A New Causal Decomposition Paradigm towards Health Equity

Xinwei Sun*
School of Data Science
Fudan University

Xiangyu Zheng
Department of Statistics, Guanghua School of Management
Peking University

Jim Weinstein*
Microsoft Corporation, Microsoft Research, Redmond, WA
Dartmouth, Tuck Business School
Northwestern University, Kellogg School of Business

Abstract

Causal decomposition has provided a powerful tool to analyze health disparity problems by assessing the proportion of disparity caused by each mediator (the variable that mediates the effect of the exposure on the health outcome). However, most of these methods lack *policy implications*, as they fail to account for all sources of disparities caused by the mediator. Besides, its identifiability needs to specify a set to be admissible to make the strong ignorability condition hold, which can be problematic as some variables in this set may induce new spurious features. To resolve these issues, under the framework of the structural causal model, we propose a new decomposition, dubbed as *adjusted* and *unadjusted* effects, which is able to include all types of disparity by adjusting each mediator's distribution from the disadvantaged group to the advantaged ones. Besides, by learning the *maximal ancestral graph* and implementing *causal discovery from heterogeneous data*, we can identify the admissible set, followed by an efficient algorithm for estimation. The theoretical correctness and efficacy of our method are demonstrated using a synthetic dataset and a common spine disease dataset.

1 Introduction

Health disparity/inequity, which is defined as the health outcome difference between socially advantaged and disadvantaged populations [Braveman et al., 2011], is a serious public health issue. To address this issue, a natural policy would be to identify all sources of disparities and develop policies to intervene and/or adjust for factors that mediate the effect of exposure (e.g., race or ethnicity) on the health outcome. To assess the effect of this policy, a natural question is: *what is the amount of disparity reduction, after making a policy that removes all disparities through the mediator of interest?*

To answer this and related questions, many causal decomposition methods have been proposed, including natural indirect effects PEARL [2001], Robins and Greenland [1992], controlled mediated effects VanderWeele [2011], path-specific effects Avin et al. [2005], Oaxaca-Blinder decomposition Oaxaca [1973], among others, and have been applied to many health inequity scenarios Ibfelt et al. [2013], Hystad et al. [2013], Jackson [2018], Blakely et al. [2018], McGuire et al. [2006]. Although the common target is quantitatively attributing the disparity to the mediator via causal inference, they adopt different intervention strategies. Particularly, the natural indirect effect intervenes on the mediator to the value it would have taken had the exposure variable changed to the advantaged group and everything else had been the same; while the controlled mediated effect intervenes on other mediators at a fixed value.

Different decomposition methods have different implications; however, when it comes to our goal, they either lack *policy implications* or have issues with *identifiability*. For instance, the natural indirect effect is more attributional than operational since it measures the effect under the original situation (*i.e.*, the value of exogenous is unchanged) instead of the effect of a policy under a new situation. Besides, its identifiability requires that the exposure not affect the mediator-outcome confounders. In many scenarios, however, this assumption may not hold (if $R \to X$ in Fig. 1). Furthermore, when this assumption is violated, the natu-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. * Correspond to sunxinwei@fudan.edu.cn, Jim.Weinstein@microsoft.com

ral indirect effect may fail to account for all sources of disparities (*e.g.*, $M' \perp\!\!\!\perp R$ when X is an unobserved selection variable in Fig. 1 (b)). Although Jackson [2020], VanderWeele et al. [2014] can alleviate this problem, their identifiability is based on the strong ignorability condition, which means admissibility for some provided covariate sets, *i.e.*, to deconfound between the mediator and the outcome. However, this condition may not hold when the provided covariate set includes some variables that perturb directed paths or induce new spurious features Pearl [2009].

Our Contributions. To resolve these issues, we propose a new decomposition paradigm under the framework of the structural causal model (SCM), namely *unadjusted and adjusted effects*, towards better policy implications and identifiability results, that account for social disparities.

To be specific, we adjust each mediator's distribution from the disadvantaged group to the advantaged one, by generating mediators with an independent copy of exogenous variables. This has better policy implications since **i)** it is more aligned with the goal of assessing the effectiveness of a policy, and **ii)** can include all sources of disparities. Moreover, guided by the back-door criterion, we are able to identify the admissible set for appropriate attribution, by leveraging Markovian equivalence conditions of maximal ancestral graphs (MAGs) and exploring the heterogeneity among different social groups. With this guarantee, we propose an efficient algorithm to estimate the adjusted and unadjusted effects of social inequities. Theoretical correctness and utility are demonstrated by the consistent estimation results on a synthetic and robust spine-disease dataset.

Organization. In Sec. 2, we review existing causal decomposition methods. In Sec. 3, we introduce some background knowledge regarding the structural causal model and causal discovery. In Sec. 4, we propose a new causal decomposition framework: *unadjusted* and *adjusted* effects. We will show that compared to the natural indirect effect, our proposed adjusted effect is able to include all sources of disparity, thus having better policy implications. In Sec. 5, we provide an identifiability analysis. Different from the strong ignorability condition that pre-specifies a covariate set to be admissible, we propose to identify the admissible set, via **i)** local causal discovery from heterogeneity data according to the exposure variable; and **ii)** exploring Markovian equivalence, which is more computationally efficient but requires additional conditions to determine equivalence. Guided by this analysis, we in Sec. 6 introduce our estimation method. In Sec. 7, we evaluate our method on a synthetic dataset and a spine-disease dataset. Sec. 8 concludes the paper.

2 Related Work

Existing methods for health inequity decompose the total amount of disparity into the proportion directly from exposure to the reported outcome, and the proportion indirectly through the mediated factor. Typical decomposition includes natural indirect effects, path-specific effects, controlled mediated effects, Oaxaca-Blinder decomposition, *etc.* Perhaps the most familiar method is natural indirect (mediation)/direct effects [PEARL, 2001, Vansteelandt and VanderWeele, 2012, Valeri and VanderWeele, 2013], which have been applied to many health inequity problems [Ibfelt et al., 2013, Hystad et al., 2013]. However, this effect is more attributional [PEARL, 2001] than operational since it assesses the change under the original situation, which may change at the time when the new policy is made. Besides, its identifiability requires no mediator-outcome confounder to be affected by the exposure, which may not be satisfied in health equity scenarios [VanderWeele et al., 2014]. Besides, when such confounders exist, the natural indirect effect may fail to account for all sources of disparities caused by the mediator of interest. Similarly, the controlled mediated effect, which assigns a fixed value to other mediated factors, is of less policy interest as it does not allow these factors to vary across individuals by default. The path-specific effect [Jackson, 2018, Avin et al., 2005, Gong and Zhu, 2021] measures the effect along some specific paths. As the intervention is implemented on the edge rather than the exposure or mediator variables, it has to integrate over the distribution of the exposure to obtain the path-specific effect; hence, it fails to measure the change after intervention (policy making). The Oaxaca-Blinder decomposition Oaxaca [1973], Quirk et al. [2006] linearly relates the outcome to other covariates and is later endowed with causal meaning [Jackson and VanderWeele, 2018] when the regression function refers to structural equations and the provided covariate set happens to deconfound the mediator and the outcome. Furthermore, the non-parametric form of this decomposition is proposed in Jackson and VanderWeele [2018]. Other types of decomposition include [Jackson, 2020, VanderWeele et al., 2014]. Similar to Oaxaca-Blinder decomposition, these methods also presume the admissibility of a provided covariate set (*i.e.*, strong ignorability), which can be problematic since the provided covariate set may induce new spurious features, violating the strong ignorability condition.

In contrast, our decomposition method, namely adjusted and unadjusted effects, can resolve the above defects in terms of *policy implications* and *identifiability*. Specifically, we define the adjusted effect as the difference between the advantaged and disadvantaged groups after equalizing the mediator's distribution, in the framework of structural causal models (SCMs) that serve an intuitive exhibition of the difference from other decomposition methods clearly. We show that our adjusted effect can well assess the disparity reduction after making an policy, by eliminating all sources of disparity. Besides, for identifiability, equipped with the back-door criterion in SCM, we establish the theory and the procedure to identify the admissible sets among covariates rather than presume by default that some pre-specified covariate sets are admissible. These results can extend to other

estimators, such as [Jackson, 2020] which lets the mediator additionally condition on other variables.

3 Preliminary

Structural Causal Model (SCM). The SCM is defined as $\mathcal{M} := \langle G, \mathcal{F}, \mathbf{U}, P(\mathbf{U})\rangle$. Here, the $G := (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph (DAG) that describes the causal structure over endogenous variables $\mathbf{V}$, where $\mathbf{E}$ denotes the edge set such that $X \to Y \in \mathbf{E}$ means a direct causal relation from X to Y. We say X causally influences Y if there is a directed path from X to Y. We refer to $\mathrm{An}(V), \mathrm{Pa}(V), \mathrm{De}(V), \mathrm{Ch}(V)$ as ancestral, parent, descendant, and child variable sets of V. The $\mathbf{U} := \{U_i\}$ denote exogenous variables, with $P(\mathbf{U})$ denoting its probability function. The $\mathcal{F} := \{f_{V_i}(Pa(i), U_i)\}_{V_i \in \mathbf{V}}$ denotes the set of structural equations. These structural equations are autonomous to each other, *i.e.*, breaking one equation will not affect others. Under the *Markovian* assumption that $\mathbf{U} := \{U_i\}$ are independent, we can factorize joint distribution into *disentangled* factors [Schölkopf et al., 2021], *i.e.*, $P(\mathbf{V}) := \Pi_{V_i \in \mathbf{V}} P(V_i|Pa(i))$. These permit us to define the interventional distribution, *i.e.*, "$P(\mathbf{O}|do(O_i = o_i))$", and study the (conditional) average causal effect (ACE): $\mathbb{E}(Y|do(T = 1), X = x) - \mathbb{E}(Y|do(T = 0), X = x)$. Graphically speaking, $do(O = o)$ means removing arrows into O and setting its value to o. $G_{\overline{V}}$ and $G_{\underline{V}}$ respectively denote the graph with arrows going into and emanating from V deleted. With the SCM, we further define the submodel $\mathcal{M}_x$ as $\mathcal{M}$ with $X \leftarrow f_X(Pa(X), U_x)$ replaced by $X = x$. The counterfactual $Y(X = x)$ is then defined as $Y(X = x)(\mathbf{U} = u) := Y_{\mathcal{M}_x}(\mathbf{U} = u)$.

The SCM translates the strong ignorability assumption, *i.e.*, $\{Y(1), Y(0)\} \perp\!\!\!\perp T|Z, X$ into a more intuitive graphical criteria [Galles and Pearl, 2013] for Z to be admissible: $P(Y|do(T = t), X) = \int P(Y|T = t, Z, X)P(Z|X)dZ$. Among these criteria, the back-door criterion is the most typical one, which refers to a subset that **i)** blocks each back-door path from T to Y and **ii)** meanwhile does not perturb directed paths or induce new spurious features. This criterion, especially **ii)**, informs us that the presumption of admissibility for some covariate sets may be problematic when it includes variables that violate **ii)**.

Causal Discovery. Equipped with the SCM that converts the strong ignorability condition into graphical criteria, the admissible sets can be easily detected via causal discovery. Typical methods for causal discovery include the PC algorithm, which first learns an undirected skeleton via conditional independence test and identifies as many directions as possible via v-structure, *e.g.*, $V_i \to V_k \leftarrow V_j$.

Maximal Ancestral Graph. Indeed, when there are unobserved variables, one can learn a *maximal ancestral graph* (MAG) over a subset node [Zhang and Spirtes, 2005], which is a mixed graph that preserves the conditional independence relation in the sense that two variables are not adjacent if and only if there is no inducing paths[1] between them. Specifically, the MAG may contain: 1) $X \to Y$, 2) $X \leftarrow Y$ and 3) $X \leftrightarrow Y$, which respectively mean 1) $X \in \mathrm{An}(Y)$, 2) $Y \in \mathrm{An}(X)$ and 3) X and Y is correlated but $X \notin \mathrm{An}(Y)$, $Y \notin \mathrm{An}(X)$. We define $\circ$ that can refer to the head or tail of an arrow. Zhang and Spirtes [2005], Spirtes and Richardson [1996] provided conditions for two MAGs to be *Markovian equivalent* (in Prop. 3.1), which can help to determine causal directions that differentiate among classes of MAGs.

Proposition 3.1 (Proposition 1 in [Spirtes and Richardson, 1996])**.** *Two MAGs are Markovian equivalent iff* ***i)*** *they have the same adjacencies;* ***ii)*** *they have the same unshielded colliders;* ***iii)*** *for any discriminating path of a vertex V, V is a collider on the path in one graph iff it is a collider on the path in the other. Here, B is an unshielded collider if in (A, B, C), $A \circ\!\!\to B \leftarrow\!\!\circ C$. A discriminating path $u := \langle X, W_1, ..., W_K, V, Y\rangle$ means u includes three edges, with X not adjacent to Y, V adjacent to Y, and any W_k (for $k \in \{1, ..., K\}$) is a collider and a parent of Y.*

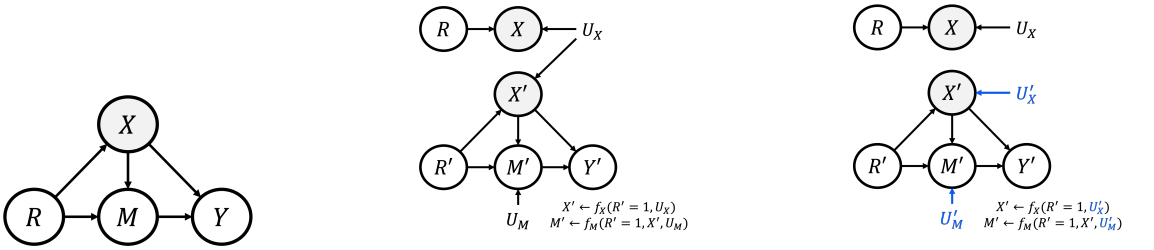
Identifying Directions from Heterogeneity. Recently, Huang et al. [2020] proposed to leverage domain index C to identify directions (beyond v-structure) among variables with changed causal mechanisms, *i.e.*, $P^c(V_i|Pa(i))$ changes across C. This can be achieved under a "faithfulness" condition at distributional level, *i.e.*, if $C \to V_i$, $C \to V_j$ and $V_i \perp\!\!\!\perp V_j|C, X$ for some X, then $\{P(V_i|X, C = c\}$ is independent of $\{P(V_j|V_i, X, C = c)\}$ due to disentanglement of generating factors $\{P(V_i|Pa(i))\}$; while $\{P(V_j|X, C = c)\}$ and $\{P(V_i|V_j, X, C = c)\}$ are dependent. One can thus determine $V_i \to V_j$ (*resp.* $V_j \to V_i$) if the Hilbert Schmidt Independence Criterion (HSIC) [Gretton et al., 2007] $\Delta_{V_i \to V_j|X} < \alpha$ (*resp.*, $\Delta_{V_j \to V_i|X} < \alpha$) for some significance level $\alpha > 0$.

4 Adjusted and Unadjusted Effect

Problem Setup. Our data consists of $\{r^{(i)}, o^{(i)}, y^{(i)}\}_{i=1}^n \sim_{i.i.d} P(R, \mathbf{O}, Y)$, where R denotes the exposure variable with $R = 1$ (*resp.* $R = 0$) denoting the advantaged (*resp.* disadvantaged) group, Y denotes the health outcome, and $\mathbf{O}$ denotes the observed covariate set. Denote $Y(R = r)$ as the counterfactual. The health inequity means $\mathbb{E}(Y(R = 1)) - \mathbb{E}(Y(R = 0)) > 0$. Our goal is to assess the disparity reduction to measure the effect of making policy on some mediator variables $M \in \mathbf{M} \subset \mathbf{O}$. Here, we call M a mediator if there exists a directed causal path from R to Y that goes through M.

In this section, we introduce a new decomposition method: the *unadjusted* and *adjusted* effects. Formally speaking, $\forall\, M \in \mathbf{M}$, we define the adjusted effect δ_M and the unad-

[1] We leave the definition of inducing path to the supplementaries



(a) Original causal graph (b) Mediation $M(R'=1)(\mathbf{u})$ (c) Adjusted (Ours) $M_{R'=1}(\mathbf{u}')$

Figure 1: R, M respectively denote exposure and mediator variables; X is unobserved; V' is a copy of V after making policy, generated by another group $R' \neq R$. Unlike the *Mediation Effect* with original exogenous $\mathbf{U}$; our *Adjusted Effect* uses $\mathbf{U}' \sim_{i.i.d} \mathbf{U}$ to generate M' (Eq. (5)), which can effectively account for all sources of disparities.

justed effect ζ_M as:

$$\delta_M := \mathbb{E}(Y(R=0, G_{M|R=1})) - \mathbb{E}(Y(R=0)), \quad (1)$$
$$\zeta_M := \mathbb{E}(Y(R=1)) - \mathbb{E}(Y(R=0, G_{M|R=1})), \quad (2)$$

where $G_{M|R=1}$ means to replace the original factor $P(M|Pa(M))$ with $P(M|R=1)$, *i.e.*, assigning the mediator M with the distribution of the advantaged group $M|R=1$. We use V'_i to denote the copy of V_i generated by the new policy from $R' \neq R$, in order to differentiate the original V_i. Then, correspondingly, structural equations for $Y(R=0, G_{M|R=1})$ are:

$$R \leftarrow 0, \quad (3)$$
$$V_i \leftarrow f_i(Pa(V_i), u_i),\ \forall V_i \in \mathbf{V} \backslash \mathrm{De}(M), \quad (4)$$
$$\mathbf{G_{M|R=1}}:\ \mathbf{M} \leftarrow \mathbf{M_{R=1}}(\mathbf{u}'),\ \mathbf{u}' \sim P(\mathbf{U}), \quad (5)$$
$$V_i \leftarrow f_i(Pa(V_i) \backslash \mathbf{M}, \mathbf{M}, u_i),\ \forall V_i \in \mathrm{Ch}(\mathbf{M}), \quad (6)$$
$$V_i \leftarrow f_i(Pa(V_i), u_i),\ \forall V_i \in \mathrm{De}(\mathbf{M}) \backslash \mathrm{Ch}(\mathbf{M}), \quad (7)$$

Here, Eq. 3, 4, (6), (7) describe the original generating process under $\mathbf{u}$; while Eq. 5 describes the regenerating process of M under the intervened model $M_{R=1}$ with $\mathbf{u}' \sim P(\mathbf{U})$ that is independent to $\mathbf{u}$. Finally, we regenerate M's children and descendants using Eq. 6, (7). Simply speaking, the adjusted effect replaces $M \leftarrow f_M(R=0, Pa(M) \setminus R, u_M)$ in the original SCM with an intervened policy on M: $M_{R=1}(\mathbf{u}' \sim P(\mathbf{U}))$ that equalizes the distribution of M between advantaged and disadvantaged groups.

Compared with the natural indirect effect with $Y(R=0, M(R=1))$, our adjusted effect generates M using another exogenous $\mathbf{U}' \sim_{i.i.d} \mathbf{U}$, which has better policy implications in terms of measuring the effect of making policy in a new situation. Besides, it can effectively account for all sources of disparity. To illustrate, consider the example in Fig. 1:

Example 4.1. *Suppose R, X (unobserved), M, Y in Fig. 1 (a) respectively denote the race, retirement income, medical insurance, and health outcome. The disadvantaged group has a lower level of retirement income and thus a lower level of medical insurance. As illustrated in Fig. 1 (b), the structural equations for natural indirect effect $Y(R=0, M(R=1))$ are:*

$$\begin{cases} R \leftarrow 0, \\ X \leftarrow f_X(R=0, U_X), \\ \mathbf{X}' \leftarrow \mathbf{f_X}(\mathbf{R}'=\mathbf{1}, \mathbf{U_X}), \mathbf{M}' \leftarrow \mathbf{f_M}(\mathbf{X}', \mathbf{R}'=\mathbf{1}, \mathbf{U_M}), \\ Y \leftarrow f_Y(X, M', U_Y), \end{cases}$$
SCM of $Y(R=0, M(R=1))$

in which U_X is the confounder between X and X' to generate M'. When X is unobserved, M' may not eliminate all sources of disparities, i.e., *$M' \not\perp\!\!\!\perp R$, either when X is the selection or latent variable. Specifically, when X is taken as the selection variable, then as the collider between the path from $R=0$ to M' (in Fig. 1 (b)) it induces the correlation between $R=0$ and M', which means M' still mixes the information of $R=0$. When X is the latent variable, it can be mistakenly taken as exogenous of M,* i.e., *$X \in \mathbf{U}_M$ since the condition that whether there is another variable affecting X is unknown. Since the exogenous is unchanged in natural indirect effect, $M' \leftarrow f_M(R'=1, X, \mathbf{U}_M \backslash X)$ is generated with X that is generated from $R=0$. In this case, the $M(R=1)(\mathbf{u})$ of the natural indirect effect fails to account for the disparity in the path of $R \rightarrow X \rightarrow M$.*

$$\begin{cases} R \leftarrow 0, \\ X \leftarrow f_X(R=0, U_X), \\ \mathbf{X}' \leftarrow \mathbf{f_X}(\mathbf{R}=\mathbf{1}, \tilde{\mathbf{U}}_\mathbf{X}), \mathbf{M} \leftarrow \mathbf{f_M}(\mathbf{X}', \mathbf{R}=\mathbf{1}, \mathbf{U_M}), \\ Y \leftarrow f_Y(X, M, U_Y). \end{cases}$$
SCM of $Y(R=0, G_{M|R=1})$

In contrast, our adjusted effect with $Y(R=0, G_{M|R'=1})$ is able to remove all disparities sourced from medical insurance M, i.e., *$M_{R'=1} \perp\!\!\!\perp R$, since it generates M' with independent exogenous $\mathbf{U}'$.*

In addition to policy implications, our adjusted effect requires weaker conditions for identifiability, compared to the natural indirect effect. Specifically, the natural indirect effect requires the deconfounding set between M and Y

(given R) to not contain the descendants of R. This assumption may not hold in our health equity scenario, as shown by example 4.1. In contrast, our $Y(R=0, G_{M|R=1})$ only requires no existence of unobserved confounders between M and Y, which is commonly made in the literature, including the natural indirect effect and other decomposition methods.

Indeed, our definition is closely related to others in the literature. Specifically, when the linear equations refer to structural equations, our $Y(R=0, G_{M|R=1})$ degenerates to Oaxaca–Blinder decomposition [Oaxaca, 1973, Quirk et al., 2006]. Besides, it has the same form as randomized intervention effects [VanderWeele et al., 2014, Jackson and VanderWeele, 2018], if the covariate set provided in their works happens to be admissible. Further, our definition can be easily extended to [Jackson, 2020] that lets M and Y additionally condition on allowable variables. *More importantly*, unlike these methods that assume without checking that the covariate set $\mathbf{O}$ happens to be admissible, our adjusted effect is endowed with a method that is provable to identify the admissible set among $\mathbf{O}$.

5 Identifiability

In this section, we discuss the identifiability property of δ_M. We first give some basic assumptions.

Assumption 5.1 (SCM with no Unobserved Confounders)**.** *We assume that $\langle G, \mathcal{F}, \mathbf{U}, P(\mathbf{U})\rangle$ is a structural causal model, with $G := (\mathbf{V}, \mathbf{E})$ and $\mathbf{O} \cup \{Y, R\} \subset \mathbf{V}$. In addition, we assume for each $M \in \mathbf{M}$, there is no unobserved confounder between M and Y.*

Note that we do not assume all variables in $\mathbf{V}$ are observable; rather, we only observe a subset of them $\{\mathbf{O}, R, Y\} \subset \mathbf{V}$. The "no unobserved confounder" condition, which does not additionally pre-specifies some covariate sets to be admissible, is weaker than the strong ignorability condition that is widely assumed in the literature of causal inference [Pearl, 2009, Spirtes et al., 2000] and other causal decomposition works Jackson [2020], VanderWeele et al. [2014]. Furthermore, we will show later that this assumption is necessary for the adjusted effect to be identifiable.

In the following, we define the *context variable*, which refers to the variable with no arrows going into them.

Definition 5.1 (Context Variable)**.** We say $\mathcal{C} \subset \mathbf{O} \cup R$ is the set of *context variables* if any $C \in \mathcal{C}$ is not causally affected by any variable in $\mathbf{V}$.

Remark. It naturally holds that R is a context variable, *i.e.*, $R \to \mathcal{C}$ since the exposure variable normally refers to the innate attributes. Similarly, other demographic variables such as gender, and ethnicity are also context variables.

Besides, as Y refers to the outcome that is temporally reported later than others, it naturally holds that Y does not causally influence other variables. Then we have $\delta_M = \int yp(y|do(m), R=0)p(m|R=1)dydm - \int yp(y|R=0)dy$, due to $(R \perp\!\!\!\perp Y)_{G_{\overline{M},\underline{R}}}$, $(R \perp\!\!\!\perp Y)_{G_{\underline{R}}}$, and $(R \perp\!\!\!\perp M)_{G_{\underline{R}}}$. To identify δ_M, all that is left is to identify $p(y|do(m), r)$.

Next, we introduce Markovian and Faithfulness conditions.

Assumption 5.2 (Markovian and Faithfulness)**.** *The* Markovian *means that the exogenous variables $\{U_i\}$ are independent. This implies $\forall$ disjoint sets $\mathbf{V}_i, \mathbf{V}_j, \mathbf{V}_k$, it holds that $\mathbf{V}_i \perp\!\!\!\perp_d \mathbf{V}_j|\mathbf{V}_k \implies \mathbf{V}_i \perp\!\!\!\perp \mathbf{V}_j|\mathbf{V}_k$. On the other hand,* faithfulness *means $\mathbf{V}_i \perp\!\!\!\perp \mathbf{V}_j|\mathbf{V}_k \implies \mathbf{V}_i \perp\!\!\!\perp_d \mathbf{V}_j|\mathbf{V}_k$, where $\perp\!\!\!\perp_d$ and $\perp\!\!\!\perp$ respectively mean d-separation and probability independence.*

This assumption, as was commonly made in the causal inference literature, allows us to implement conditional independence tests for causal discovery, which can provide a graphical criterion to identify $p(y|do(m), r)$. In the following, we show that it is almost necessary to assume no unobserved confounder in assump. 5.1 to identify $p(y|do(m), r)$.

Proposition 5.1. *Suppose assumption 5.2 holds in the SCM $\langle G, \mathcal{F}, \mathbf{U}, P(\mathbf{U})\rangle$, we have that the $p(y|do(m), r)$ is identifiable for all forms of $P(\mathbf{U})$ and $\mathcal{F}$, if and only if there exists an admissible set $\mathbf{B}_M$,* i.e.*, $(M \perp\!\!\!\perp Y|\mathbf{B}_M, R)_{G_{\underline{M}}}$.*

Remark. If there exists an unobserved confounder such that its path to Y or M is not blocked by $\mathbf{B}_M$, then $p(y|do(m), r)$ cannot be identified generally. Note that under the linear model, [Kasy, 2014, Frölich and Huber, 2017] leverages instrumental variables for identifiability with unobserved confounders, however, assumed that $M(Z=z, u_M)$ increases with respect to the instrumental variable Z for all individuals u_M. This condition may not hold in our health equity scenario. For example, if $Z := R$ denotes race and M denotes the income level, then it is unreasonable to assume that any person in the advantaged group has a higher income level than all persons in $R=0$.

The Prop. 5.1 informs us to identify an admissible set, of which each O_i is selected according to the type of the maximal ancestral graph (MAG) over (O_i, M, Y). Specifically, we first list a broader family of types of MAGs in Fig. 2, which are equivalent according to Prop. 3.1. We show that any set that contains variables belonging to type (a) and does not contain any variables in types (d,e,f) is admissible. To explain, note that the variable with the type (a) is sufficient to block each back-door path between M and Y as there is no unobserved confounder between M and Y. Besides, the variable with types (b) and (c) does not perturb directed paths or induce new spurious features. Therefore, the inclusion of these variables can still make the set admissible. Types (d,e,f) are other members of the equivalence class.

Formally speaking, we define

$$\underline{\mathbf{B}}_M := \{O_i : \mathrm{MAG}(O_i, M, Y) \text{ is (a).}\},$$
$$\overline{\mathbf{B}}_M := \{O_i : \mathrm{MAG}(O_i, M, Y) \text{ belongs to (a,b,c).}\},$$

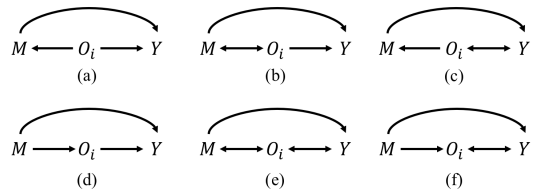


Figure 2: MAG of admissible sets: (a,b,c) and their equivalence classes: (d,e,f).

and show that any set that contains $\underline{\mathbf{B}}_M$ and belongs to $\overline{\mathbf{B}}_M$ is an admissible set:

Theorem 5.1. *Under assump. 5.1, 5.2, any* $\mathbf{B}_M$ *such that* $\underline{\mathbf{B}}_M \subset \mathbf{B}_M \subset \overline{\mathbf{B}}_M$ *is admissible.*

According to Thm. 5.1, it is sufficient to identify any intermediate sets between $\underline{\mathbf{B}}_M$ and $\overline{\mathbf{B}}_M$. To achieve this goal, we first iteratively learn $\mathrm{MAG}(O_i, M, Y)$ for each $O_i \in \mathbf{O}$ via the FCI-JCI algorithm, which is provable to be sound and complete Zhang [2008b] to identify the equivalence class of MAG. As members (d),(e),(f) are equivalent to types (a),(b),(c), we need to discriminate variables with types (a-c) from those with types (d-f).

To achieve this goal, we propose to leverage the context variables in Def. 5.1. Specifically, inspired by conditions (ii),(iii) in Prop. 3.1, we introduce the *unshielded*-condition and the *discriminate*-condition on the context variable:

Definition 5.2. For the mediator M, we say $\mathcal{C}$ satisfies the **MAG-Equi** condition for O_i and M, if either one of the following conditions holds:

- *Unshielded*-condition: there exists $C_M^i \in \mathcal{C}$ such that $\mathrm{MAG}(C_M^i, O_i, M)$ is $C_M^i \to O_i \circ\!\!-\!\!\circ M$;
- *Discriminate*-condition: there exists a pair of context variables $(C_M^{i,1}, C_M^{i,2})$ such that **i)** $\mathrm{MAG}(C_M^{i,1}, O_i, M)$ is $C_M^{i,1} \to M \circ\!\!-\!\!\circ O_i$; and **ii)** $C_M^{i,2}$ is adjacent to M and not adjacent to Y in $\mathrm{MAG}(C_M^{i,2}, O_i, M, Y)$.

As the name suggests, the *unshielded*-condition and *discriminate*-condition respectively exploit the unshielded collider criterion and the discriminating path criterion in Prop. 3.1 to identify different classes of MAGs, with the assistance of context variable. In the following, we will show that if $\mathcal{C}$ satisfies the **MAG-Equi** condition, we can discriminate types (a,b,c) from (d,e,f) in Fig. 2.

Identify $\overline{B}_M$ when the MAG-Equi condition holds. The following lemma summarizes this result.

Lemma 5.1. *Suppose assumptions 5.1, 5.2 hold and* $\mathcal{C}$ *satisfies the* ***MAG-Equi*** *condition. Then if* $\mathcal{C}$ *satisfies the* ***MAG-Equi*** *condition, we can discriminate types (a,b,c) from (d,e,f) in Fig. 2. That is, we can determine whether* $O_i \in \overline{\mathbf{B}}_M$.

Proof-sketch. Indeed, both *unshielded*-condition and *discriminate*-condition utilize "unshielded colliders" criterion in Prop. 3.1. Specifically, $C_M^i \to O_i \circ\!\!-\!\!\circ M$ eliminates types (d,e,f) with $M \circ\!\!\to O_i$. While for *discriminate*-condition, the $C_M^{i,1} \to M \circ\!\!-\!\!\circ O_i$ first eliminates types (d,f) with $M \leftarrow\!\!\circ O_i$, in which O_i is an unshielded collider. Further, the MAG structure over $(C_M^{i,2}, M, O_i, Y)$ explores the discriminating path $\langle C_M^{i,2}, M, O_i, Y\rangle$ for O_i to eliminate type (e) with $M \leftrightarrow O_i \leftrightarrow Y$. Please refer to the appendix for details. □

The **MAG-Equi** condition requires the existence of a context variable that does not simultaneously affect O_i and M. This condition can hold for some O_i when there are multiple context variables, such as gender, ethnicity, race, age, religious affiliation, and marital status. For instance, if O_i, M, Y respectively denote education level, working style and treatment outcome, it is reasonable and hence probable for "religious affiliation" to determine the working style M but not affect the education level O_i and Y. In this regard, the *discriminate*-condition is satisfied when $C_M^{i,1} = C_M^{i,2}$ and denotes "religious affiliation". Similarly, the *unshielded*-the condition can be satisfied when C_M^i denotes "race/ethnicity" as it may not affect the working style M given the education level O_i.

However, this **MAG-Equi** condition may uniformly hold for all O_i, especially when there are few context variables. In this case, lemma 5.1 cannot be applied to determine whether $O_i \in \overline{\mathbf{B}}_M$. To address this problem, inspired by Thm. 5.1, we turn to discriminate type (a) from others, *i.e.*, whether $O_i \in \underline{\mathbf{B}}_M$, by leveraging heterogeneity induced by the exposure variable R to identify the causal directions $O_i \to M, Y$ that characterizes type (a).

Identify $\underline{\mathbf{B}}_M$ when the MAG-Equi condition is violated. We split data into multiple domains according to the value of R. The heterogeneity among these domains can help us to determine causal directions by exploiting changed causal modules Huang et al. [2020]. This requires the following assumption, namely *faithfulness* at the distributional level, which was firstly proposed by Huang et al. [2020].

Assumption 5.3 ([Huang et al., 2020])**.** *If* $O_i \to O_j$ *and at least* $R \to O_i$ *or* $R \to O_j$ *holds, then* $\{P^r(O_i|O_j, \mathbf{X})\}_r$ *is dependent to* $\{P^r(O_j|\mathbf{X})\}_r$*, where* $\mathbf{X}$ *is the minimal deconfounding set* [2].

Under this assumption, we have the following result:

Lemma 5.2. *Under assump. 5.1-5.3, we have* $O_i \in \underline{\mathbf{B}}_M$ *iff there exists* $\mathbf{X}_M^i$ *such that* ***i)*** $\{P(M, Y|O_i, \mathbf{X}_M^i, r)\}_r$ *and* $\{P(O_i|\mathbf{X}_M^i, r)\}_r$ *are independent;* ***ii)*** $\{P(M|\mathbf{X}_M^i, r)\}_r$ *and* $\{P(O_i|M, \mathbf{X}_M^i, r)\}_r$ *are dependent; and* ***iii)*** $\{P(Y|\mathbf{X}_M^i, r)\}_r$ *and* $\{P(O_i|Y, \mathbf{X}_M^i, r)\}_r$ *are dependent.*

To implement these independence tests, we correspondingly examine whether $\Delta_{O_i, Y \leftarrow M|\mathbf{X}_M^i} < \alpha$, $\Delta_{O_i \to M|\mathbf{X}_M^i} \geq \alpha$ and

[2] X is a deconfounding set between O_i and O_j if $O_i \perp\!\!\!\perp O_j|X$ and $X \cap (\mathrm{De}(O_i) \cup \mathrm{De}(O_j)) = \emptyset$.

$\Delta_{Y\to M|\mathbf{X}_M^i} \geq \alpha$ with some $\alpha > 0$ to respectively determine whether **i)**, **ii)** and **iii)** hold, derived from Hilbert-Schmidt Independence Criterion (HSIC) norm Gretton et al. [2007] to measure the dependency between two sets of distributions Huang et al. [2020]. Compared to lemma 5.1, the lemma 5.2 does not require the **MAG-Equi** condition, however, is more computationally expensive as it requires searching over all subsets of all O_i with its type of MAGs belonging to (a-f) in Fig. 2, *i.e.*, $\mathbf{B}_M^{\text{all}} := \{O_i : \text{MAG}(O_i, M, Y) \text{ belongs to (a-f).}\}$. We summarize lemma 5.1, 5.2 into the following theorem.

Theorem 5.2 (Main Theorem)**.** *Under assump. 5.1-5.3, there exists an identifiable admissible set* $\mathbf{B}_M$ *with* $\underline{\mathbf{B}}_M \subset \mathbf{B}_M \subset \overline{\mathbf{B}}_M$*. Particularly,* $\mathbf{B}_M$ *respectively degenerates to* $\overline{\mathbf{B}}_M$ *and* $\underline{\mathbf{B}}_M$ *if the* ***MAG-Equi*** *condition in lemma 5.1 holds and does not hold for all* $O_i \in \mathbf{O}$*.*

Remark. This analysis can easily extend to the form in [Jackson, 2020] that additionally conditions on some "allowable variables". As these variables are assumed to not be affected by R and M, conditioning on these variables will not induce new spurious features or block directed paths.

With Thm. 5.2, we can identify the adjusted effect δ_M and unadjusted effect ζ_M via $\mathbf{B}_M$ as:

$$\delta_M = \mathbb{E}_{m|R=1}\left[\mathbb{E}_{\mathbf{b}_M|R=0}\left[\mathbb{E}[Y|m, \mathbf{b}_M, R=0]\right]\right] - \mathbb{E}^0[Y],$$
$$\zeta_M = \mathbb{E}^1[Y] - \mathbb{E}_{m|R=1}\left[\mathbb{E}_{\mathbf{b}_M|R=0}\left[\mathbb{E}[Y|m, \mathbf{b}_M, R=0]\right]\right],$$

where $\mathbb{E}^r[Y]$ denotes $\mathbb{E}[Y|R=r]$ for $r = 0, 1$.

6 Estimation

Algorithm 1 summarizes the whole procedure to calculate δ_M, ζ_M for each M. Roughly speaking, it contains **i)** detect all mediators $\mathbf{M}$; then for each $M \in \mathbf{M}$, **ii)** identify $\mathbf{B}_M$ and **iii)** estimate δ_M, ζ_M by estimating $\mathbb{E}(Y|m, \mathbf{b}_M, R=0)$ and $p(\mathbf{b}_M|R=r)$.

For $\mathbb{E}(Y|m, \mathbf{b}_M, R=0)$, we can choose several off-the-shelf predictors such as tree-based or generalized linear models for fitting. To estimate $p(\mathbf{b}_M|R=r)$ with $\mathbf{B}_M = \{B_i\}_{i=1}^s$, we write $p(\mathbf{b}_M|R=r) = \Pi_{i=1}^s p(b_i|b_{1:i-1}, R=r)$ and implement kernel density estimation (KDE) to estimate $p(b_i|b_{1:i-1}, R=r)$ if b_i is continuous and use softmax if b_i is discrete. If s is large, we use Markov chain Monte Carlo to generate data from $p(\mathbf{b}_M|R=r)$.

Computational Complexity. Under the worst case, *i.e.*, the **MAG-Equi** condition is violated for all O_i, one has to search over $S \subset \mathbf{B}_M^{\text{all}}$. If we follow an ascending order in terms of $|S|$, the overall complexity to estimate δ_M, ζ_M for each M is with order $(p * |\mathcal{C}|)^3 (\mathbf{B}_M^{\text{all}})^{|\mathbf{B}_M|}$ with $p := |\mathbf{O}|$, where p^3, $|\mathcal{C}|^3$ are respectively spent for calculating $\{\text{MAG}(O_i, M, Y)\}_{O_i}$ and $\{\text{MAG}(C_i, O_i, M)\}_{C_i}$, $\{\text{MAG}(C_i, M, O_i, Y)\}_{C_i}$. When the **MAG-Equi** condition holds for all O_i, the complexity is only with order $(p*|\mathcal{C}|)^3$.

Algorithm 1 Estimate δ_M, ζ_M for $M \in \mathbf{M}$.

INPUT: $\{r^{(i)}, o^{(i)}, y^{(i)}\}_{i=1}^n$ and $\alpha > 0$.
OUTPUT: $\hat{\delta}_M$ for each $M \in \mathbf{M}$.

Detect $\mathbf{M} := \{O_i | O_i \not\perp R, Y, O_i \not\perp R | Y, O_i \not\perp Y | R\}$.
For each $M \in \mathbf{M}$,
  Initialize $\mathbf{B}_M = \emptyset$, $\mathbf{B}_M^{\text{all}} = \emptyset$. Denote $\mathbf{O}_M := \mathbf{O} \backslash M$.
  Run FCI-JCI to obtain $\{\text{MAG}(O_i, M, Y)\}_{O_i \in \mathbf{O}_M}$.
  Detect $\mathbf{B}_M^{\text{all}}$ with types (a-f) in Fig. 2.
  For each $O_i \in \mathbf{B}_M^{\text{all}}$,
    $\forall C_i \in \mathcal{C}$, run FCI-JCI to obtain $\text{MAG}(C_i, O_i, M)$ and $\text{MAG}(C_i, M, O_i, Y)$.
    If **MAG-Equi** holds, add $\mathbf{B}_M = \mathbf{B}_M \cup O_i$.
    Else, search over $\mathbf{S} \subset \mathbf{B}_M^{\text{all}}$ such that $\Delta_{M, Y\leftarrow O_i|\mathbf{S}} < \alpha$, $\Delta_{O_i \leftarrow M|\mathbf{S}} > \alpha$ and $\Delta_{Y\leftarrow M|\mathbf{S}} > \alpha$, add $\mathbf{B}_M = \mathbf{B}_M \cup O_i$; else continue.
  Estimate $\mathbb{E}(Y|m, \mathbf{b}_M, R=0)$, $\mathbb{E}(Y|R=r)$ and $p(\mathbf{b}_M|R=r)$ for $r = 0, 1$.
  Estimate $\hat{\delta}_M$ and $\hat{\zeta}_M$.

7 Experiment

In this section, we evaluate our method on a synthetic dataset and a spine disease dataset, namely *Spine Patient Outcomes Research Trial* (SPORT) Birkmeyer et al. [2002], Weinstein et al. [2008, 2006b,a], Pearson et al. [2012].

7.1 Simulation

Data Generation. We follow the causal graph in Fig. 3 to generate data via: $R \sim Bern(1, 0.5)$; $X \leftarrow f_X(\varepsilon_x)$ ($\varepsilon_x \sim \mathcal{N}(\mu_x, \sigma_x^2)$); $M_1 \leftarrow f_{M_1}(x, r) := \alpha_0 + \alpha_R R + \alpha_X X + \varepsilon_{M_1}$; $M_2 \leftarrow f_{M_2}(x, r, m_1) := \beta_0 + \beta_1 M_1 + \beta_R R + \beta_X X + \varepsilon_{M_2}$; $Y \leftarrow f_Y(x, r, m_1, m_2) := \rho_0 + \rho_1 M_1 + \rho_2 M_2 + \rho_X X + \rho_R R + \varepsilon_Y$. Our goal is to estimate adjusted and unadjusted effects for M_1, M_2, and $\{M_1, M_2\}$, respectively.

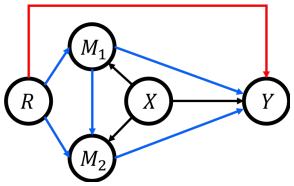


Figure 3: Causal graph to generate data.

Set $(\alpha_0, \alpha_R, \alpha_X) := (0, 2, 3)$, $(\beta_0, \beta_1, \beta_R, \beta_X) := (0, 2, 3, 4)$ and $(\rho_0, \rho_1, \rho_2, \rho_R, \rho_X) := (0, 2, 3, 4, 5)$, we have $\delta_{M_1} = \rho_1 \mathbb{E}(M_1(1) - M_1(0)) + \rho_2 \beta_1 \mathbb{E}(M_1(1) - M_1(0)) = (\rho_1 + \rho_2\beta_1)\alpha_1 = 16$; $\delta_{M_2} = \rho_2 (\beta_1 \mathbb{E}(M_1(1) - M_1(0)) + \beta_2) = \rho_2(\beta_1\alpha_1 + \beta_2) = 21$; $\delta_{M_1, M_2} = (\rho_1 + \rho_2\beta_1)\mathbb{E}(M_1(1) - M_1(0)) + \rho_2\beta_2 = (\rho_1 + \rho_2\beta_1)\alpha_1 + \rho_2\beta_2 = 25$; $\zeta_{M_1} = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) - \delta_{M_1} = 14$; $\zeta_{M_2} = 9$ and $\zeta_{M_1, M_2} = 5$.

Implementation. We follow algorithm 1 wih $\alpha = 0.05$ and linear model to fit $\mathbb{E}(Y|m, \mathbf{b}_M, R=0)$.

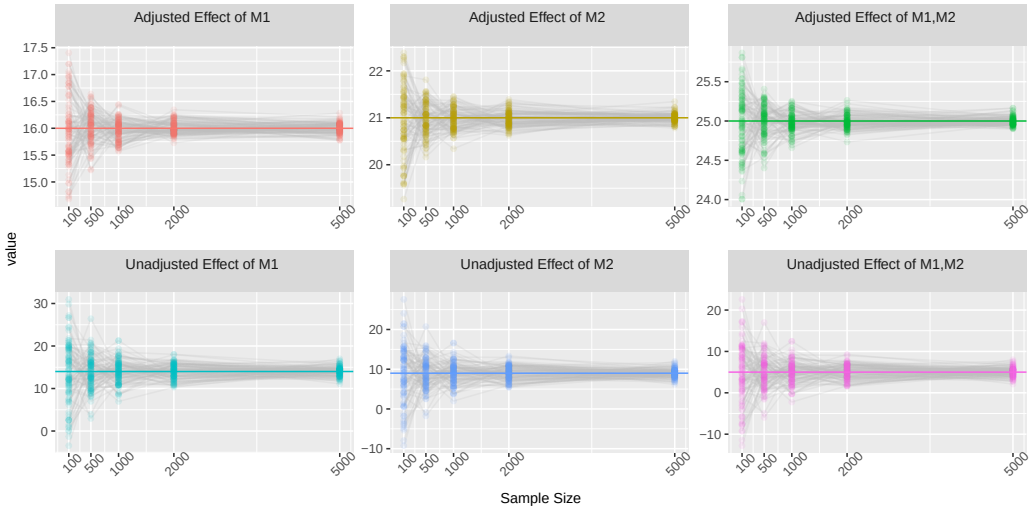


Figure 4: Estimation Results on Synthetic Dataset.

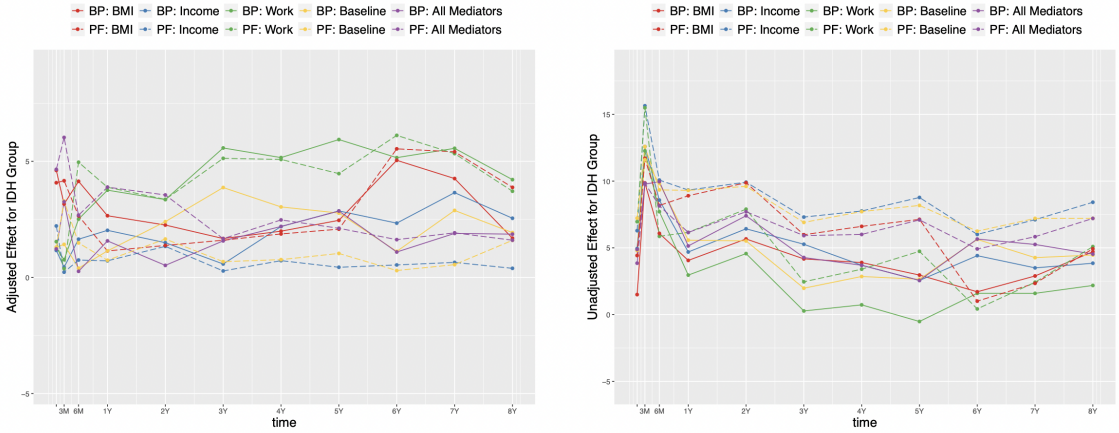


Figure 5: Adjusted/Unadjusted Effects on IDH data.

Results. Fig. 4 shows that our estimations can asymptotically approximate the ground-truth value as n grows.

7.2 Spine Patient Outcomes Research Trial (SPORT)

Dataset. The Spine Patient Outcomes Research Trial (SPORT) dataset was designed to investigate the effectiveness of spine surgery for the three most common reasons for low back pain (LBP) surgical disease. Low back pain-related conditions remain one of the most controversial diseases, as to, whether surgery is more effective than non-surgical treatments Birkmeyer et al. [2002]. As reported in Vos et al. [2017], LBP is a top-5 leading cause of disability globally, which has brought heavy burdens economically and socially to health systems globally. The SPORT study enrolled 2,505 patients in the United States from March 2000 to February 2005. There were three groups with associated LBP: 1) intervertebral disc herniation (IDH), 2) spinal stenosis (SPS), and 3) degenerative spondylolisthesis (DS). The race fell into three main groups: Asian Americans, African Americans, and White–Hispanic and non-Hispanic Americans. To investigate the long-term surgical effects (as LBP is a chronic disease), SPORT studied outcomes using validated measures. Primary outcomes were (i.e., Y) bodily pain (BP) and physical functioning (PF). Both are scaled from 0 to 100 (the higher, the better). Longitudinal data were collected at various intervals over 8 years after implementing surgical and non-surgical treatments (initially 6 weeks, 12 weeks, 6 months, 1 year, and yearly thereafter). For each patient, SPORT recorded the baseline demographic characteristics, social-economic factors such as income, working style, education years, and other commodities such as hypertension, diabetes, weight, smoking history, *etc.* Weinstein et al. [2008], Birkmeyer et al. [2002].

Implementation. We observe large differences between the white and non-white groups (which are mainly Asian and African Americans) in terms of BP and PF scores, as

shown in Fig. 8 in the appendix. We follow algorithm 1 with $\alpha = 0.05$ and use a random forest predictor with 50 estimators to estimate $\mathbb{E}(Y|\mathbf{B}_M, M, R)$. The detected mediator set $\mathbf{M}$ contains four features: income and working styles (social economics status) Body mass index (BMI) and the outcome score at baseline (health status). We then calculate the adjusted effects for the white and non-white groups ($R = 1$ means white) for DS, SPS, and IDH with $n = 601, 634$, and $1,195$ patients; and $p = 18$ covariates. More implementation details can be found in the appendix.

Results Analysis. Due to space limits, we only show the results on the IDH disease and placed other results in the appendix. As shown in Fig. 5, both adjusted and unadjusted effects are non-ignorable, which suggests more compensations for minorities (income, working status), healthier daily habits (BMI), and more personalized health care (the outcome at baselines) such as better medications. Particularly, for the working style and the income which are more of practical interest for policy-making, if we assign them to the advantaged group's distribution, the disparity in both BP and PF scores can be decreased by a noticeable margin. To explain, we observed that the proportion of the non-white group with no or low-level income is 5.4% more than that of the white patient subgroup and, that the proportion of employment of the black patient group was nearly 18% less than that of the white subgroup. This suggests the possibility that alleviating disparities in outcomes can be achieved by focusing on one's employment status and income. Besides, it is interesting to note that the BMI has significant adjustment effects, which help explain the disparities in terms of outcomes after surgical treatments. Finally, one can observe that making policies on all mediators (purple curve) does not necessarily have better effects, since the formula of δ_M does not necessarily increase as $|M|$ increases. To illustrate, consider the group $R = 0$ that does not have good living or dietary habits, adjusting all factors may let them treat their disease lightly (e.g., choosing non-surgical treatment), which may in turn result in worse outcomes.

8 Conclusion

We define the adjusted and unadjusted effects under the structural-causal model framework. Compared to existing methods, our method has better policy implications, by eliminating all sources of disparity to measure the effect of a policy on the mediator of interest. Additionally, unlike traditional methods that rely on the strong ignorability condition for identifiability, we can identify the admissible set via causal discovery. Our methods are efficient and easy to implement. In the SPORT dataset, we identified important, interpretable socioeconomic variables that significantly impact disparities in terms of short and long-term treatment effects. For limitations, our methods require no unobserved confounder conditions. That said we plan to explore the 'relaxation' of this condition for future work.

References

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 357–363, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

N. J. Birkmeyer, J. N. Weinstein, A. N. Tosteson, T. D. Tosteson, J. S. Skinner, J. D. Lurie, R. Deyo, and J. E. Wennberg. Design of the spine patient outcomes research trial (sport). *Spine*, 27(12):1361, 2002.

T. Blakely, G. Disney, L. Valeri, J. Atkinson, A. Teng, N. Wilson, and L. Gurrin. Socioeconomic and tobacco mediation of ethnic inequalities in mortality over time: repeated census-mortality cohort studies, 1981 to 2011. *Epidemiology (Cambridge, Mass.)*, 29(4):506, 2018.

P. A. Braveman, S. Kumanyika, J. Fielding, T. LaVeist, L. N. Borrell, R. Manderscheid, and A. Troutman. Health disparities and health equity: the issue is justice. *American journal of public health*, 101(S1):S149–S155, 2011.

M. Frölich and M. Huber. Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1645–1666, 2017.

D. Galles and J. Pearl. Testing identifiability of causal effects. *arXiv preprint arXiv:1302.4948*, 2013.

H. Gong and K. Zhu. Path-specific effects based on information accounts of causality. *arXiv preprint arXiv:2106.03178*, 2021.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

B. Huang, K. Zhang, J. Zhang, J. D. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

P. Hystad, R. M. Carpiano, P. A. Demers, K. C. Johnson, and M. Brauer. Neighbourhood socioeconomic status and individual lung cancer risk: evaluating long-term exposure measures and mediating mechanisms. *Social science & medicine*, 97:95–103, 2013.

E. Ibfelt, S. Kjaer, C. Høgdall, M. Steding-Jessen, T. Kjaer, M. Osler, C. Johansen, K. Frederiksen, and S. Dalton. 'socioeconomic position and survival after cervical cancer: influence of cancer stage, comorbidity and smoking among danish women diagnosed between 2005 and 2010. *British journal of cancer*, 109(9):2489–2495, 2013.

J. W. Jackson. On the interpretation of path-specific effects in health disparities research. *Epidemiology*, 29(4):517–520, 2018.

J. W. Jackson. Meaningful causal decompositions in health equity research: definition, identification, and estimation through a weighting framework. *Epidemiology*, 32(2): 282–290, 2020.

J. W. Jackson and T. J. VanderWeele. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology (Cambridge, Mass.)*, 29(6):825, 2018.

M. Kasy. Instrumental variables with unrestricted heterogeneity and continuous treatment. *The Review of Economic Studies*, 81(4):1614–1636, 2014.

T. G. McGuire, M. Alegria, B. L. Cook, K. B. Wells, and A. M. Zaslavsky. Implementing the institute of medicine definition of disparities: an application to mental health care. *Health services research*, 41(5):1979–2005, 2006.

R. Oaxaca. Male-female wage differentials in urban labor markets. *International economic review*, pages 693–709, 1973.

J. PEARL. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pages 411–420. Morgan Kaufman, 2001.

J. Pearl. *Causality*. Cambridge university press, 2009.

A. Pearson, J. Lurie, T. Tosteson, W. Zhao, W. Abdu, and J. Weinstein. Who should have surgery for spinal stenosis?: Treatment effect predictors in sport. *Spine*, 37(21): 1791, 2012.

J. Quirk, B. Harbors, and P. Design. Administrative draft existing conditions report. 2006.

J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

P. Spirtes and T. Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

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

L. Valeri and T. J. VanderWeele. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. *Psychological methods*, 18(2):137, 2013.

T. J. VanderWeele. Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian Journal of Statistics*, 38(3):551–563, 2011.

T. J. VanderWeele, S. Vansteelandt, and J. M. Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.)*, 25(2):300, 2014.

S. Vansteelandt and T. J. VanderWeele. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*, 68(4):1019–1027, 2012.

T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo, S. F. Abera, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1211–1259, 2017.

J. N. Weinstein, J. D. Lurie, T. D. Tosteson, J. S. Skinner, B. Hanscom, A. N. Tosteson, H. Herkowitz, J. Fischgrund, F. P. Cammisa, T. Albert, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the spine patient outcomes research trial (sport) observational cohort. *Jama*, 296(20):2451–2459, 2006a.

J. N. Weinstein, T. D. Tosteson, J. D. Lurie, A. N. Tosteson, B. Hanscom, J. S. Skinner, W. A. Abdu, A. S. Hilibrand, S. D. Boden, and R. A. Deyo. Surgical vs nonoperative treatment for lumbar disk herniation: the spine patient outcomes research trial (sport): a randomized trial. *Jama*, 296(20):2441–2450, 2006b.

J. N. Weinstein, T. D. Tosteson, J. D. Lurie, A. N. Tosteson, E. Blood, B. Hanscom, H. Herkowitz, F. Cammisa, T. Albert, S. D. Boden, et al. Surgical versus nonsurgical therapy for lumbar spinal stenosis. *New England Journal of Medicine*, 358(8):794–810, 2008.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008a. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2008.08.001. URL https://www.sciencedirect.com/science/article/pii/S0004370208001008.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008b.

J. Zhang and P. Spirtes. A transformational characterization of markov equivalence for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 667–674, 2005.

A Theoretical Analysis

Our theoretical analysis is based on Prop. 3.1 and some MAG-related definitions. These definitions can be found in Zhang [2008b]. We will also introduce them here for completeness.

We begin with the *mixed graph*, which is a graph with two kinds of edges: directed ($\rightarrow$), and bi-directed ($\leftrightarrow$). Note that this definition is slightly different from that in Zhang [2008b] as we do not require it to contain an undirected graph ($-$), which means two variables are dependent due to latent selection variables. According to our problem setting, the observed variables are not assumed to be dependent on some unobserved selection variables, so it is unnecessary for our definition to contain the undirected graph. We first introduce the definition of *ancestral graph*.

Definition A.1 (Ancestral Graph)**.** An ancestral graph G is a mixed graph with no directed and almost directed cycles. Here, the almost directed cycle occurs when $A \leftrightarrow B$ and $B \in \mathrm{An}_G(A)$.

Next, we introduce the definition of inducing path, which can further define *maximal ancestral graph*.

Definition A.2 (Inducing Path)**.** A path p from X to Y in G is an inducing path if every non-endpoint vertex on p is a collider and is either $\in \mathrm{An}_G(X)$ or $\in \mathrm{An}_G(Y)$.

Definition A.3 (Maximal Ancestral Graph)**.** An ancestral graph is Maximal Ancestral Graph if, for any two non-adjacent variables, there is no inducing path between them.

Next, we introduce the definition of *unshielded collider* and *discriminating path*, which form two main principles in Prop. 3.1.

Definition A.4 (Unshielded Collider)**.** A triple $\langle X, Y, Z\rangle$ is called an unshielded collider if the edge between (X, Y) and the one between (Y, Z) are pointed to Y.

Definition A.5 (Discriminating path)**.** In a MAG, a path $p = \langle X, V_1, ..., V_p, W, Y\rangle$, is a discriminating path for W if **i)** p includes at least three edges; **ii)** W is a non-endpoint on p, and is adjacent to Y. **iii)** X is not adjacent to Y, and each V_i for $i \in \{1, ..., p\}$ is a collider on p and $\in \mathrm{Pa}(Y)$.

Proof of Prop. 5.1. The Galles and Pearl [2013] provided four conditions for $p(y|do(m), r)$ to be identifiable, and the theorem 4.3.2 in Pearl [2009] proved that at least one of these conditions holds. These four conditions are:

- There is no-back door path from M to Y.
- There is no directed path from M to Y in G.
- There exists a set of nodes $\mathbf{B}$ that blocks all back-door paths from M to Y.
- There exists sets of nodes Z_1 and Z_2 such that:
  1. Z_1 blocks every directed path from M to Y;
  2. Z_2 blocks all back-door paths between Z_1 and Y;
  3. Z_2 blocks all back-door paths between M and Z_1;
  4. Z_2 does not activate any back-door paths between M and Y.

Since there is a directed edge $M \rightarrow Y$, therefore the second and the fourth conditions are violated. As we condition on R, so the first condition can be modified to no back-door path from M to Y, given R. In this regard, the $\mathbf{B}$ in the third condition degenerates to $\emptyset$. So we have $p(y|do(m), r)$ is identifiable if and only if there exists $\mathbf{B}$ that block all back-door paths from M to Y. If this exists unobserved confounder between M to Y, this sufficient and necessary condition will not hold, making $p(y|do(m), r)$ unidentifiable. □

Proof of Theorem. 5.1. It is sufficient to show that the **i)** $\underline{\mathbf{B}}_M$ contain all mediator-outcome confounders and meanwhile **ii)** the variable in $\overline{\mathbf{B}}_M$ does not perturb existing directed paths from M to Y and not induce new spurious features, according to Pearl [2009]. For **i)**, note that for each mediator-outcome confounder, it causally influences both M and Y, which means this there exists directed paths from O_i to M and O_i to Y. Therefore, the maximal ancestral graph over (M, O_i, Y) is $\{M \leftarrow O_i \rightarrow Y\} \cup \{M \rightarrow Y\}$. Therefore, we have $O_i \in \underline{\mathbf{B}}_M$. For **ii)**, for each $O_i \in \overline{\mathbf{B}}_M$, if it lies in the directed path from M to Y, then the learned MAG must be $\{M \rightarrow O_i \rightarrow Y\} \cup \{M \rightarrow Y\}$, which is not contained in $\overline{\mathbf{B}}_M$. For O_i with

the MAG (in addition to the edge $M \to Y$) with$M \leftarrow O_i \leftrightarrow Y$ or $M \leftrightarrow O_i \to Y$. Since there is no unobserved confounder between M and Y, then there respectively exists an observed confounder that blocks the path between O_i and Y and the path between M and O_i. Besides, such a confounder is in $\underline{\mathbf{B}}_M$. Therefore, such an O_i will not induce new spurious features, since even there is a collider induces this path, there exists a variable in $\underline{\mathbf{B}}_M$ that can block this path. In this regard, any set that contains $\underline{\mathbf{B}}_M$ and belong to $\overline{\mathbf{B}}_M$ will satisfy both **i)** and **ii)**. Therefore, this set is admissible. □

Proof of Lemma 5.1. It is easy to note that we have $O_i \in \overline{\mathbf{B}}_M$ if we can discriminate the associated types of MAGs from those in $\tilde{\mathbf{B}}_{\mathrm{MAG}}(M)$. Specifically, the MAGs in $\tilde{\mathbf{B}}_{\mathrm{MAG}}(M)$ (we omit $M \to Y$ as it is included in all MAGs) are: i) $M \leftarrow O_i \to Y$; ii) $M \leftarrow O_i \leftrightarrow Y$; iii) $M \leftrightarrow O_i \to Y$; iv) $M \to O_i \to Y$; v) $M \to O_i \leftrightarrow Y$; vi) $M \leftrightarrow O_i \leftrightarrow Y$. Therefore, our goal is to discriminate i), ii), iii) from others. Note that when the first condition in Lemma 5.1 is satisfied, we can discriminate i) and ii) using unshielded collider condition in Prop. 3.1, as the $M \to O_i$ or $M \leftrightarrow O_i$ can make $\langle O_i, M, C_M^i \rangle$ unshielded collier, therefore we have $O_i \in \overline{\mathbf{B}}_M$. If the second condition is satisfied, we can first discriminate i), ii), iii), vi) from iv) and v) using unshielded collider condition; then to eliminate vi) which can induce spurious path for instance $M \leftarrow S_1 \to O_i \leftarrow S_2 \to Y$ with S_1, S_2 unobserved, we note that $\langle C_M^j, M, O_i, Y \rangle$ form a discriminating path for O_i in i), ii), iii) and vi), with the difference that the O_i in vi) is a collider on the path $M \leftrightarrow O_i \leftrightarrow Y$ in vi) but is not a collider on the same path in i), ii) and iii). Therefore, we can discriminate i), ii), iii) from vi). In this regard, we can determine whether $O_i \in \overline{\mathbf{B}}_M$ if either one of the condition holds in Lemma 5.1. □

Proof of Lemma 5.2. For the O_i such that $M \leftarrow O_i \to Y$ in the MAG over (O_i, M, Y), we have that the deconfounding sets between O_i and M, and that between M and Y, can be observed, thus we there exists $\mathbf{X}_i$ which contains both deconfounding sets, such that $\{P^r(M|O_i)\}$, $\{P^r(Y|O_i)\}$ are independent to $\{P^r(O_i)\}$, because they have disentangled causal mechanisms. Besides, we have $\{P^r(O_i|M)\}$, $\{P^r(O_i|Y)\}$ are respectively dependent to $\{P^r(M)\}$ and $\{P^r(Y)\}$ under assumption 5.3, since they are entangled. On the other hand, we will show that this conclusion does not hold for O_i with other types of MAG. Similarly, the MAGs in $\tilde{\mathbf{B}}_{\mathrm{MAG}}(M)$ (we omit $M \to Y$ as it is included in all MAGs) are: i) $M \leftarrow O_i \to Y$; ii) $M \leftarrow O_i \leftrightarrow Y$; iii) $M \leftrightarrow O_i \to Y$; iv) $M \to O_i \to Y$; v) $M \to O_i \leftrightarrow Y$; vi) $M \leftrightarrow O_i \leftrightarrow Y$. For the MAG with $M \leftrightarrow O_i$, then for any set $\mathbf{A}$, the independent relations for ($\{P^r(M|\mathbf{A}, O_i)\}$,$\{P^r(O_i|\mathbf{A})\}$) and ($\{P^r(M|\mathbf{A})\}$,$\{P^r(O_i|M, \mathbf{A})\}$) simultaneously hold or not hold, determined by whether A is the deconfounding set between O_i and M. This conclusion similarly holds for the MAG with $O_i \leftrightarrow Y$. Therefore, we can discriminate MAGs ii), iii),v), vi) from others. For the MAG with $M \to O_i \to Y$, if there exists a set $\mathbf{A}$ such that $\{P^r(M|\mathbf{A}, O_i)\}$ is $\{P^r(O_i|\mathbf{A})\}$ independent, then $\mathbf{A}$ must block the directed paths between M and O_i. Suppose otherwise there exists an unblocked directed path from M to O_i; in this regard, both distributions are entangled since they are affected by the mechanism $P^r(O_i|\mathbf{A}, M)$. Then if $\mathbf{A}$ blocks the directed paths between M and O_i, the independence between $\{P^r(O_i|\mathbf{A}, M)\}$ and $\{P^r(M|\mathbf{A})\}$ also holds. This can help us further eliminate type iv). As a result, we have $O_i \in \underline{\mathbf{B}}_M$. □

Proof of Theorem 5.2. Under assumptions 5.1, 5.2, and 5.3, we can detect $\underline{\mathbf{B}}_M$. Besides, if condition in Lemma 5.1 holds for O_i, we can determine whether $O_i \in \overline{\mathbf{B}}_M$. Therefore, we have the detected $\mathbf{B}_M$ is

$$\underline{\mathbf{B}}_M \cup \{O_i | \text{the condition in Lemma 5.1 for } O_i \text{ holds and } O_i \in \overline{\mathbf{B}}_M\}, which \supset \underline{\mathbf{B}}_M and \subset \overline{\mathbf{B}}_M.$$

The proof is completed. □

B Experiment on SPORT

B.1 Implementation Details

Our implementation is composed of three steps: **i)** detect the mediator set $\mathbf{M}$; **ii)** identify admissible sets; and **iii)** Estimation.

Detection of $\mathbf{M}$. We follow the FCI-JCI algorithm Zhang [2008a] to identify adjustment sets among social-economic factors. In total, M contains four features: income and working styles (social economics status) Body mass index (BMI), and the outcome score at baseline (health status).

Identification of admissible sets. In addition to the race variable, there are also other context variables such as gender, age, *etc.* We discovered that there is at least a context variable that is non-adjacent to M for each adjusted factor M. As a result, we can directly identify $\mathbf{B}_M$ by iteratively implementing the FCI-JCI algorithm to obtain MAG over (R, M, O, Y) for each $O \in \mathbf{B}_M^{\mathrm{all}}$. Specifically, the $\mathbf{B}_M$ is (age, gender, education years, working style) when M denotes income; is (age, gender, education years, back pain at baseline (bp_{BA})) when M denotes working style; (gender, diabetes) when M denotes BMI and finally is $\emptyset$ when $M := bp_{BA}$.

Estimation. To fit $\mathbb{E}(Y|\mathbf{B}_M, M, R)$, we train a random forest predictor with 50 estimators.

B.2 Results on SPS, DS diseases when R denotes "Race"

In this section, we show our estimated adjusted and unadjusted effects of SPS and DS diseases in the SPORT dataset. Fig. 6 and Fig. 7 respectively show the adjusted and unadjusted effects of the SPS and DS diseases. Fig. 8 shows the total effects.

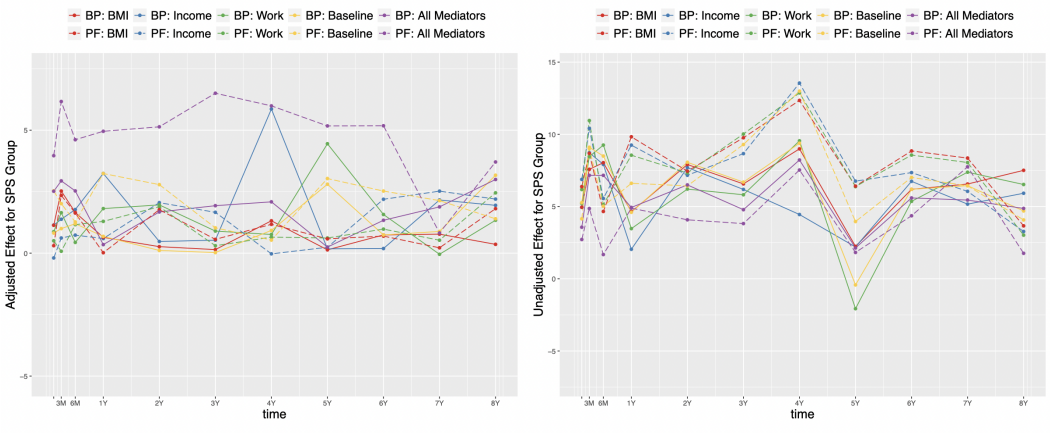


Figure 6: Adjusted and Unadjusted effects of the SPS disease when R denotes race.

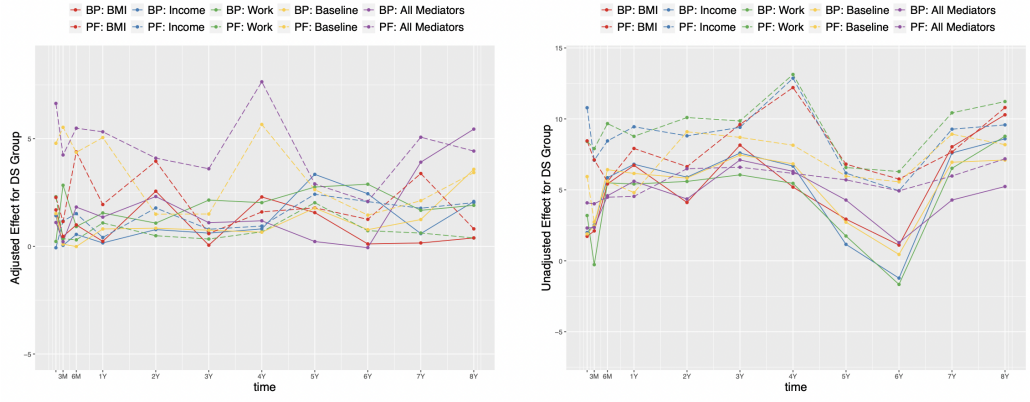


Figure 7: Adjusted and Unadjusted effects of the DS disease when R denotes race.

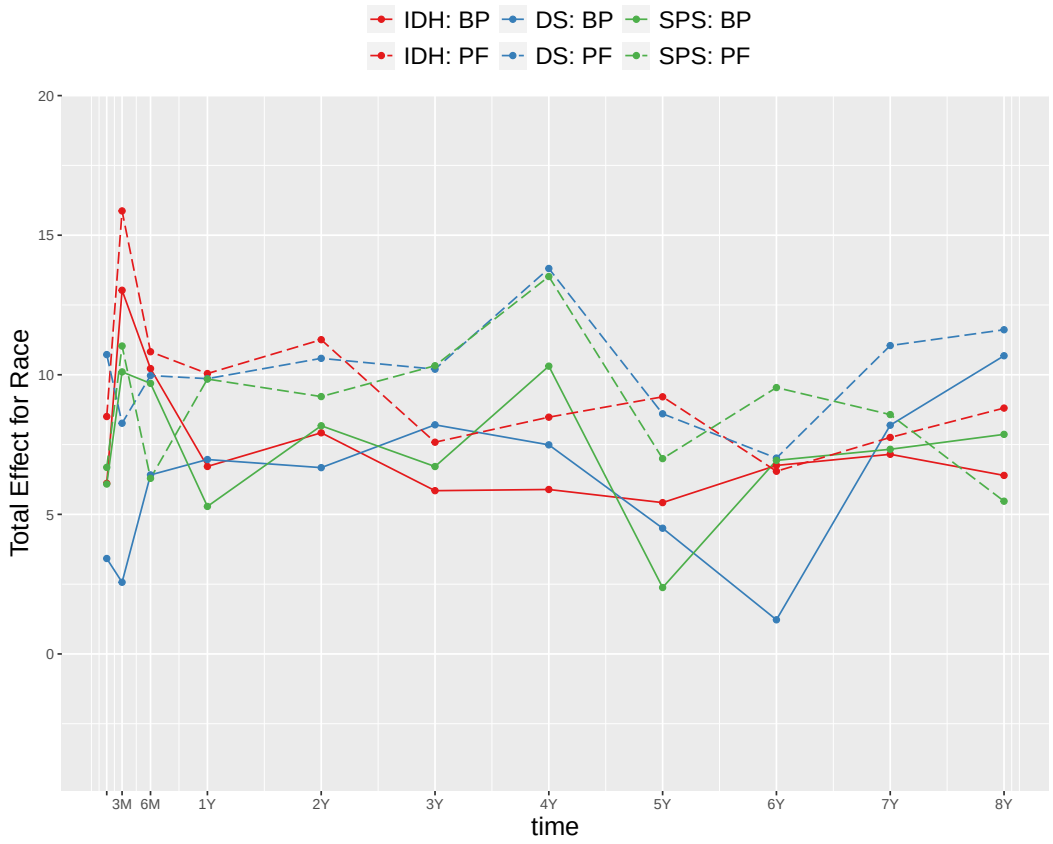


Figure 8: Total Effects with Y denoting BF and PF scores when R denotes race.

B.3 Results on when R denotes "Gender"

We also conduct experiments when R denotes "Gender", with $R = 1$ (or $R = 0$) meaning the male (female) group. Fig. 8 shows the total effects of $\mathbb{E}(Y|R = 1) - \mathbb{E}(Y|R = 0)$, which suggests a significant difference in SPS and DS diseases; and a minor but unignorable difference in IDH disease. The selected mutable variable set $\mathbf{M}$ contains "income", "working styles" and "BMI". The $\mathbf{B}_M$ is (age, gender, education years, working style) when M denotes income; is (age, gender, education years, back pain at baseline (bp_{BA})) when M denotes working style; (gender, diabetes) when M denotes BMI. The whole implementation procedure is the same to the one when R denotes "Race". Fig. 10, 11 and Fig. 12 respectively shows the adjusted and unadjusted effects of the IDH, SPS, and DS diseases.

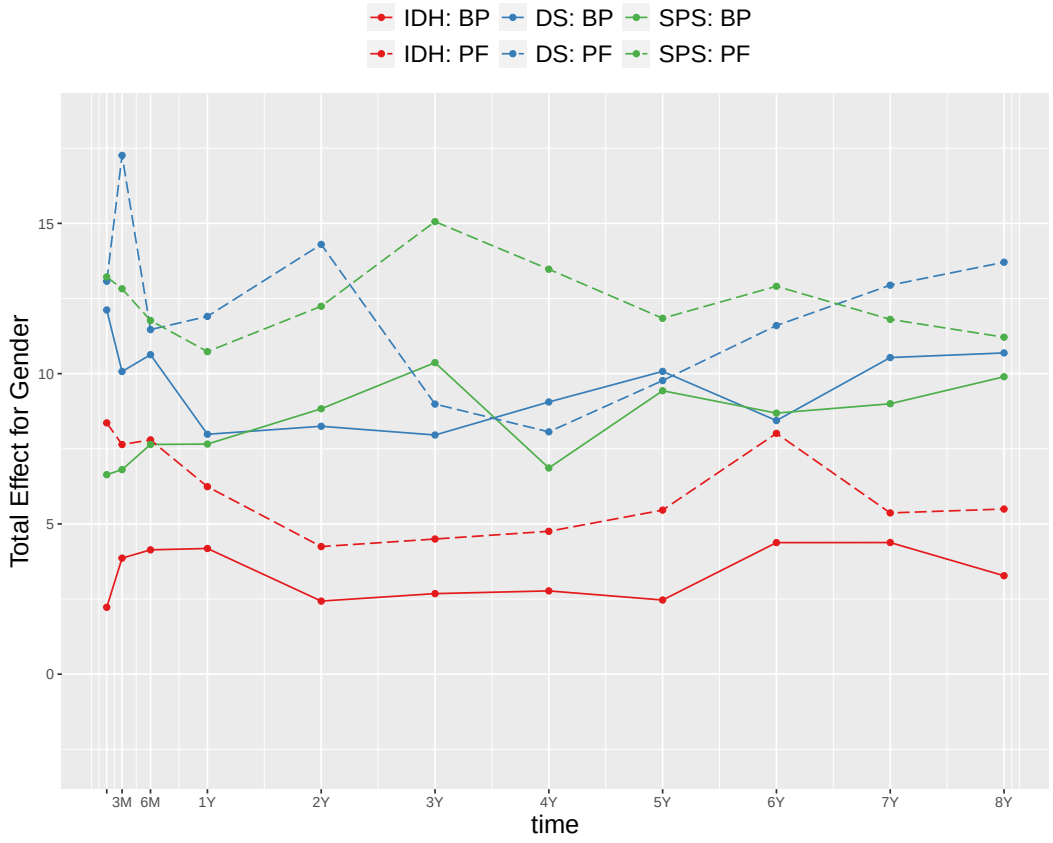


Figure 9: Total Effects with R denotes gender.

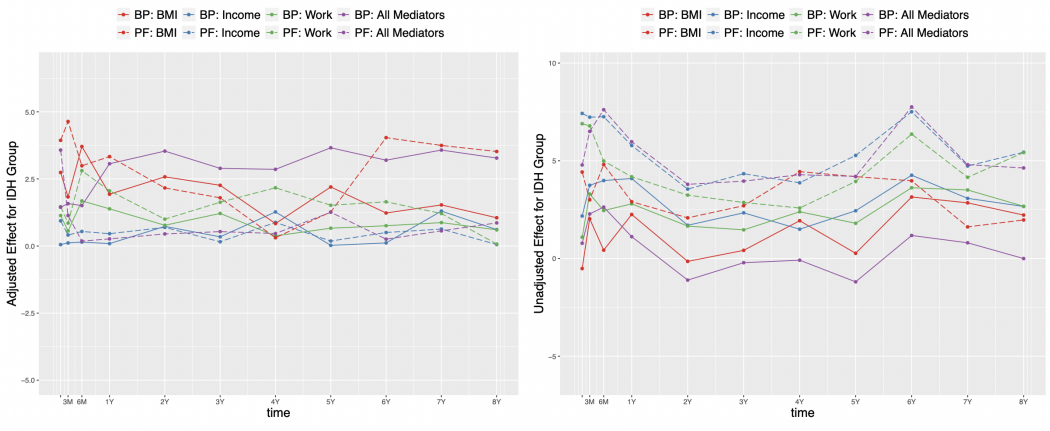


Figure 10: Adjusted and Unadjusted effects of the IDH diseases when R denotes gender.

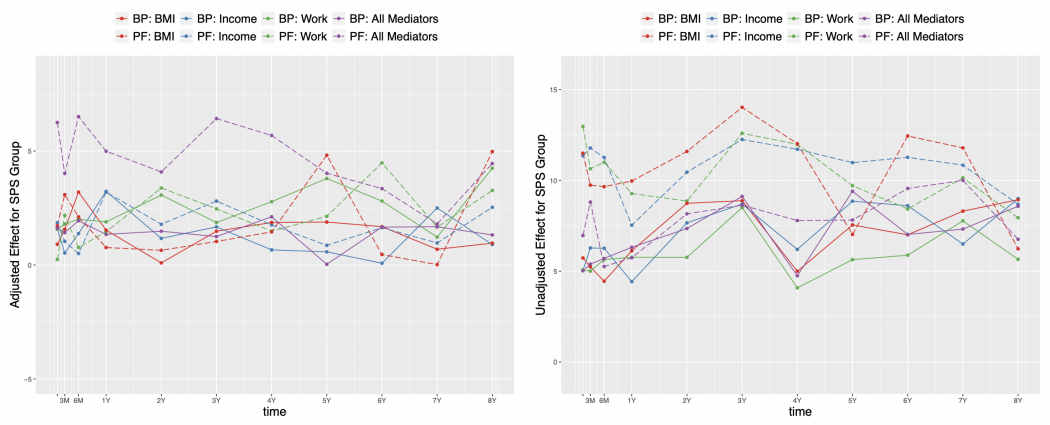


Figure 11: Adjusted and Unadjusted effects of the SPS disease when R denotes gender.

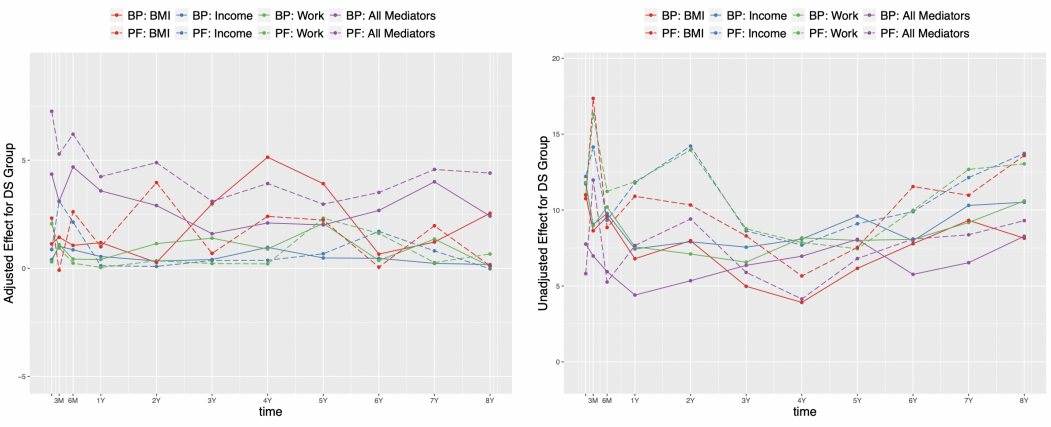


Figure 12: Adjusted and Unadjusted effects of the DS disease when R denotes gender.