate the need of investigating from a causality perspective, rather than a purely IML perspective. We initiate this rationale by illustrating the presence of Simpson's paradox [134] within the dataset and identifying potential spurious correlations highlighted in our prior research, MInf1 [84]. In MInf1, 'scchdh' emerged as the most recurrent variable in our IML explanations. We present the relationship between 'scchdh' and dementia in Appendix B.1, revealing a conspicuous non-linear association; however, upon stratification by age (depicted in Appendix B.2), significant variations in the relationship surface, challenging its validity when age is factored in. Notably, a discernible trend emerges whereby the correlation between 'scchdh' and dementia fluctuates as age advances, suggesting age to be a confounder.

Figure 3.4a showcases potential relationships between variables within the dataset - specifically, the illustration underscores correlations between age, income, and hours worked. However, relying solely on correlations fails to explain the directional causality between these variables, as depicted in Figure 3.4b, where only an undirected graph of relations is discernible. To unveil the underlying causal mechanisms, a more sophisticated approach is imperative. Indeed, the genuine relation between these variables is discernible from the ground truth in Figure 4.2a (right-most branch).



(a) The distribution between the variables denoting the age of participants, their income (eqtotinc_bu_s), and the number of hours they work in their jobs.



(b) Undirected graph showing the relationship we can infer from simply observing correlations, as in Figure 3.4a

Figure 3.4: Figures showing the existence of a correlation between variables in the dataset, and what we can infer from observing these correlations.

Chapter 4

Experiments

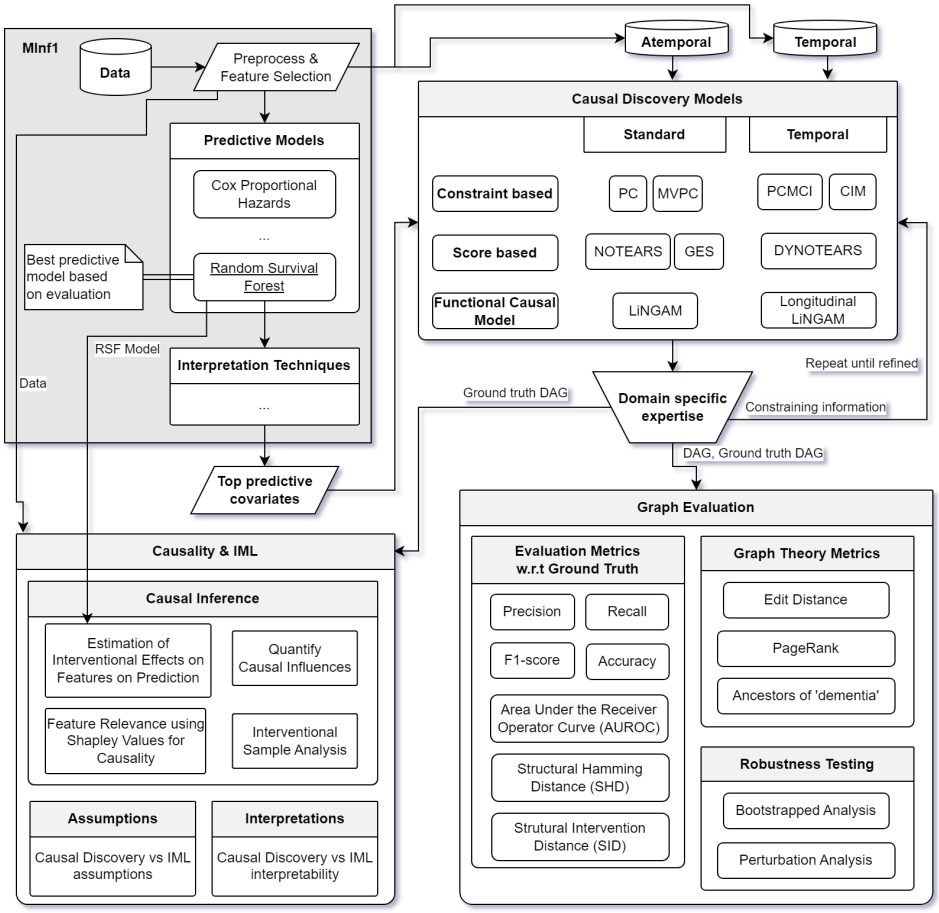
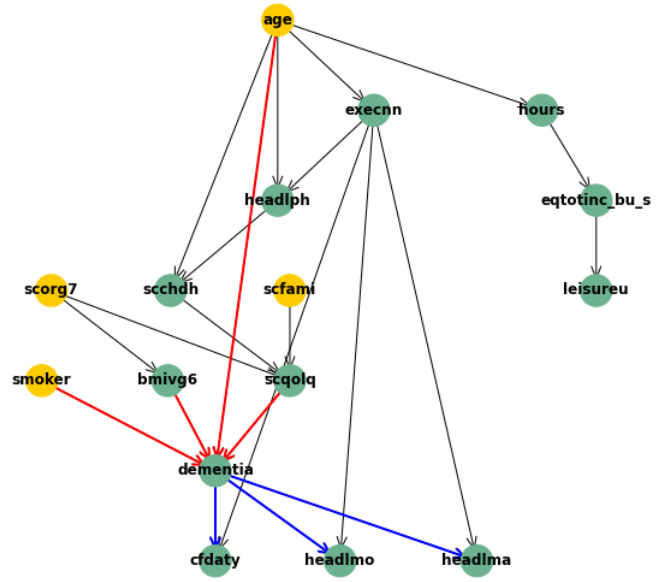


Figure 4.1: A pipeline showing the general methodologies followed in the research.

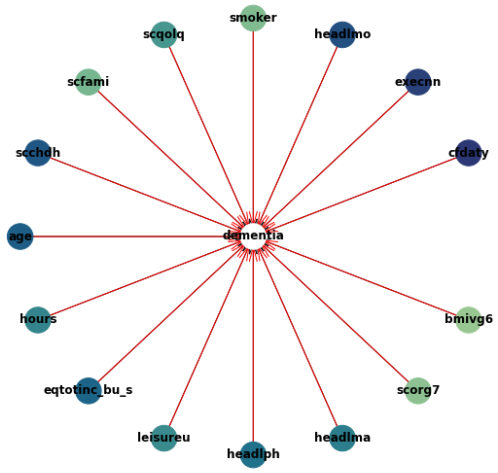
In this chapter, we outline the experiments conducted and present their results, setting the stage for a detailed discussion in Section 5 where we correlate these findings with our research questions. The pipeline illustrates the process of feature selection based on the outcomes of MInf1 (top left, discussed in Section 3.1), the application of SCM

methods (top right, discussed in Section 4.2), the methodology for evaluation (bottom right, discussed in Section 4.2), and the techniques enabling inference based on the results (bottom left, discussed in Section 4.3 and 4.4).

4.1 Defining Ground Truth



(a) The ground truth DAG, as created with the assistance of a domain expert.



(b) Understanding used by IML techniques, as in MInf1. The darker the colour, the more predictive the covariate, as per the IML approach followed in Figure 3.1

.

Figure 4.2: The ground truth causal DAG curated with a domain expert in comparison to the causal understanding of feature importance methods underlying machine learning models, as in MInf1 [84]. The variable definitions can be found in Appendix A.

The ground truth causal graph comprises nodes delineated in yellow to denote source nodes—those without incoming connections. Notably, in our model, these variables lack direct causes, although in practise we acknowledge the existence of latent variables [22] affecting them; however practical constraints prevent modelling every potential latent variable [21]. The graph acknowledges the potential influence of confounding factors on relationships, including transitive associations stemming from indirect relationships. Expert domain knowledge guided the curation of this ground truth, specifically tailored within the context of our dataset and variable of interest (dementia).

To validate its coherence, comparison with existing studies employing causal DAGs in dementia research is imperative. For instance, [76, 59] elucidate causal DAGs outlining hypothesised causal effects among study variables, revealing similarities with our graph. Notably, both depict age as a direct cause of dementia while also affecting intermediary factors like 'execnn' before culminating in dementia. Similarly, depression's causal link to dementia in [76] mirrors our depiction of 'scqolq' (life satisfaction) impacting dementia, amongst other similarities.

However, it is crucial to note that while such comparisons provide valuable insights, they do not constitute a robust method for forming a causal diagram. This is because they often fail to account for the underlying data generation process. For example, some variables in our study are self-reported, whereas other studies may not assume this. Therefore, we have meticulously accounted for such nuances in our ground truth graph, ensuring its fidelity to the data at hand.

4.2 Evaluating Causal Discovery Models

Reisach et al. [100] highlight that existing benchmarking setups for causal structure learning, particularly for additive noise models (ANMs), may have unintended patterns or regularities that can be exploited by certain algorithms. Specifically, the paper shows that in commonly used simulation schemes for ANMs, the marginal variances of the variables tend to increase along the causal order, a property termed "varsortability" [100]. This pattern can be leveraged by some algorithms, leading to unexpectedly strong performance on the benchmarks, which may not transfer to real-world scenarios. We use the sortnregress model as a benchmark, as proposed by [100], which accounts for the varsortability in the data, and is claimed to be a superior benchmark for that reason.

We devised a bootstrapping algorithm aimed at mitigating the issue of excessive edge presence within a CD generated graph. The challenge arises from the complexity introduced by numerous edges, which can obscure meaningful relationships and hinder interpretation [33]. To address this challenge, we developed an algorithm that constructs a consensus graph by generating multiple bootstrap samples from the original dataset. This consensus graph retains only edges that manifest in a significant majority (at least 60%) of the bootstrap samples, thereby reducing the prevalence of excessive edges, which often contribute to a high false positive rate; this occurs when the model identifies numerous edges that do not exist in the ground truth, resulting in a low precision score. Bootstrapping is crucial as it mitigates the risk of spurious relationships and enhances the generalisability of the model by focusing on the most consistently occurring relationships. The psuedocode in Algorithm 3 captures the general notion of how we implemented this bootstrapping procedure.

Algorithm 3 Pseudocode for Bootstrapping Algorithm

Require: $model, data, numSamples, sampleFraction, pruneThreshold$

$sampleSize \leftarrow \lceil len(data) \times sampleFraction \rceil$
$edgeFrequency \leftarrow \{\}$
for $i \leftarrow 1$ to $numSamples$ **do**
    $sample \leftarrow$ draw random sample of size $sampleSize$ from $data$ with replacement
    $predictedGraph \leftarrow$ model prediction on $sample$
    $updateEdgeFrequency(predictedGraph, edgeFrequency)$
end for
$consensusGraph \leftarrow$ filter edges in $predictedGraph$ based on $edgeFrequency$ and $pruneThreshold$
return $consensusGraph, edgeFrequency$

In order to obtain the results in Table 4.1, we performed the bootstrap process outlined earlier with 10 samples for each model, comparing the output graph with our ground truth graph to compute the metrics shown in the table. From the table, it is evident that PCMCI, NOTEARS, and DYNOTEARS perform notably worse than other models by a significant margin; qualitative evaluation of their graphs corroborates these findings. The poor performance of PCMCI and DYNOTEARS can be attributed to their tailored focus on temporal datasets [74]. Unlike our longitudinal dataset with thousands of samples from different participants over time, these methods are designed for following a single participant over a prolonged duration. Consequently, a considerable amount of data remains unused by the model, leading to a loss of information and explaining their subpar performance compared to the benchmark. The poor performance of NOTEARS is discussed in detail in Section 5.4.

Our results align with those reported by Strobl et al. [119], indicating that CIM outperforms other constraint-based algorithms across two real datasets with known ground truths. Both GES and CIM exhibit strong performance when bootstrapped, with each excelling in a particular metric (highlighted in bold). However, GES tends to produce an excessive number of edges, leading to a relatively lower precision score and consequently a lower F1-score.

Notably, MVPC demonstrates superior performance compared to PC with MICE imputation, consistent with the findings of Tu et al. [126]. Additionally, longitudinal variants of algorithms, such as Longitudinal LiNGAM and CIM, generally outperform their counterparts across most metrics. Further discussions on these observations are provided in Section 5.

Model Class	**Temporal**	**Model Name**	**SHD**	**SID**	**Precision**	**Recall**	**F1-score**
Benchmark	*N/A*	Sortnregress	65	141	0.114	0.255	0.158
Constraint	✗	PC	48	123	0.184	0.409	0.254
		MVPC	44	108	0.212	0.400	0.277
		PCMCI	62	84	0.078	0.156	0.104
	✓	CIM	**31**	68	**0.427**	0.406	**0.416**
Score	✗	NOTEARS	61	140	0.107	0.273	0.154
		GES	49	**56**	0.259	**0.682**	0.375
	✓	DYNOTEARS	64	157	0.098	0.151	0.119
Functional	✗	LiNGAM	71	71	0.143	0.5	0.222
	✓	Longitudinal LiNGAM	45	65	0.351	0.395	0.372

Table 4.1: Performance metrics achieved by CSD models in comparison to the ground truth (with bootstrap, without basic knowledge). The values in **bold** highlight the best performing model for that metric.

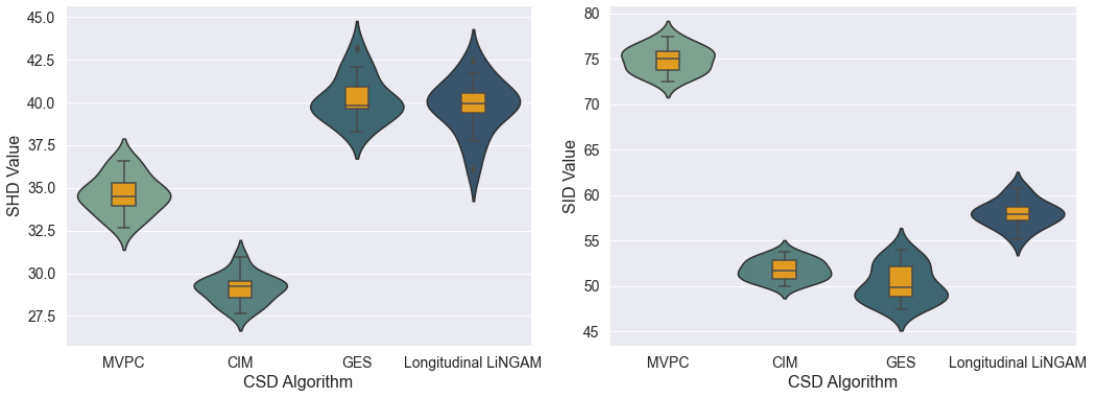
4.2.1 Infusing Prior Knowledge

Upon qualitative analysis of the graphical outputs produced by each method, it became apparent that enforcing a rule to blacklist certain edges could enhance the performance of the models and yield more interpretable graphs. This process, known as applying prior knowledge to our causal model [30], involves restricting certain edges from appearing in the resulting causal graph [32].

Shen et al. [111] argue that providing CSD algorithms with ample prior knowledge leads to optimal results. However, we acknowledge that this approach may undermine the essence of causal discovery itself. If prior knowledge were readily available, the need for causal discovery would be unnecessary. Therefore, we opt for a basic form of prior knowledge by enforcing the rule that $G' = (u,v) \in E(G) | v \neq$ "*age*", where G' denotes the resulting graph after incorporating knowledge and $E(G)$ represents the set of all edges in the original DAG G. This information does not necessitate input from a domain expert, aligning with the premise of our posed Question 3 - further discussion on this topic is presented in our proposed framework in Section 4.4.

Figure 4.3 illustrates the distance metric values between the ground truth and four of the best-performing causal models selected across 20 runs. It is evident that CIM consistently achieves the best SHD value, with a mean of 29.2 and the smallest variance. However, for SID, GES demonstrates slightly superior performance, exhibiting a lower mean value of approximately 3 units but with higher variance. The higher variance in GES's SID could be attributed to its nature as a score-based algorithm that operates in a greedy fashion. It employs strategies such as adding, removing, or reversing edges to maximize a scoring function [112]. This approach may lead to the convergence on local optima that do not accurately reflect the true causal structure, consequently impacting the SID metric.

Overall, when considering both SHD and SID metrics, CIM emerges as the top-performing model.



(a) SHD metric for CSD models

(b) SID metric for CSD models

Figure 4.3: Performance metrics achieved by select CSD models in comparison to the ground truth (with bootstrap, with basic knowledge) over 20 runs.

4.2.2 Robustness Testing

To assess the robustness of our causal discovery approach, we followed a methodology similar to that outlined by Volkova et al. [132]. This method can be described by the Formula 4.1, where robustness is a measure of the consistency of edges at different levels of perturbation. Our evaluation began with a small sample size of 10% of the data and incremented the sample size to 10%, to analyse the algorithm's sensitivity to varying sampling proportions.

$$R = \frac{1}{n} \sum_{i=1}^{n} \frac{f_i}{N} \tag{4.1}$$

Where n is the total number of directed edges under consideration. f_i is the frequency of the i^{th} directed edge appearing in the graphs. N is the total number of graphs.

Our findings, as depicted in Figure 4.4, indicate that CIM and GES consistently demonstrate higher stability compared to MVPC and Longitudinal LiNGAM across all sample portions. This observation contrasts with the results reported in [132], where the ensemble approach exhibited lower stability in comparison to individual algorithms across all sample sizes. Notably, CIM stands out as the only ensemble approach utilised in our study, and it appears to be the most robust. We hypothesise that this discrepancy may stem from the utilisation of simulated datasets in the aforementioned study, and as documented in [100, 116], are known to be unrepresentative of real-world datasets, as is the case in our investigation.

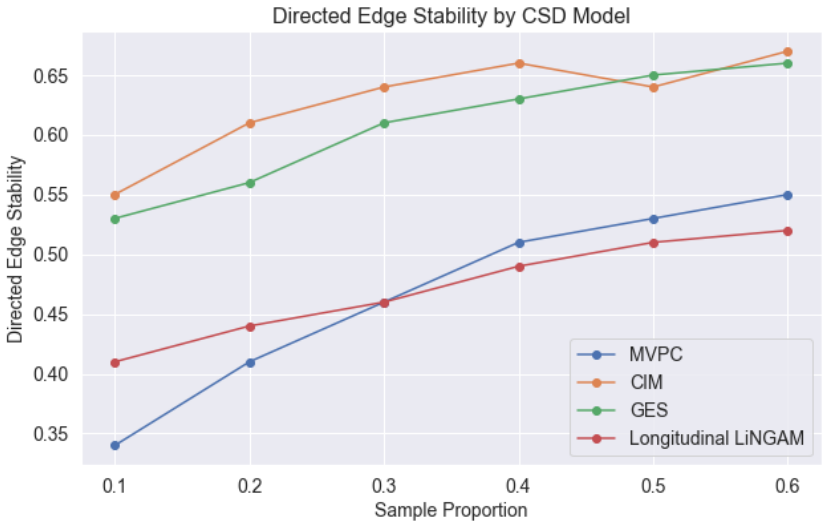


Figure 4.4: Robustness tests of the different CSD algorithms as a function of the portion of sampled data.

4.2.3 Further Evaluation on CIM model

From hereon, we refer to the CIM CSD model exclusively, given that we have shown it to be the superior model on our dataset based on the tests that we have done.

The figure in Appendix B.3 illustrates the outcomes of an algorithm we devised to iteratively evaluate the impact of random data removal on the metrics of the CIM model. Over three iterations, the algorithm randomly removes portions of the dataset, initially removing 5% of the data and incrementally increasing this percentage by 2.5% in each subsequent iteration. For each pruning level (i.e. portion of data to be removed), the algorithm calculates metrics of the CIM model using the pruned data and stores these metrics for analysis. Subsequently, it computes the average of these metrics, pooling together ten metric values obtained for each metric type across all pruning levels. We observe both the precision and recall begin to drop significantly after pruning 15% of the data, and the SHD spikes upwards after pruning 20% of the data; which suggests that the model's performance deteriorates notably beyond 15% pruning. Although it should be noted that these values vary considerably across different runs, which underscores the necessity for multiple iterations and further analysis.

PageRank (see Section 2.3.3) is a link analysis algorithm that can be used as a metric to compare causal models - Lin et al. [77] utilize PageRank for root cause analysis, while Nicole et al. [40] leverage it for making inferences from the graphical structure of causal graphs, similar to our approach. However, we extend the work of Nicole et al. [40] by applying PageRank to both our ground truth graph and our CIM model to quantitatively assess differences between the graphs.

The PageRank algorithm aims to gauge the importance of each node in a network, with higher scores indicating greater importance. Upon comparing PageRank scores in Table 4.2 from the ground truth with those from the CIM model, discrepancies emerge. For instance, 'dementia' remains the feature with the highest PageRank in both models, but its score decreases slightly in the CIM model. Conversely, 'scqolq' sees an increase in its PageRank score in the CIM model, suggesting that the model may attribute more importance to this feature than the ground truth does. While there are some variations in the PageRank scores between the ground truth and the CIM model, the overall ranking of features by importance is fairly consistent. This consistency is crucial for validating the CIM model's ability to approximate the ground truth in terms of feature importance within the network.

Ground Truth		CIM Model	
Feature	**PageRank**	**Feature**	**PageRank**
dementia	0.185	dementia	0.168
scqolq	0.116	scqolq	0.149
cfdaty	0.086	headlmo	0.091
headlmo	0.086	cfdaty	0.079
headlma	0.086	headlma	0.079

Table 4.2: The top 5 PageRank scores between the ground truth and CIM model.

We also perform perturbation analysis, whereby we see how the distance scoring metrics (SID, SHD) change when data pertaining to a certain node is removed. The algorithm initialises by bootstrapping the provided model with the data to create a reference graph, incorporating basic domain knowledge. It then iterates over each node, generating

a perturbed graph by excluding the current node from the dataset and bootstrapping the model. Simultaneously, it creates an ideal perturbed graph by duplicating the reference graph and removing the current node. Next, it calculates distances, including edit distance, structural Hamming distance (SHD), and structural intervention distance (SID), between the ideal and actual perturbed graphs. Finally, it stores the perturbed graph, ideal perturbed graph, and computed distances in a dictionary, using the current node as the key.

This analysis differs from PageRank, as we specifically re-train our underlying causal discovery model (CIM) without the variable that we are perturbating over, generating a completely new graph, which we then compare to our pruned ground truth graph.

Figure 4.5 illustrates the resultant distance scoring metrics from the perturbation analysis; notably, 'scqolq' and 'age' exhibit substantial distance measures, suggesting significant influence in the generation of the inferred graph. Conversely, 'leisureu' and 'headlph' demonstrate the smallest impact on distance measures, indicating minimal influence on determining the underlying causal graph. These findings align with expectations, as referring back to our ground truth in Figure 4.2a, 'headlph' acts as a mediating variable, implying its absence may not alter inferred causal relations, while 'leisureu' serves as a leaf node of limited branch significance, thus its omission minimally affects the graph structure. Overall, these results align well with the ground truth causal graph, as qualitative comparisons elucidate the coherence of distance metrics with the ground truth graph.

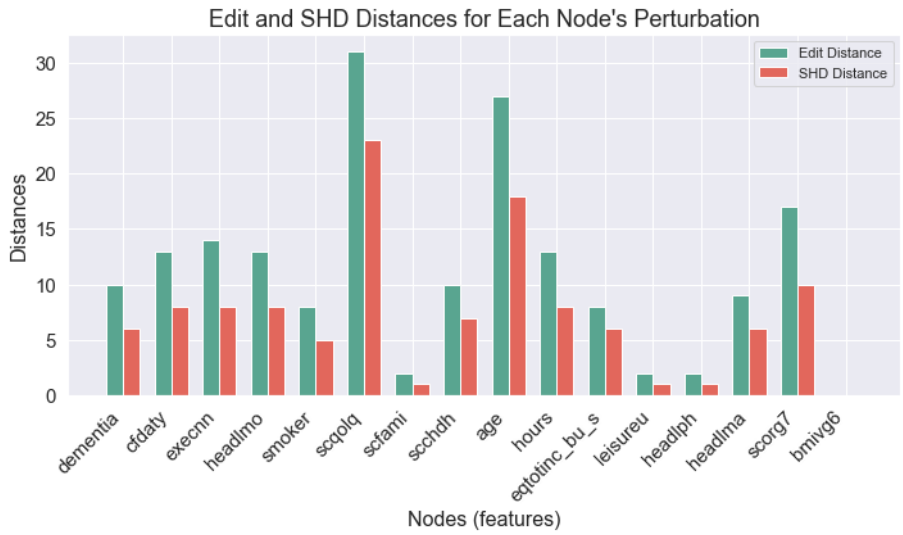


Figure 4.5: Distance metrics when specific variables are removed from the training data for the CIM model.

4.3 Feature Importance

4.3.1 Global Explanations

The study conducted by Janzing et al. [62] tackles the challenge of quantifying the causal impact between variables within a DAG framework. The authors propose a set of axioms that any measure of causal strength should adhere to and introduce a novel approach to quantify causal influence based on interventions. This method focuses on measuring the direct influence of one variable on another, excluding indirect pathways through intermediary variables; contrasting the variance in the target variable when the direct causal link from a parent node is ”removed” to the variance when the link is present. Mathematically, the causal strength is delineated as the Kullback-Leibler divergence between the original distribution P and the intervened distribution P_S (see Section 2.1.3).

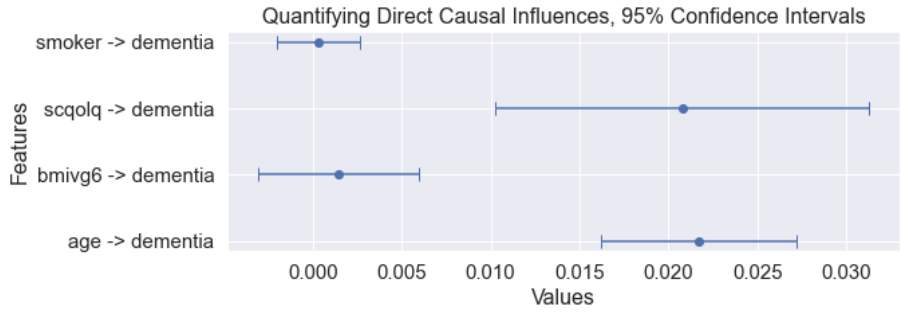


Figure 4.6: Quantified direct causal influences of various factors on dementia, based on ground truth trained on our dataset

From Figure 4.6, we observe that the direct influence from age to dementia (approximately 0.021) is notably stronger (roughly 21 times) than the direct influence from bmivg6 to dementia (approximately 0.001). In simpler terms, removing the arrow from ’age’ to dementia leads to an increase in the variance of dementia by approximately 0.021 units, while removing ’bmivg6’ results in an increase of approximately 0.001 units in the variance of dementia. By default, the scalar values for arrow strengths are measured in variance for a continuous real-valued target and in bits for a categorical target, typically represented by KL divergence [5]. From the figure, we observe that the presence of both ’scqolq’ and ’age’ linked to dementia significantly reduces the variance of dementia; we could interpret this as suggesting that these two variables hold greater feature importance, as highlighted by Janzing et al. [62]. Additionally, considering the narrower confidence interval for age and its slightly higher mean value compared to scqolq, we can infer that the feature importance for age is the greatest, followed by scqolq, bmivg6, and then smoking.

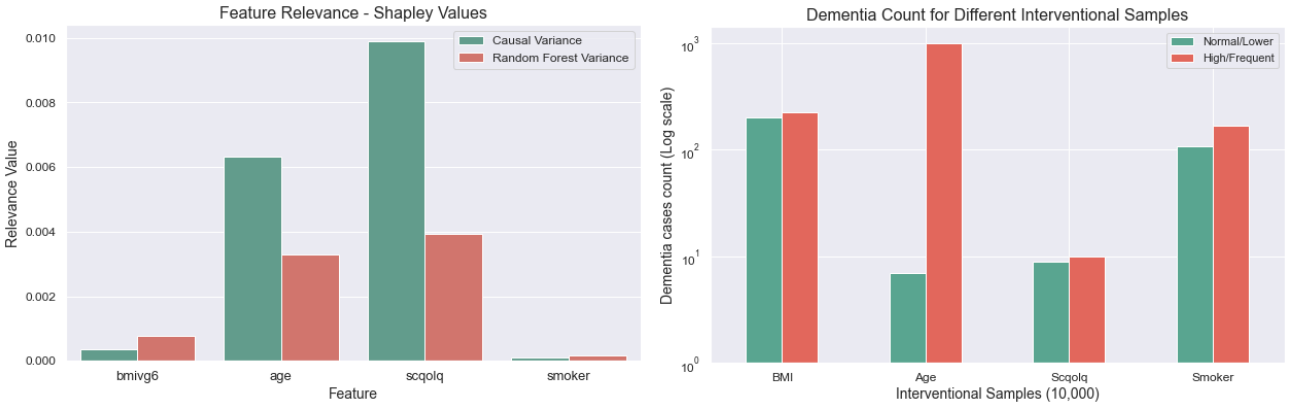
The study conducted by [63] discusses feature relevance quantification in IML [84] and argues that the standard SHAP approach (Section 2.5) of using conditional expectations to define ”simplified functions” when marginalising out unused features is conceptually flawed. Instead, the authors propose utilising unconditional or marginal expectations $E[f(x_T, X_{\bar{T}})]$ rather than conditional expectations $E[f(x_T, X_{\bar{T}})|X_T = x_T]$ when defining

the simplified functions $f_T(x)$ [63]. This suggestion aligns with the causal perspective, emphasising hypothetical interventions on input features X_T while keeping the remaining features $X_{\bar{T}}$ unaffected and sampling from their natural joint distribution.

Using this approach, Figure 4.7a illustrates the resulting SHAP values for direct causal influences on dementia, as derived from our ground truth graph. The green lines represent leveraging the causal structure directly to estimate global relevance. Conversely, the red lines depict utilizing this method with a black box predictor [6]; in this instance, a random forest model was employed. Notably, both the green and red values exhibit the same Shapley value ordering among the variables: 'scqolq', 'age', 'bmivg6', 'smoker'. The prominence of the variable 'scqolq' with the largest SHAP value suggests its significance as the most important feature according to this methodology [88].

In Figure 4.7b, we delve into addressing a causal inquiry: "what will happen to the variable dementia if I intervene on Y?", where Y represents one of the four direct causes of dementia in our ground truth. This is accomplished by simulating interventions and generating hundreds of interventional samples [8] for various scenarios. The scenarios are distinguished by green and red lines, with red indicating extremes, such as the following intervention tuples: (Age=55, Age=85), (smoker=0, smoker=1). Therefore, our focus lies in observing the disparity between the red and green lines for each variable, depicting the difference in dementia count across different values for each of the variables.

From the figure, it's discernible that age exhibits a substantial disparity in dementia outcomes, whereas the remaining variables show differences, albeit smaller. Notably, in instances where extreme values are selected for each variable (those in red), they consistently yield higher dementia outcomes than the green ones, indicating that they all contribute to some extent to the dementia outcome. However, it's important to exercise caution when interpreting the graph, as the y-axis is log-scaled, thereby accentuating differences. Thus, while the disparity in the BMI variable, for instance, appears lesser than that in Scqolq, this may not be the case in absolute terms.



(a) Feature importance using SHAP for causality. (b) Feature importance based on interventional samples.

Figure 4.7: Methods to infer feature importance, as a function of the ground truth graph and the longitudinal dataset.

4.3.2 Local Explanations

To facilitate the assessment of local explanations, we revisit the two individuals from our dataset that we had chosen in MInf1, shown in Figure 4.8. From this figure, individual 6545 represents a subject with a consistently low probability of developing dementia, whereas individual 1692 exhibits a progressively increasing likelihood of dementia during the study period. By employing the same individuals as in MInf1, we enable a direct comparison between local explanations derived from an interpretable machine learning (IML) approach, as in MInf1, and those derived from a causal approach in this study. This comparison is elaborated further in Section 5.3.

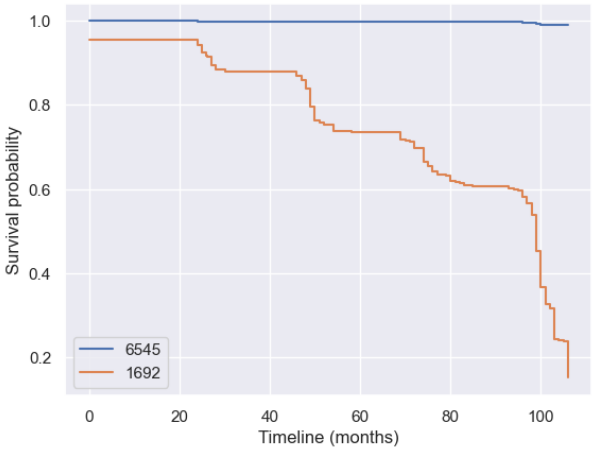


Figure 4.8: Figure from MInf1 [84] showing the survival curve of chosen individuals. A lower survival probability denotes a higher probability of dementia onset.

We can extend the SHAP analysis from Figure 4.7a, where we conducted a global analysis, to instead focus on local SHAP analysis [6]. Using Figure 4.9, we delve into individual observations from each of the two individuals described in Figure 4.8, aiming to explain why the high-risk individual exhibits a heightened risk of dementia based on the direct causes outlined in our ground truth graph.

From the figure, it becomes apparent that age emerges as the primary contributor to this individual's elevated risk of dementia, while scqolq (life satisfaction, see Table A.1) assumes significance in potentially mitigating their risk of dementia onset. This aligns logically with the given scenario, as the high-risk individual is aged 82 and reports high life satisfaction. Conversely, for the lower-risk participant, aged 52, one of the youngest individuals in the study (focused on the 50+ age group), we observe a negative relevance value.

Additionally, in both cases, the absence of smoking is evident, rendering smoking irrelevant in these scenarios. However, it's noteworthy that while smoking does not appear to decrease the risk of dementia in these instances, our analysis of other participants suggests that smokers tend to exhibit a higher risk, consistent with existing studies [27]. Interestingly, our model discerns that smoking individuals have an increased risk, yet does not infer a causative relationship in the opposite direction, suggesting a non-linear relationship between smoking and dementia.

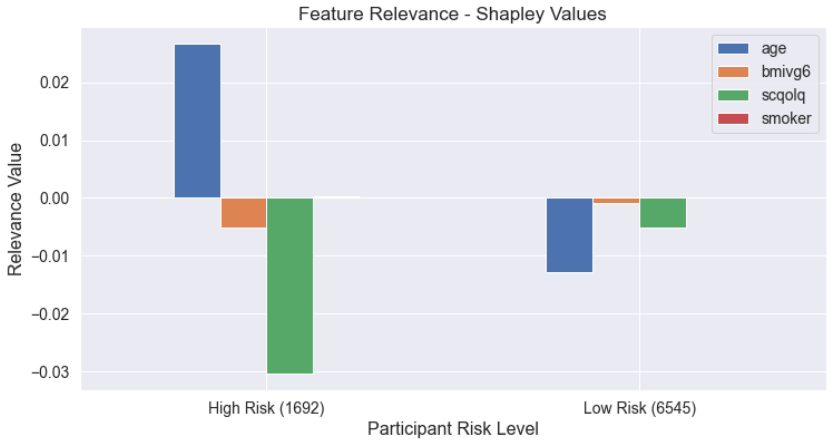


Figure 4.9: Local feature importance using SHAP for causality, as a function of the ground truth and the longitudinal dataset.

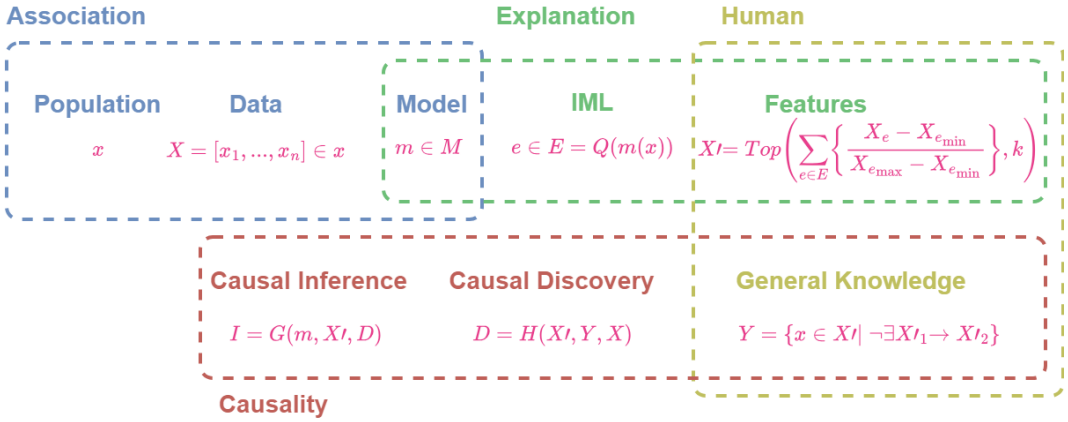
Interventions involve the simulation of future outcomes through the random sampling of noise, whereas counterfactuals estimate alternative pasts by reconstructing specific noise values [3]. The former approach necessitates fewer modelling assumptions and accommodates all data types, while the latter relies on stronger assumptions [93].

In Table 4.3, we present examples illustrating the minimum adjustments required in variables to potentially alter the dementia outcome for each of the two participants. Notably, for the participant deemed to have a higher risk, no such adjustments to variables were identified, potentially indicating the overriding influence of age. This inference aligns with the findings of our feature importance analysis, where age consistently emerges as a significant factor, a phenomenon also corroborated in existing literature [118, 129]. Conversely, for the participant categorised with a lower risk, adjustments such as modifying BMI to its extreme values (i.e., from 20-25 to 40+) lead to a change in the dementia outcome. It should be noted that these instances represent elementary modifications targeting single features; more intricate counterfactual scenarios could be devised by adjusting multiple variables - the intent was to show the feasibility of such analyses within our established causal framework.

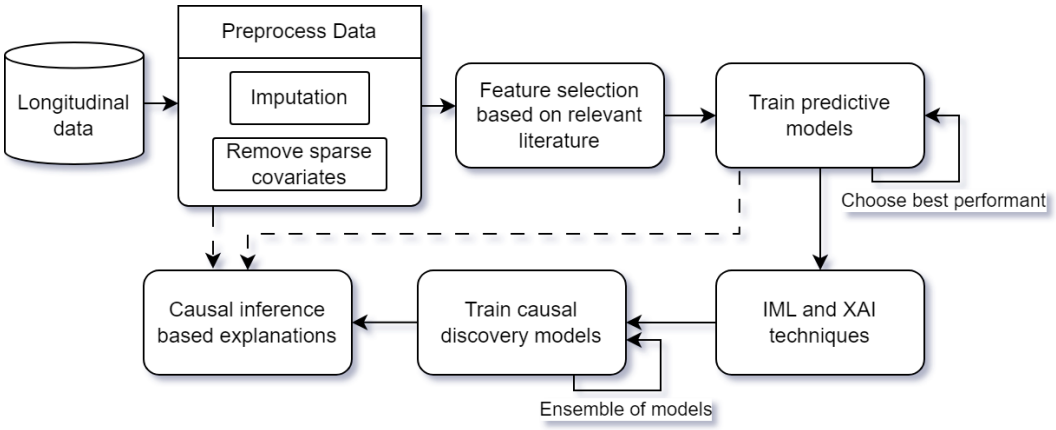
Person ID	**Feature Values to Change**				**Dementia Outcome**
1692	*Original*	-	-	-	1
	Proposed	-	-	-	0
6545	*Original*	bmivg6 = 2.0	scqolq = 2.0	scfami = 5.52	0
	Proposed	bmivg6 = 5.0	scqolq = 0.1	scfami = 0.09	1

Table 4.3: Counterfactual cases performed on our two participants, illustrating the adjustments necessary to the features to modify the dementia outcome.

4.4 Proposed Causal Methodology Utilising IML & CSD



(a) A theoretical guide to our proposed pipeline.



(b) A practical guide to our proposed pipeline.

Figure 4.10: A proposed pipeline that practitioners can follow, enabling the combination of IML & CSD techniques.

The objective of this section is to formalise and justify a methodology facilitating inference from longitudinal data with a large number of features, incorporating an IML and causal approach. We propose this framework under the premise that integrating causality and explainability within a unified model enhances its robustness beyond what can be achieved by employing either concept in isolation [99, 56].

As discussed in Section 1.2, current methodologies aim to show either how associative IML processes can be understood as causation using human knowledge [67], or attempt to derive inferences separately through the utilisation of IML and causal techniques as distinct entities [99]. Neither approach endeavours to integrate the two paradigms to facilitate inference from the data. We advocate for a framework that harnesses the predictive capabilities of state-of-the-art (SOTA) machine learning models in conjunction with IML techniques for feature selection (aimed at understanding associations within the data). Subsequently, in the absence of domain-specific knowledge, we propose

leveraging causal discovery models (augmented with what we term 'general knowledge'—basic prior knowledge devoid of domain specificity, as elucidated in Section 4.2.1, if available) to enable the resolution of causal inquiries and to enhance feature importance assessments based on the discovered causal graph.

Figure 4.10a draws inspiration from Figure 1 in [67] (also depicted in Appendix B.4), illustrating the progression from correlation to causation by leveraging human interpretations of IML explanations. We extend and adapt this framework to incorporate an intermediate causal discovery step, facilitating the inference of the causal structure of the model in the absence of a ground truth graph. Furthermore, we utilize the IML step for feature selection to input variables into the causal discovery model, considering the challenges associated with handling large graphs (comprising more than 50 nodes), as highlighted by [142].

4.4.1 Evaluating the Framework

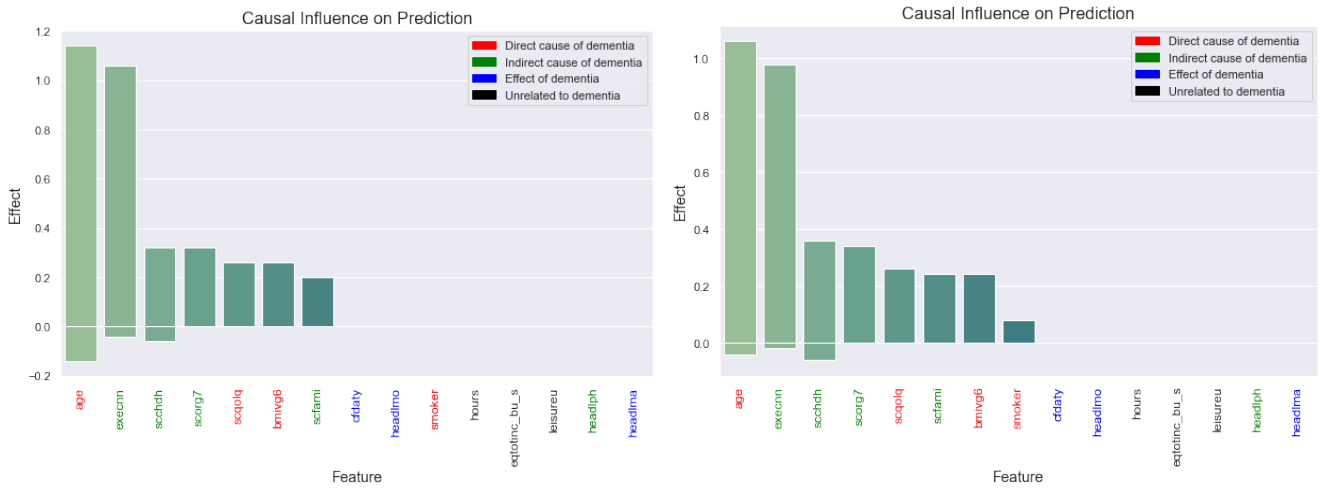
Comparison between the causal inferences derived from our CSD model's graph and the ground truth graph offers valuable insight into the efficacy of the proposed framework. Within the context of the pipeline in Fig. 4.10a, we used IML techniques and performed feature selection in Section 3.1, we infused general knowledge in Section 4.2.1, we performed causal discovery in Section 4.2. Therefore, comparing the inferences between our CSD model and the ground truth serves as a useful evaluation of the framework albeit in the constraints of our dataset.

The framework proposed in [25] presents a novel approach that bridges the gap between predictive models and causal reasoning. Traditionally, predictive models primarily focus on generating accurate predictions based on input features, often overlooking the underlying causal relationships between these features and the target variable. The framework outlined in the aforementioned paper enables the integration of the causal structure inherent in the data generation process with the causal structure implied by the predictive model (in our case, a random forest model). This integration facilitates the identification of the input feature exerting the greatest causal influence on the predicted output.

Figure 4.11 showcases the feature importance based on the aforementioned method. Specifically, Fig. 4.11a illustrates the results using the ground truth model, while Fig. 4.11b illustrates the results using the CIM causal discovery model - this allows for a comparison of the feature importances derived from the two causal graphs. Notably, the top five features selected by both models are identical, and they are arranged in the same order, moreover, the feature importance scores between the two models exhibit remarkable similarity for each of these features, suggesting that similar inferences can be made from both the ground truth and the CSD model. Interestingly, the CIM model assigns some importance to smoking, which is not evident in our ground truth model.

Each of the plots also shows the variables' causal relation to dementia with respect to the ground truth graph; in both models, only the direct and indirect causes of dementia are accorded importance. This observation is encouraging, as it shows the effectiveness of the method in describing the feature importance for all variables with respect to our

underlying causal graph. Unlike other methods discussed, which typically focus solely on direct causes, this method also highlights the significance of indirect causes while excluding effects or unrelated variables to dementia.



(a) Ground truth as the underlying causal model.

(b) CIM as the underlying causal model.

Figure 4.11: Estimate interventional effects on features on prediction model. This is a function of the ground truth graph, the longitudinal dataset, and a predictive model. The legend denotes the causal connection to dementia with respect to our ground truth graph.

While acknowledging that a significant limitation of this framework is the potential propagation of errors across the various statistical methods employed, we contend that such errors may already be present at each stage due to other inherent biases. For instance, the feature selection process necessitates human intervention to select a subset of covariates for examination or relies on alternative statistical methods, each carrying its own set of assumptions [102]. Unlike these methods, our machine learning model enables predictions to be made, enhancing its utility. Similarly, the causal discovery stage operates on substantial assumptions (as discussed in Section 5.4), compounded by the introduction of human biases during the incorporation of prior knowledge [30]. Rather than propagating these biases, our framework merely shifts their nature, replacing human bias during causal discovery with biases from our IML model for example.

In summary, the above methodology enables us to harnesses the predictive capabilities of SOTA machine learning models in conjunction with IML techniques for feature selection into causal discovery techniques. This approach instils greater confidence in the predictive model constructed, as it leverages the diverse array of methods discussed in this paper to derive inferences. Another advantage of the framework is its ability to make assumptions more transparent. By incorporating both IML knowledge and causal graph knowledge, along with various causal inference techniques, the framework allows for a clearer understanding of the underlying assumptions.

Chapter 5

Discussion

In this chapter, we revisit our original research questions (see Section 1.3) and provide insights into each of them based on the experiments conducted in the preceding chapter. Subsequently, we delve into an examination of the underlying assumptions inherent in our methodology and discuss their potential implications on the practical applicability of our findings across diverse datasets.

5.1 Feasibility of CSD algorithms

This section addresses our research question 1, exploring the extent to which CSD algorithms can capture the true causal graph.

The NOTEARS algorithm, has been critiqued for its lack of scale-invariance [66]; this deficiency is crucial as causal relationships should remain invariant regardless of the units or scales of measurement. The algorithm's sensitivity to scaling implies that mere rescaling of variables can yield different causal graphs, indicating its unreliability in uncovering causal structures. Recent work by [100] shows that the remarkable performance of some continuous structure learning algorithms can be attributed to high varsortability. In our study context, these models include NOTEARS and DYNOTEARS, both utilising ANMs [26]. We measured the mean varsortability of our dataset to be 0.59, which is close to the chance level at 0.5, potentially explaining the mediocre performance of both of these models. [100] suggests that current linear ANM structure learning algorithms may perform comparably to naive baselines on real-world datasets, consistent with our results shown in Table 4.1, where they match or under-perform relative to the baseline.

Additionally, we note that models that incorporate longitudinal temporality in the dataset outperform their counterparts. For instance, Longitudinal LiNGAM outperforms LiNGAM, and CIM outperforms PC, as evidenced in Table 4.1. However, models that do not specifically focus on longitudinal temporality, such as DYNOTEARS, perform worse than their counterparts (NOTEARS), likely due to their emphasis on single observations over time rather than multiple observations over time, as seen in longitudinal studies. Shen et al. (2020) [111] found that incorporating longitudinal data

even in standard causal models like PC enhances model performance. Although our attempt to use 'temporal data' instead of 'atemporal data' (see 4.1) resulted in minor improvements, they did not match the performance of the CIM model.

Our results from Figure 4.3 demonstrate that the CIM model outperforms both PC and MVPC across all metrics. The standard PC algorithm can address confounders to some extent, yet it has limitations; it typically assumes the presence of a single latent confounder that impacts all variables in the model [45]. However, this assumption can be restrictive as there may be multiple confounders affecting different aspects of the model [45]. In contrast, the CIM model is designed to handle confounders more effectively - it does not necessitate the assumption of a single latent confounder and can accommodate multiple latent variables as well as selection bias [119]. This renders the CIM model more flexible and potentially more accurate in inferring causal relationships from observational data where confounders are present.

In summary, temporal-focused models outperform their non-temporal-focused counterparts, and the use of longitudinal temporal data, even in non-temporal models, can confer performance benefits, echoing the findings of [111]. Overall, CIM emerged as the best-performing model, as further evaluated in Section 4.2.3 using perturbation and PageRank tests, as well as in Figure 4.11. Across all evaluations, we observed that the inferences drawn from the ground truth and the CIM model exhibit similarity, especially for highly predictive features like Age.

5.2 Primary Causal Factors Contributing to Dementia

This section aims to address our research question 2, focusing on identifying the primary intervenable causal factors contributing to dementia onset.

This is elaborated in Section 4.3, where we analyse both global and local explanations using methods from causal inference literature to integrate information from our ground truth causal graph and the longitudinal dataset. With the exception of the analysis in Figure 4.7b, each of our feature relevance models identifies age and scqolq as the top two major contributing factors to dementia onset by a significant margin. However, considering our research question from the perspective of modifiable or intervenable features, we exclude age for the purposes of this analysis. Scqolq, representing how often a participant feels satisfied in life (see Appendix A.1), can technically be included.

The next most important feature is BMI (bmivg6), consistently ranked higher than smoking in influencing dementia across all analyses. Table 4.3 provides counterfactual examples, hypothetically modifying the BMI of a low-risk participant to be morbidly obese [4], thereby increasing their likelihood of dementia. Conversely, all analyses indicate that smoking has the least causal influence on dementia, with no significant change observed in the dementia outcome solely by modifying smoking status. However, feature importance based on interventional samples in Figure 4.7b presents smoking as a significant factor. It's essential to note that this graph depicts the effect of interventions on dementia count, not the overall risk of dementia. Thus, while smoking may seem influential in this context, it does not necessarily imply that it is the most significant

risk factor for dementia overall. There may be a potential hidden confounder between smoking and dementia, although not modelled in our ground truth graph.

In summary, the general consensus for direct causes of dementia (as modelled in our ground truth) follows the order: $[Age, Scqolq, BMI, Smoker]$. While we cannot precisely verify whether this order aligns with other studies, each of these variables is known in the literature to be causally linked with dementia. Age stands out as the most prominent factor, supported by ample literature [14, 71]. Scqolq, although a specific variable, can be inferred to relate to depression, which is known to be linked to dementia [20]. Similarly, both BMI and smoking have established associations with dementia [12, 42].

It is noteworthy that Figure 4.11a extends beyond direct causes of dementia, also considering indirect causes. Comparing the set of features identified in this figure with those in Section 6.1 of MInf1 [84], the only similarity is 'scchdh'. This discrepancy arises because most of the other variables identified from MInf1 are effects of dementia rather than causes, as modelled in our ground truth. An interesting observation is that although 'headlph' is modeled in our ground truth as an indirect cause, Figure 4.11a indicates it to be a mediating variable, with no feature importance attributed to it.

This disparity highlights the importance of delving deeper into the associations from IML models to draw causal inferences, as our findings differ significantly from those of MInf1. We contend that our research yields more robust and practical insights, as it offers guidance on potential interventions rather than solely highlighting feature importance for model prediction.

5.3 Application of IML & Casuality

This section aims to address our research question 3, regarding the potential of integrating IML and CSD to yield superior insights, as briefly discussed in Section 5.2.

This refers back to proposed framework in Section 4.4. This framework is particularly useful as it allows for the extraction of actionable insights from complex data, enabling researchers to identify modifiable factors that could potentially delay or prevent the onset of dementia. However, caution is advised when implementing this framework due to the inherent complexities of causal inference. The assumptions made by causal models can significantly influence the results, and if these assumptions are incorrect or oversimplified, they can lead to erroneous conclusions. For instance, the framework assumes that the causal graph is faithful to the observed data, which may not always hold true in other real-world scenarios. Additionally, the potential for human bias in interpreting the outcomes of the model cannot be overlooked, as it can affect the validity of the causal relationships inferred, as highlighted in Section 4.4.

The framework stands out as a valuable tool for conducting feature importance analysis, particularly for modifiable risk factors. Its utility is underscored by the explicitness of its assumptions, which are made more transparent by the known underlying causal graph modelled from domain expertise or CSD. This transparency allows for a clearer understanding of the causal mechanisms at play and facilitates the identification of key factors that can be modified to potentially alter the course of dementia. Further,

it enables causal questions to be asked, such as counterfactual questions, which have found to be a very useful tool in practise to help with decision making [136].

In conclusion, while it is essential to approach the application of this framework with caution, this framework aims to provide a clearer pathway for researchers to navigate the intricate web of factors contributing to their outcome of interest, in this case, dementia.

5.4 Further Assumptions & Limitations

		Assumptions				
		Faithfulness	Distribution	Handles Confounders	Additional	Output
Constraint based	**PC**	Yes	No	No	No	MEC
	MVPC	Yes	No	No	Yes	
	CIM	Yes	No	No	No	
Score based	**NOTEARS**	Some	No	No	Yes	DAG
	GES	Some	Yes	No	No	MEC
Functional based	**LiNGAM**	No	Yes	No	No	DAG
	Longitudinal LiNGAM	No	Yes	No	Yes	

Table 5.1: Assumptions of our CSD models {MEC: Markov Equivalence Class}

In evaluating the performance of various causal discovery models, understanding the underlying assumptions is paramount due to their potential impact on outcomes. The CIM model, which demonstrated superior performance in our study, relies on the assumption of faithfulness, meaning it presumes that observed statistical relationships in the data accurately reflect the underlying causal structures. However, it lacks mechanisms for handling confounding variables or specifying distributions, potentially limiting its utility in scenarios where these factors are prevalent.

The GES and Longitudinal LiNGAM models, which follow CIM in performance, operate under distinct assumptions. GES assumes a degree of faithfulness and imposes distributional assumptions, albeit without addressing confounders explicitly; this may explain why it did not outperform CIM, as sensitivity to distributional assumptions could hinder its effectiveness. Longitudinal LiNGAM, while not assuming faithfulness, assumes Gaussian distributions [45], which may contribute to its robustness in this particular data analysis but also might restrict its use in more complex causal structures.

The selection of covariates holds pivotal importance in causal inference [49], and our study adopts a restricted set, presuming them to be the most relevant. However, incorporating alternative covariates could potentially alter outcomes, as models operate under the assumption that the chosen ones adequately mitigate confounding variables. Moreover, experts may interpret the structure of causal graphs differently, leading to varied perspectives on causality and variable significance. Our causal graph represents just one of many conceivable interpretations, each with its own set of assumptions that influence the conclusions drawn from the study. Hence, it is imperative to consider a spectrum of expert viewpoints and refrain from asserting a definitive causal framework [109].

Chapter 6

Conclusion & Future Work

6.1 Conclusion

The present study aimed to investigate the utility of causal structure discovery (CSD) methods in identifying the causal factors contributing to dementia onset from longitudinal data. We evaluated the performance of various temporal and atemporal CSD models in recovering the true causal graph, guided by a ground truth causal diagram curated with the assistance of domain expertise.

In addressing Question 1, our findings indicate that recent advancements in constraint-based CSD algorithms, notably the CIM model, outperform established baselines and demonstrate robust performance in uncovering the underlying causal structure. Incorporating temporal information into these models further enhanced their accuracy, emphasizing the importance of leveraging longitudinal data for causal discovery endeavours. However, certain temporal models tailored for single-participant time series data exhibited sub-optimal results when applied to our multi-participant longitudinal dataset.

Regarding Question 2, quantitative analyses and interventional studies guided by the ground truth causal graph revealed age and life satisfaction (scqolq) as primary direct causal factors influencing dementia onset, followed by BMI and smoking. Additionally, we identified significant indirect causes, such as executive index (execnn), frequency of communication with children (scchdh), and participation in sports/gym activities (scorg7), consistent with existing literature. These findings highlight modifiable risk factors that could inform preventive interventions before the onset of symptoms.

Lastly, in response to Question 3, we propose a novel framework that integrates interpretable machine learning (IML) techniques with causal structure discovery. By leveraging IML for feature selection and CSD models for causal inference, augmented with general prior knowledge, our framework facilitates the development of predictive models with causal explanations. Evaluations demonstrate this approach's potential to derive robust inferences aligning closely with the ground truth, paving the way for enhanced decision-making and actionable insights.

While our study contributes novel methodologies and insights, we acknowledge certain

limitations inherent to the assumptions underlying both IML and causal discovery algorithms, as well as the arguments for these errors to propagate. Nonetheless, our work highlights the potential of combining IML and CSD techniques, offering a promising avenue for advancing causal reasoning in complex domains such as dementia research.

6.2 Future Work

A key area for future investigation lies in the theoretical formalisation and analysis of error propagation [123] within the proposed framework. While the current study demonstrates the framework's ability to derive robust inferences aligning with the ground truth, a more rigorous mathematical treatment is warranted to quantify and characterise the propagation of errors across the various stages of the pipeline.

Specifically, a comprehensive theoretical model could be developed to capture the interplay between the inherent biases and assumptions underlying each component of the framework, such as the feature selection process, the causal discovery algorithms, and the predictive models employed. By formulating a unified mathematical framework that explicitly accounts for these error sources and their dependencies, researchers could gain deeper insights into the conditions under which the framework's inferences remain reliable. Moreover, the theoretical model could integrate methodologies from sensitivity analysis and uncertainty quantification [114], enabling researchers to systematically explore how variations in individual components impact the overall framework's performance. This approach could facilitate the identification of strategies to mitigate error propagation, such as employing ensemble techniques [48] or adopting probabilistic causal models [104].

Ultimately, the development of a comprehensive theoretical foundation would not only improve the interpretability and reliability of the proposed framework but also promote its wider acceptance and integration into decision-making processes [124]. By quantifying the uncertainties associated with the causal inferences drawn, stakeholders could make more informed decisions, weighing the potential implications of their actions against the confidence levels provided by the framework.

Recent research has also explored data fusion, exemplified by Josey et al. [64], who propose a method for integrating data from various sources in observational studies to assess intervention effects across diverse populations, thereby improving generalisability. Their introduced calibration method aids in mitigating biases inherent in non-randomised observational studies, like ours. Integrating this calibration approach into data fusion for causal discovery holds promise for enhancing the estimation of causal effects by reducing the influence of confounding variables and sampling bias. This could lead to more precise causal models, facilitating the development of effective interventions and policies based on robust causal evidence. However, careful consideration of the outlined assumptions and limitations is crucial for the proper implementation and interpretation of results.

Bibliography

[1] ADI - Dementia statistics — alzint.org. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/. [Accessed 25-03-2024].

[2] Brain Health and risk reduction - Dementia Statistics Hub — dementiastatistics.org. https://dementiastatistics.org/about-dementia/brain-health/. [Accessed 25-03-2024].

[3] Computing counterfactuals; dowhy documentation — pywhy.org. https://www.pywhy.org/dowhy/v0.11.1/user_guide/causal_tasks/what_if/counterfactuals.html#understand-method-counter. [Accessed 05-04-2024].

[4] Defining morbid obesity bmi at upmc hamot - erie pa — upmc.com. https://www.upmc.com/locations/hospitals/hamot/services/bariatric/defining-obesity#:~:text=A%20BMI%20above%2040%20indicates,other%20medical%20problems%20listed%20below. [Accessed 06-04-2024].

[5] Direct effect: Quantifying arrow strength; dowhy documentation — pywhy.org. https://www.pywhy.org/dowhy/v0.11.1/user_guide/causal_tasks/quantify_causal_influence/quantify_arrow_strength.html. [Accessed 03-04-2024].

[6] Feature relevance; dowhy documentation — pywhy.org. https://www.pywhy.org/dowhy/v0.11.1/user_guide/causal_tasks/root_causing_and_explaining/feature_relevance.html. [Accessed 05-04-2024].

[7] Recorded Dementia Diagnoses, August 2022 - NHS England Digital — digital.nhs.uk. https://digital.nhs.uk/data-and-information/publications/statistical/recorded-dementia-diagnoses/august-2022. [Accessed 25-03-2024].

[8] Simulating the impact of interventions; dowhy documentation — pywhy.org. https://www.pywhy.org/dowhy/v0.11.1/user_guide/causal_tasks/what_if/interventions.html. [Accessed 05-04-2024].

[9] Joaquín Abellán, Manuel Gómez-Olmedo, Serafín Moral, et al. Some variations on the pc algorithm. In *Probabilistic graphical models*, pages 1–8. Citeseer, 2006.

[10] Nans Addor, Marco Rohrer, Reinhard Furrer, and Jan Seibert. Propagation of biases in climate models from the synoptic to the regional scale: Implications for bias adjustment. *Journal of Geophysical Research: Atmospheres*, 121(5):2075–2089, 2016.

[11] Juan I Alonso-Barba, Jose A Gámez, Jose M Puerta, et al. Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *International journal of approximate reasoning*, 54(4):429–451, 2013.

[12] Ibrar Anjum, Muniba Fayyaz, Abdullah Wajid, Wafa Sohail, and Asad Ali. Does obesity increase the risk of dementia: a literature review. *Cureus*, 10(5), 2018.

[13] Ali Arab, Gregory J Christie, Mehrdad Mansouri, Maryam Ahmadzadeh, Andrew Sixsmith, Martin Ester, and Sylvain Moreno. Moderate-intensity physical activity, music and art activities preserved cognitive health in older adults: an argument for social prescribing solution. *Frontiers in Aging Neuroscience*, 13:693791, 2021.

[14] Miriam K Aronson, Wee L Ooi, Dalia L Geva, David Masur, Alan Blau, and William Frishman. Dementia: age-dependent incidence, prevalence, and mortality in the old old. *Archives of Internal Medicine*, 151(5):989–992, 1991.

[15] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[16] J. Banks, G. David Batty, J. Breedvelt, K. Coughlin, R. Crawford, M. Marmot, J. Nazroo, Z. Oldfield, N. Steel, A. Steptoe, M. Wood, and P. Zaninotto. English longitudinal study of ageing: Waves 0-10, 1998-2023. [data collection]. 40th Edition. UK Data Service, 2024. SN: 5050, DOI: `http://doi.org/10.5255/UKDA-SN-5050-27`.

[17] Deepika Bansal, Rita Chhikara, Kavita Khanna, and Poonam Gupta. Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia computer science*, 132:1497–1502, 2018.

[18] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.

[19] Loredana Bellantuono, Alfonso Monaco, Nicola Amoroso, Antonio Lacalamita, Ester Pantaleo, Sabina Tangaro, and Roberto Bellotti. Worldwide impact of lifestyle predictors of dementia prevalence: An explainable artificial intelligence analysis. *Frontiers in big Data*, 5:1027783, 2022.

[20] Sophia Bennett and Alan J Thomas. Depression and dementia: cause, consequence or coincidence? *Maturitas*, 79(2):184–190, 2014.

[21] Maia Berkane. *Latent variable modeling and applications to causality*, volume 120. Springer Science & Business Media, 2012.

[22] Christopher M Bishop. Latent variable models. In *Learning in graphical models*, pages 371–403. Springer, 1998.

[23] Fernando Blanco and Helena Matute. The illusion of causality: A cognitive bias underlying pseudoscience. *Pseudoscience: The conspiracy against science*, pages 45–75, 2018.

[24] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

[25] Patrick Blöbaum and Shohei Shimizu. Estimation of interventional effects of features on prediction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[26] Weilin Chen, Jie Qiao, Ruichu Cai, and Zhifeng Hao. On the role of entropy-based loss for learning causal structure with continuous optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[27] Daein Choi, Seulggie Choi, and Sang Min Park. Effect of smoking cessation on the risk of dementia: a longitudinal study. *Annals of clinical and translational neurology*, 5(10):1192–1199, 2018.

[28] Patricia Cohen, Jacob Cohen, Jeanne Teresi, Margaret Marchi, and C Noemi Velez. Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement*, 14(2):183–196, 1990.

[29] Michael H Connors, Katrin Seeher, Armando Teixeira-Pinto, Michael Woodward, David Ames, and Henry Brodaty. Dementia and caregiver burden: a three-year longitudinal study. *International journal of geriatric psychiatry*, 35(2):250–258, 2020.

[30] Anthony C Constantinou, Zhigao Guo, and Neville K Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434, 2023.

[31] Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. *arXiv preprint arXiv:1301.6686*, 2013.

[32] Martijn de Jongh and Marek J Druzdzel. Evaluation of rules for coping with insufficient data in constraint-based search algorithms. In *Probabilistic Graphical Models: 7th European Workshop, PGM 2014, Utrecht, The Netherlands, September 17-19, 2014. Proceedings 7*, pages 190–205. Springer, 2014.

[33] Kevin Debeire, Andreas Gerhardus, Jakob Runge, and Veronika Eyring. Bootstrap aggregation and confidence measures to improve time series causal discovery. In *Causal Learning and Reasoning*, pages 979–1007. PMLR, 2024.

[34] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6(1):21689, 2016.

[35] Anthony Devaux, Cécile Proust-Lima, and Robin Genuer. Random forests for time-fixed and time-dependent predictors: The dynforest r package. *arXiv preprint arXiv:2302.02670*, 2023.

[36] Dorit Dor and Michael Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*, page 45, 1992.

[37] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3:81–91, 2017.

[38] Sacha Epskamp. sachaepskamp.com. `http://sachaepskamp.com/files/NA2014/d-separation2013.pdf`. [Accessed 01-04-2024].

[39] David Faraoni and Simon Thomas Schaefer. Randomized controlled trials vs. observational studies: why not just live together? *BMC anesthesiology*, 16:1–4, 2016.

[40] Nicole Ferraro and Adam Lavertu. Development and comparison of tissue-specific causal gene regulatory networks. `https://snap.stanford.edu/class/cs224w-2017/projects/cs224w-10-final.pdf`. [Accessed 03-04-2024].

[41] Ronja Foraita, Juliane Friemel, Kathrin Günther, Thomas Behrens, Jörn Bullerdiek, Rolf Nimzyk, Wolfgang Ahrens, and Vanessa Didelez. Causal discovery of gene regulation with incomplete data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(4):1747–1775, 2020.

[42] Valentina Gallo, Paolo Vineis, Mariagrazia Cancellieri, Paolo Chiodini, Roger A Barker, Carol Brayne, Neil Pearce, Roel Vermeulen, Salvatore Panico, Bas Bueno-de Mesquita, et al. Exploring causality of the association between smoking and parkinson's disease. *International journal of epidemiology*, 48(3):912–925, 2019.

[43] Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Dan Geiger, Thomas Verma, and Judea Pearl. d-separation: From theorems to algorithms. In *Machine intelligence and pattern recognition*, volume 10, pages 139–148. Elsevier, 1990.

[45] Clark Glymour and Kun Zhang. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:418407, 2019.

[46] Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, and Jingping Bi. Causal discovery from temporal data: An overview and new perspectives. *arXiv preprint arXiv:2303.10112*, 2023.

[47] Jason Grossman and Fiona J Mackenzie. The randomized controlled trial: gold standard, or merely standard? *Perspectives in biology and medicine*, 48(4):516–534, 2005.

[48] Pei Guo, Yiyi Huang, and Jianwu Wang. Scalable and flexible two-phase ensemble algorithms for causality discovery. *Big Data Research*, 26:100252, 2021.

[49] Isabelle Guyon, Constantin Aliferis, et al. Causal feature selection. In *Computational methods of feature selection*, pages 79–102. Chapman and Hall/CRC, 2007.

[50] York Hagmayer and Michael R Waldmann. How temporal assumptions influence causal judgments. *Memory & Cognition*, 30:1128–1137, 2002.

[51] Sergiu Hart. Shapley value. In *Game theory*, pages 210–216. Springer, 1989.

[52] Taher Haveliwala et al. Efficient computation of pagerank. Technical report, Citeseer, 1999.

[53] Yulei He, Alan M Zaslavsky, MB Landrum, DP Harrington, and P Catalano. Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research*, 19(6):653–670, 2010.

[54] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[55] Andreas Holzinger. Explainable ai and multi-modal causability in medicine. *i-com*, 19(3):171–179, 2021.

[56] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[57] Roger A Horn. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169, 1990.

[58] Richard Howey, Alexander D Clark, Najib Naamane, Louise N Reynard, Arthur G Pratt, and Heather J Cordell. A bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships. *PLoS Genetics*, 17(9):e1009811, 2021.

[59] Hengyi Hu and Larry Kerschberg. Improved causal models of alzheimer's disease. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 274–283. IEEE, 2021.

[60] Zixin Hu, Rong Jiao, Panpan Wang, Yun Zhu, Jinying Zhao, Phil De Jager, David A Bennett, Li Jin, and Momiao Xiong. Shared causal paths underlying alzheimer's dementia and type 2 diabetes. *Scientific reports*, 10(1):4107, 2020.

[61] Takashi Ikeuchi, Mayumi Ide, Yan Zeng, Takashi Nicholas Maeda, and Shohei Shimizu. Python package for causal discovery based on lingam. *Journal of Machine Learning Research*, 24(14):1–8, 2023.

[62] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. 2013.

[63] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[64] Kevin P Josey, Fan Yang, Debashis Ghosh, and Sridharan Raghavan. A calibration approach to transportability and data-fusion with observational data. *Statistics in medicine*, 41(23):4511–4531, 2022.

[65] Kento Kadowaki, Shohei Shimizu, and Takashi Washio. Estimation of causal structures in longitudinal data using non-gaussianity. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2013.

[66] Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3):1587–1595, 2022.

[67] Lukas Klein, Mennatallah El-Assady, and Paul F Jäger. From correlation to causation: Formalizing interpretable machine learning as a statistical process. *arXiv preprint arXiv:2207.04969*, 2022.

[68] Weirui Kong, Hyeju Jang, Giuseppe Carenini, and Thalia Field. A neural model for predicting dementia from language. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 270–286. PMLR, 09–10 Aug 2019.

[69] Maxim Kovalev, Lev Utkin, Frank Coolen, and Andrei Konstantinov. Counterfactual explanation of machine learning survival models. *Informatica*, 32(4):817–847, 2021.

[70] Jisca S Kuiper, Marij Zuidersma, Richard C Oude Voshaar, Sytse U Zuidema, Edwin R van den Heuvel, Ronald P Stolk, and Nynke Smidt. Social relationships and risk of dementia: A systematic review and meta-analysis of longitudinal cohort studies. *Ageing research reviews*, 22:39–57, 2015.

[71] Elżbieta Kuźma, Eilis Hannon, Ang Zhou, Ilianna Lourida, Alison Bethel, Deborah A Levine, Katie Lunnon, Jo Thompson-Coon, Elina Hyppönen, and David J Llewellyn. Which risk factors causally influence dementia? a systematic review of mendelian randomization studies. *Journal of Alzheimer's Disease*, 64(1):181–193, 2018.

[72] David A Lagnado and Steven A Sloman. Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3):451, 2006.

[73] Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Achieving diversity in counterfactual explanations: a review and discussion. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1869, 2023.

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

[75] John F Lemmer. The causal markov condition, fact or artifact? *ACM SIGART Bulletin*, 7(3):3–16, 1996.

[76] Joanne M Li, Malaz A Boustani, and Dustin D French. Social determinants of health in community-dwelling dementia patients aged 65 and over: analysis of the 2019 national health interview survey. *Gerontology and Geriatric Medicine*, 9:23337214231190244, 2023.

[77] Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. Root cause analysis in microservice using neural granger causal discovery. *arXiv preprint arXiv:2402.01140*, 2024.

[78] Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509, 2020.

[79] Robert F Ling. Correlation and causation., 1982.

[80] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[81] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[82] Jarmo Mäkelä, Laila Melkas, Ivan Mammarella, Tuomo Nieminen, Suyog Chandramouli, Rafael Savvides, and Kai Puolamäki. Incorporating expert domain knowledge into causal structure discovery workflows. *Biogeosciences*, 19(8):2095–2099, 2022.

[83] Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

[84] Suryansh Manocha. Interpretable machine learning for dementia risk prediction. 2023.

[85] Maritza Mera-Gaona, Ursula Neumann, Rubiel Vargas-Canas, and Diego M López. Evaluating the impact of multivariate imputation by mice in feature selection. *Plos one*, 16(7):e0254720, 2021.

[86] Nandita Mitra, Jason Roy, and Dylan Small. The future of causal inference. *American journal of epidemiology*, 191(10):1671–1676, 2022.

[87] Ismail Bin Mohamad and Dauda Usman. Research article standardization and its effects on k-means clustering algorithm. *Res J Appl Sci Eng Technol*, 6(17):3299–3303, 2013.

[88] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[89] Kaoru Mulvihill. The three layer causal hierarchy. `https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf`. [Accessed 27-03-2024].

[90] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[91] John B Nezlek, Christianne P Hampton, and Glenn D Shean. Clinical depression and day-to-day social interaction in a community sample. *Journal of abnormal psychology*, 109(1):11, 2000.

[92] John B Nezlek, Mark Imbrie, and Glenn D Shean. Depression and everyday social interaction. *Journal of personality and social psychology*, 67(6):1101, 1994.

[93] Judea Pearl. Causation, action, and counterfactuals. In *Logic and Scientific Methods: Volume One of the Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995*, pages 355–375. Springer, 1997.

[94] Judea Pearl. Causal inference in statistics: An overview. 2009.

[95] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

[96] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

[97] Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

[98] Melania Pintilie. An introduction to competing risks analysis. *Revista Española de Cardiología (English Edition)*, 64(7):599–605, 2011.

[99] Atul Rawal, Adrienne Raglin, Brian M Sadler, and Danda B Rawat. Explainability and causality for robust, fair, and trustworthy artificial reasoning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, volume 12538, pages 493–500. SPIE, 2023.

[100] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[101] Christian Reiser. Causal discovery for time series with latent confounders. *arXiv preprint arXiv:2209.03427*, 2022.

[102] Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in biology and medicine*, 112:103375, 2019.

[103] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[104] Alberto Riva and Riccardo Bellazzi. Learning temporal probabilistic causal models from longitudinal data. *Artificial Intelligence in Medicine*, 8(3):217–234, 1996.

[105] Louise Robinson, Eugene Tang, and John-Paul Taylor. Dementia: timely diagnosis and early intervention. *Bmj*, 350, 2015.

[106] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[107] Judi Scheffer. Dealing with missing data. 2002.

[108] Sema K. Sgaier, Vincent Huang, and Grace Charles. The case for causal ai. *Stanford Social Innovation Review*, pages z–z, 2023. [Online].

[109] Eyal Shahar and Doron J Shahar. Causal diagrams, information bias, and thought bias. *Pragmatic and Observational Research*, pages 33–47, 2010.

[110] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[111] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 10(1):2975, 2020.

[112] Xinwei Shen, Shengyu Zhu, Jiji Zhang, Shoubo Hu, and Zhitang Chen. Reframed ges with a neural conditional dependence measure. In *Uncertainty in Artificial Intelligence*, pages 1782–1791. PMLR, 2022.

[113] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[114] Christian Soize. *Uncertainty quantification*. Springer, 2017.

[115] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[116] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer, 2016.

[117] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[118] Jonathan Stone, Daniel M Johnstone, John Mitrofanis, and Michael O'Rourke. The mechanical cause of age-related dementia (alzheimer's disease): the brain is destroyed by the pulse. *Journal of Alzheimer's Disease*, 44(2):355–373, 2015.

[119] Eric V Strobl. Causal discovery with a mixture of dags. *Machine Learning*, 112(11):4201–4225, 2023.

[120] Eric V Strobl, Shyam Visweswaran, and Peter L Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6:47–62, 2018.

[121] Elizabeth A Stuart, Melissa Azur, Constantine Frangakis, and Philip Leaf. Multiple imputation with large data sets: a case study of the children's mental health initiative. *American journal of epidemiology*, 169(9):1133–1139, 2009.

[122] Paulo Jeng Chian Suen, Pedro Starzynski Bacchi, Lais Razza, Leonardo Afonso Dos Santos, Daniel Fatori, Izio Klein, Ives Cavalcante Passos, Jordan W Smoller, Sarah Bauermeister, Alessandra Carvalho Goulart, et al. Examining the impact of the covid-19 pandemic through the lens of the network approach to psychopathology: Analysis of the brazilian longitudinal study of health (elsa-brasil) cohort over a 12-year timespan. *Journal of anxiety disorders*, 85:102512, 2022.

[123] Joel Tellinghuisen. Statistical error propagation. *The Journal of Physical Chemistry A*, 105(15):3917–3921, 2001.

[124] Paul Thagard. Causal inference in legal decision making: Explanatory coherence vs. bayesian networks. *Applied Artificial Intelligence*, 18(3-4):231–249, 2004.

[125] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

[126] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. Pmlr, 2019.

[127] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[128] Stef Van Buuren and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.

[129] Wiesje M van der Flier and Philip Scheltens. Epidemiology and risk factors of dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 5):v2–v7, 2005.

[130] Tyler J VanderWeele, Miguel A Hernán, and James M Robins. Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, 19(5):720–728, 2008.

[131] Burak Varici, Dmitriy Katz, Dennis Wei, Prasanna Sattigeri, and Ali Tajer. Separability analysis for causal discovery in mixture of dags. *Transactions on Machine Learning Research*, 2023.

[132] Svitlana Volkova, Dustin Arendt, Emily Saldanha, Maria Glenski, Ellyn Ayton, Joseph Cottam, Sinan Aksoy, Brett Jefferson, and Karthnik Shrivaram. Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. *Computational and Mathematical Organization Theory*, 29(1):220–241, 2023.

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

[134] Clifford H Wagner. Simpson's paradox in real life. *The American Statistician*, 36(1):46–48, 1982.

[135] Yuhao Wang, Vlado Menkovski, Hao Wang, Xin Du, and Mykola Pechenizkiy. Causal discovery from incomplete data: a deep learning approach. *arXiv preprint arXiv:2001.05343*, 2020.

[136] Greta Warren, Mark T Keane, and Ruth MJ Byrne. Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai. *arXiv preprint arXiv:2204.10152*, 2022.

[137] Eric W Weisstein. Complete graph. *https://mathworld. wolfram. com/*, 2001.

[138] Janine Witte, Ronja Foraita, and Vanessa Didelez. Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data. *Statistics in Medicine*, 41(23):4716–4743, 2022.

[139] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

[140] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[141] Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *arXiv preprint arXiv:2307.16405*, 2023.

[142] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

Appendix A

Variable Definitions

Variable	Variable label
cfdaty	Whether Correct Year Given
execnn	Index Of Executive (Non-Numeracy) Function (0-20)
scqolq	How Often Feels Satisfied With The Way Their Life Has Turned Out
scfami	How Often The Respondent Writes To Or Emails Other Relatives
scchdh	How Often The Respondent Speaks On The Phone To Their Children
hours	Hours Of Work Main Job (Employed Or Self Employed)
eqtotinc_bu_s	Bu Equivalised Total Income - Summary Var
leisureu	Money Spent On Leisure Wkly Upper Bound (Holeis/Holeisb)
headlph	Iadl Difficulty Making Telephone Calls
headlmo	Iadl Difficulty Managing Money, Eg Paying Bills,Keeping Track Expenses
headlma	Iadl Difficulty Using Map To Figure Out How To Get Around Strange Place
scorg7	Organisational Membership Member Of A Sports Clubs, Gym, Or Exercise Class
bmivg6	BMI, numbered from 1-6, translating to (grouped:$\leq 20, 20-25, 25-30, 30-35, 35-40, 40+$) respectively.
scchdh	How Often The Respondent Speaks On The Phone To Their Children

Table A.1: Table showing the list of variables that are discussed and used in the models presented in this work. This list is not an exhaustive list of variables that are used in the feature selection process; an exhaustive list can be found from our previous work [84].

Appendix B

Relevant Figures

B.1 Simpson's Paradox Example

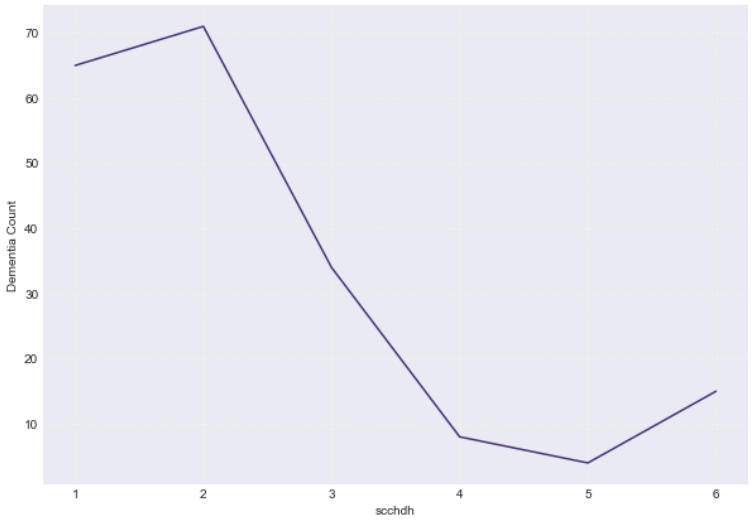


Figure B.1: The relationship between dementia and 'scchdh'.

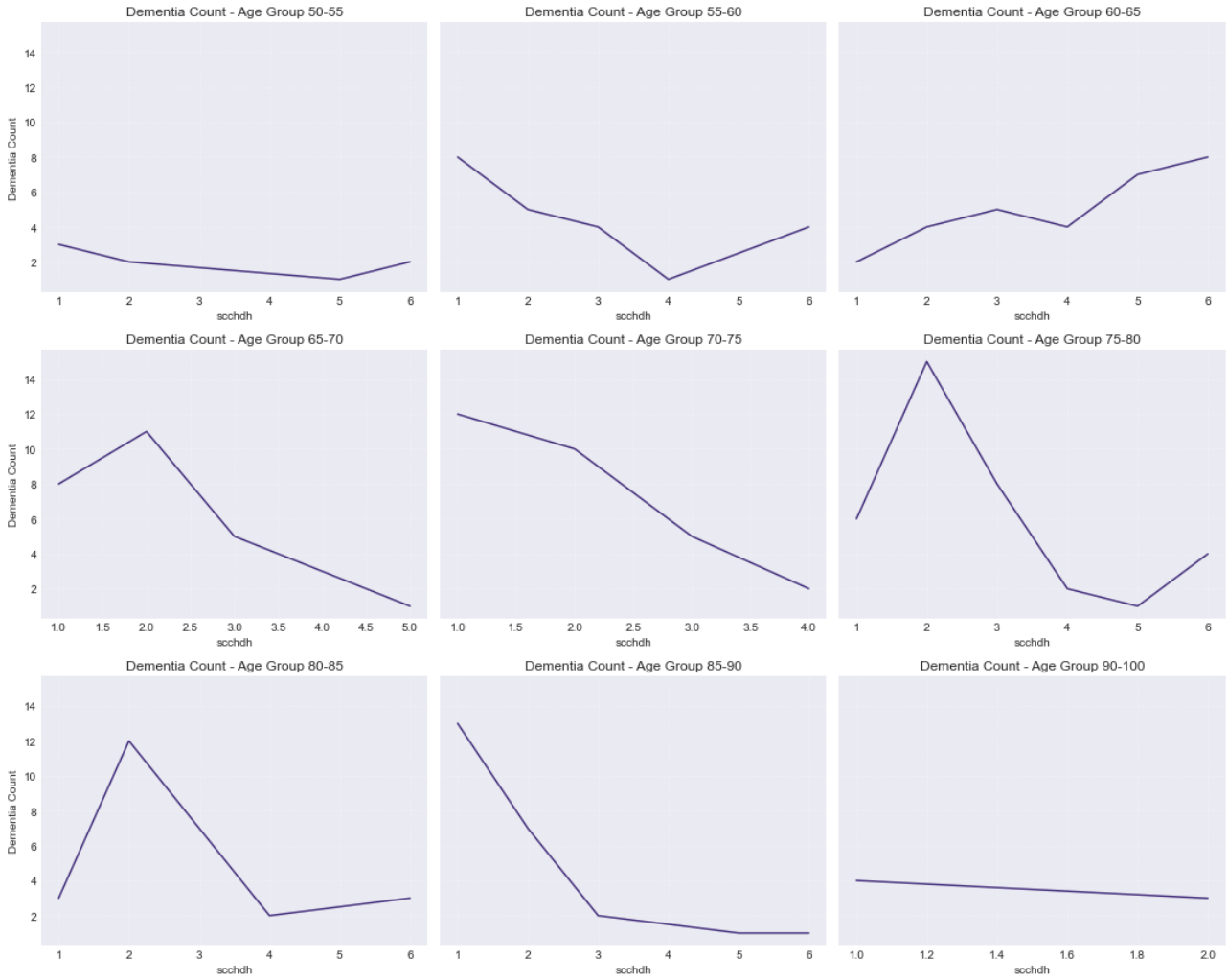


Figure B.2: The relationship between dementia and 'scchdh' stratified by 'age', which is suspected to be a confounding variable.

B.2 Robustness Analysis

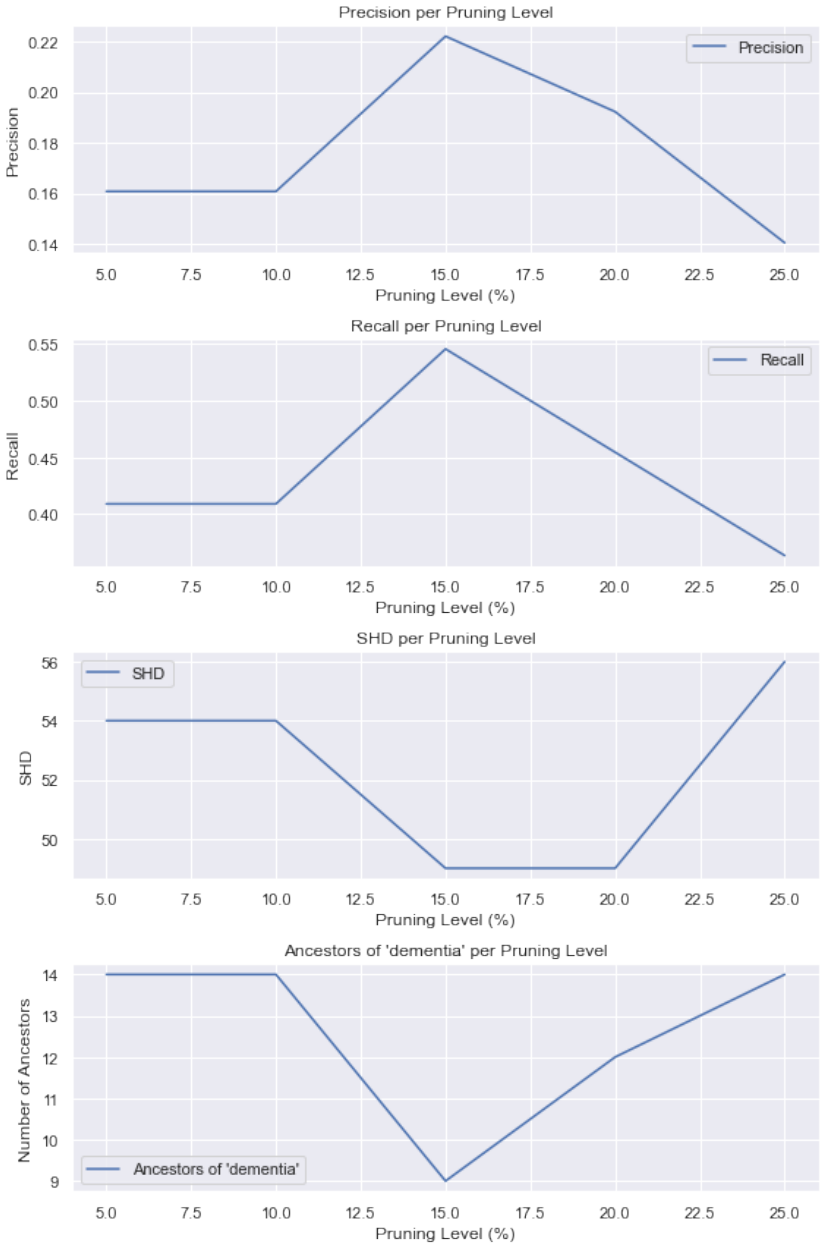


Figure B.3: Figure showing the robustness analysis performed for the CIM algorithm

B.3 IML Framework

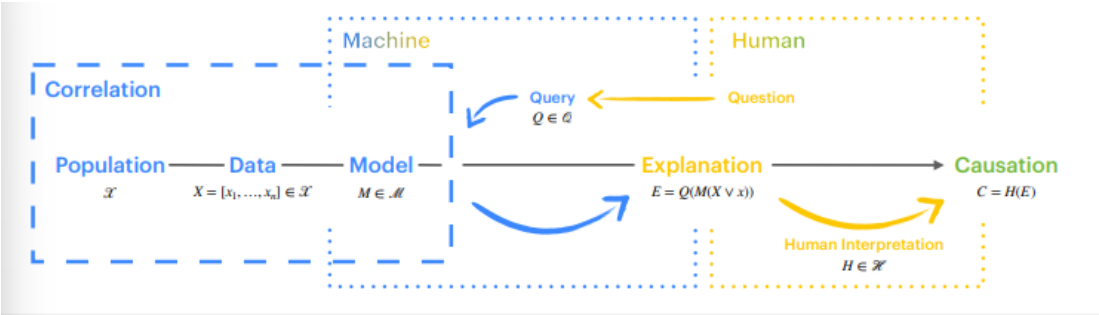


Figure B.4: Figure showing the IML to causation framework by [67]

Appendix C

Causal Discovery Graph Output Examples

C.1 PC

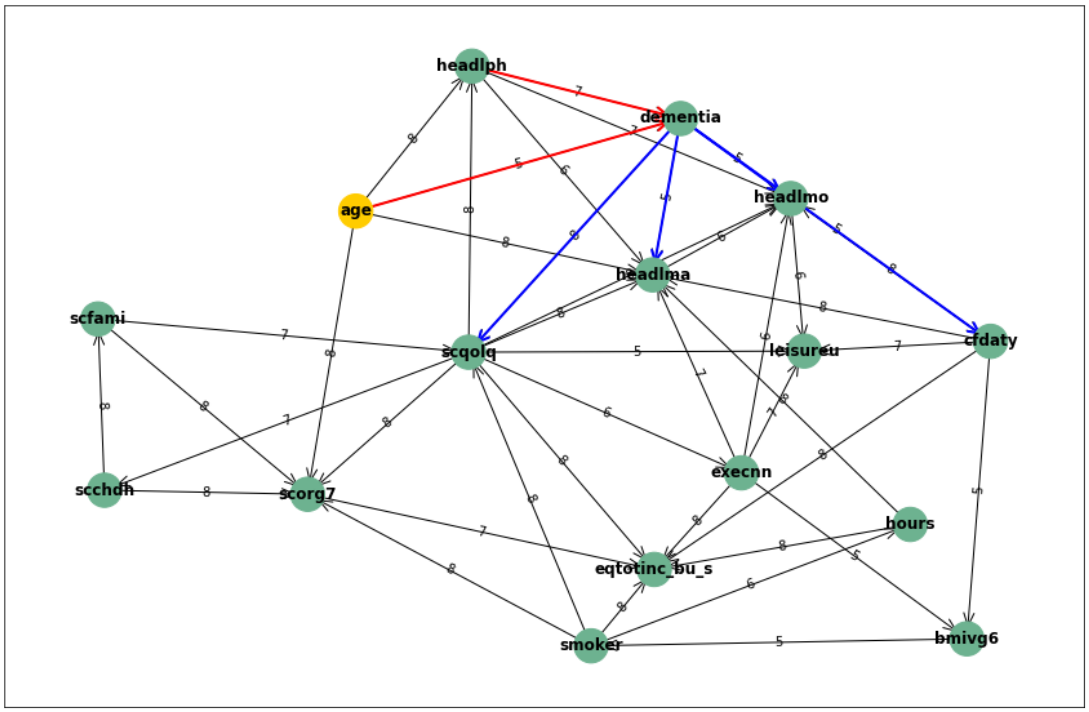


Figure C.1: Figure showing the PC graphical structure, with bootstrapping. The edges represent the occurrence of each edge in the bootstrapped samples.

C.2 NOTEARS

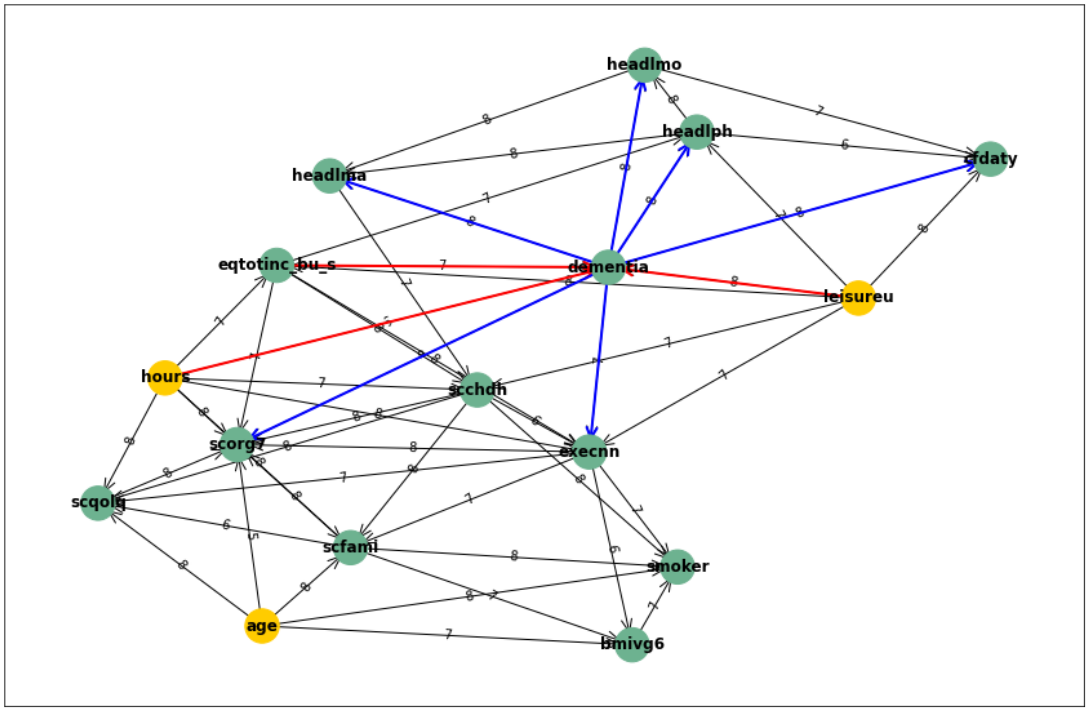


Figure C.2: Figure showing the NOTEARS graphical structure, with bootstrapping. The edges represent the occurrence of each edge in the bootstrapped samples.

C.3 DYNOTEARS

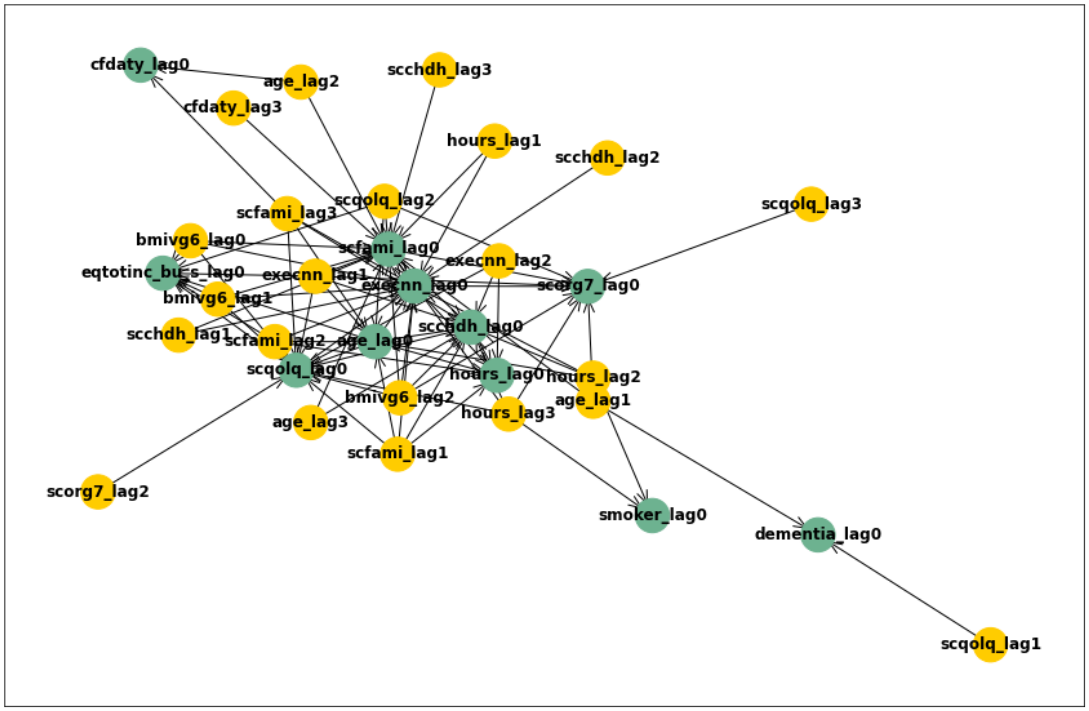


Figure C.3: The raw output from the DYNOTEARS algorithm, whereby lags correspond to points in time. To compare the final graph, we collate lagged variables into a single variable.

C.4 Longitudinal LiNGAM

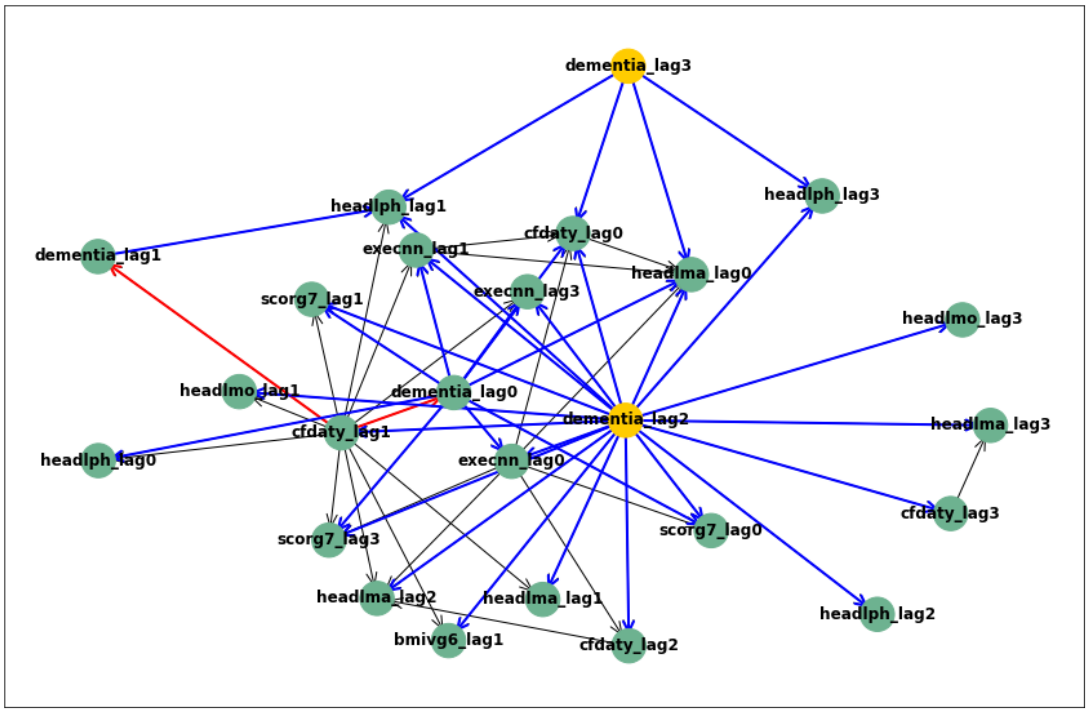


Figure C.4: The raw output from the Longitudinal LiNGAM algorithm, whereby lags correspond to points in time. To compare the final graph, we collate lagged variables into a single variable.

Appendix D

Reproducibility

D.1 Packages used

For modelling causal relations, and performing causal tasks we use the DoWhy python library [110, 24]. To perform causal discovery, we utilised a plethora of packages, including but not limited to [61, 141, 18].