On Heterogeneous Treatment Effects in Heterogeneous Causal Graphs

Richard Watson [* 1]   **Hengrui Cai** [* 2]   **Xinming An** [3]   **Samuel Mclean** [3]   **Rui Song** [1]

Abstract

Heterogeneity and comorbidity are two interwoven challenges associated with various healthcare problems that greatly hampered research on developing effective treatment and understanding of the underlying neurobiological mechanism. Very few studies have been conducted to investigate *heterogeneous causal effects* (HCEs) in graphical contexts due to the lack of statistical methods. To characterize this heterogeneity, we first conceptualize *heterogeneous causal graphs* (HCGs) by generalizing the causal graphical model with confounder-based interactions and multiple mediators. Such confounders with an interaction with the treatment are known as *moderators*. This allows us to flexibly produce HCGs given different moderators and explicitly characterize HCEs from the treatment or potential mediators on the outcome. We establish the theoretical forms of HCEs and derive their properties at the individual level in both linear and nonlinear models. An interactive structural learning is developed to estimate the complex HCGs and HCEs with confidence intervals provided. Our method is empirically justified by extensive simulations and its practical usefulness is illustrated by exploring causality among psychiatric disorders for trauma survivors. Code implementing the proposed algorithm is open-source and publicly available at: `https://github.com/richard-watson/ISL`.

1. Introduction

During the last several decades, little progress has been made in the understanding of the underlying neurobiological mechanism and in developing effective treatments for psychiatric disorders due to several unique challenges [18]. Accumulating evidence from existing research suggests that *Heterogeneity* is a common issue for psychiatric disorders based on traditional classification and diagnoses. Most of the existing studies that address heterogeneity based on symptom severity levels [8] are less likely to provide new insights about underlying neurobiological mechanisms. On the other hand, studying heterogeneous causal relationships of different psychiatric disorders has the potential to improve understanding, risk prediction, and treatment selection. However, to date, no such studies have been conducted due to the lack of appropriate analysis techniques. To fill in this gap and advance research for psychiatric disorders, we developed an innovative analysis approach to investigate and test heterogeneous causal relationships and applied it to study psychiatric disorders after trauma exposure using data from a large cohort study.

Causal discovery, the task of discovering the causal relations between variables in a dataset, has interesting and important applications in many areas, such as epidemiology [9], medicine [10], economics [24], etc. Under a general causal graph, an event or a treatment may have a direct effect on the outcome of interest as well as an indirect effect regulated by a set of *mediators* (which are variables that are affected by the treatment and in turn affect the outcome as M does in Figure 1) [see an overview in 26]. Yet, due to the heterogeneity of individuals in response to different events/treatments, there may not exist a uniform causal graph for everyone. This implies the existence of heterogeneous causal graphs under different *moderators* which are pre-treatment confounders that also have an interaction term with the treatment as X does in Figure 1). Learning such heterogeneity is one of the most crucial obstacles that continue to hamper advances in many fields such as psychiatric disorders [17]. With the launch of the Advancing Understanding of RecOvery afteR traumA (AURORA) study [18], where thousands of trauma survivors were recruited after traumatic experiences and followed to collect a broad range of bio-behavioral data, discovering heterogeneous causal patterns thus becomes a timely issue.

Despite the popularity of causal discovery methods [e.g., 36; 35; 13; 3; 30; 45; 43; 47], none of these works considers heterogeneous causal graphs to account for the heterogeneity among different subjects. On the other hand, all existing
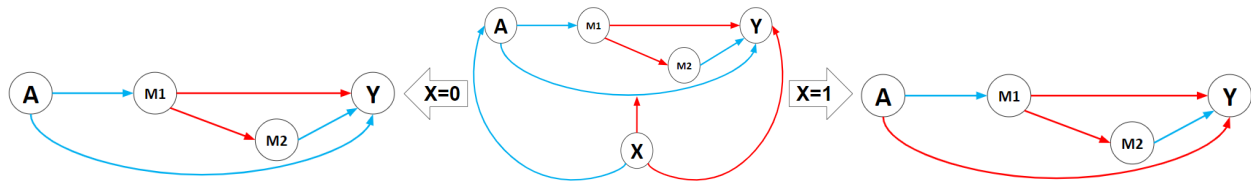
---

[*]Equal contribution [1]Department of Statistics, North Carolina State University, USA [2]Department of Statistics, University of California Irvine, USA [3]Department of Anesthesiology, University of North Carolina Chapel Hill, USA. Correspondence to: Rui Song <songray@gmail.com>.

Figure 1: The middle panel is the whole causal graph, and the left and right graphs are heterogeneous causal graphs for different values of X. Here X confounds the relationship between A (the treatment or event) and Y (the outcome of interest), and modifies the direct effect of A on Y dependent on the value of X. M_1 and M_2 are two different mediators that mediate the indirect effect of A onto Y. The red and blue arrows represent whether the causal relation is positive or negative, respectively.

works of learning heterogeneity in causal inference literature [e.g., 42; 15; 22] either assume a known homogeneous causal graph or completely ignore the graphical structure, and thus fail to capture the heterogeneous causal pathways. In this paper, we focus on learning heterogeneous causal graphs and effects with multiple (possibly sequentially ordered) mediators. Our **contributions** are three-fold.

- To our knowledge, this is the first work that considers heterogeneity in terms of causal graphs. We first conceptualize *heterogeneous causal graphs* (HCGs), by incorporating the information of moderators and their interactions into the causal graphical model. This introduces several unique challenges in presenting the interacted nodes in the graphical model and handling such interactions in a consistent self-generated system. We address these difficulties by proposing a novel structural equation model with interactions to flexibly produce HCGs from a hybrid causal graph.

- We propose *heterogeneous causal effects* (HCEs) to systematically quantify the impact of treatments and mediators on the outcome of interest given different values of moderators. Based on the proposed graphical model with interactions, we establish the explicit forms of the defined HCEs and derive their theoretical properties at the individual level, in either **linear** or **non-linear** structural equation model.

- We develop a comprehensive procedure for extracting heterogeneous causal reasoning. Specifically, we propose an interactive structural learning algorithm to learn the complex HCGs, estimate HCEs, and compute bootstrap confidence intervals for these estimates via a debiasing process. Our method is general enough to address most causal structures.

1.1. Related Works

Conditional average treatment effect. A vast number of approaches have been proposed for estimating the heterogeneity via the conditional average treatment effect (CATE) that quantifies the effect size of treatment on the outcome of interest given different confounders [see recent advances in 1; 33; 42; 15; 22; 7]. Here, confounders used in CATE usually play the role of the moderators [14] which modify the impact of the treatment on the outcome as a presence of

confounder-treatment interaction in modeling the outcome. All these works neglect the underlying causal structure among potential mediators in regulating the treatment effects and thus lead to a less interpretable causal mechanism.

Traditional mediation analysis. There is an abundance of literature for traditional mediation analysis [see a review in 39, and the reference therein] quantifying treatment effects in the presence of multiple mediators. Due to multiple causally dependent mediators, Robins [31] found that causal effects are not identifiable unless no treatment-mediator interaction is imposed, with more discussions on the identification of causal effects later on [11; 38; 41]. Yet, all of these works primarily assumed a known causal graph with only a few mediators included for analysis. In addition, the moderators in traditional mediation analysis [19] interact with the treatment but are independent of the treatment. In this work, we do not enforce such an independence requirement.

Causal structural learning. Plentiful causal structural learning (CSL) methods have been proposed to learn the unknown causal structure within a class of directed acyclic graphs from observed data. The large literature can be categorized into three types. The testing-based methods [e.g., 36, for the well-known PC algorithm] rely on the conditional independence tests to find the causal skeleton and edge orientations under the linear structural equation model (SEM). Based on additional and proper assumptions on noises and models, the functional-based methods handle both linear SEM [e.g., 35] and non-linear SEM [e.g., 3]. Recently, the score-based methods formulate the CSL problem into optimization by certain score functions, for both linear SEM [e.g., 30; 45] and non-linear SEM [e.g., 43; 47; 46; 32]. Yet, all these works neglect causal contexts (i.e., treatments, mediators, and outcomes) and thus cannot develop HCGs.

Causal mediation analysis with CSL. To visualize causes and counterfactuals, Pearl et al. [26] proposed to use the causal graphical model and the 'do-operator' to quantify the causal effects. A number of follow-up works [e.g., 16; 20; 5] have been developed recently to estimate direct and indirect causal effects that are regulated by mediators in the linear SEM. These studies relied on the PC algorithm [36] which requires strong assumptions of graph sparsity and noise nor-

mality due to computational limits. To overcome these difficulties, Cai et al. [4] proposed to leverage score-based CSL methods [e.g., 30; 45; 43; 47] with background causal knowledge to estimate mediation effects. These works assumed no moderator in their linear SEMs, such that the causal graph or effect learned is on the population level, and thus cannot access the heterogeneity.

2. Framework

Causal graph terminology. A directed acyclic graph (DAG) is defined as a directed graph that does not contain directed cycles. Specifically, let a DAG $\mathbb{G} = (\boldsymbol{Z}, \boldsymbol{E})$ characterize the causal relationship among a set of random variables $\boldsymbol{Z}$ and an edge set $\boldsymbol{E}$. Here, $\boldsymbol{Z} = \{Z_1, \ldots, Z_w\}$ represents a random vector of $w = |\boldsymbol{Z}|$ nodes and an edge $Z_i \rightarrow Z_j \in \boldsymbol{E}$ means that Z_i is a direct cause of Z_j. A variable node Z_i is said to be a parent of Z_j if there is a directed edge from Z_i to Z_j. The set of all parents of node Z_j in $\mathbb{G}$ is denoted as $PA_{\mathbb{G}}(Z_j)$. Let $\boldsymbol{B} = \{b_{i,j}\}_{1 \leq i \leq w, 1 \leq j \leq w}$ be a $w \times w$ matrix, where $b_{i,j}$ is the weight of the edge $Z_i \rightarrow Z_j \in \boldsymbol{E}$, and $b_{i,j} = 0$ otherwise. Then, we say that $\mathbb{G} = (\boldsymbol{Z}, \boldsymbol{B})$ is a weighted DAG with the variable/node set $\boldsymbol{Z}$ and the weighted adjacency matrix $\boldsymbol{B}$.

Notations and assumptions. Let $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]^\top$ be a p-dimensional vector of baseline information prior to the event or the treatment A, $\boldsymbol{M} = [M_1, M_2, \ldots, M_s]^\top$ be a s-dimensional vector of post-treatment possible mediators, and Y be the final outcome of interest. We define $\boldsymbol{X}$ generally as a vector of possible moderators that precede the treatment with no independence restriction. There may be no interaction between X_1 and A, in which case it is not a moderator, but a confounder. Further X_2 may only affect Y, in which case it is a covariate. We allow $\boldsymbol{X}$ to be any combination of covariates, confounders, and moderators. As commonly imposed in CSL works [e.g., 37; 28], we assume the Markov and faithfulness assumptions (see explicit definitions and further discussion in Appendix A.1). Let $Y^* \equiv Y^*(A = a, \boldsymbol{M} = \boldsymbol{m})$ when $A = a$ and $\boldsymbol{M} = \boldsymbol{m}$ as the potential outcome and $M_i^* \equiv M_i^*(A = a, M_j = m_j)$ when $A = a$ and $M_j = m_j$ as the potential mediator. For any mediator M_i and any of its parent $M_j \in PA_{\mathbb{G}}(M_i) \backslash \{\boldsymbol{X}, A\}$, we assume no unmeasured confounders for:

(A1). the treatment on the outcome and mediators:
$$Y^* \perp\!\!\!\perp A | \boldsymbol{X}; \qquad M_i^* \perp\!\!\!\perp A | \boldsymbol{X};$$

(A2). the mediators on the outcome:
$$Y^* \perp\!\!\!\perp M_j | \{A, \boldsymbol{X}\}; \qquad Y^* \perp\!\!\!\perp M_i | \{A, \boldsymbol{X}, M_j\};$$

(A3). the potential mediators:
$$Y^* \perp\!\!\!\perp M_j^* | \boldsymbol{X}; \quad M_i^* \perp\!\!\!\perp M_j^* | \boldsymbol{X}; \quad Y^* \perp\!\!\!\perp M_i^* | \{M_j^*, \boldsymbol{X}\}.$$

Assumptions (A1-A3) are concerned with the completeness of the data as common in causal inference [e.g., 26; 20], where we assume that all variables causally related to any

variable in the data are included in data. Also note that we **make no assumptions on the structure of the mediators** which allows our method to take into account the fact that not all potential mediators are true mediators. Finally, the causal graph, in general, is only identifiable to its Markov Equivalence Class (MEC), however if, for example, the data is assumed to be Gaussian with equal variance, it can be identified from its MEC [27] (see details in Appendix A.2).

3. Heterogeneous Causal Graphs

We conceptualize the heterogeneous causal graph (HCG), by generalizing the causal graphical model with moderation.

3.1. Structural Equation Model with Interactions

Considering interactions among variables introduces several unique challenges in the causal graphical model, for example, presenting the interacted nodes in the graph and modeling heterogeneous interactions in a consistent self-generated system. To address these issues, we first generalize the linear SEM for the causal graphical model with interactions, which yields good interpretation to easily link the parameters with causal effects to be specified (see the functional SEM version of our method to handle complex relationships in Section 6). For simplicity of exposition, we use the product of $\boldsymbol{X}$ and A, $\boldsymbol{X}A$, as the interaction term (see discussions on the importance of such an interaction term in Remark 3.1). $\boldsymbol{D} = [\boldsymbol{X}, A, \boldsymbol{X}A, \boldsymbol{M}, Y]^\top$ is a $(2p + s + 2)$-dimensional data vector for an individual. Suppose there exists a weighted adjacency matrix $\boldsymbol{B}$ such that $\boldsymbol{D}$ and $\boldsymbol{B}$ together form a DAG, denoted as $\mathbb{G} = (\boldsymbol{D}, \boldsymbol{B})$. This DAG presents the causality among interested $w = (2p + s + 2)$ variables, where the node set $\boldsymbol{D}$ can be described using a linear SEM with an unknown $\boldsymbol{B}$ and some noise vector $\boldsymbol{\epsilon}$ as

$$\boldsymbol{D} = \boldsymbol{B}^\top \boldsymbol{D} + \boldsymbol{\epsilon}: \quad \begin{matrix} p \\ 1 \\ p \\ s \\ 1 \end{matrix} \begin{bmatrix} \boldsymbol{X} \\ A \\ \boldsymbol{X}A \\ \boldsymbol{M} \\ Y \end{bmatrix} = \boldsymbol{B}^\top \begin{bmatrix} \boldsymbol{X} \\ A \\ \boldsymbol{X}A \\ \boldsymbol{M} \\ Y \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon_X} \\ \epsilon_A \\ \boldsymbol{\epsilon_{XA}} \\ \boldsymbol{\epsilon_M} \\ \epsilon_Y \end{bmatrix}, \quad (1)$$

where $\boldsymbol{\epsilon_X}, \epsilon_A, \boldsymbol{\epsilon_M}, \epsilon_Y$ are random error variables associated with $\boldsymbol{X}, A, \boldsymbol{M}$, and Y, respectively, and are jointly independent with mean zero. Note that the noise is distribution-free, so our method can handle any arbitrary distribution without breaking the linear SEM. In addition, since $\boldsymbol{X}A$ is determined given $\boldsymbol{X}$ and A, in order to accommodate the interaction setting for the linear SEM, we define $\boldsymbol{\epsilon_{XA}} \equiv \boldsymbol{X}A$ with a constraint that $\boldsymbol{B}^\top \boldsymbol{D}$ produces a zero vector for elements corresponding to $\boldsymbol{X}A$. Alternatively, one can let $\boldsymbol{\epsilon_{XA}}$ be a zero vector, with a constraint that $\boldsymbol{B}^\top \boldsymbol{D}$ produces exact $\boldsymbol{X}A$. For simplicity, we use the first formulation. We next explicitly characterize the weighted adjacency matrix $\boldsymbol{B}$ that satisfies Model (1) based on causal knowledge among $\boldsymbol{X}, A, \boldsymbol{M}$, and Y, as well as the interaction

term $\boldsymbol{X}A$. Specifically, according to HCGs as illustrated in Figure 1 and assumptions in Section 2, we have:

1. $\boldsymbol{X}$ has no parents, i.e.,
   $g_1(\boldsymbol{B}) = \sum_{j=1}^{p} \sum_{i=1}^{2p+s+2} |b_{i,j}| = 0$;
2. the only parents of A are $\boldsymbol{X}$, i.e.,
   $g_2(\boldsymbol{B}) = \sum_{i=p+1}^{2p+s+2} |b_{i,p+1}| = 0$;
3. Y has no descendants, i.e.,
   $g_3(\boldsymbol{B}) = \sum_{i=1}^{2p+s+2} |b_{2p+s+2,i}| = 0$; and
4. the interaction $\boldsymbol{X}A$ also does not have parents, i.e.,
   $g_4(\boldsymbol{B}) = \sum_{j=p+2}^{2p+1} \sum_{i=1}^{2p+s+2} |b_{i,j}| = 0$.

The conditions in $g_1(\boldsymbol{B})$ to $g_4(\boldsymbol{B})$ yield the following matrix $\boldsymbol{B}^\top$ consisting of unknown parameters whose sparsity is due to prior causal information:

$$\boldsymbol{B}^\top = \begin{bmatrix} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times 1} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times s} & \mathbf{0}_{p\times 1} \\ \boldsymbol{\delta_X} & 0 & \mathbf{0}_{1\times p} & \mathbf{0}_{1\times s} & 0 \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times 1} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times s} & \mathbf{0}_{p\times 1} \\ \boldsymbol{B_X^\top} & \boldsymbol{\beta_A} & \boldsymbol{B_{XA}^\top} & \boldsymbol{B_M^\top} & \mathbf{0}_{s\times 1} \\ \boldsymbol{\gamma_X} & \gamma_A & \boldsymbol{\gamma_{XA}} & \boldsymbol{\gamma_M} & 0 \end{bmatrix},$$

where $\mathbf{0}_{a\times b}$ is a $a \times b$ zero matrix/vector, and the parameters $\boldsymbol{\delta_X}, \boldsymbol{B_X^\top}$, and $\boldsymbol{\gamma_X}$ represent the influence of $\boldsymbol{X}$, on the treatment A, the mediators $\boldsymbol{M}$, and the outcome Y, respectively. Likewise, $\boldsymbol{\beta_A}$ and γ_A represent the influence of A on $\boldsymbol{M}$ and Y, respectively, and $\boldsymbol{\gamma_M}$ represent the influence of $\boldsymbol{M}$ on Y. $\boldsymbol{B_M^\top}$ represents the influence of the mediators on other mediators. If $\boldsymbol{B_M^\top} = \mathbf{0}_{s\times s}$ then we say that mediators are *parallel*, otherwise they are *sequentially ordered*. Finally, $\boldsymbol{B_{XA}^\top}$ and $\boldsymbol{\gamma_{XA}}$ represent the influences of the interaction between the possible moderators and the treatment, $\boldsymbol{X}A$, on $\boldsymbol{M}$ and Y. This gives the proposed model the capability to characterize moderation to allow heterogeneity and multiple sequentially ordered mediators to allow complexity.

Remark 3.1. *We extend the linear SEM by directly integrating moderation into the casual graph. This is done for computational purposes and easier interpretation. Moderation is essential in learning HCGs and quantifying heterogeneity in causal effects. Specifically, the interaction $\boldsymbol{X}A$ in Model (1) allows causal effects to depend on the value of $\boldsymbol{X}$. As will be seen in Theorem 4.1, without interactions in the model, causal effects would be the same for all subjects within a population. Our model is thus flexible enough to handle heterogeneity and is general enough to cover other frameworks such as moderated mediation and mediated moderation as detailed in Appendix A.3. We also allow a fully heterogeneous causal relationship among mediators, including changes in causal direction, by adding $\boldsymbol{X}\boldsymbol{M}$ interactions to the model such that $\boldsymbol{X}$ would moderate the way mediators affect each other (see Section A.5). Also see further discussion on HCGs in Appendix A.4.*

Remark 3.2. *The structural constraints in g_1 to g_4 can be modified or more can be added to account for prior knowledge. The more information is known about the variables, the more structural constraints can be added to improve the estimation of the unknown causal graph [4].*

3.2. Generating Heterogeneous Causal Graphs

We next show how to produce an HCG given matrix $\boldsymbol{B}^\top$ and the value of moderators based on Model (1). To this end, we utilize the $do(\cdot)$ operator studied in Pearl [25], which simulates an intervention such that a variable can take a specific value irrespective of parent variable effects. Suppose we are interested in a group of subjects with $\boldsymbol{X} = \boldsymbol{x}$. By setting the value of $\boldsymbol{X}$ as $\boldsymbol{x}$ via $do(\cdot)$ operator, all the edges that go to $\boldsymbol{X}$ will be eliminated, which yields a new linear SEM:

$$\boldsymbol{D}_{do(\boldsymbol{X}=\boldsymbol{x})} = \boldsymbol{B}_{do(\boldsymbol{X}=\boldsymbol{x})}^\top \boldsymbol{D}_{do(\boldsymbol{X}=\boldsymbol{x})} + \boldsymbol{\epsilon}' \rightarrow \quad (2)$$

$$\begin{bmatrix} A \\ \boldsymbol{M} \\ Y \end{bmatrix} = \begin{bmatrix} \boldsymbol{\delta_X}\boldsymbol{x} \\ \boldsymbol{B_X^\top}\boldsymbol{x} \\ \boldsymbol{\gamma_X}\boldsymbol{x} \end{bmatrix} + \boldsymbol{B}_{do(\boldsymbol{X}=\boldsymbol{x})}^\top \begin{bmatrix} A \\ \boldsymbol{M} \\ Y \end{bmatrix} + \begin{bmatrix} \epsilon_A \\ \epsilon_M \\ \epsilon_Y \end{bmatrix},$$

where $\boldsymbol{B}_{do(\boldsymbol{X}=\boldsymbol{x})}^\top = \begin{bmatrix} 0 & \mathbf{0}_{1\times s} & 0 \\ \boldsymbol{\beta_A} + \boldsymbol{x}\boldsymbol{\beta_{XA}} & \boldsymbol{B_M^\top} & \mathbf{0}_{s\times 1} \\ \gamma_A + \boldsymbol{x}\boldsymbol{\gamma_{XA}} & \boldsymbol{\gamma_M} & 0 \end{bmatrix}$,

characterizes the heterogeneous causal graph with respect to $\boldsymbol{X} = \boldsymbol{x}$, and the additional intercept is a constant effect due to the fixed baseline. Thus, any HCG given estimated parameters for $\boldsymbol{B}^\top$ and $\boldsymbol{X} = \boldsymbol{x}$ can be obtained without extra model training, as illustrated in Figure 1.

Our proposed method first estimates a causal graph on the population level by including the treatment-and-moderator interaction in modeling to reflect heterogeneity, which allows us to project the estimated causal graph on the subgroup level to a certain value of moderator $\boldsymbol{X}$ through the do-operator as in equation 2. Under different values of $\boldsymbol{X} = \boldsymbol{x}$, the corresponding projected causal graphs are thus produced as the HCGs. Mathematically, it can be understood as $E(\widehat{\mathcal{G}}) = E_{\boldsymbol{X}}\{E(\widehat{\mathcal{G}}|\boldsymbol{X} = \boldsymbol{x})\} = E_{\boldsymbol{X}}\{E(\widehat{\mathcal{G}}|do(\boldsymbol{X} = \boldsymbol{x})\}$, where $\widehat{\mathcal{G}}$ is the estimated causal graph on the population level, the first equation is due to the formula of iterative expectation, and the second equation is owing to the no unmeasured confounder assumption. Thus, the entire causal graph at the population level can be viewed as the weighted average of HCGs and the weights are the density of $\boldsymbol{X}$. In the case that the subgroup of interest is characterized only by a subset of variables, we can set the values of those variables as appropriate and average across the rest as done in the population-level case.

4. Heterogeneous Causal Effects

In this section, we officially define heterogeneous causal effects (HCEs) and derive their theoretical properties.

4.1. Heterogeneous Causal Effects of Treatment

We focus on HCEs of treatment first. Following Pearl [25], we have the heterogeneous total effect (HTE) to be the total effect of the treatment on the outcome, the natural heterogeneous direct effect (HDE) to be the effect of the

treatment on the outcome that is not mediated by mediators, and the natural heterogeneous indirect effect (HIE) to be the effect of the treatment on the outcome that is regulated by mediators, given the baseline information.

Definition 1. *Heterogeneous Causal Effects of Treatment:*

$HTE(\boldsymbol{x}) = \partial E\{Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}\}/\partial a,$

$HDE(\boldsymbol{x}) = \partial E\{Y|do(A = a, \boldsymbol{M} = \boldsymbol{m}^{(a')}), \boldsymbol{X} = \boldsymbol{x}\}/\partial a,$

$HIE(\boldsymbol{x}) = \partial E\{Y|do(A = a', \boldsymbol{M} = \boldsymbol{m}^{(a)}), \boldsymbol{X} = \boldsymbol{x}\}/\partial a.$

Here, $\boldsymbol{m}^{(a)}$ is the value of $\boldsymbol{M}$ by setting $A = a$, and a' is some fixed value.

In Definition 1, the effects are functions of $\boldsymbol{x}$ as they depend on the particular individual's baseline information. Specifically, $HTE(\boldsymbol{x})$ is the change in the outcome of interest if we could have increased the treatment by one unit. Similarly, $HDE(\boldsymbol{x})$ can be interpreted as the change of the outcome due to one unit increase of the treatment when holding all mediators fixed, as the direct edge between A and Y in Figure 1 given different $\boldsymbol{x}$. In contrast, $HIE(\boldsymbol{x})$ captures the indirect effect of the treatment on the outcome regulated by mediators as two indirect paths (i.e., $A \rightarrow M_1 \rightarrow Y$ and $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) in Figure 1, under different $\boldsymbol{x}$. Similar causal effects, dubbed interventional effects, were proposed by Vansteelandt & Daniel [40], yet, under a very specific causal graph with only two mediators, without considering $\boldsymbol{X}$ as possible moderators. In contrast, the above HCEs are defined for a much more general scenario with multiple mediators. We develop the explicit form of the proposed HCTs under Model (1) in the following theorem.

Theorem 4.1. *Under assumptions (A1-A3) and the model described in Equation (1), we have:*
1). $HDE(\boldsymbol{x}) = \gamma_A + \boldsymbol{\gamma}_{\boldsymbol{X}A}\boldsymbol{x}$;
2). $HIE(\boldsymbol{x}) = \boldsymbol{\gamma}_{\boldsymbol{M}}(\boldsymbol{I}_s - \boldsymbol{B}_{\boldsymbol{M}}^{\top})^{-1}(\boldsymbol{\beta}_A + \boldsymbol{B}_{\boldsymbol{X}A}^{\top}\boldsymbol{x})$;
3). $HTE(\boldsymbol{x}) = HDE(\boldsymbol{x}) + HIE(\boldsymbol{x})$, where $\boldsymbol{I}_s$ is a $s \times s$ identity matrix and $\boldsymbol{x}$ is the value of $\boldsymbol{X}$.

The proof of Theorem 4.1 is provided in Section F in Appendix. We make a few remarks. First, results in 1) and 2) provide the exact form of HCEs as functions of a given $\boldsymbol{x}$, where the functions are determined by the unknown parameters in the weighted adjacency matrix $\boldsymbol{B}$ in Model (1). Therefore, learning $\boldsymbol{B}$ is not only the key to recovering the underlying causal structure but also the essential middle step of estimating the HCEs of interest. Second, results in 3) produce an exact decomposition of the heterogeneous total causal effect as the summation of the natural heterogeneous direct and indirect effects. This result thus generalizes Section 5.1.3 in Pearl et al. [26] by considering extra baseline information. This can be seen more clearly by setting $\boldsymbol{X} = 0$, in which case Pearl's equations will be generated. Third, it can be shown based on results in 1) to 3) that when $\boldsymbol{\gamma}_{\boldsymbol{X}A}$ and $\boldsymbol{B}_{\boldsymbol{X}A}^{\top}$ are all zeros, the HCTs do not depend on $\boldsymbol{x}$ anymore. In other words, without the interaction

terms, the causal impacts for different sub-populations are homogeneous. This supports the statements in Remark 3.1.

4.2. Heterogeneous Causal Effects of Mediators

We next define the HCEs of a particular mediator on the outcome. To this end, we define some intermediate quantities that deliver the partial causal impact of potential mediators on the outcome. Specifically, denote the effect of the treatment on mediator M_i given $\boldsymbol{X} = \boldsymbol{x}$ as

$\Delta_i(\boldsymbol{x}) \equiv \partial E\{M_i|do(A = a), \boldsymbol{X} = \boldsymbol{x}\}/\partial a,$

which quantifies how the treatment causally influences a single mediator given $\boldsymbol{X} = \boldsymbol{x}$. Next, we conceptualize the conditional mean outcome when increasing mediator M_i by 1 while fixing all other mediators as

$\Phi_i(a, \boldsymbol{x}) \equiv E\{Y|do(A = a, M_i = m_i^{(a)} + 1, \boldsymbol{\Omega}_i = \boldsymbol{o}_i^{(a)}), \boldsymbol{X} = \boldsymbol{x}\},$

where $\boldsymbol{\Omega}_i$ is a vector of mediators that do not include mediator M_i and $\boldsymbol{o}_i^{(a)}$ is the value of $\boldsymbol{\Omega}_i$ by setting $A = a$. With these intermediate definitions, we define the natural heterogeneous direct mediation effect (HDM), the natural heterogeneous indirect mediation effect (HIM), and the natural heterogeneous total mediation effect (HTM) as follows for each mediator on the outcome of interest, given the baseline.

Definition 2. *Heterogeneous Causal Effects of Mediator M_i:*

$HDM_i(\boldsymbol{x}) = \big[\Phi_i(a, \boldsymbol{x}) - E\{Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}\}\big] \times \Delta_i(\boldsymbol{x}),$

$HIM_i(\boldsymbol{x}) = \big[E\{Y|do(A = a, M_i = m_i^{(a)} + 1), \boldsymbol{X} = \boldsymbol{x}\}$
$\qquad - \Phi_i(a, \boldsymbol{x})\big] \times \Delta_i(\boldsymbol{x}),$

$HTM_i(\boldsymbol{x}) = \big[E\{Y|do(M_i = m_i + 1), \boldsymbol{X} = \boldsymbol{x}\}$
$\qquad - E\{Y|do(M_i = m_i), \boldsymbol{X} = \boldsymbol{x}\}\big] \times \Delta_i(\boldsymbol{x}).$

Similar to Definition 1, the mediation effects are functions of $\boldsymbol{x}$. The product inside Definition 2 is used to combine the causal effects of mediator M_i on the outcome with the total effect of the treatment on this mediator M_i, i.e., Δ_i. Here, $HDM_i(\boldsymbol{x})$ indicates the direct change of the outcome of interest if we could have increased a particular mediator by one unit while holding all its descendent/children mediators fixed. Likewise, $HIM_i(\boldsymbol{x})$ corresponds to the indirect effect of a particular mediator on the outcome regulated by its descendent/children mediators. Finally, $HTM_i(\boldsymbol{x})$ characterizes the total influence of a particular mediator regardless of the format. Take mediator M_1 in Figure 1 as an instance. The effect associated with the direct path $A \rightarrow M_1 \rightarrow Y$ is $HDM_1(\boldsymbol{x})$ and with the indirect path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ is $HIM_1(\boldsymbol{x})$, where the summation of these two paths representing $HTM_1(\boldsymbol{x})$. Our definition also generalizes the natural causal effects for an individual mediator proposed by Cai et al. [4] and Chakrabortty et al. [5] by incorporating baseline information and handling heterogeneity. Based on these definitions, we establish the theoretical forms of the mediation effects below proved in Appendix F.

Algorithm 1 Interactive Structural Learning

Input: data $\boldsymbol{D}$, baseline dimension p, dimension of mediators s, tolerance for estimating causal skeleton δ_b.

Initialize: $\widehat{\boldsymbol{B}}^0 \leftarrow \boldsymbol{0}$; $\widehat{\boldsymbol{B}} \leftarrow \boldsymbol{0}$

Estimate $\widehat{\boldsymbol{B}}^0$ using a causal discovery method of choice by incorporating h_2 into the loss function.

Compute $\widehat{\boldsymbol{B}}^b$, where $\widehat{\boldsymbol{B}}^b_{i,j} = \widehat{\boldsymbol{B}}^0_{i,j} > \delta_b$

for $i = 1$ **to** $2p + s + 2$ **do**

    Estimate direct children of ith node, denoted D_i, by $\{D_j | \widehat{\boldsymbol{B}}^b_{i,j} \neq 0\}$

    Update ith row of $\widehat{\boldsymbol{B}}$ using coefficients of fitted LASSO model using $\{D_j | \widehat{\boldsymbol{B}}^b_{i,j} \neq 0\}$ as the predictors and D_i as the response

end for

Compute the HCEs and desired CIs using $\widehat{\boldsymbol{B}}$

Return $\widehat{\boldsymbol{B}}$, HCEs, and CIs

Theorem 4.2. *Under assumptions (A1-A3) and Model (1),*
1a). $HDM_i(\boldsymbol{x}) = \{\boldsymbol{\gamma_M}\}_i \{(\boldsymbol{I}_s - \boldsymbol{B}_M^\top)^{-1}(\boldsymbol{\beta}_A + \boldsymbol{B}_{XA}^\top \boldsymbol{x})\}_i$;
1b). $\sum_{i=1}^s HDM_i(\boldsymbol{x}) = HIE(\boldsymbol{x})$;
2). $HIM_i(\boldsymbol{x}) = HTM_i(\boldsymbol{x}) - HDM_i(\boldsymbol{x})$;
3). $HTM_i(\boldsymbol{x}) = HIE(\boldsymbol{x}) - HIE_{\mathbb{G}(-i)}(\boldsymbol{x})$,
where $\{\cdot\}_i$ *is the ith element of a vector and* $HIE_{\mathbb{G}(-i)}$ *is the HIE under the causal graph* $\mathbb{G}(-i)$ *in which the ith mediator is removed from the original causal graph* $\mathbb{G}$.

We make a few remarks for Theorem 4.2. First, results in 1a) give the explicit form of the proposed HDM, where each multiplier corresponds to the ith element of the vector multiplier in results 2) of Theorem 4.1 for HIE. This immediately reveals the relationship between HIE and HDMs, i.e., the summation of all heterogeneous direct mediation effects equals to the heterogeneous indirect effect of the treatment, shown by results 1b) in Theorem 4.2. Second, the total mediation effect for a mediator M_i, HTM_i, can be interpreted as the effect of the treatment A on the outcome Y regulated by mediator M_i, or inversely as the change in total treatment effect that is due to M_i being removed from the causal graph. This provides a feasible way to obtain HTMs as indicated in results 3). Third, according to Definition 2, HTM_i can be further decomposed into direct effect that goes directly to the outcome through mediator M_i from the treatment (i.e., HDM_i) and indirect effect that is regulated by other descendant mediators of M_i as HIM_i. Such decomposition gives an alternative method to calculate HIM_i, as implied in results 2).

5. Main Algorithm

In this section, we detail the procedure used to estimate HCGs and HCEs with three components.

1. Causal discovery with structural constraints. In order to estimate the weighted adjacency matrix $\boldsymbol{B}$, we follow the

approach taken by Cai et al. [4] and combine the background structural knowledge with the score-based CSL [e.g., 30; 45; 43; 47]. This requires finding the best $\boldsymbol{B}$ given that it must be a DAG and satisfy the structural constraints posed in Section 3.1. In order for $\boldsymbol{B}$ to be a DAG, it can be shown that it must satisfy the following continuous acyclicity constraint typical of score-based CSL [e.g., 43; 47]:

$$h_1(\boldsymbol{B}) = \text{tr}[(\boldsymbol{I}_{2p+s+2} + t\boldsymbol{B} \odot \boldsymbol{B})^{2p+s+2}] - (2p+s+2) = 0,$$

where tr is the trace of a matrix, t is a hyper-parameter that depends on an estimation of the largest eigenvalue of the matrix $\boldsymbol{B}$, and $\odot$ denotes the element-wise square. In order for $\boldsymbol{B}$ to satisfy structural constraints, $g_1(\boldsymbol{B})$ to $g_4(\boldsymbol{B})$ (see Section 3.1), it must satisfy: $h_2(\boldsymbol{B}) = \sum_{i=1}^4 g_i(\boldsymbol{B}) = 0$. As remarked earlier, more structural constraints can be added and any added would be included in $h_2(\boldsymbol{B})$. Combining the two constraints above (h_1 and h_2) yields the following objective loss by an augmented Lagrangian [4],

$$\begin{aligned} L(\boldsymbol{B}, \theta) = & f(\boldsymbol{B}, \theta) + \lambda_1 h_1(\boldsymbol{B}) \\ & + \lambda_2 h_2(\boldsymbol{B}) + c|h_1(\boldsymbol{B})|^2 + d|h_2(\boldsymbol{B})|^2, \end{aligned}$$

where $f(\boldsymbol{B}, \theta)$ is some loss function such as the Kullback-Leibler (KL) divergence in DAG-GNN by Yu et al. [43] (see details and choices of the base learner in Appendix B.1), where model parameter θ, λ_1 and λ_2 are Lagrange multipliers, and c and d are tuning parameters to ensure a hard constraint on h_1 and h_2. With these methods we can generate an estimate for $\boldsymbol{B}$, denoted $\widehat{\boldsymbol{B}}^0$.

2. Debiasing with regularized regression. To reduce the bias in $\widehat{\boldsymbol{B}}^0$ introduced during causal discovery, we apply a refitting procedure [34] in accommodating Model (1). First, we extract the estimated causal skeleton from $\widehat{\boldsymbol{B}}^0$ by the binary matrix $\widehat{\boldsymbol{B}}^b$ where $\widehat{\boldsymbol{B}}^b_{i,j} = \mathbb{I}\{|\widehat{\boldsymbol{B}}^0_{i,j}| > \delta\}$. Here, δ is a pre-determined threshold to prune edges in the estimated causal graph with weak signals. Applying a *tolerance level* is a standard technique used in the causal discovery literature [see more details in 45; 43; 47]. In practice, we choose the best δ to minimize the loss (such as the mean squared error) between the original data and the data generated by $\widehat{\boldsymbol{B}}^0 \odot \mathbb{I}\{|\widehat{\boldsymbol{B}}^0| > \delta\}$. For each node D_i, we obtain its parent set as $PA_{\widehat{\mathbb{G}}}(D_i)$, where $\widehat{\mathbb{G}}$ is the estimated causal graph based on $\widehat{\boldsymbol{B}}^b$, and regress, using LASSO, D_i on $PA_{\widehat{\mathbb{G}}}(D_i)$ to obtain the refitted coefficients $\widehat{b}_{i,j}$ if $D_j \in PA_{\widehat{\mathbb{G}}}(D_i)$ and $\widehat{b}_{i,j} = 0$ otherwise. Denote the refitted matrix as $\widehat{\boldsymbol{B}}$, which is used to estimate causal effects based on Theorems 4.1 and 4.2. While we agree that the regularized regression is biased, yet, with sufficient sample size, the bias term is negligible. Such a debiasing idea is commonly used for statistical inference in high-dimensional settings [12; 44; 23].

3. Construction of confidence intervals (CIs). We construct CIs of the estimated effects based on bootstrap [6] and false discovery control [2]. Specifically, we resample n data points with replacements for independent K times.

In each bootstrap sample, we estimate the causal graph and compute the desired effect. For distribution-free noises, one can apply quantile-based CI by using $\alpha/2$-quantile of the bootstrap estimates as the lower bound and the $(1 - \alpha/2)$-quantile as the upper bound, where α is the significance level. In the case that the data is Gaussian in nature, the Gaussian bootstrap CI can be applied. We name the combined procedure Interactive Structural Learning (ISL). The full algorithm is included in Appendix B.1, however, an abbreviated version can be found in Algorithm 1, with the time complexity provided in Appendix B.2.

6. Extension to Functional SEM

We can extend our methodology to the *nonlinear* setting by generalizing the nature of the relationship between the nodes and their parents. Specifically, for each node $Z_i \in \mathbf{Z}$, we model the relationship between this node and one or more of its parents as $\alpha f_{ij}\{PA_{\mathbb{G}}^j(Z_i)\}$ where $PA_{\mathbb{G}}^j(Z_i)$ is the jth parent of Z_i and α is a scalar, vector, or matrix coefficient that is appropriately shaped serving the same purpose as the coefficients in $\mathbf{B}^\top$ from Model (1). Keep in mind, that the "combined" variable of $\mathbf{X}$ and A is considered a single distinct from $\mathbf{X}$ and A such that the heterogeneity of the data can be fully captured and interpreted. For example, the relationship between Y and A is modeled by $\gamma_A f_{Y,a}(A)$ and the modification of A's relationship with $\mathbf{M}$ is modeled by $\mathbf{B}_{\mathbf{XA}}^\top \mathbf{f}_{M,xa}(\mathbf{X}, A)$. Putting these all together, we have the following functional SEM:

$$\begin{cases} \mathbf{X} = \boldsymbol{\epsilon}_{\mathbf{X}}, \\ A = \boldsymbol{\delta}_{\mathbf{X}} \mathbf{f}_{A,x}(\mathbf{X}) + \epsilon_A, \\ \mathbf{M} = \mathbf{B}_{\mathbf{X}}^\top \mathbf{f}_{M,x}(\mathbf{X}) + \boldsymbol{\beta}_A^\top \mathbf{f}_{M,a}(A) \\ \quad\quad + \mathbf{B}_{\mathbf{XA}}^\top \mathbf{f}_{M,xa}(\mathbf{X}, A) + \mathbf{B}_{\mathbf{M}}^\top \mathbf{M} + \boldsymbol{\epsilon}_{\mathbf{M}}, \\ Y = \boldsymbol{\gamma}_{\mathbf{X}} \mathbf{f}_{Y,x}(\mathbf{X}) + \gamma_A f_{Y,a}(A) \\ \quad\quad + \boldsymbol{\gamma}_{\mathbf{XA}} f_{Y,xa}(\mathbf{X}, A) + \boldsymbol{\gamma}_{\mathbf{M}} \mathbf{f}_{Y,m}(\mathbf{M}) + \epsilon_Y. \end{cases} \quad (3)$$

We take a functional approach here, rather than a fully nonlinear approach, in order to have a closed form for the causal effects and allow for the model to be interpreted in a similar manner to the linear Model (1). This also allows us to directly extend the ISL algorithm in Section 5 to the functional case, by replacing the base learner with a learner proven to work efficiently in the nonlinear setting and generalizing the regression step to be appropriate for any functional model. The base learner we have chosen to use is the one recently introduced by Rolland et al. [32] which uses the data distribution's score function to prune the full DAG to estimate the causal graph. For more information see Appendix B.4. After the causal graph has been estimated, we can use it to estimate the coefficients in Model (3) by first positing a functional model and fitting it, given the estimated parents for each of the nodes. Here is where background information, like structural constraints, can be incorporated. While structural constraints can also be applied in the above ordering

---

Algorithm 2 Interactive Structural Learning Extended

Input: data $\mathbf{D}$, baseline dimension p, dimension of mediators s, tolerance for estimating causal skeleton δ_b.
Initialize: $\widehat{\mathbf{B}}^0 \leftarrow \mathbf{0}$; $\widehat{\mathbf{B}} \leftarrow \mathbf{0}$
Estimate the skeleton of $\widehat{\mathbf{B}}^0$ using Rolland et al. [32]'s SCORE method.
Fit chosen functional model using estimated skeleton while incorporating structural constraints to estimate $\widehat{\mathbf{B}}^0$
Compute $\widehat{\mathbf{B}}^b$, where $\widehat{\mathbf{B}}_{i,j}^b = \widehat{\mathbf{B}}_{i,j}^0 > \delta_b$
for $i = 1$ **to** $2p + s + 2$ **do**
 Estimate direct children of ith node, denoted D_i, by $\{D_j | \widehat{\mathbf{B}}_{i,j}^b \neq 0\}$
 Update ith row of $\widehat{\mathbf{B}}$ using coefficients of fitted LASSO model using $\{D_j | \widehat{\mathbf{B}}_{i,j}^b \neq 0\}$ as the predictors and D_i as the response
end for
Compute the HCEs and desired CIs using $\widehat{\mathbf{B}}$
Return $\widehat{\mathbf{B}}$, HCEs, and CIs

---

process, we have found that in practice it is more efficient to do so afterwards during the fitting process. Background information can also be incorporated in the form of the chosen functional models used for each relationship. Without background information, misspecification can occur, leading to inaccurate causal effect estimation.

Once the functional model has been fitted, analogous theorems to those posed in Section 4 can be used to calculate the causal effects. We emphasize here that the quantities defined in Definitions 1 and 2 in Section 4 are still applicable in the functional case, though the explicit form of heterogeneous causal effects derived in Theorems 4.1 and 4.2 are somewhat different in the functional model setting. According to Definition 1, we can develop the explicit form of the proposed HCEs for the treatment under Model (3) as follows.

Theorem 6.1. *Under assumptions (A1-A3) and Model (3),*
1). $HDE(\boldsymbol{x}, a) = \gamma_A \frac{\delta f_{Ya}(a)}{\delta a} + \boldsymbol{\gamma_{XA}} \frac{\delta f_{Yxa}(\boldsymbol{x}, a)}{\delta a}$;
2). $HIE(\boldsymbol{x}, a) = \boldsymbol{\gamma_M} \frac{\delta \boldsymbol{f_{Ym}}(\boldsymbol{m}^{(a)})}{\delta \boldsymbol{m}^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a}$;
3). $HTE(\boldsymbol{x}, a) = HDE(\boldsymbol{x}, a) + HIE(\boldsymbol{x}, a)$;
where $\boldsymbol{m}^{(a)} = E[\boldsymbol{M} | do(A = a), \boldsymbol{X} = \boldsymbol{x}]$.

Similarly, the form of HCEs for mediators can be developed.

Theorem 6.2. *Under assumptions (A1-A3) and Model (3),*
1a). $HDM_i(\boldsymbol{x}, a) = \{\boldsymbol{\gamma_M}\}_i \frac{\delta \boldsymbol{f_{Ym}}(\boldsymbol{m}^{(a)})}{\delta m_i^{(a)}} \frac{\delta m_i^{(a)}}{\delta a}$;
1b). $\sum_{i=1}^s HDM_i(\boldsymbol{x}, a) = HIE(\boldsymbol{x}, a)$;
2). $HIM_i(\boldsymbol{x}, a) = HTM_i(\boldsymbol{x}, a) - HDM_i(\boldsymbol{x}, a)$;
3). $HTM_i(\boldsymbol{x}, a) = HIE(\boldsymbol{x}, a) - HIE_{\mathbb{G}(-i)}(\boldsymbol{x}, a)$.

Proofs are presented for both theorems in the appendix. The main results under Theorems 6.1 and 6.2 can be interpreted analogously as in Theorems 4.1 and 4.2. The extended functional ISL is given in Algorithm 2.

7. Simulation Studies

We test the ability of our method on estimating HCGs and HCEs in a variety of situations. The computing infrastructure used is a virtual machine at Google Colab with the Pro tier for the majority of the computations, and a cluster server with 30 processor cores with 15GB of random access memory for computing bootstrap estimates in parallel.

Scenario generation. We generate 8 scenarios. Scenario 1 (S1) is the simple case in which only X, A, and Y have non-zero weights; Scenario 2 (S2) has parallel mediators ($B_M^\top = 0_{s \times s}$); and Scenario 3 (S3) has sequentially ordered mediators ($B_M^\top \neq 0_{s \times s}$). The true causal graphs for S1 and S2 are randomly generated given structural constraints, while the causal graph for S3 is generated using the Erdős-Reńyi (ER) model with a degree of 2. The weights in graphs are randomly chosen from $\{-1, 0, 1\}$. Each of these scenarios sets $p = 2$ and $s = 6$ for a total of 12 nodes in the graph (including treatment, interaction, and outcome). In addition, we require that there is at least one non-zero interaction term to ensure that causal graphs are heterogeneous as discussed in Remark 3.1. However, we also study two alternative cases: (1) a variant of S3 in which there is no interaction term (S3nx) and (2) a variant of S3 in which X is a moderator as defined by Kraemer et al. [14] (S3mod), that is, X is independent of A and there is an interaction term. Scenario 4 (S4), 5 (S5), and 6 (S6) are higher-dimensional versions of S3, with $s = 38$ in S4, $p = 18$ in S5, and $s = 10$ plus $p = 22$ in S6, where the underlying graphs are generated by the ER model with a degree of 4. The rest settings are the same as in S3. The sample size, n, is chosen from $\{500, 1000\}$. For S4-S6, the sample size is fixed at $n = 1000$. We also set the noise to be standard gaussian for X, A, M, and Y. Given the noise and the edge weights, Model 1 can be directly used to generate the data by setting X as its noise and using X to generate A with the edge weights and associated noise. X and A can then be used with the edge weights and associated noise to generate M which is in turn used to generate Y in a similar manner. After this process is complete a baseline value of 1.0 is added to the response. Finally, the column-wise mean value for the dataset is subtracted from the dataset. The seeds used to generate the 100 datasets used for each simulation scenario are 1 through 100.

Simulation results. We apply the proposed method to estimate the heterogeneous causal graphs and effects. The averaged estimated matrix $\widehat{B}^\top$ over 100 replicates under different sample sizes with its true causal graph are summarized in Figure D.1 to D.6 in Appendix for S1 to S6 to evaluate the proposed method. Each plot has a yellow box that identifies the inner weighted adjacency matrix for mediators and intersecting orange lines to emphasize the direct effects on or from the treatment. The corresponding accuracy metrics, here we use the false discovery rate (FDR), the true positive rate (TPR), and the structural Hamming
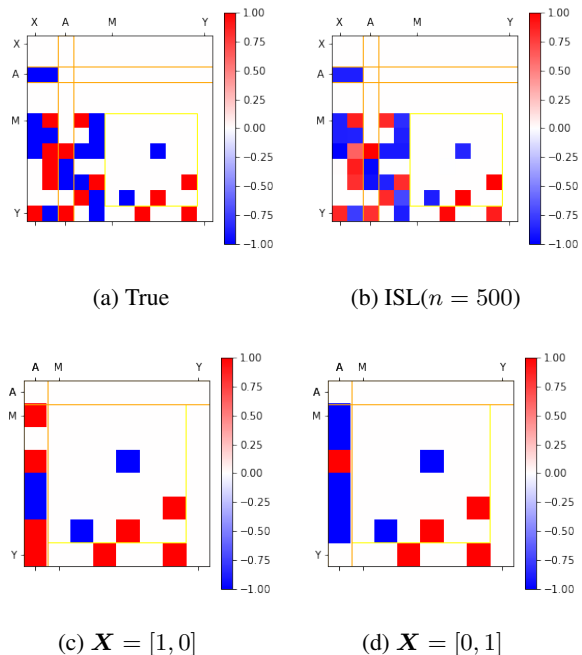

(a) True (b) ISL($n = 500$)

(c) $X = [1, 0]$ (d) $X = [0, 1]$

Figure 2: Results for S3 with subgraphs for different X.

distance (SHD), with their standard errors are provided in Tables 1 and 2. The bias and standard error of the estimated heterogeneous causal effects for S1-S3mod are summarized in Tables D.1-D.2 in the Appendix. The effects for S4-S6 are omitted for brevity. In addition, we compare our method with NOTEARS proposed by Zheng et al. [45] and DAG-GNN proposed by Yu et al. [43] for S1 to S3 with $n = 500$, as summarized in Tables D.3 and D.4. The penalty term in NOTEARS is set at $l_1 = 0.03$ the settings for DAG-GNN are left as default to achieve optimal performance. Here, for all estimated causal graphs, we set a threshold of 0.4 as the maximum absolute value that an edge weight needs to be to be considered nonzero. We also conduct a bootstrap simulation using 100 runs of 1000 bootstrap resamples to create 100 bootstrap CIs for each causal effect associated with S3. The average coverage probabilities are reported in Table D.5 with $\alpha = 0.05$. Here we note that in estimating graphs for use in constructing these confidence intervals, thresholding is used to accurately determine the causal skeleton. As will be seen in the real data section, thresholding can sometimes lead to cases in which all bootstrap samples return the same estimate yielding a confidence interval that only contains one value. While this is not ideal, it should be noted that this is not an error, but a consequence of using accuracy-improving techniques, such as thresholding.

As can be seen in Table 1 and 2, our method can estimate the weighted adjacency matrix of a causal graph with high accuracy which in turn leads to highly accurate estimates for the causal effects as seen in Tables D.1-D.2. The accuracy

Table 1: Accuracy metrics (standard errors) for $n = 500$.

	FDR	TPR	SHD
S1	0.00(0.00)	1.00(0.00)	0.00(0.00)
S2	0.01(0.02)	1.00(0.03)	0.25(0.86)
S3	0.00(0.01)	1.00(0.01)	0.04(0.24)
S3nx	0.00(0.01)	1.00(0.01)	0.03(0.22)
S3mod	0.00(0.00)	1.00(0.01)	0.02(0.14)

Table 2: Accuracy metrics (standard errors) for $n = 1000$.

	FDR	TPR	SHD
S1	0.00(0.00)	1.00(0.00)	0.00(0.00)
S2	0.00(0.00)	1.00(0.01)	0.01(0.10)
S3	0.00(0.00)	1.00(0.00)	0.00(0.00)
S3nx	0.00(0.00)	1.00(0.00)	0.00(0.00)
S3mod	0.00(0.00)	1.00(0.00)	0.00(0.00)
S4	0.03(0.01)	0.90(0.02)	14.67(3.45)
S5	0.00(0.01)	1.00(0.02)	0.43(2.53)
S6	0.00(0.01)	0.99(0.02)	2.39(6.26)

also increases marginally as the sample size increases. Our method can also be seen, in Figures D.1 to D.3 and Table D.3, to perform better than NOTEARS. Furthermore, our method can be seen to perform well in scenarios without interaction and in the case in which moderators that are independent of the treatment are used. In the higher-dimension case, it can be seen in Table 2 and D.6 that the accuracy of the proposed method remains quite high. S5 had the best accuracy metrics of the three and S4 had the worst, with S6 in the middle. It can then be concluded that in general, it is easier to estimate a graph with more potential moderators than with potential mediators. Here, SHD should be interpreted given the total possible errors a model could make. Since S4 has 44 nodes, $44^2/2 = 968$ errors are possible. So an SHD of 14.67 for S4 would be equivalent to an accuracy of 98.5%.
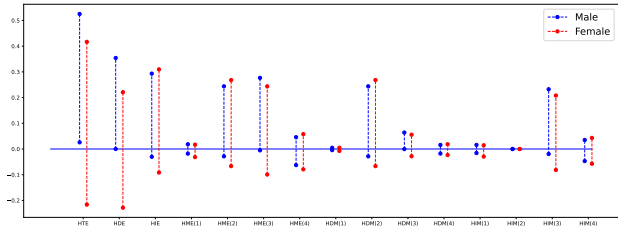

Figure 3: 95% Confidence intervals for effects of male and female subgraphs, produced using 1000 bootstrap samples.

8. Real Data Analysis

To demonstrate the practical usefulness of our method, we apply it to investigate the causal relationship of psychiatric disorders for trauma survivors using data collected from the AURORA study. The response of interest is post-traumatic stress disorder (PTSD) measured three months after trauma. The event of interest is pre-trauma insomnia of trauma survivors. Peri-traumatic (PT) PTSD, stress, acute distress (ASD), and depression measured 2 weeks after the trauma event are included as potential mediators. Moderators used for this study include age, gender, race, education, pre-trauma physical and mental health, perceived stress level, neuroticism, and childhood trauma. Before applying our method, categorical variables were one-hot encoded, numerical variables were centered, and missing data were removed. Results are summarized in causal graphs in Figures E.7a - E.10b with higher resolution for easier viewing. Motivations and overall effects are summarized in Appendix C.

Heterogeneous causal effects for this real data example are characterized by the interactions of pre-trauma insomnia with moderators, such as gender and race. These interaction terms can affect the effect of pre-trauma insomnia to response, or to potential mediators. Results are summarized in Figures E.2 - E.6. Results in Figure E.3 suggest the interaction of pre-trauma insomnia with Gender, education, and lifetime chronic stress are marginally significant for 3-month PTSD which indicates that the direct effect of pre-trauma insomnia to 3-month PTSD is moderated by gender,

education level, and level of chronic stress. In Figure E.2 and E.4 to E.6, we can also observe that the chronic stress level moderate (increase) the effect of pretrauma insomnia to peritraumatic distress, stress, PTSD, and depression. On the other hand, education level decreases the effect of pretrauma insomnia on the four peritraumatic disorders. Finally, Figure E.4 suggest that gender moderate the effect of pretrauma insomnia to peritraumatic stress but not the other three peritraumatic disorders. Heterogeneous causal effects for male and female trauma survivors are shown in Figure 3. Results suggest that the total, direct and indirect effects of pretrauma insomnia to 3-month PTSD are (marginally) significant for male trauma survivors but not for females, and the effect of pretrauma insomnia on 3-month PTSD are mainly mediated by peritrauma stress and PTSD. These results also indicate that preventive intervention of 3-month PTSD after trauma exposure that focuses on reducing pertrauma stress and PTSD is more likely to be effective for male than for female trauma survivors. The causal effects for the male and female subgraph are summarized in Tables E.2 and E.3, respectively.

References

[1] Athey, S. and Imbens, G. W. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

[2] Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

[3] Bühlmann, P., Peters, J., Ernest, J., et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6): 2526–2556, 2014.

[4] Cai, H., Song, R., and Lu, W. Anoce: Analysis of causal effects with multiple mediators via constrained structural learning. In *International Conference on Learning Representations*, 2020.

[5] Chakrabortty, A., Nandy, P., and Li, H. Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*, 2018.

[6] Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

[7] Farrell, M. H., Liang, T., and Misra, S. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[8] Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends in cognitive sciences*, 23(7):584–601, 2019.

[9] Hernán, M. A. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004.

[10] Hernán, M. Á., Brumback, B., and Robins, J. M. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pp. 561–570, 2000.

[11] Imai, K. and Yamamoto, T. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

[12] Jessica, M., Tian, L., and Cai, T. A perturbation method for inference on regularized regression estimates. *J Am Stat Association*, 106(496):1371–1382, 2011.

[13] Kalisch, M. and Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 (Mar):613–636, 2007.

[14] Kraemer, H. C., Wilson, T. G., Fairburn, C. G., and Agras, S. W. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59(10):877–883, 2002.

[15] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[16] Maathuis, M. H., Kalisch, M., Bühlmann, P., et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133–3164, 2009.

[17] Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):433–447, 2016.

[18] McLean, S. A., Ressler, K., Koenen, K. C., Neylan, T., Germine, L., Jovanovic, T., Clifford, G. D., Zeng, D., An, X., Linnstaedt, S., et al. The aurora study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. *Molecular psychiatry*, 25(2):283–296, 2020.

[19] Muller, D., Judd, C. M., and Yzerbyt, V. Y. When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6), 2005.

[20] Nandy, P., Maathuis, M. H., Richardson, T. S., et al. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674, 2017.

[21] Neylan, T. C., Kessler, R. C., Ressler, K. J., Clifford, G., Beaudoin, F. L., An, X., Stevens, J. S., Zeng, D., Linnstaedt, S. D., Germine, L. T., et al. Prior sleep problems and adverse post-traumatic neuropsychiatric sequelae of motor vehicle collision in the aurora study. *Sleep*, 44(3):zsaa200, 2021.

[22] Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2021.

[23] Ning, Y. and Liu, H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

[24] Panizza, U. and Presbitero, A. F. Public debt and economic growth: is there a causal effect? *Journal of Macroeconomics*, 41:21–41, 2014.

[25] Pearl, J. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[26] Pearl, J. et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[27] Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[28] Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. 2014.

[29] Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

[31] Robins, J. M. Semantics of causal dag models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*, pp. 70–81, 2003.

[32] Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. Score matching enables causal discovery of nonlinear additive noise models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

[33] Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

[34] Shi, C. and Li, L. Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, 2021.

[35] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct):2003–2030, 2006.

[36] Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., and Wimberly, F. Constructing bayesian network models of gene expression networks from microarray data. 2000.

[37] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

[38] Tchetgen, E. J. and Vanderweele, T. J. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25(2):282–291, 2014.

[39] Vanderweele, T. J. Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1): 17–32, 2016.

[40] Vansteelandt, S. and Daniel, R. M. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(2):258, 2017.

[41] Vansteelandt, S., Linder, M., Vandenberghe, S., Steen, J., and Madsen, J. Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Statistics in Medicine*, 38(24):4828–4840, 2019.

[42] Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

[43] Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098*, 2019.

[44] Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76:217–242, 2014.

[45] Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.

[46] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

[47] Zhu, S. and Chen, Z. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

Appendix

Table of Contents

A. More Details on Framework

A.1. Complete Definitions for Assumptions and further discussion

In order to understand the Causal Markov and Faithfulness assumptions, one must first understand the concept of D-separation.

Definition A.1 (D-separation). *Nodes, X and Y, are d-separated by a set of nodes, Z, if and only if for every path, p, there exists a node, $m \in Z$, that extends p ($i \rightarrow m \rightarrow j$) or forks p ($i \leftarrow m \rightarrow j$) and for any node, c, along p that is a so-called collider ($i \rightarrow m \leftarrow j$), c and all descendents of c are not in Z [26].*

Given that Z d-separates X and Y and X preceeds Y causally, the implication of d-seperation is that $X \perp\!\!\!\perp Y | Z$.

Definition A.2 (Causal Markov assumption). *For a given causal graph, $G = \{Z, \boldsymbol{E}\}$, the set of independences among the nodes, Z, contains the set of independences implied by d-separation.*

Definition A.3 (Faithfulness assumption). *For a given causal graph, $G = \{Z, \boldsymbol{E}\}$, the set of independences among the nodes, Z, is **exactly** described by the set of independences implied by applying d-separation to G.*

It should be noted that the assumptions made in this paper are common in the literature of causal inference. Please refer to [25; 26; 1; 20; 42; 15; 22] for discussions of these assumptions and their impact. There are a few future extensions to relax or diagnose these assumptions. For instance, a full sensitivity analysis of the assumptions would be useful to the field when it is hard to include all variables causally related to any variable in the data in practice. In addition, utilizing the instrumental variables in the contexts of causal graphs with multiple mediators may be beneficial to address no unmeasured confounders. We leave these directions for future work.

A.2. Further Discussion Regarding the Identifiability of the Causal Graph

Under the linear SEM, when the noises are Gaussian distributed, the model yields the class of standard linear-Gaussian model that has been studied in Spirtes et al. [36]; Peters et al. [29]. If the noises have equal variances, the DAG is uniquely identifiable from the observational data according to Peters & Bühlmann [27]. If the functions are linear but the noises are non-Gaussian, the true DAG can be uniquely identified under favorable conditions described in Shimizu et al. [35]. Also, the DAG can be naturally identified from the observational data if the corresponding MEC contains only one DAG, however, this cannot be known. The recent score-based causal discovery methods [45; 43; 47; 4] usually consider synthetic datasets that are generated from fully identifiable models so that it is practically meaningful to evaluate the estimated graph with respect to the true DAG. The non-linear case shares similar identifiability statements, with more details provided in Sections 2.1 and 3.3 in [46].

A.3. Comparison between Moderated Mediation and Mediated Moderation

The interaction seen in Model (1) is an example of moderation in which the relationship between two variables depends on another. In this case, the moderator would be $\boldsymbol{X}$, and the relationships being moderated would be A and $\boldsymbol{M}$ and A and Y. For those familiar with moderation, our setting may seem similar to the moderated mediation and/or mediated moderation settings. In Figure A.1, we have an example of moderation, moderated mediation, and mediated moderation taken from Muller et al. [19]'s paper on the topic. Here X is the treatment, Me is the mediator, Y is the response and Mo is the moderator, not shown in the figure. Here it can be seen that Mo moderates the relationship between X and Y, which is the example of moderation described earlier. Mo also moderates the relationship between X and Me and Me and Y which is an example of mediated moderation and moderated mediation respectively.
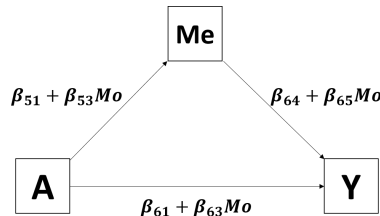

Figure A.1: Illustration of moderation, moderated mediation, and mediated moderation.

In their paper, Muller et al. [19] give equations similar to the linear SEM equations above to model both of these settings:

$$Me = \beta_{50} + \beta_{51}X + \beta_{52}Mo + \beta_{53}XMo + \epsilon_5,$$

$$Y = \beta_{60} + \beta_{61}X + \beta_{62}Mo + \beta_{63}XMo + \beta_{64}Me + \beta_{65}MeMo + \epsilon_6.$$

If we restate this as a linear SEM, the difference between our setting and the moderated settings can be seen:

$$
\begin{bmatrix} Mo \\ X \\ XMo \\ Me \\ MeMo \\ Y \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ \beta_{50} \\ 0 \\ \beta_{60} \end{bmatrix}
+
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\beta_{52} & \beta_{51} & \beta_{53} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\beta_{62} & \beta_{61} & \beta_{63} & \beta_{64} & \beta_{65} & 0
\end{bmatrix}
\begin{bmatrix} Mo \\ X \\ XMo \\ Me \\ MeMo \\ Y \end{bmatrix}
+ \boldsymbol{\epsilon}.
\tag{A.1}
$$

Comparing this model to our model, we can see that our model is more general. The moderator in this model is allowed to directly affect the treatment in our model, that is to say, we allow the confounder to moderate relations between other variables.

A.4. Further Discussion on Heterogeneous Causal Graphs

By heterogeneity, we are referring to the differences that may exist between subgroups within a population. We pose these differences via the way in which moderators affect and interact with the treatment variable in the causal graph. This leads to the possibility that different subgroups may have different causal graphs. This set of causal graphs is what we refer to as heterogeneous causal graphs (HCGs). Note that we assume here for practical reasons that X is a set of pre-treatment confounders and the topological order of A is lower than that of Y, i.e. the causal direction of the edge between A and Y is assumed. Furthermore, the existence of any other edge is not assumed and may be different among the HCGs depending on how X moderates A. The model presented in this work, given the topological assumption from earlier, will not lead to HCGs with the same edge that differ in causal direction, except that under a certain value of x the edge from A to M and A to Y may disappear based on the results in section 3.2. Such changes in causal edges from the treatment are owing to the XA interactions we considered.

Following a similar logic, theoretically, we could allow a fully heterogeneous causal relationship among mediators, including changes in causal direction, by adding XM interactions to the model such that X would moderate the way mediators affect each other (see Section A.5). A "fully heterogeneous" model may not be practically necessary since it is very uncommon to expect the causal direction to be heterogeneous. The slightly restricted version we proposed in this study should be able to address the majority of applied research questions with respect to heterogeneous causal relationships.

A.5. Adding XM Interaction to Model 1

In this section, we introduce an extension of Model 1 that includes XM interaction. Note that this interaction only includes a term for the response and does not include a term for mediators, i.e. we assume that X does not moderate how mediators affect other mediators. This is done mainly for three practical reasons: Firstly, we believe practitioners may not find value in modeling this relationship. Secondly, modeling this relationship dramatically increases the number of quantities that need estimating and thus increases the estimation error. Thirdly, including this term would not allow for a closed-form formula for the effects as we have done previously for Model 1 which reduces its interpretability.

$$
\begin{matrix} p \\ 1 \\ p \\ s \\ s*p \\ 1 \end{matrix}
\begin{bmatrix} \boldsymbol{X} \\ A \\ \boldsymbol{X}A \\ \boldsymbol{M} \\ \boldsymbol{X} \otimes \boldsymbol{M} \\ Y \end{bmatrix}
=
\begin{bmatrix}
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{\delta_X} & 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & 0 \\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{B_X^\top} & \boldsymbol{\beta_A} & \boldsymbol{B_{XA}^\top} & \boldsymbol{B_M^\top} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{\gamma_X} & \gamma_A & \boldsymbol{\gamma_{XA}} & \boldsymbol{\gamma_M} & vec(\boldsymbol{\Gamma_{XM}})^\top & 0
\end{bmatrix}
\begin{bmatrix} \boldsymbol{X} \\ A \\ \boldsymbol{X}A \\ \boldsymbol{M} \\ \boldsymbol{X} \otimes \boldsymbol{M} \\ Y \end{bmatrix}
+
\begin{bmatrix} \epsilon_X \\ \epsilon_A \\ \boldsymbol{X}A \\ \epsilon_M \\ \boldsymbol{X} \otimes \boldsymbol{M} \\ \epsilon_Y \end{bmatrix}.
$$

Theorem A.1. *Under the modified assumptions and model above:*

- *The natural direct effect*

$$DE = \gamma_A + \boldsymbol{\gamma_{XA}}\boldsymbol{x};$$

- *The natural indirect effect is*

$$IE = (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top\boldsymbol{\Gamma_{XM}})(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top}\boldsymbol{x});$$

- *The total effect of A on Y given X is*

$$TE = \gamma_A + \boldsymbol{\gamma_{XA}}\boldsymbol{x} + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top\boldsymbol{\Gamma_{XM}})(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top}\boldsymbol{x}),$$

where $\boldsymbol{\Gamma_{XM}}$ is a p by s coefficient matrix and $vec(\boldsymbol{\Gamma_{XM}})^\top$ is the vectorization of this matrix.

As one might expect, when taking $\boldsymbol{XM}$ interaction into account, the indirect effect of A changes accordingly.

B. More Details of Algorithms

B.1. ISL Algorithm Specifics

In this section, we provide more details for the proposed method we call Interactive Structural Learning (ISL). We start with the extension of the base learner DAG-GNN Yu et al. [43] in our proposed framework with confounder-based interaction, with the corresponding implementation details below. The complete pseudocode is also provided below. Other score-based algorithms are also applicable to our framework such as causal discovery with RL [47] expect the NOTEARS method [45] which requires differentiable constraints to support their optimization algorithm.

We use Yu et al. [43]'s deep generative approach for causal structural learning as an example to detail the estimation of $\boldsymbol{B}$. First, we convert Model (1) and treat the random error as independent latent variables such that

$$\boldsymbol{D} = (\boldsymbol{I_{2p+s+2}} - \boldsymbol{B}^\top)^{-1}\boldsymbol{\epsilon}.$$

We adopt the variational autoencoder model in Yu et al. [43] to generate $\boldsymbol{D}$ using two multilayer perceptrons (MLPs) as the encoder and the decoder with mean zero noise $\boldsymbol{\epsilon}$. The goal is to minimize the evidence lower bound f such that the generated data is close to the observed data as follows,

$$f(\boldsymbol{B}, \theta|\boldsymbol{D}) = \sum_{i=1}^{2p+s+2} \Delta_{KL}\{q(\boldsymbol{\epsilon}|\boldsymbol{D}_i)\|p(\boldsymbol{\epsilon})\}/(2p+s+2) - E_{q(\boldsymbol{\epsilon}|\boldsymbol{D}_i)}\{\log p(\boldsymbol{D}_i|\boldsymbol{\epsilon})\},$$

where $\boldsymbol{D}_i$ is the i-th variable in $\boldsymbol{D}$, θ is a vector of weights for MLPs, Δ_{KL} refers to the KL divergence, $q(\boldsymbol{\epsilon}|\boldsymbol{D}_i)$ is the reconstructed posterior distribution, $p(\boldsymbol{\epsilon})$ is the prior distribution, and $p(\boldsymbol{D}_i|\boldsymbol{\epsilon})$ is the likelihood function. This objective can be minimized using any preferred black-box stochastic optimization.

The pseudocode of Interactive Structural Learning (ISL) is provided in Algorithm 3.

Remark B.1. *Although ISL is able to identify causal heterogeneity, to have a complete examination of such causal heterogeneity, one has to apply the method to all possible configurations of the moderators. This may be a limitation in an extremely high dimensional setting, though we demonstrate a reasonably good performance of our method in Section 7 with $p = 22$ and more than 50 nodes in total. We leave the variable selection in causal graphs as a future investigation.*

B.2. Time Complexity of ISL Algorithm

Our method uses another causal discovery method as a base in order to find the estimated topological ordering of the nodes. Due to this, the time complexity of our method is based on the base method used. Let the time complexity of the base method be $g(N, n)$ where N is the data sample size, and the number of nodes, $n = 2p + s + 2$ where p is the number of confounders/moderators and s is the number of mediators. Let the time complexity of LASSO be $f(N, n)$. Computing the HCEs requires inverting a size n matrix and a size $n - 1$ matrix s times. Putting this together, the time complexity of our method is $O(g(N, n) + f(N, n) + n^2)$ for $s << n$ and $O(g(N, n) + f(N, n) + sn^2)$ otherwise.

B.3. Full Algorithms

Algorithm 3 Interactive Structural Learning

Input: data $\boldsymbol{D}$, dimension of baseline information p, dimension of mediators s, number of epochs H, max number of iterations K, parameter upper bound U, original learning rate r_0, tolerance for constraint δ, tuning parameters ρ and τ, tolerance for estimating causal skeleton δ_b.

Initialize: $\lambda_1 \leftarrow 0$; $\lambda_2 \leftarrow 0$; $c \leftarrow 1$; $d \leftarrow 1$; $r \leftarrow r_0$; $\widehat{\boldsymbol{B}}^0 \leftarrow \boldsymbol{0}$; $\widehat{\boldsymbol{B}} \leftarrow \boldsymbol{0}$; $h_1^{old} \leftarrow \infty$; $h_2^{old} \leftarrow \infty$; Encoder weights W^1, W^2 and Decoder weights W^3, W^4 with Gaussian initialization.

for $k = 0$ **to** $K - 1$ **do**
  repeat
    for $i = 0$ **to** $H - 1$ **do**
      1. Compute the distribution parameters for the noise using the encoder:

$$(\boldsymbol{\mu}_\epsilon, \boldsymbol{\sigma}_\epsilon) \leftarrow (\boldsymbol{I}_{2p+s+2} - (\widehat{\boldsymbol{B}}^0)^\top) MLP\{\boldsymbol{D}, W^1, W^2\}$$

      2. Compute the distribution parameters for the data using the decoder:

$$(\boldsymbol{\mu}_{\boldsymbol{D}}, \boldsymbol{\sigma}_{\boldsymbol{D}}) \leftarrow MLP\{(\boldsymbol{I}_{2p+s+2} - (\widehat{\boldsymbol{B}}^0)^\top)^{-1}\boldsymbol{\epsilon}, W^3, W^4\}$$

      3. Calculate structural constraints, $h_1^{new} \leftarrow h_1(\widehat{\boldsymbol{B}}^0)$, $h_2^{new} \leftarrow h_2(\widehat{\boldsymbol{B}}^0)$, and loss $\ell \leftarrow L(\widehat{\boldsymbol{B}}^0, W^1, W^2, W^3, W^4)$
      3. Use backward propagation to update $\widehat{\boldsymbol{B}}^0$ and network weights $\{W^1, W^2, W^3, W^4\}$
      4. Update learning rate r
    end for
    if $h_1^{new} > \rho h_1^{old}$ & $h_2^{new} > \rho h_2^{old}$ **then**
      $c \leftarrow c \times \tau$; $d \leftarrow d \times \tau$
    else if $h_1^{new} > \rho h_1^{old}$ & $h_2^{new} < \rho h_2^{old}$ **then**
      $c \leftarrow c \times \tau$
    else if $h_1^{new} < \rho h_1^{old}$ & $h_2^{new} > \rho h_2^{old}$ **then**
      $d \leftarrow d \times \tau$
    else
      Break
    end if
    Update: $h_1^{old} \leftarrow h_1^{new}$; $h_2^{old} \leftarrow h_2^{new}$; $\lambda_1 \leftarrow \lambda_1 \times h_1^{new}$; $\lambda_2 \leftarrow \lambda_2 \times h_2^{new}$;
    if $h_1^{new} < \delta$ & $h_2^{new} > \delta$ **then**
      Break
    end if
  until $c \times d < U$
end for
Compute $\widehat{\boldsymbol{B}}^b$, where $\widehat{\boldsymbol{B}}^b_{i,j} = \widehat{\boldsymbol{B}}^0_{i,j} > \delta_b$
for $i = 1$ **to** $2p + s + 2$ **do**
  Estimate direct children of ith node, denoted D_i, by $\{D_j | \widehat{\boldsymbol{B}}^b_{i,j} \neq 0\}$
  Update ith row of $\widehat{\boldsymbol{B}}$ using coefficients of fitted LASSO model using $\{D_j | \widehat{\boldsymbol{B}}^b_{i,j} \neq 0\}$ as the predictors and D_i as the response
end for
Return $\widehat{\boldsymbol{B}}$

B.4. More Information on Rolland et al. [32]'s SCORE Method and Extended ISL

Rolland et al. [32]'s SCORE Method is based on their finding that leaf nodes can be found by using the Jacobian of the data distribution's score function or the variance of the Jacobian. This fact is then used to estimate a topological ordering for the variables by iteratively finding a leaf node, adding it to the top of the current topological order, removing it from the graph, and repeating until all nodes have been ordered [32]. The causal graph is then estimated by using any pruning method to prune the full graph associated with the estimated topological order [32]. The SCORE algorithm is summarized below in Algorithm 4. Note that for practical reasons, the authors identify the leaf node as the node corresponding as the smallest

element of the variance of the Jacobian [32]. For further information on the computation of the score function, please see Rolland et al. [32]'s paper.

Algorithm 4 SCORE method

    Input: data X with d nodes and sample size N.
    Initialize: topological order $\pi = []$ and nodes$= \{1, ..., d\}$
    for $i = 1$ **to** d **do**
        Estimate score function s_nodes
        Estimate variance $V_j = \text{Var}_{X_{\text{nodes}}}\left[\frac{\delta s_j(X)}{\delta x_j}\right]$
        Find the leaf node: $l \leftarrow \text{nodes}[\arg\min_j V_j]$
        Add leaf node to topological order: $\pi \leftarrow [l, \pi]$
        Remove l from the node set and data matrix
    end for
    Prune the full DAG associated with π using desired method
    Return the pruned DAG

C. More Information about the Real Data Analysis

Heterogeneity is common for many types of disorders in terms of phenotype, risk factors, treatment effect, and underlying neuro-biological mechanism. This is especially true for psychiatric disorders [17], such as depression and post-traumatic stress disorder (PTSD). Heterogeneity is one of the important obstacles that continue to hamper research for psychiatric disorders [17]. To help overcome these challenges and advance research for psychiatric disorders, the National Institutes of Mental Health, joined by several institutions and foundations, launched the Advancing Understanding of RecOvery afteR traumA (AURORA) study [18], where thousands of trauma survivors were recruited at the emergency departments after trauma exposures, and followed for one year to collect a broad range of bio-behavioral data.

One of the aims of the AURORA study is to address the heterogeneity issue by identifying homogeneous sub-types or subgroups of patients after trauma exposure (i.e., homogeneous prognosis patterns among patients with PTSD). Currently, most of the researchers for psychiatric disorders focus on stratifying patients into subgroups using different clustering algorithms [17] based on cross-sectional or longitudinal patterns using self-reported data and neuroimaging data. On the other hand, a large number of studies have also been conducted to investigate the associations among pre-trauma characteristics, peritraumatic symptoms, and post-traumatic disorders. However, due to the lack of appropriate statistical tools, very few studies have been conducted to explore heterogeneous causal relationships for psychiatric disorders.

To develop preventive interventions for psychiatric disorders, such as post-traumatic stress disorder (PTSD), in the immediate aftermath of trauma exposure, it is essential to identify risk factors that can be targeted to prevent or alleviate the corresponding disorder. For instance, a recent study by Neylan et al. [21] found that pre-trauma sleep disorders (e.g., nightmare and insomnia) are associated with 8-week PTSD, and these associations are mediated by peritraumatic outcomes (e.g., acute stress disorder) measured at 2 weeks. However, as the authors have discussed, the associations identified in this study do not support causal conclusions. Furthermore, the possible causal effect of pre-trauma sleep disorder on PTSD is very likely moderated by characteristics such as gender, race, pre-trauma mental health, etc. As a result, it is valuable to understand the heterogeneous causal pathways of pre-trauma characteristics with PTSD and other psychiatric disorders.

Under our linear model setting, the marginal effect of each covariate is directly characterized by the coefficient corresponding to the interaction term of each covariate with treatment A. Our model can help identify real moderators from a set of potential covariates which can be potentially very useful for applied studies since prior knowledge about real moderators is often not available.

We worked with domain experts to try to take into account as many possible confounders as possible based on previous research, but it is impossible to claim that all potential confounders have been identified and included in the model. Results from this real data analysis indicate that overall the total, direct and indirect effects of pre-trauma insomnia are all marginally significant (HTE, HDE, and HIE in Figure E.1), and total mediation effects from peritraumatic acute stress and peritraumatic PTSD are also marginally significant (HME(2) and HME(3) in Figure E.1). The mediation effect from peritraumatic acute stress is mainly from the direct mediation effect (HDM(2) in Figure E.1). On the other hand, the mediation effect from peritraumatic PTSD is from indirect mediation effects (HIM(3) in Figure E.1). The causal effects for the whole population

graph are summarized in Table E.1.

D. Additional Simulation Figures and Tables

Table D.1: Bias (standard error) of causal effects for S1-S3.

	$n = 500$			$n = 1000$		
	S1	S2	S3	S1	S2	S3
HDE	-0.01(0.05)	-0.25(0.23)	-0.13(0.07)	-0.02(0.04)	-0.29(0.09)	-0.13(0.04)
HIE	0.00(0.00)	0.40(0.17)	0.29(0.23)	0.00(0.00)	0.42(0.11)	0.25(0.06)
HDM_1	0.00(0.00)	0.53(0.14)	0.00(0.00)	0.00(0.00)	0.52(0.06)	0.00(0.00)
HDM_2	0.00(0.00)	0.00(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.03)	0.00(0.00)
HDM_3	0.00(0.00)	0.55(0.08)	0.13(0.20)	0.00(0.00)	0.56(0.05)	0.10(0.05)
HDM_4	0.00(0.00)	-0.55(0.09)	0.00(0.02)	0.00(0.00)	-0.54(0.05)	0.00(0.00)
HDM_5	0.00(0.00)	-0.12(0.07)	0.00(0.00)	0.00(0.00)	-0.11(0.05)	0.00(0.00)
HDM_6	0.00(0.00)	0.00(0.04)	0.16(0.07)	0.00(0.00)	0.00(0.02)	0.15(0.04)
HIM_1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_4	0.00(0.00)	0.00(0.00)	0.01(0.06)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_5	0.00(0.00)	0.00(0.00)	0.12(0.09)	0.00(0.00)	0.00(0.00)	0.10(0.03)
HIM_6	0.00(0.00)	0.00(0.00)	0.01(0.05)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Table D.2: Bias (standard error) of causal effects for S3nx and S3mod.

	$n = 500$		$n = 1000$	
	S3nx	S3mod	S3nx	S3mod
HDE	-0.11(0.07)	0.02(0.08)	-0.12(0.05)	0.03(0.06)
HIE	0.13(0.18)	0.73(0.15)	0.11(0.05)	0.72(0.12)
HDM_1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HDM_2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HDM_3	-0.05(0.18)	0.49(0.12)	-0.07(0.05)	0.48(0.09)
HDM_4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HDM_5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HDM_6	0.19(0.05)	0.25(0.09)	0.18(0.02)	0.24(0.07)
HIM_1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_4	0.00(0.04)	0.00(0.00)	0.00(0.00)	0.00(0.00)
HIM_5	0.19(0.06)	0.22(0.09)	0.18(0.02)	0.21(0.07)
HIM_6	0.00(0.04)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Table D.3: Accuracy metrics for S1-S3 using NOTEARS and a sample size of 500 averaged across 100 simulated datasets.

	$n = 500$		
	FDR	TPR	SHD
S1	0.00(0.00)	1.00(0.00)	0.00(0.00)
S2	0.09(0.09)	0.90(0.11)	3.06(3.14)
S3	0.10(0.09)	0.94(0.08)	2.52(1.97)

Table D.4: Accuracy metrics for S1-S3 using DAG-GNN and a sample size of 500 averaged across 100 simulated datasets.

	$n = 500$		
	FDR	TPR	SHD
S1	0.03(0.09)	0.97(0.08)	0.14(0.42)
S2	0.05(0.04)	0.97(0.07)	1.43(1.51)
S3	0.03(0.05)	0.98(0.04)	0.70(1.05)

(a) True   (b) NOTEARS($n = 500$)   (c) DAG-GNN($n = 500$)   (d) ISL($n = 500$)   (e) ISL($n = 1000$)
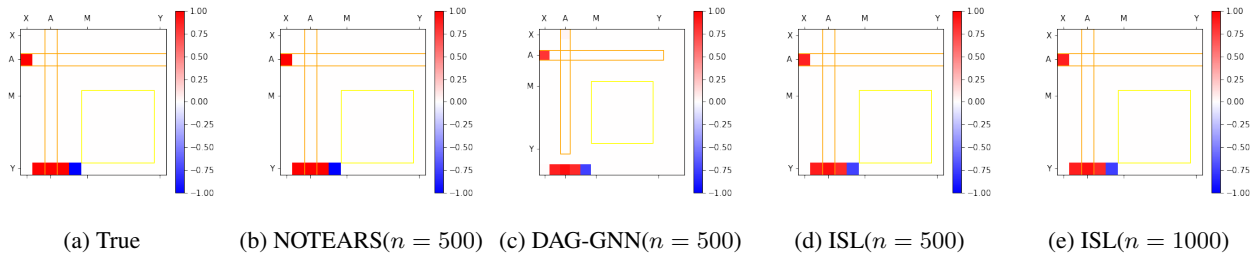
Figure D.1: Estimated Causal Graph for Scenario 1 without mediator effects for $n = 500$ and 1000 with NOTEARS and DAG-GNN shown for comparison.

(a) True   (b) NOTEARS($n = 500$)   (c) DAG-GNN($n = 500$)   (d) ISL($n = 500$)   (e) ISL($n = 1000$)
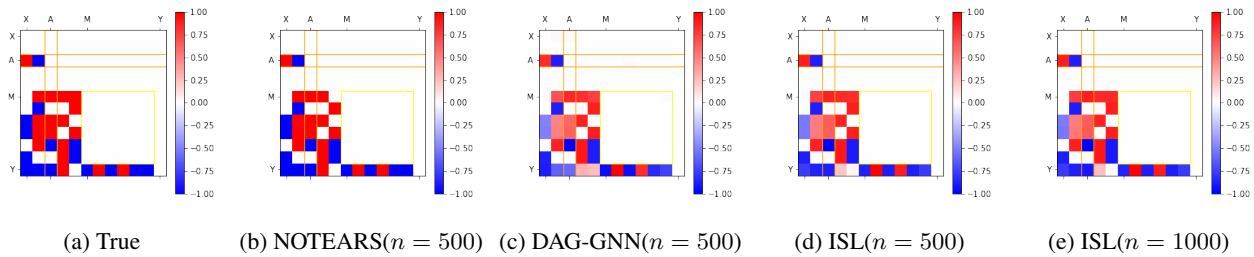
Figure D.2: Estimated Causal Graph for Scenario 2 with parallel mediators for $n = 500$ and 1000 with NOTEARS and DAG-GNN shown for comparison.

(a) True   (b) NOTEARS($n = 500$)   (c) DAG-GNN($n = 500$)   (d) ISL($n = 500$)   (e) ISL($n = 1000$)
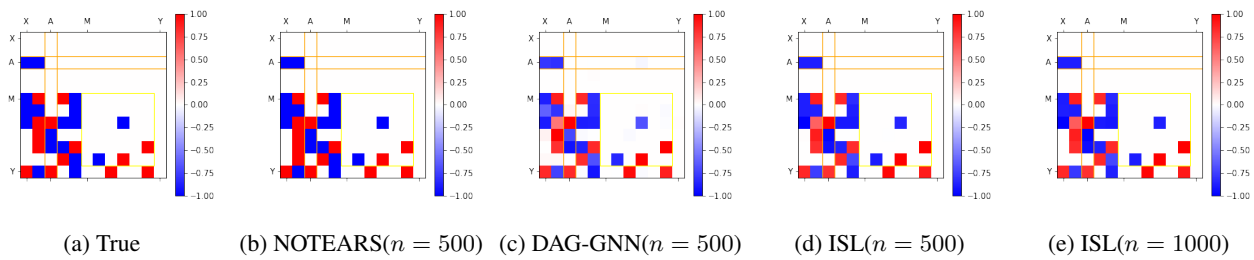
Figure D.3: Estimated Causal Graph for Scenario 3 with sequentially ordered mediators or $n = 500$ and 1000 with NOTEARS and DAG-GNN shown for comparison.

(a) True (b) $n = 500$ (c) $n = 1000$
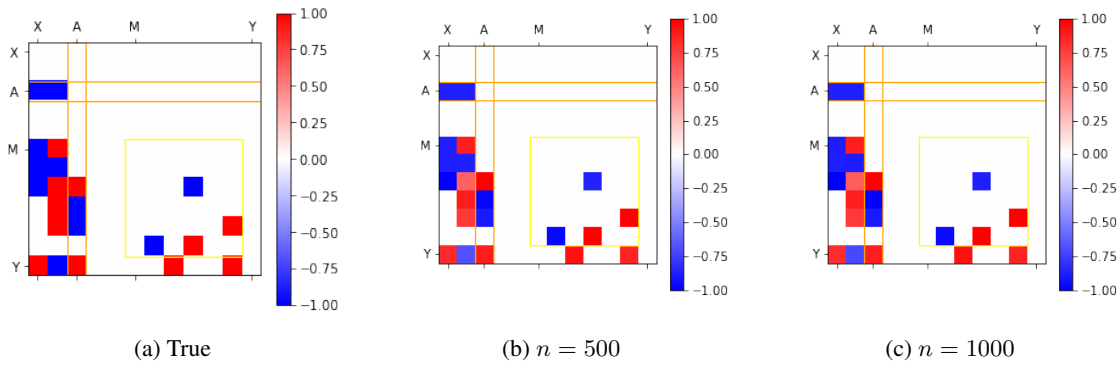
Figure D.4: Estimated Causal Graphs for Scenario 3 when X is a moderator (d to f) with sample sizes of $n = 500$ and 1000.

(a) True (b) $n = 500$ (c) $n = 1000$
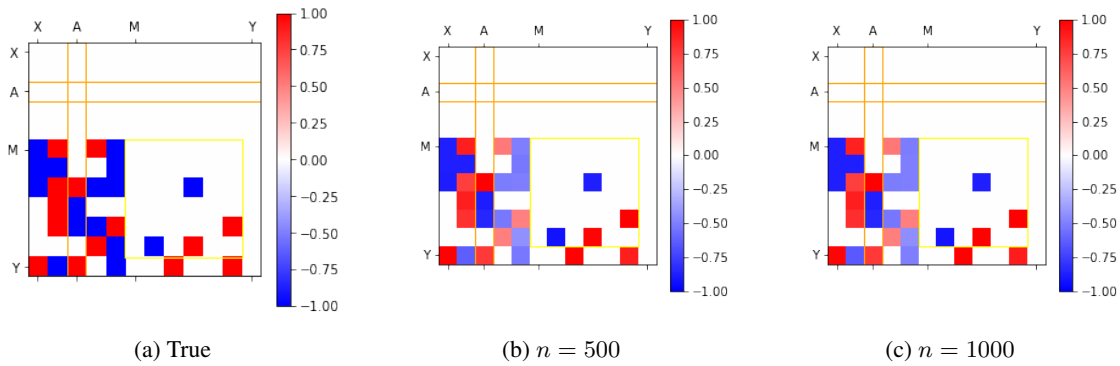
Figure D.5: Estimated Causal Graphs for Scenario 3 with no interaction (a to c) with sample sizes of $n = 500$ and 1000.

Table D.5: Coverage for causal effects and interaction edge weights ($XA \rightarrow Y$) for S3 averaged across 100 simulated datasets using 1000 bootstrap resamples each with the significance level of 0.05.

	Percentile	Gaussian
HDE	0.94	0.88
HIE	0.76	0.90
HDM_1	1.00	1.00
HDM_2	1.00	0.99
HDM_3	0.96	0.96
HDM_4	1.00	1.00
HDM_5	1.00	1.00
HDM_6	1.00	0.99
HIM_1	1.00	1.00
HIM_2	1.00	1.00
HIM_3	1.00	1.00
HIM_4	1.00	1.00
HIM_5	0.97	0.93
HIM_6	1.00	0.99
IX_1	1.00	1.00
IX_2	0.84	0.86

(a) True S4 (b) S4 estimate (n=1000) (c) S4 estimate (n=1500)

(d) True S5 (e) S5 estimate (n=1000) (f) S5 estimate (n=1500)

(g) True S6 (h) S6 estimate (n=1000) (i) S6 estimate (n=1500)
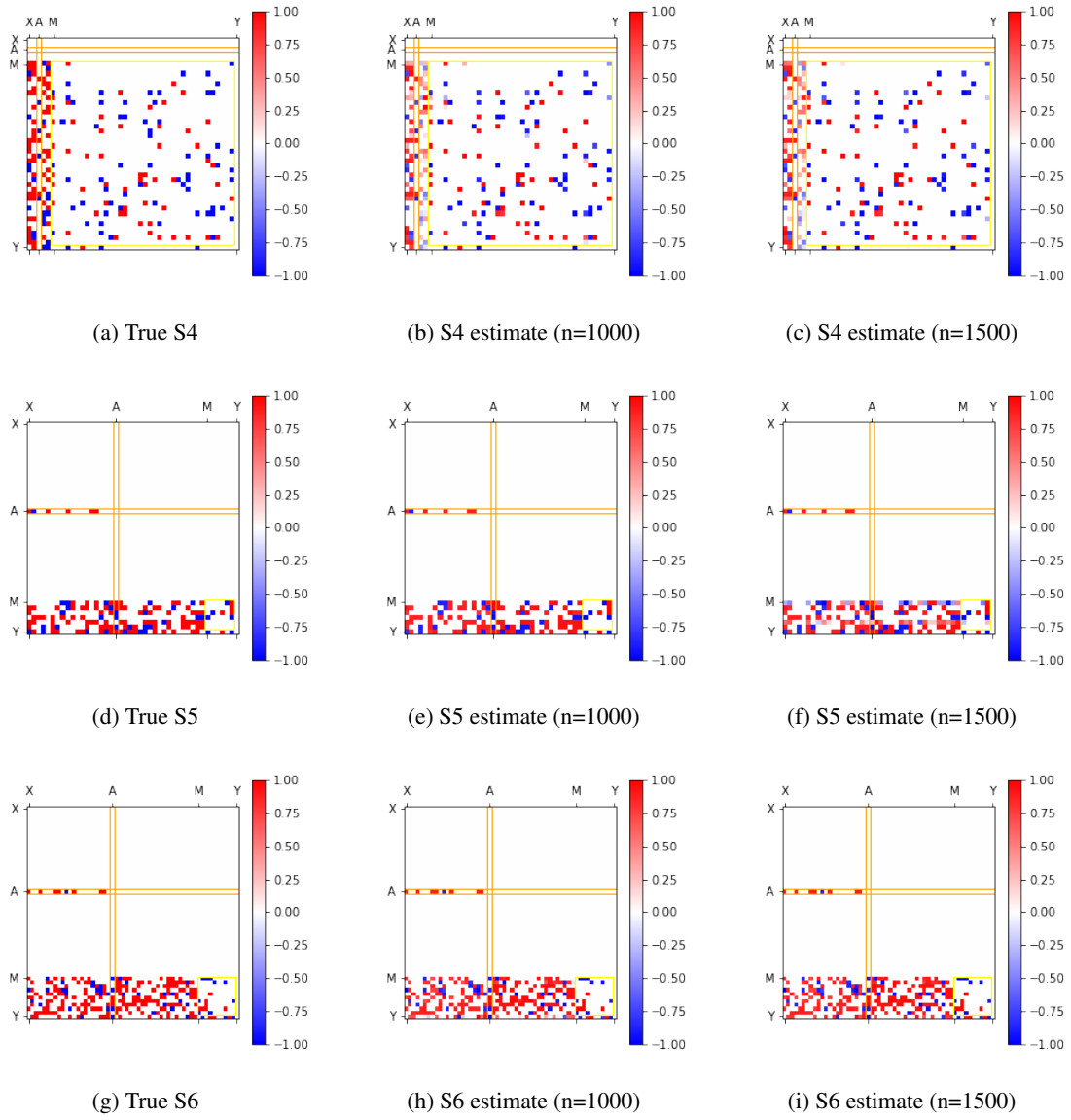
Figure D.6: Results for Scenario 4 with more mediators, Scenario 5 with more moderators, and Scenario 6 with more mediators and moderators with sample sizes of $n = 1000$ and 1500.

Table D.6: Accuracy metrics for S4-S6 with $n = 1500$.

	$n = 1500$		
	FDR	TPR	SHD
S4	0.01(0.01)	0.91(0.02)	12.28(3.07)
S5	0.03(0.03)	0.96(0.04)	6.99(7.44)
S6	0.00(0.02)	1.00(0.02)	1.08(6.09)

E. Figures and Tables from the Real Data Analysis

Table E.1: Estimated effects for the whole population.

	Insomnia	PT distress	ASD	PT PTSD	PT depression
Total	0.1931	-0.0036	0.0572	0.0802	-0.0059
Direct	0.1160	-0.0007	0.0572	0.0227	-0.0021
Indirect	0.0771	-0.0029	0	0.0576	-0.0038

Table E.2: Estimated causal effects for males.

	Total	Direct	Indirect
Insomnia	0.2418	0.1729	0.0690
PT distress	-0.0023	-0.0004	-0.0019
ASD	0.0498	0.0498	0
PT PTSD	0.0831	0.0235	0.0597
PT Depression	-0.0108	-0.0039	-0.0069

Table E.3: Estimated causal effects for females.

	Total	Direct	Indirect
Insomnia	0.102	0.0099	0.0921
PT distress	-0.006	-0.0011	-0.0048
ASD	0.0709	0.0709	0
PT PTSD	0.0748	0.0211	0.0537
PT Depression	0.0034	0.0012	0.0021
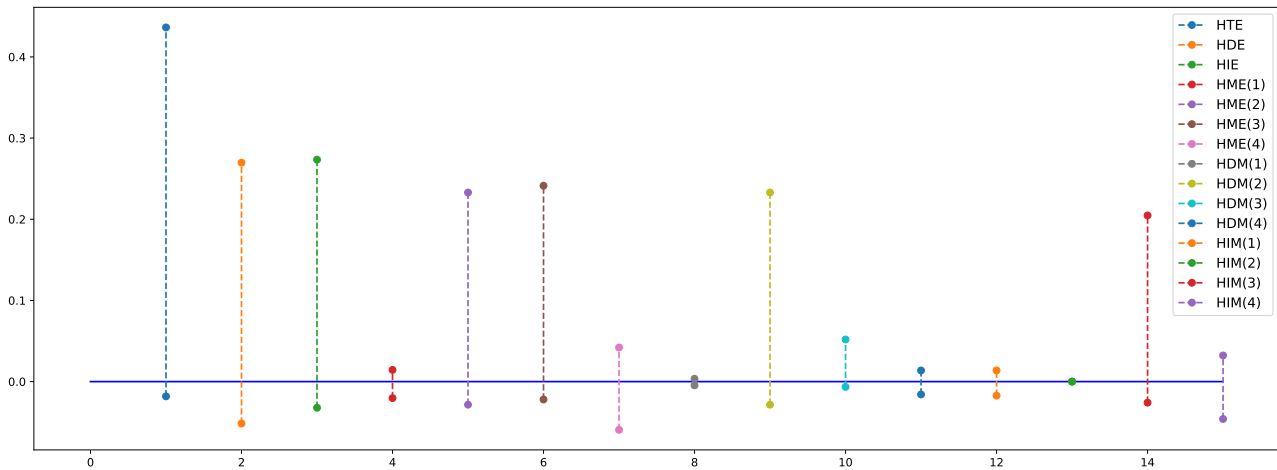
Figure E.1: 95% Confidence intervals for causal effects on the population level graph, produced using 1000 bootstrap samples.
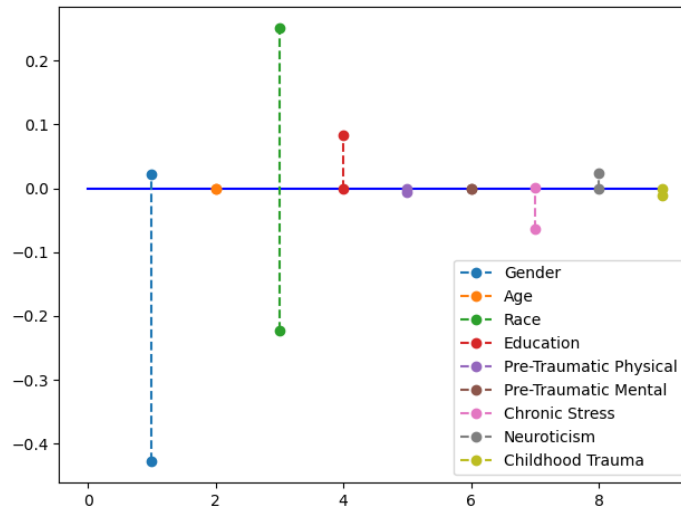
Figure E.2: 95% Confidence intervals for interaction effects on response computed over 1000 bootstrap resamples.
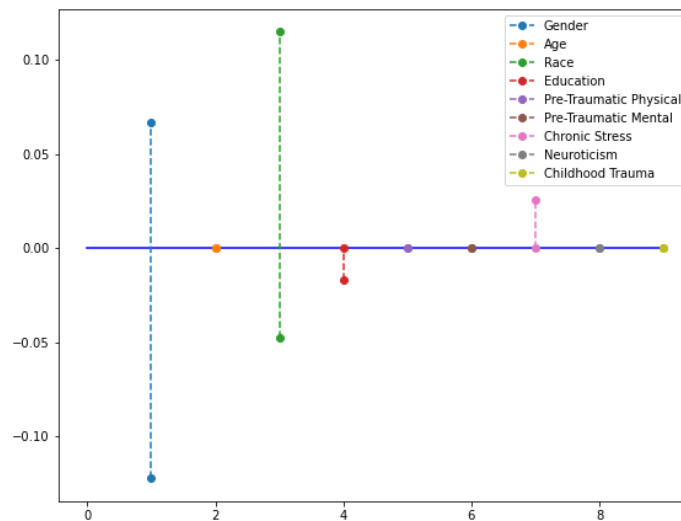
Figure E.3: 95% Confidence intervals for interaction effects on PT Distress computed over 1000 bootstrap resamples.
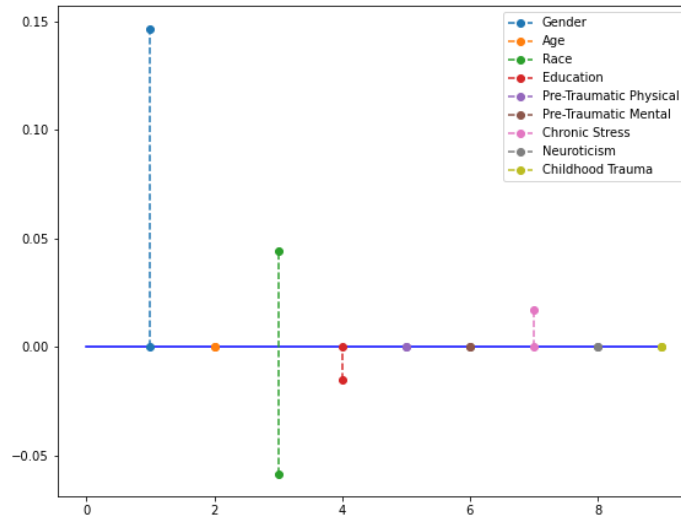
Figure E.4: 95% Confidence intervals for interaction effects on ASD computed over 1000 bootstrap resamples.
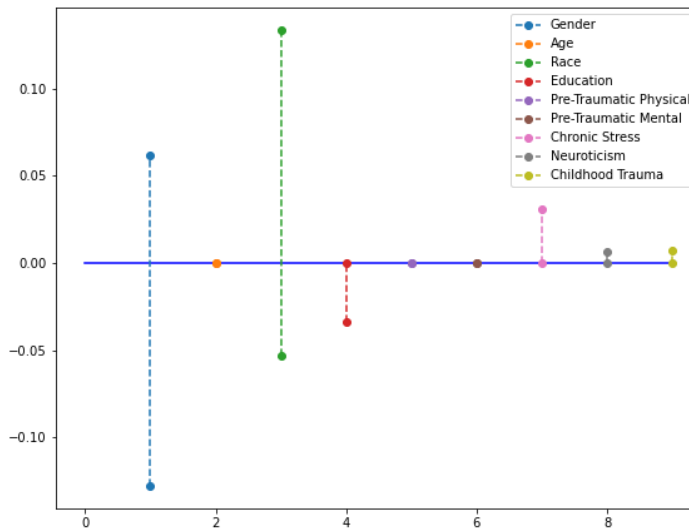
Figure E.5: 95% Confidence intervals for interaction effects on PT PTSD computed over 1000 bootstrap resamples.
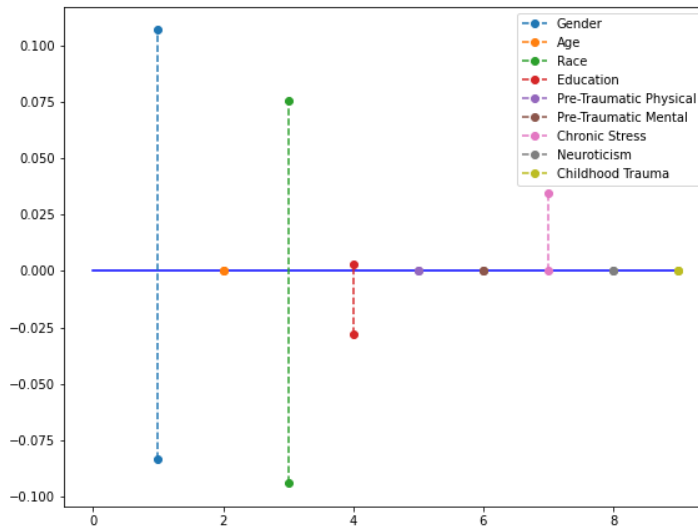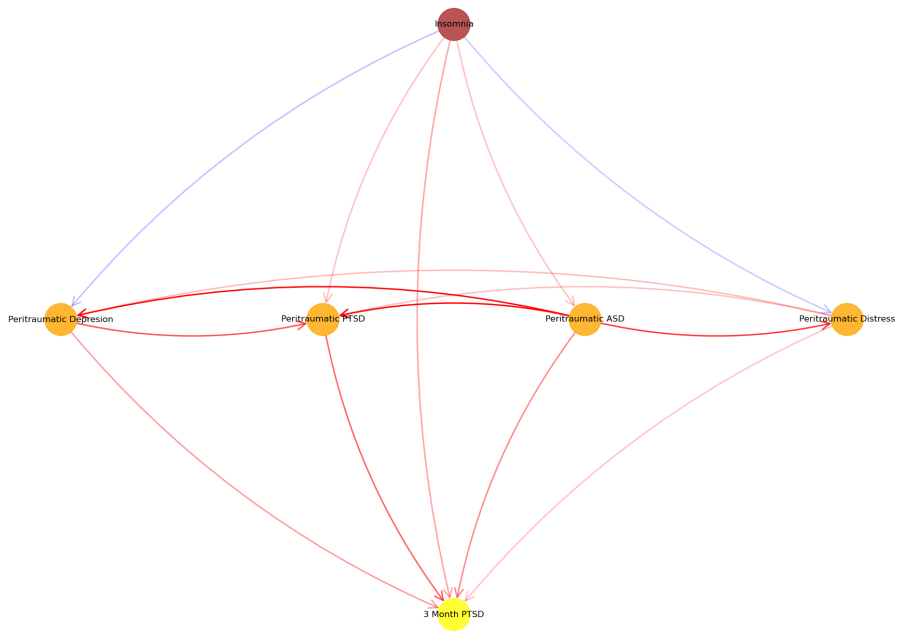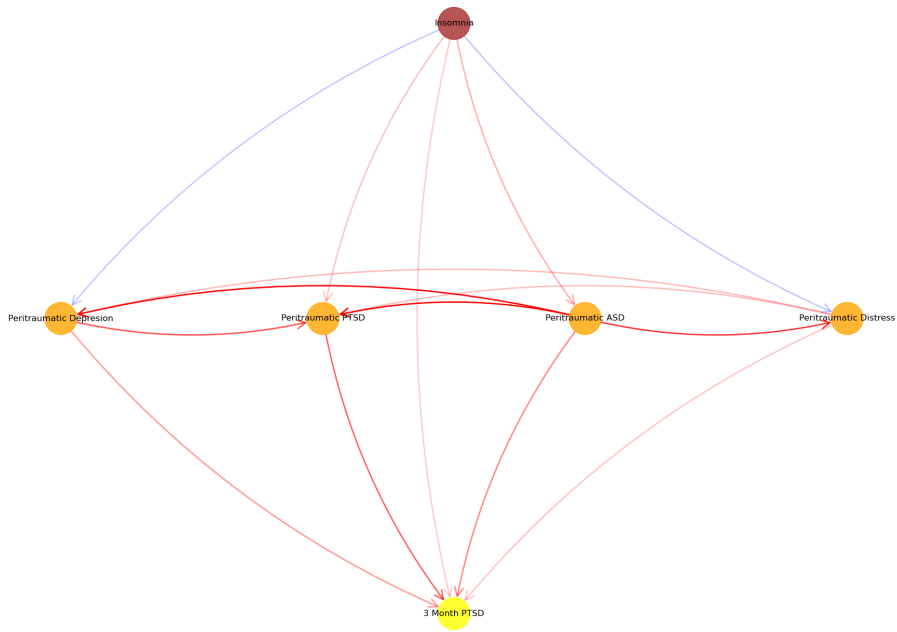
Figure E.6: 95% Confidence intervals for interaction effects on PT Depression computed over 1000 bootstrap resamples.

(a) Estimated causal graph for males

(b) Estimated causal graph for females

Figure E.7: Causal graphs for the male and female demographic. The red node is the event of interest. Orange nodes are mediators. The yellow node is the outcome.
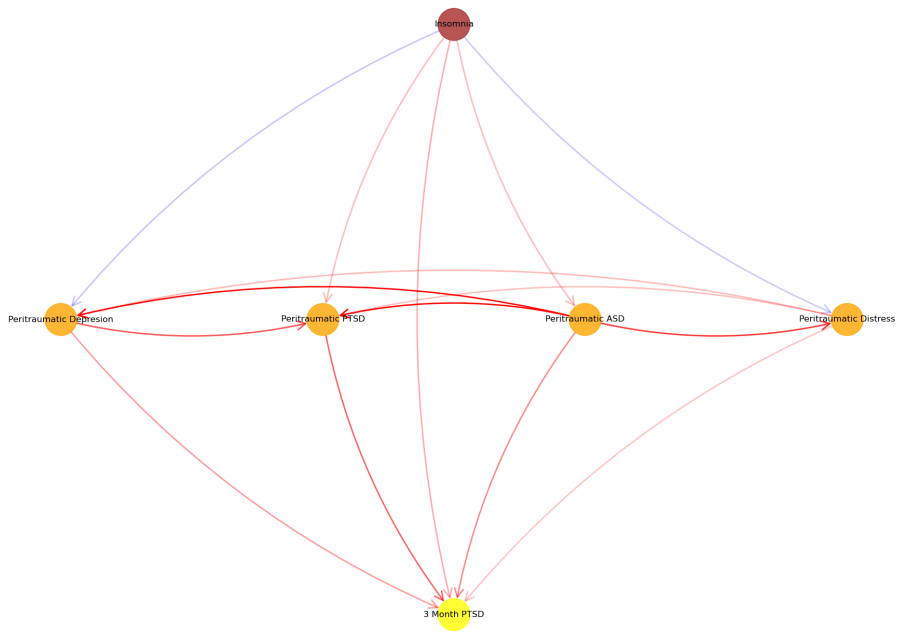
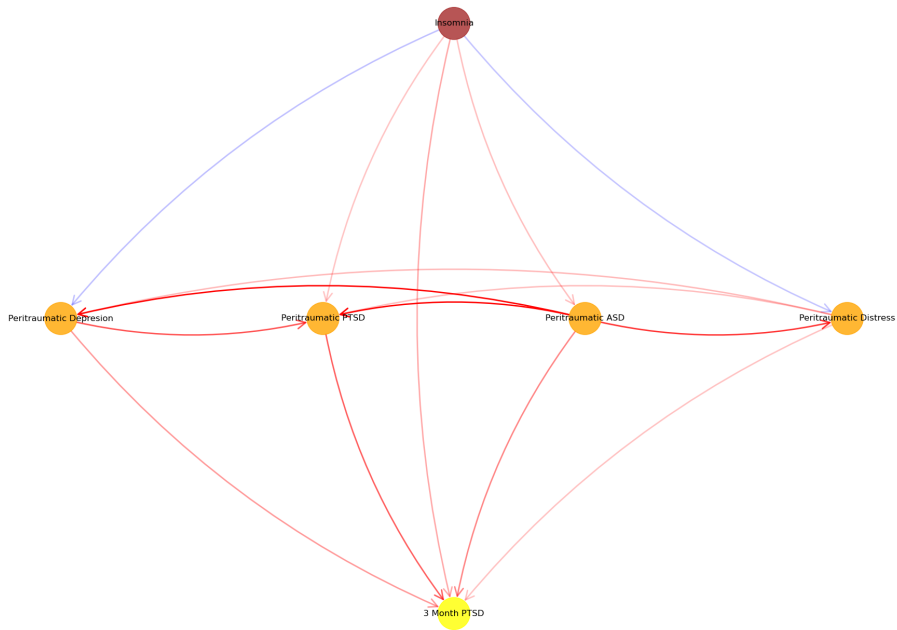(a) Estimated causal graph for the white population.
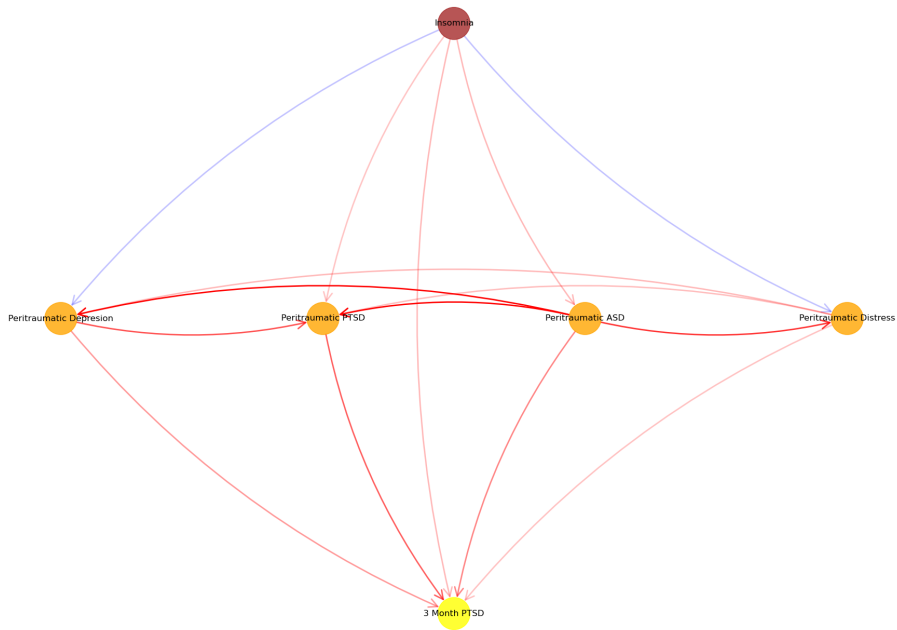
(b) Estimated causal graph for the black population

Figure E.8: Causal graphs for the white and black demographic. The red node is the event of interest. Orange nodes are mediators. The yellow node is the outcome.
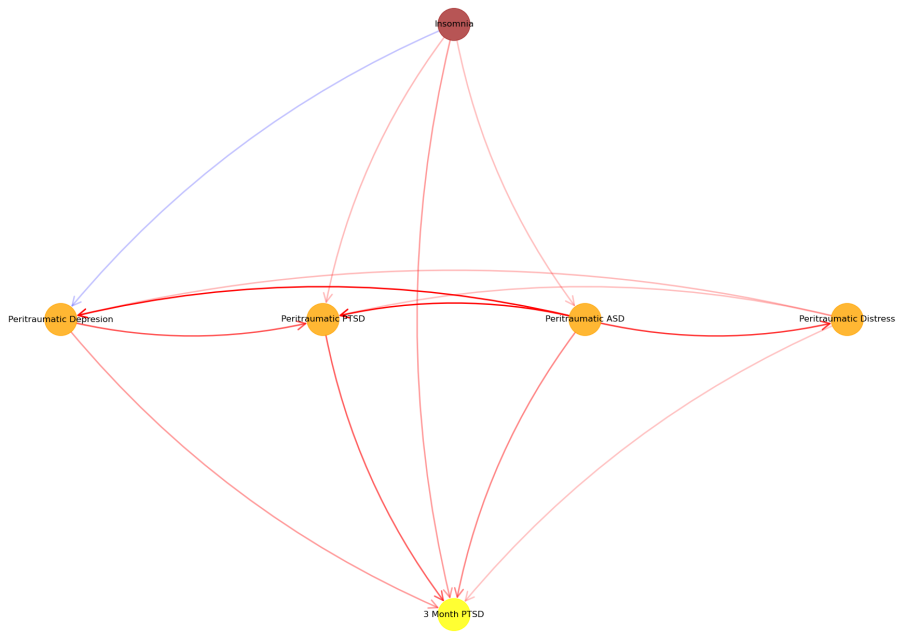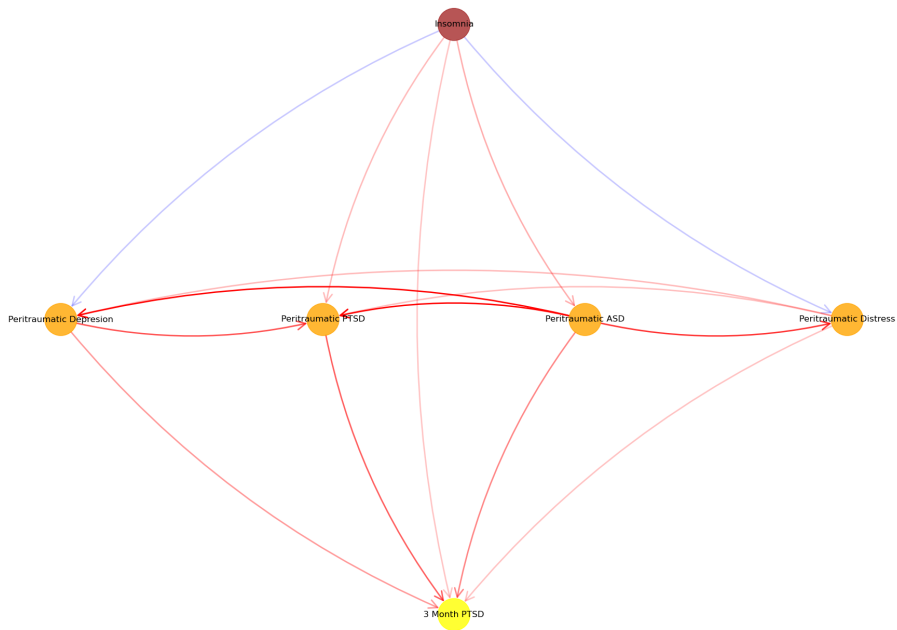
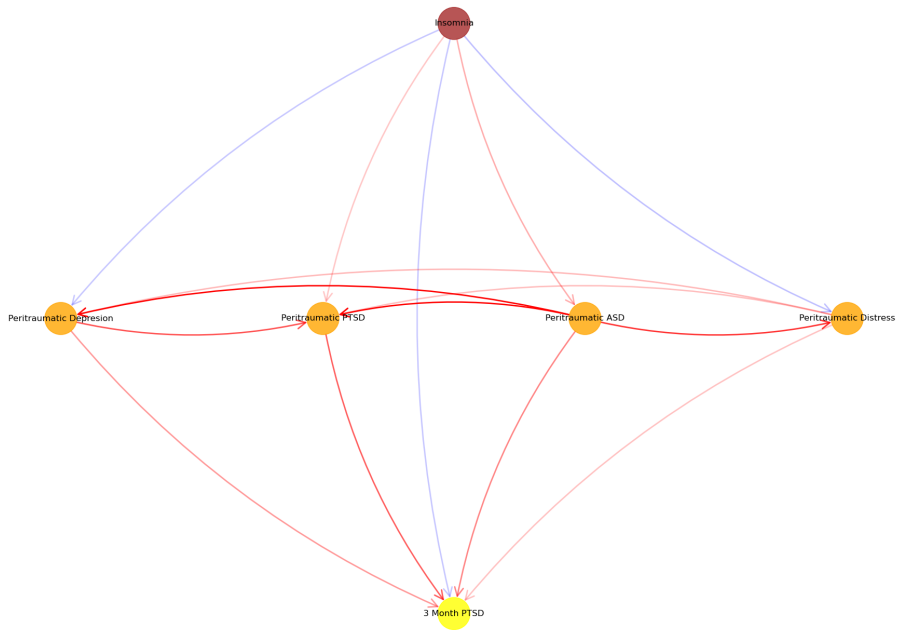(a) Estimated causal graph for the white male population.

(b) Estimated causal graph for the black male population.

Figure E.9: Causal graphs for the white and black male demographic. The red node is the event of interest. Orange nodes are mediators. The yellow node is the outcome.

(a) Estimated causal graph for the white female population.

(b) Estimated causal graph for the black female population.

Figure E.10: Causal graphs for the white and black female demographic. The red node is the event of interest. Orange nodes are mediators. The yellow node is the outcome.

F. Proofs

F.1. Proof of Theorem 4.1

Theorem F.1. *Under assumptions (A1-A3) and Model (1), we have:*

- $HDE(\boldsymbol{x}) = \gamma_A + \boldsymbol{\gamma_{XA}x}$;

- $HIE(\boldsymbol{x}) = \boldsymbol{\gamma_M}(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top x})$;

- $HTE(\boldsymbol{x}) = HDE(\boldsymbol{x}) + HIE(\boldsymbol{x})$,

where $\boldsymbol{I_s}$ is a $s \times s$ identity matrix and $\boldsymbol{x}$ is is the value of $\boldsymbol{X}$.

Proof. We begin with the systems of equations described by Model (1) as

$$\begin{cases} \boldsymbol{X} &= \boldsymbol{\epsilon_X}, \\ A &= \boldsymbol{\delta_X X} + \epsilon_A, \\ \boldsymbol{M} &= \boldsymbol{B_X^\top X} + \boldsymbol{\beta_A} A + \boldsymbol{B_{XA}^\top X} A + \boldsymbol{B_M^\top M} + \boldsymbol{\epsilon_M}, \\ Y &= \boldsymbol{\gamma_X X} + \gamma_A A + \boldsymbol{\gamma_{XA} x} A + \boldsymbol{\gamma_M M} + \epsilon_Y. \end{cases} \tag{F.1}$$

Under the assumptions, we have the following:

$$E[\boldsymbol{M}|do(A=a), \boldsymbol{X} = \boldsymbol{x}] = E[\boldsymbol{M}|A=a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A=a), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A=a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A=a, \boldsymbol{M} = \boldsymbol{m}), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A=a, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{X} = \boldsymbol{x}].$$

We can then use Equation F.1 to compute each of these expectations:

$$E[\boldsymbol{M}|A=a, \boldsymbol{X} = \boldsymbol{x}] = (\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top X} + \boldsymbol{\beta_A} A + \boldsymbol{B_{XA}^\top X} A), \tag{F.2}$$

$$\begin{aligned} E[Y|A=a, \boldsymbol{X} = \boldsymbol{x}] &= E[\boldsymbol{\gamma_X X} + \gamma_A A + \boldsymbol{\gamma_{XA} x} A + \boldsymbol{\gamma_M M} + \epsilon_Y | A=a, \boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M} E[\boldsymbol{M}|A=a, \boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M}(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta_A} a + \boldsymbol{B_{XA}^\top x} a), \end{aligned} \tag{F.3}$$

and

$$E[Y|A=a, \boldsymbol{M} = \boldsymbol{m}^{(a)}, \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M m}^{(a)}. \tag{F.4}$$

From Equations F.3 and F.4, we can compute the HDE and HIE as follows,

$$\begin{aligned} HDE &= E[Y|do(A=a+1, \boldsymbol{M} = \boldsymbol{m}^{(a)}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A=a)|\boldsymbol{X} = \boldsymbol{x}] \\ &= E[Y|A=a+1, \boldsymbol{M} = \boldsymbol{m}^{(a)}, \boldsymbol{X} = \boldsymbol{x}] - E[Y|A=a|\boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A(a+1) + \boldsymbol{\gamma_{XA} x}(a+1) + \boldsymbol{\gamma_M m}^{(a)} - (\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M m}^{(a)}) \\ &= \gamma_A + \boldsymbol{\gamma_{XA} x}, \end{aligned}$$

and

$$\begin{aligned} HIE &= E[Y|do(A=a, \boldsymbol{M} = \boldsymbol{m}^{(a+1)}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A=a)|\boldsymbol{X} = \boldsymbol{x}] \\ &= E[Y|A=a, \boldsymbol{M} = \boldsymbol{m}^{(a+1)}, \boldsymbol{X} = \boldsymbol{x}] - E[Y|A=a|\boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M m}^{(a+1)} - (\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M m}^{(a)}) \\ &= \boldsymbol{\gamma_M}\left(\left[(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta_A}(a+1) + \boldsymbol{B_{XA}^\top x}(a+1)) \right] \right. \\ &\quad \left. - \left[(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta_A} a + \boldsymbol{B_{XA}^\top x} a) \right] \right) \\ &= \boldsymbol{\gamma_M}(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top x}). \end{aligned}$$

And finally, we have the HTE as

$$HTE = HDE + HIE = \gamma_A + \gamma_{XA}x + \gamma_M(I_s - B_M^\top)^{-1}(\beta_A + B_{XA}^\top x).$$

$\square$

F.2. Proof of Theorem 4.2

Theorem F.2. *Under assumptions (A1-A3) and Model (1), we have:*

- $HDM_i(x) = \{\gamma_M\}_i\{(I_s - B_M^\top)^{-1}(\beta_A + B_{XA}^\top x)\}_i;$

- $\sum_{i=1}^s HDM_i(x) = HIE(x);$

- $HIM_i(x) = HTM_i(x) - HDM_i(x);$

- $HTM_i(x) = HIE(x) - HIE_{\mathbb{G}(-i)}(x)$

where $\{\cdot\}_i$ is the ith element of a vector and $HIE_{\mathbb{G}(-i)}$ is the HIE under the causal graph $\mathbb{G}(-i)$ in which the ith mediator is removed from the original causal graph $\mathbb{G}$.

Proof. We begin by demonstrating that the HTM_i can be decomposed into HDM_i and HIM_i using the definitions. As the second term in the product,

$$E\{M_i|do(A = a + 1), X = x\} - E\{M_i|do(A = a), X = x\},$$

is included in all mediation effects, it is only necessary to show that

$$E\{Y|do(M_i = m_i + 1), X = x\} - E\{Y|do(M_i = m_i), X = x\} \tag{F.5}$$

can be decomposed into

$$E[Y|do(A = a, M_i = m_i^{(a)} + 1, \Omega_i = o_i^{(a)}), X = x] - E[Y|do(A = a), X = x] \tag{F.6}$$

and

$$E[Y|do(A = a, M_i = m_i^{(a)} + 1), X = x] - E[Y|do(A = a, M_i = m_i^{(a)} + 1, \Omega_i = o_i^{(a)}), X = x]. \tag{F.7}$$

Adding F.6 and F.7, we have

$$E[Y|do(A = a, M_i = m_i^{(a)} + 1), X = x] - E[Y|do(A = a), X = x]$$

which can be shown is equivalent to F.5 on the interventional level in the following manner:

$$E\{Y|do(M_i = m_i + 1), X = x\} - E\{Y|do(M_i = m_i), X = x\}$$
$$= E\{Y|do(M_i = m_i^{(a)} + 1), X = x\} - E\{Y|do(M_i = m_i^{(a)}), X = x\}$$
$$= E\{Y|do(A = a, M_i = m_i^{(a)} + 1), X = x\} - E\{Y|do(A = a, M_i = m_i^{(a)}), X = x\}$$
$$= E[Y|do(A = a, M_i = m_i^{(a)} + 1), X = x] - E[Y|do(A = a), X = x],$$

where the first equality is true because the mediator value chosen, m_i, is arbitrary and the effect of increasing this value by a unit is assumed to be constant. The latter equalities are true by definition. Thus we have that

$$HTM_i(x) = HDM_i(x) + HIM_i(x). \tag{F.8}$$

This means that if we can compute the HDM for the ith mediator, then we can use the corresponding HTM to calculate the HIM for that mediator. From Equation F.2, we have

$$E\{M_i|do(A = a + 1), X = x\} - E\{M_i|do(A = a), X = x\}$$
$$= \{(I_s - B_M^\top)^{-1}(B_X^\top x + \beta_A(a + 1) + B_{XA}^\top x(a + 1))\}_i$$
$$\quad - \{(I_s - B_M^\top)^{-1}(B_X^\top x + \beta_A a + B_{XA}^\top xa)\}_i$$
$$= \{(I_s - B_M^\top)^{-1}(\beta_A + B_{XA}^\top x)\}_i.$$

Likewise, from Equation F.4, we have

$$E[Y|do(A = a, M_i = m_i^{(a)} + 1, \boldsymbol{\Omega}_i = \boldsymbol{o}_i^{(a)}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}]$$
$$= \big(\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M}(\boldsymbol{m}^{(a)} + \boldsymbol{e_i})\big) - \big(\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x} a + \boldsymbol{\gamma_M m}^{(a)}\big)$$
$$= \boldsymbol{\gamma_M e_i} = \{\boldsymbol{\gamma_M}\}_i,$$

where $\boldsymbol{e_i}$ is the ith basis vector in $\boldsymbol{R}^s$. Putting these all together, we can compute the HDM of the i-th mediator:

$$HDM_i(\boldsymbol{x}) = \Big[E[M_i|do(A = a+1), \boldsymbol{X} = \boldsymbol{x}] - E[M_i|do(A = a), \boldsymbol{X} = \boldsymbol{x}]\Big]$$
$$\times \Big[E[Y|do(A = a, M_i = m_i^{(a)} + 1, \boldsymbol{\Omega}_i = \boldsymbol{o}_i^{(a)}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}]\Big]$$
$$= \{\boldsymbol{\gamma_M}\}_i \times \{(\boldsymbol{I_s} - \boldsymbol{B_M}^\top)^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top x})\}_i,$$

It is then easy to see that

$$\sum_{i=1}^s HDM_i(\boldsymbol{x}) = \sum_{i=1}^s \{\boldsymbol{\gamma_M}\}_i \times \{(\boldsymbol{I_s} - \boldsymbol{B_M}^\top)^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top x})\}_i$$
$$= \boldsymbol{\gamma_M}(\boldsymbol{I_s} - \boldsymbol{B_M}^\top)^{-1}(\boldsymbol{\beta_A} + \boldsymbol{B_{XA}^\top x}) = HIE(\boldsymbol{x})$$

By definition we have that the HIM for the i-th mediator is

$$HIM_i(\boldsymbol{x}) = HTM_i(\boldsymbol{x}) - HDM_i(\boldsymbol{x}).$$

Thus, all that is left is the HTM_i which can be interpreted as the effect of the treatment A on the outcome Y that is mediated through mediator M_i, or inversely as the change in total treatment effect that is due to M_i being removed from the causal graph. Thus it can be calculated as follows,

$$HTM_i(\boldsymbol{x}) = HTE(\boldsymbol{x}) - HTE_{(i)}(\boldsymbol{x})$$
$$= \big\{HDE(\boldsymbol{x}) + HIE(\boldsymbol{x})\big\} - \big\{HDE_{\mathbb{G}(-i)}(\boldsymbol{x}) + HIE_{\mathbb{G}(-i)}(\boldsymbol{x})\big\}$$
$$= \big\{HDE(\boldsymbol{x}) - HDE_{\mathbb{G}(-i)}(\boldsymbol{x})\big\} + \big\{HIE(\boldsymbol{x}) - HIE_{\mathbb{G}(-i)}(\boldsymbol{x})\big\}$$
$$= \big\{(\gamma_A + \boldsymbol{\gamma_{XA} x}) - (\gamma_A + \boldsymbol{\gamma_{XA} x})\big\} + \big\{HIE(\boldsymbol{x}) - HIE_{\mathbb{G}(-i)}(\boldsymbol{x})\big\}$$
$$= HIE(\boldsymbol{x}) - HIE_{\mathbb{G}(-i)}(\boldsymbol{x}).$$

$\square$

F.3. Proof of Theorem 6.1

Proof. Similar to before, We begin with the systems of equations described Model 3

$$\begin{cases} \boldsymbol{X} &= \boldsymbol{\epsilon_X}, \\ A &= \boldsymbol{\delta_X f_A}(\boldsymbol{X}) + \epsilon_A, \\ \boldsymbol{M} &= \boldsymbol{B_X^\top f_{Mx}}(\boldsymbol{X}) + \boldsymbol{\beta_A^\top f_{Ma}}(A) + \boldsymbol{B_{XA}^\top f_{Mxa}}(\boldsymbol{X}, A) + \boldsymbol{B_M^\top M} + \boldsymbol{\epsilon_M}, \\ Y &= \boldsymbol{\gamma_X f_{Yx}}(\boldsymbol{X}) + \gamma_A f_{Ya}(A) + \boldsymbol{\gamma_{XA} f_{Yxa}}(\boldsymbol{X}, A) + \boldsymbol{\gamma_M f_{Ym}}(\boldsymbol{M}) + \epsilon_Y. \end{cases}$$

As before, Under the assumptions, we have the following:

$$E[\boldsymbol{M}|do(A = a), \boldsymbol{X} = \boldsymbol{x}] = E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A = a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A = a, \boldsymbol{M} = \boldsymbol{m}), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A = a, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{X} = \boldsymbol{x}].$$

Applying Model 3, we can compute each of these expectations:

$$E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}] = (\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}\big(\boldsymbol{B_X^\top f_{Mx}}(\boldsymbol{x}) + \boldsymbol{\beta_A^\top f_{Ma}}(a) + \boldsymbol{B_{XA}^\top f_{Mxa}}(\boldsymbol{x}, a)\big), \tag{F.9}$$

$$E[Y|A = a, \boldsymbol{X} = \boldsymbol{x}] = E[\boldsymbol{\gamma_X f_{Yx}}(\boldsymbol{X}) + \gamma_A f_{Ya}(A) + \boldsymbol{\gamma_{XA} f_{Yxa}}(\boldsymbol{X}, A) + \boldsymbol{\gamma_M f_{Ym}}(\boldsymbol{M}) + \epsilon_Y|A = a, \boldsymbol{X} = \boldsymbol{x}]$$
$$= \boldsymbol{\gamma_X f_{Yx}}(\boldsymbol{x}) + \gamma_A f_{Ya}(a) + \boldsymbol{\gamma_{XA} f_{Yxa}}(\boldsymbol{x}, a) + \boldsymbol{\gamma_M} E[\boldsymbol{f_{Ym}}(\boldsymbol{M})|A = a, \boldsymbol{X} = \boldsymbol{x}], \tag{F.10}$$

$$E[Y|A = a, \boldsymbol{M} = \boldsymbol{m}^{(a)}, \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{\gamma_X} \boldsymbol{f_{Yx}}(\boldsymbol{x}) + \gamma_A f_{Ya}(a) + \boldsymbol{\gamma_{XA}} f_{Yxa}(\boldsymbol{x}, a) + \boldsymbol{\gamma_M} \boldsymbol{f_{Ym}}(m^{(a)}). \qquad \text{(F.11)}$$

Where $\boldsymbol{m}^{(\boldsymbol{a})}$ is the value M would take given $A = a$. From Equations F.10 and F.11, we can compute the HDE and HIE as follows,

$$
\begin{aligned}
HDE(\boldsymbol{x}, a) &= \frac{\delta E[Y|do(A = a, \boldsymbol{M} = \boldsymbol{m}^{(a)}), \boldsymbol{X} = \boldsymbol{x}]}{\delta a} \\
&= \frac{\delta E[Y|A = a, \boldsymbol{M} = \boldsymbol{m}^{(a)}, \boldsymbol{X} = \boldsymbol{x}]}{\delta a} \\
&= \frac{\delta\left[\boldsymbol{\gamma_X} \boldsymbol{f_{Yx}}(\boldsymbol{x}) + \gamma_A f_{Ya}(a) + \boldsymbol{\gamma_{XA}} f_{Yxa}(\boldsymbol{x}, a) + \boldsymbol{\gamma_M} \boldsymbol{f_{Ym}}(m^{(a)})\right]}{\delta a} \\
&= \gamma_A \frac{\delta f_{Ya}(a)}{\delta a} + \boldsymbol{\gamma_{XA}} \frac{\delta f_{Yxa}(\boldsymbol{x}, a)}{\delta a},
\end{aligned}
$$

and

$$
\begin{aligned}
HIE(\boldsymbol{x}, a) &= \frac{\delta E[Y|do(A = a, \boldsymbol{M} = \boldsymbol{m}^{(a)}), \boldsymbol{X} = \boldsymbol{x}]}{\delta \boldsymbol{m}^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a} \\
&= \frac{\delta E[Y|A = a, \boldsymbol{M} = \boldsymbol{m}^{(a)}, \boldsymbol{X} = \boldsymbol{x}]}{\delta \boldsymbol{m}^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a} \\
&= \frac{\delta\left[\boldsymbol{\gamma_X} \boldsymbol{f_{Yx}}(\boldsymbol{x}) + \gamma_A f_{Ya}(a) + \boldsymbol{\gamma_{XA}} f_{Yxa}(\boldsymbol{x}, a) + \boldsymbol{\gamma_M} \boldsymbol{f_{Ym}}(m^{(a)})\right]}{\delta \boldsymbol{m}^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a} \\
&= \boldsymbol{\gamma_M} \frac{\delta \boldsymbol{f_{Ym}}(m^{(a)})}{\delta \boldsymbol{m}^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a}.
\end{aligned}
$$

where

$$\frac{\delta \boldsymbol{m}^{(a)}}{\delta a} = \frac{\delta E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}]}{\delta a} = (\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1} \frac{\delta[\boldsymbol{\beta_A^\top} \boldsymbol{f_{Ma}}(a) + \boldsymbol{B_{XA}^\top} \boldsymbol{f_{Mxa}}(\boldsymbol{x}, a)]}{\delta a}.$$

$\square$

F.4. Proof of Theorem 6.2

Proof. As it was shown in the proof of theorem 4.2 that the computation of the HIM and HTM is model-independent and does not have an explicit form, we only give the computation of the HDM here. From Equation F.9 and F.11 and Model 6.1, we have the following:

$$
\begin{aligned}
HDM_i(\boldsymbol{x}) &= \frac{\delta E[Y|do(A = a, M_i = m_i^{(a)}, \boldsymbol{\Omega}_i = \boldsymbol{o}_i^{(a)}), \boldsymbol{X} = \boldsymbol{x}]}{\delta m_i^{(a)}} \times \frac{\delta E[M_i|do(A = a), \boldsymbol{X} = \boldsymbol{x}]}{\delta a} \\
&= \frac{\delta E[Y|A = a, M_i = m_i^{(a)}, \boldsymbol{\Omega}_i = \boldsymbol{o}_i^{(a)}, \boldsymbol{X} = \boldsymbol{x}]}{\delta m_i^{(a)}} \times \frac{\delta E[M_i|A = a, \boldsymbol{X} = \boldsymbol{x}]}{\delta a} \\
&= \frac{\delta\left[\boldsymbol{\gamma_X} \boldsymbol{f_{Yx}}(\boldsymbol{x}) + \gamma_A f_{Ya}(a) + \boldsymbol{\gamma_{XA}} f_{Yxa}(\boldsymbol{x}, a) + \boldsymbol{\gamma_M} \boldsymbol{f_{Ym}}(\boldsymbol{m}^{(\boldsymbol{a})})\right]}{\delta m_i^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a} \\
&= \{\boldsymbol{\gamma_M}\}_i \frac{\delta E[\boldsymbol{f_{Ym}}(\boldsymbol{m}^{(\boldsymbol{a})})]}{\delta m_i^{(a)}} \frac{\delta \boldsymbol{m}^{(a)}}{\delta a}.
\end{aligned}
$$

$\square$

F.5. Proof of Theorem A.1

Proof. We begin with the systems of equations described in Section A.5:

$$\begin{cases} \boldsymbol{X} &= \boldsymbol{\epsilon_X}, \\ A &= \boldsymbol{\delta_X X} + \epsilon_A, \\ \boldsymbol{M} &= \boldsymbol{B_X^\top X} + \boldsymbol{\beta}_A A + \boldsymbol{B_{XA}^\top} XA + \boldsymbol{B_M^\top M} + \boldsymbol{\epsilon_M}, \\ Y &= \boldsymbol{\gamma_X X} + \gamma_A A + \boldsymbol{\gamma_{XA}} XA + \boldsymbol{\gamma_M M} + \boldsymbol{X}^\top \boldsymbol{\Gamma_{XM} M} + \epsilon_Y. \end{cases} \tag{F.12}$$

Noting that $\boldsymbol{X}^\top \boldsymbol{\Gamma_{XM} M} = vec(\boldsymbol{\Gamma_{XM}})^\top (\boldsymbol{X} \otimes \boldsymbol{M})$. Under the assumptions, we have the following:

$$E[\boldsymbol{M}|do(A = a), \boldsymbol{X} = \boldsymbol{x}] = E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A = a), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A = a, \boldsymbol{X} = \boldsymbol{x}],$$
$$E[Y|do(A = a, \boldsymbol{M} = \boldsymbol{m}), \boldsymbol{X} = \boldsymbol{x}] = E[Y|A = a, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{X} = \boldsymbol{x}].$$

We can then use Equation F.12 to compute each of these expectations:

$$E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}] = (\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta}_A a + \boldsymbol{B_{XA}^\top xa}), \tag{F.13}$$

$$\begin{aligned} E[Y|A = a, \boldsymbol{X} = \boldsymbol{x}] &= E[\boldsymbol{\gamma_X X} + \gamma_A A + \boldsymbol{\gamma_{XA} x}A + \boldsymbol{\gamma_M M} + \boldsymbol{X}^\top \boldsymbol{\Gamma_{XM} M} + \epsilon_Y|A = a, \boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})E[\boldsymbol{M}|A = a, \boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta}_A a + \boldsymbol{B_{XA}^\top xa}), \end{aligned} \tag{F.14}$$

and further

$$E[Y|A = a, \boldsymbol{M} = \boldsymbol{m^{(a)}}, \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\boldsymbol{m^{(a)}}. \tag{F.15}$$

Where $\boldsymbol{m^{(a)}}$ is the value M would take given $A = a$. From Equations F.14 and F.15, we can compute the HDE and HIE as follows,

$$\begin{aligned} HDE &= E[Y|do(A = a + 1, \boldsymbol{M} = \boldsymbol{m^{(a)}}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A = a)|\boldsymbol{X} = \boldsymbol{x}] \\ &= E[Y|A = a + 1, \boldsymbol{M} = \boldsymbol{m^{(a)}}, \boldsymbol{X} = \boldsymbol{x}] - E[Y|A = a|\boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A(a + 1) + \boldsymbol{\gamma_{XA} x}(a + 1) + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\boldsymbol{m^{(a)}} \\ &\quad - (\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\boldsymbol{m^{(a)}}) \\ &= \gamma_A + \boldsymbol{\gamma_{XA} x}, \end{aligned}$$

and

$$\begin{aligned} HIE &= E[Y|do(A = a, \boldsymbol{M} = \boldsymbol{m^{(a+1)}}), \boldsymbol{X} = \boldsymbol{x}] - E[Y|do(A = a)|\boldsymbol{X} = \boldsymbol{x}] \\ &= E[Y|A = a, \boldsymbol{M} = \boldsymbol{m^{(a+1)}}, \boldsymbol{X} = \boldsymbol{x}] - E[Y|A = a|\boldsymbol{X} = \boldsymbol{x}] \\ &= \boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\boldsymbol{m^{(a+1)}} - (\boldsymbol{\gamma_X x} + \gamma_A a + \boldsymbol{\gamma_{XA} x}a + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\boldsymbol{m^{(a)}}) \\ &= (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})\left(\left[(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta}_A(a + 1) + \boldsymbol{B_{XA}^\top x}(a + 1))\right] \\ &\quad - \left[(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{B_X^\top x} + \boldsymbol{\beta}_A a + \boldsymbol{B_{XA}^\top xa})\right]\right) \\ &= (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta}_A + \boldsymbol{B_{XA}^\top x}). \end{aligned}$$

And finally, we have the HTE as

$$HTE = HDE + HIE = \gamma_A + \boldsymbol{\gamma_{XA} x} + (\boldsymbol{\gamma_M} + \boldsymbol{x}^\top \boldsymbol{\Gamma_{XM}})(\boldsymbol{I_s} - \boldsymbol{B_M^\top})^{-1}(\boldsymbol{\beta}_A + \boldsymbol{B_{XA}^\top x}).$$

$\square$