Biometrics 000, 000–000 DOI: 000
000 0000

Non-linear Mediation Analysis with High-dimensional Mediators whose Causal Structure is Unknown

Wen Wei Loh[1,*], Beatrijs Moerkerke[1], Tom Loeys[1], and Stijn Vansteelandt[2,3]

[1] Department of Data Analysis, Ghent University, Gent, Belgium

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[3] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, United Kingdom

email: WenWei.Loh@UGent.be

Summary:

With multiple possible mediators on the causal pathway from a treatment to an outcome, we consider the problem of decomposing the effects along multiple possible causal path(s) through each distinct mediator. Under Pearl's path-specific effects framework (Pearl, 2001; Avin et al., 2005), such fine-grained decompositions necessitate stringent assumptions, such as correctly specifying the causal structure among the mediators, and no unobserved confounding among the mediators. In contrast, interventional direct and indirect effects for multiple mediators (Vansteelandt and Daniel, 2017) can be identified under much weaker conditions, while providing scientifically relevant causal interpretations. Nonetheless, current estimation approaches require (correctly) specifying a model for the joint mediator distribution, which can be difficult when there is a high-dimensional set of possibly continuous and non-continuous mediators. In this article, we avoid the need to model this distribution, by developing a definition of interventional effects previously suggested by VanderWeele and Tchetgen Tchetgen (2017) for longitudinal mediation. We propose a novel estimation strategy that uses non-parametric estimates of the (counterfactual) mediator distributions. Non-continuous outcomes can be accommodated using non-linear outcome models. Estimation proceeds via Monte Carlo integration. The procedure is illustrated using publicly available genomic data (Huang and Pan, 2016) to assess the causal effect of a microRNA expression on the three-month mortality of brain cancer patients that is potentially mediated by expression values of multiple genes.

Key words: Collapsibility; Direct and indirect effects; Effect decomposition; Marginal and conditional effects; Multiple mediation analysis; Path analysis

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Mediation analysis is commonly used to study the effect of a treatment or exposure (A) on an outcome (Y) that may be transmitted through intermediate variable(s) on the causal pathway from A to Y. Counterfactual-based definitions of *natural direct and indirect effects* (Robins and Greenland, 1992; Pearl, 2001) permit decomposing the total effect of a treatment on an outcome into a direct and an indirect effect for a single mediator, without relying on any specific statistical models for the mediator and outcome. However, multiple possible mediators often exist in substantive research. For example, different mediators may be posited in trying to understand the different causal pathways from A to Y, or confounders of the mediator-outcome relation for a mediator of interest are themselves affected by treatment and thus perceived as competing mediators, or interventions may be designed to affect outcome by simultaneously changing different mediators on the causal pathway from A to Y. Extensions of natural indirect effects for a single mediator to the multiple mediator setting are therefore complicated by the complex (possibly unknown) confounding patterns among the different mediators. In particular, when one mediator exerts a causal effect on another, so that the former is a confounder of the mediator-outcome relation for the latter (henceforth termed *post-treatment* confounding), the assumptions needed to (non-parametrically) identify fine-grained *path-specific* effects via certain mediators may be violated (Avin et al., 2005). Consider the following example in Figure 1. The path-specific effect along the causal path $A \rightarrow M_1 \rightarrow M_3 \rightarrow Y$ cannot be identified because M_1 is a post-treatment confounder of the $M_3 \rightarrow Y$ relation. More stringent assumptions may be imposed so that natural indirect effects defined using certain combinations of separately (un)identifiable path-specific effects can be identified (Daniel et al., 2015; Steen et al., 2017). But one such empirically untestable assumption that can prohibit identification when violated is no unobserved confounding among the mediators. For example, the path-specific

effect along the assumed path $A \to M_1 \to M_2 \to Y$ cannot be identified because U is an unobserved confounder of M_1 and M_2.

[Figure 1 about here.]

Current methods for assessing natural effects with multiple mediators are therefore restricted to situations where either the mediators can be causally ordered so that (combinations of) path-specific effects can be identified in the presence of post-treatment confounding (VanderWeele and Vansteelandt, 2014; Daniel et al., 2015; Steen et al., 2017; Albert et al., 2019), or the mediators do not exert causal effects on each other, and are independent given treatment and covariates (Lange et al., 2013; Taguri et al., 2018). Hence a limitation, shared by all the above approaches, is that they are predicated on (correct) a priori assumptions about the detailed causal relations between the mediators. These restrictions impede applications in most realistic settings, when the causal structure among the mediators is unknown or cannot be (correctly) specified based on sound scientific knowledge or empirical evidence, or the mediators are manifestations of a latent process or variable, and thus correlated.

In contrast, *interventional direct and indirect effects*, first introduced by Didelez et al. (2006) and VanderWeele et al. (2014) for a single mediator, can be identified under much weaker assumptions than natural or path-specific effects, especially when there is post-treatment confounding of the mediator(s)-outcome relation(s). Interventional effects consider population-level (stochastic) interventions that set the value of the mediator to a random draw from its (counterfactual) distribution; whereas natural effects are defined in terms of individual-level (deterministic) interventions on the mediator, which can lack scientific and practical meaning when the treatment cannot be manipulated at the individual level (VanderWeele, 2013). Vansteelandt and Daniel (2017) extended the definitions to the multiple mediator setting so that the total effect of a treatment on an outcome can be decomposed into a direct effect and a joint indirect effect via the mediators. They further decomposed the joint

indirect effect into separate indirect effects via each distinct mediator, and an indirect effect via the mediators' mutual dependence. Interventional indirect effects are defined in terms of the underlying (possibly unknown) causal effects among the mediators. They therefore possess valid interpretations by construction, and can be identified without prior (correct) specification of the mediators' causal structure, and even when mediators share hidden common causes. Recent work discussing interventional (in)direct effects include Moreno-Betancur and Carlin (2018), Lok (2019), and Quynh Nguyen et al. (2019), among others, for comparisons with natural (in)direct effects; VanderWeele and Tchetgen Tchetgen (2017) for a single longitudinal mediator; Lin and VanderWeele (2017) for multiple mediators with a known causal ordering; and Loh et al. (2019) for comparisons with prevalent "product-of-coefficients" methods (MacKinnon, 2000; Preacher and Hayes, 2008) assuming linear models for all variables.

In this article, we generalize the interventional effects framework for multiple mediators to high-dimensional mediators. A key complication when estimating current definitions of the interventional effects is the need to correctly specify a (parametric) model for the joint distribution for all mediators conditional on treatment and all observed confounders. However, VanderWeele and Tchetgen Tchetgen (2017) suggested the possibility of interventional effects defined using (counterfactual) mediator distributions that depend only on treatment, and are unconditional on the baseline confounders of the mediator(s)-outcome relation(s). Such definitions avoid specifying models for the mediators in terms of the baseline confounders, and can thus be particularly advantageous when there is a high-dimensional set of (non-)continuous mediators. All current high-dimensional mediation methods for assessing indirect effects through each distinct mediator are restricted to continuous mediators, with almost all additionally considering only continuous outcomes, so that the product-of-coefficients method under assumed linear models for (transformations of the) mediators and outcomes

can be employed; see e.g., Chakrabortty et al. (2018), Huang and Pan (2016), Zhang et al. (2016), Zhao and Luo (2016), Zhao et al. (2019), Derkach et al. (2020), and Zhao et al. (2020) among many others. While most methods allow for either correlated errors in the joint mediator model (thus allowing for certain forms of hidden confounding among the mediators), or mediators to influence one another, only Zhao and Luo (2016) and Zhao et al. (2019) allow for both within the same mediation model. In most realistic and important settings, high-dimensional mediators that are closely linked either have complex but unknown causal relations, or share unmeasured common causes, or both. Even in the single mediator setting, indirect effects using the traditional product-of-coefficients method may not correspond to those derived using the causal mediation framework, due to misspecification of non-linear models for non-continuous mediator and outcome (MacKinnon et al., 2018). With the increasing demand for high-dimensional mediation methods in biology, medical and public health research, our proposal is, to the best of our knowledge, the first to allow for both continuous and non-continuous mediators to simultaneously exist on the causal pathway between treatment and outcome, where they can concurrently causally affect one another, and share hidden common causes.

In this article, we therefore build on the suggestion by VanderWeele and Tchetgen Tchetgen (2017) for a single (longitudinal) mediator to develop interventional direct and indirect effects for high-dimensional mediators. Our proposed definitions of interventional indirect effects differ from existing definitions by Vansteelandt and Daniel (2017), Lin and VanderWeele (2017), and Loh et al. (2019) in two respects. The first is that the decomposition of the joint indirect effect into separate indirect effects via each mediator is invariant to the presumed (possibly arbitrary) ordering of the mediator indices. Existing definitions assume different hypothetical treatment levels for different subsets of the mediators, thus allowing the possibility of different decompositions (depending on the mediator indices). The second and more pertinent differ-

ence is that the interventional indirect effects developed in this article rely on (counterfactual) mediator distributions that depend only on treatment, whereas existing definitions depend on treatment and all baseline covariates. Hence estimating existing definitions demands specifying a correct model for the joint distribution for all mediators that is congenial with models for the marginal distribution for each mediator, all of which are conditional on the covariates. The difficulties in correctly specifying such (parametric) mediator models in practice are exacerbated in high-dimensional mediation settings. The definitions proposed in this article therefore permit a novel estimation strategy that requires specifying only a mean model for the potential outcomes, and no models for the mediators. The mediators and outcome can be continuous or noncontinuous, with non-continuous outcomes accommodated using non-linear outcome models. Non-parametric estimates of the (counterfactual) mediator distributions under each treatment level are used. Estimation proceeds via Monte Carlo integration.

The remainder of this article is as follows. In Section 2 notation is introduced, the interventional direct and indirect effects are defined, and the identification assumptions are stated. In Section 3 the novel estimation strategy that requires no models for the mediators, and only a mean model for the outcome, is proposed. For pedagogic purposes, we will focus on randomized studies with fewer mediators than observations in the theoretical development of our proposed procedure. Extensions to high-dimensional mediators, and observational studies when treatment is not randomly assigned, are then described. In Section 4 the proposed estimation procedures are assessed via extensive simulation studies. In Section 5 the estimation strategy is illustrated using publicly-available data from a previous high-dimensional mediation analysis (Huang and Pan, 2016) that investigated the causal effect of a microRNA expression (miR-223) on a dichotomous three-month survival status among patients suffering from an aggressive form of brain cancer, that is potentially mediated by

expression values of different genes. In Section 6 we describe settings in which the proposed and existing definitions of interventional indirect effects are (approximately) equivalent. Under such (common) settings, the proposed estimation procedure simplifies current estimation approaches (Vansteelandt and Daniel, 2017; Loh et al., 2019), by eliminating the need to specify a model for the joint distribution of the mediators. A brief discussion is provided in Section 7.

2. Definition and identification of interventional (in)direct effects

Consider the setting with an exposure or a treatment A, multiple possible mediators $M_1, \ldots, M_p$, and an outcome Y. In this article, we adopt the perspective that all post-treatment confounders of a mediator-outcome relation for a mediator in question are themselves competing mediators, and therefore included in the set of possible mediators. Here and throughout subscripts in the notation for the mediators are merely used to arbitrarily index the different mediators, and not to indicate any assumed causal (or temporal) ordering of the mediators; e.g., M_1 need not precede M_2 causally (or temporally). Let $Y_{am_1\cdots m_p}$ denote the potential outcome for Y if, possibly counter to fact, A is set to a, when each mediator M_s is set to the value $m_s, s = 1, \ldots, p$. Let $M_{sa^{(s)}}$ denote the potential outcome for M_s if, possibly counter to fact, A is set to $a^{(s)}$. Let $Y_{a^{(0)}\{\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(1)}}\}}$ denote the potential outcome for Y under treatment $A = a^{(0)}$, when the mediator values are set to a random draw from the *joint* (counterfactual) distribution under (hypothetical) treatment $A = a^{(1)}$, i.e., $\{\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(1)}}\} \sim F(M_{1a^{(1)}}, \ldots, M_{pa^{(1)}})$, where $F(X)$ denotes a cumulative distribution function for X. Potential outcomes where the mediators are set to random draws from the joint counterfactual distribution are henceforth denoted by (curly) brackets in the subscripts. Let $Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(p)}}}$ denote the potential outcome under treatment $A = a^{(0)}$, when the value of each mediator is set to a random draw from the *marginal* (counterfactual) distribution that does not depend on any other mediator, i.e., $\tilde{M}_{sa^{(s)}} \sim F(M_{sa^{(s)}}), s = 1, \ldots, p$. Let L denote all observed baseline (i.e., unaffected by treatment) covariates that may affect

any of $(A, M_1, \ldots, M_p, Y)$. Here and throughout, the joint and all marginal distributions for the (counterfactual) mediators are unconditional on L. The average potential outcomes (hereafter termed "estimands") are respectively defined as:

$$\mathrm{E}\left(Y_{a^{(0)}\{\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(1)}}\}}\right) = \int \mathrm{E}(Y_{a^{(0)}m_1\cdots m_p})\, dF_{M_{1a^{(1)}},\cdots,M_{pa^{(1)}}}(m_1,\cdots,m_p), \quad \text{and} \tag{1}$$

$$\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(p)}}}\right) = \int \mathrm{E}(Y_{a^{(0)}m_1\cdots m_p})\, dF_{M_{1a^{(1)}}}(m_1)\cdots dF_{M_{pa^{(p)}}}(m_p). \tag{2}$$

2.1 *Interventional (in)direct effects for multiple mediators*

The interventional effects comparing estimands (1) and (2) under different hypothetical treatment levels for a binary treatment A are defined as follows. Let g denote a user-specified link function, such as the log link $g(x) = \log(x)$, or the "logit" link $g(x) = \log\{x/(1-x)\}$. Define the total effect as:

$$g\left\{\mathrm{E}\left(Y_{1\{\tilde{M}_{11}\cdots\tilde{M}_{p1}\}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{10}\cdots\tilde{M}_{p0}\}}\right)\right\},$$

which can be decomposed into the direct effect, defined as:

$$g\left\{\mathrm{E}\left(Y_{1\{\tilde{M}_{11}\cdots\tilde{M}_{p1}\}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{11}\cdots\tilde{M}_{p1}\}}\right)\right\}, \tag{3}$$

and the (joint) indirect effect, defined as:

$$g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{11}\cdots\tilde{M}_{p1}\}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{10}\cdots\tilde{M}_{p0}\}}\right)\right\}. \tag{4}$$

The indirect effect via each mediator $M_s, s = 1, \ldots, p$, is defined as:

$$g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{s-1,0}\,\tilde{M}_{s1}\,\tilde{M}_{s+1,0}\cdots\tilde{M}_{p0}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{p0}}\right)\right\}. \tag{5}$$

Changing only the *marginal* counterfactual distribution of M_s from treatment to control implies an "overall" effect of treatment on M_s that is unconditional on any other mediators. This effect therefore marginalizes over all (underlying) causal effects along (unknown) paths from treatment to the mediator in question that may intersect other causally antecedent mediators. Hence the interventional indirect effect through a mediator can be readily interpreted as the combined causal effect along all (underlying) paths from treatment to the mediator in question, then directly to the outcome. Continuing the example shown in Figure 1, the

indirect effect through M_3 is the combined effect along the paths $A \to M_3 \to Y$ and $A \to M_1 \to M_3 \to Y$. Further interpretations are deferred to the simulation studies in Section 4 and the illustration in Section 5.

Note that in the first term of (5), only the mediator in question M_s is drawn from its distribution under treatment $a^{(s)} = 1$; all other mediators are drawn from their distributions under control, i.e., $a^{(k)} = 0$ for $k = 1, \ldots, p, k \neq s$. This definition of the indirect effect (5) is especially amenable to settings with high-dimensional mediators because it does not require fixing the hypothetical treatments for the other mediators at different levels, and is hence invariant to the chosen (possibly arbitrary) ordering of the mediator indices. In contrast, existing definitions (Vansteelandt and Daniel, 2017; Loh et al., 2019) set $a^{(k)} = 1$ for $k = 1, \ldots, s-1$, and $a^{(k)} = 0$ for $k = s+1, \ldots, p$, so that different definitions are possible depending on the mediator indices, although the conceptual interpretations remain the same.

Lastly, the difference between the joint indirect effect (4) and the sum of the separate indirect effects (5) for all mediators can be further partitioned into:

$$\begin{aligned} &\left[g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{11}\cdots\tilde{M}_{p1}\}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{11}\cdots\tilde{M}_{p1}}\right)\right\}\right] \\ &\quad - \left[g\left\{\mathrm{E}\left(Y_{0\{\tilde{M}_{10}\cdots\tilde{M}_{p0}\}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{p0}}\right)\right\}\right], \end{aligned} \tag{6}$$

and

$$\begin{aligned} &\left[g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{11}\cdots\tilde{M}_{p1}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{p0}}\right)\right\}\right] \\ &\quad - \sum_{s=1}^{p}\left[g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{s-1,0}\,\tilde{M}_{s1}\,\tilde{M}_{s+1,0}\cdots\tilde{M}_{p0}}\right)\right\} - g\left\{\mathrm{E}\left(Y_{0\tilde{M}_{10}\cdots\tilde{M}_{p0}}\right)\right\}\right]. \end{aligned} \tag{7}$$

Following Vansteelandt and Daniel (2017), we refer to (6) as the indirect effect via the mediators' *mutual dependence*. This indirect effect describes how treatment affects the relationships between the mediators, which subsequently affects the outcome, and thus cannot be attributed to any single mediator. We term the indirect effect in (7) as the "remainder" effect after removing the indirect effect via the mutual dependence (6) from the difference between the joint indirect effect (4) and the sum of the separate indirect effects (5). Interpretations

and closed-form expressions for the indirect effects (6) and (7), under the setting with two mediators and (non-)linear models for the mediators (and outcome), are provided in the Web Appendix. For example, when the mean model for the outcome is linear and the mediators are normally distributed, the indirect effect via the mediators' mutual dependence (6) is non-zero if and only if the covariance of the mediators is affected by treatment and the mediator-mediator interaction effect on the outcome is non-zero. When the mean model for the outcome is log-linear and the mediators are normally distributed, this indirect effect is non-zero if and only if the covariance of the mediators is affected by treatment and the main effects of the mediators on the outcome are both non-zero.

2.2 *Assumptions for identification*

Identification of the interventional effects requires the following assumptions, where "$\perp\!\!\!\perp$" denotes conditional independence:

$$Y_{am_1\cdots m_p} \perp\!\!\!\perp A|L \quad \forall\, a, m_1, \ldots, m_p; \tag{8}$$

$$Y_{am_1\cdots m_p} \perp\!\!\!\perp \{M_1 \cdots M_p\}|(A = a, L) \quad \forall\, a, m_1, \ldots, m_p; \tag{9}$$

$$\{M_{1a} \cdots M_{pa}\} \perp\!\!\!\perp A|L \quad \forall\, a. \tag{10}$$

Assumptions (8) and (10) state that the effect of treatment A on outcome Y, and the effects of treatment A on all mediators, are unconfounded conditional on L. Assumption (9) states that there is sufficient information observed in L so that the association between any of the mediators $(M_1, \ldots, M_p)$ and outcome Y is unconfounded within levels of the covariates L. Because this assumption is not empirically testable, all baseline measurements of the mediators and the outcome (prior to treatment being received) should be adjusted for in practice, even when treatment is randomly assigned. We refer readers to Smith and VanderWeele (2019) for implications when this assumption is violated, and recommended sensitivity analyses.

3. Estimation of interventional (in)direct effects

In this section we develop a novel estimation strategy that requires only an outcome model, and no models for the mediators. The strategy exploits the proposed definitions of the interventional effects using (counterfactual) mediator distributions that are unconditional on any covariates. For pedagogic purposes, we will first assume that there are sufficient observations for an unpenalized outcome model, conditional on treatment, all mediators, and covariates, to be fitted to the observed data. We will further assume that treatment is randomly assigned so that both independence assumptions (8) and (10) are satisfied unconditionally on L. Extensions to high-dimensional mediators, and observational studies with non-randomly assigned treatments, are presented in later sections.

3.1 *Randomly assigned treatment with an unpenalized outcome model*

Estimators of the proposed interventional (in)direct effects are obtained as follows:

A0. Fit an outcome model, conditional on treatment, mediators, and covariates, to the observed data, e.g., $\mathrm{E}(Y|A, M_1, \ldots, M_p, L)$. The outcome model can be expressed as a function of its inputs, e.g., $\mathrm{E}(Y|A = a, M_1 = m_1, \ldots, M_p = m_p, L) = h(a, m_1, \ldots, m_p, L)$, where $h(\cdot)$ is a user-specified function. The observed values of the covariates L for each individual are assumed to be fixed and invariant to different values of the treatment a and counterfactual mediator values (denoted by $m_1, \ldots, m_p$). Denote the estimated function by $\hat{h}(a, m_1, \ldots, m_p, L)$.

A1. Construct the duplicated data for each individual as shown in Table 1. The hypothetical treatment levels $a^{(0)}$ and $a^{(1)}$ are chosen so that the interventional direct effect (3) is the difference between the (transformed) estimands in the last and penultimate rows, and the joint indirect effect (4) is the difference between the (transformed) estimands in the penultimate and first rows.

[Table 1 about here.]

A2. For each row of Table 1, randomly draw the counterfactual mediator values $\{\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(1)}}\}$

jointly from the observed treatment group $A = a^{(1)}$. Because treatment is randomly assigned, there are no confounders of the treatment and the (counterfactual) mediators, and assumption (10) is satisfied unconditionally on L. Randomly select an individual whose observed treatment is $A = a^{(1)}$, then set the counterfactual mediators $\{\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(1)}}\}$ to the selected individual's observed values $\{M_1 \cdots M_p\}$.

A3. Impute the expected potential outcomes as predictions $\hat{h}(a^{(0)}, \tilde{M}_{1a^{(1)}}, \ldots, \tilde{M}_{pa^{(1)}}, L)$ from the fitted outcome model in step A0.

A4. Repeat steps A2 and A3 to account for the variability in the (counterfactual) mediator values, thereby obtaining the (Monte Carlo averaged) imputed potential outcomes $\mathrm{E}\left(Y_{a^{(0)}\{\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(1)}}\}} | L\right)$ for each individual.

A5. For each unique value of $\{a^{(0)}, a^{(1)}\}$ in Table 1, calculate the average imputed potential outcome $\mathrm{E}\left(Y_{a^{(0)}\{\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(1)}}\}}\right)$ across all individuals in the observed sample, which in doing so, averages over the empirical distribution of the covariates L. The estimators of the direct effect (3) and joint indirect effect (4) are obtained by plugging in the sample averages for the unknown (population) quantities.

Next, we estimate the separate indirect effects via each mediator. In general, the potential outcome $Y_{a^{(0)}\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(p)}}}$ is unobservable, even when the hypothetical treatments all equal the observed treatment (i.e., $a^{(0)} = a^{(1)} = \ldots = a^{(p)} = A$). This is because each counterfactual mediator $\tilde{M}_{sa^{(s)}}$, even under the observed treatment $a^{(s)} = A$, has to be drawn from its marginal (counterfactual) distribution that does not depend on any other mediators by definition. In contrast, the observed values $\{M_1 \cdots M_p\}$ are jointly distributed according to some (unknown) distribution for all the mediators. Estimating the indirect effects via each mediator therefore requires randomly sampling (counterfactual) mediator values from their marginal distributions, and proceeds as follows:

B1. Construct the duplicated data for each individual as shown in Table 2. Set all hypothetical

treatments in the first row to 0; i.e., $a^{(0)} = a^{(1)} = \ldots = a^{(p)} = 0$. For $s = 1, \ldots, p$, set the hypothetical treatments in row $s+1$ to those in the first term of the interventional indirect effect via mediator M_s as defined in (5); i.e., $a^{(s)} = 1$ in row $s+1$, and 0 otherwise. The hypothetical treatment levels are chosen so that the interventional indirect effect via each mediator M_s corresponds to the difference between the (transformed) estimands in rows $s+1$ and 1. In the last row, set $a^{(0)} = 0$ and $a^{(1)} = \ldots = a^{(p)} = 1$. The interventional indirect effect via the mediators' mutual dependence (6) is thus the difference between (i) the difference in (transformed) estimands in the penultimate and first rows of Table 1, and (ii) the difference in (transformed) estimands in the last and first rows of Table 2.

[Table 2 about here.]

B2. In each row of Table 2, for column $s = 1, \ldots, p$, randomly sample the counterfactual mediator $\tilde{M}_{sa^{(s)}}$, unconditionally on the other mediators, from the observed treatment group $A = a^{(s)}$. This can be carried out by randomly selecting an individual whose observed treatment is $A = a^{(s)}$, then setting the counterfactual mediator $\tilde{M}_{sa^{(s)}}$ to the selected individual's observed value M_s. Because the randomly assigned treatment is jointly independent of all the (counterfactual) mediators when assumption (10) holds, it implies that the treatment is marginally independent of each (counterfactual) mediator, unconditionally on L.

B3. Impute the expected potential outcomes as a prediction $\hat{h}(0, \tilde{M}_{1a^{(1)}}, \ldots, \tilde{M}_{pa^{(p)}}, L)$ from the fitted outcome model in step A0.

B4. Repeat steps B2 and B3 to account for the variability in the (counterfactual) mediator values, thereby obtaining the (Monte Carlo averaged) imputed potential outcomes $\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(1)}}}|L\right)$ for each individual.

B5. For each unique value of $\{a^{(0)}, a^{(1)}, \ldots, a^{(p)}\}$ in Table 2, calculate the average imputed potential outcome $\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(p)}}}\right)$ across all individuals in the observed sample. The estimators of the interventional indirect effect via each mediator M_s (5) are obtained by

plugging in the sample averages for the unknown (population) quantities. The estimator of the indirect effect via the mediators' mutual dependence (6) is similarly obtained using the sample averages in the second and first rows of Table 1, and the last and first rows of Table 2.

3.1.1 *Remarks.* Unbiased estimation of the interventional effects therefore depends on correctly specifying an outcome model conditional on treatment, mediators, and covariates that is unbiased for the mean model for the potential outcomes; i.e., $\hat{h}(a, m_1, \ldots, m_p, L = l)$ converges in probability to $\mathrm{E}\left(Y_{am_1\cdots m_p}|L = l\right)$ for each (observed) value of l. The consistency of the estimator under the identifying assumptions (8)–(10) is shown in Web Appendix A. Standard errors can be estimated using a non-parametric percentile bootstrap procedure (Efron and Tibshirani, 1994) that randomly resamples observations with replacement, then repeating steps A0 – A5 and B1 – B5 for each bootstrap sample.

In principle, when given the mediators' causal ordering, the mediators' joint probability density function can be factorized into a product of the individual mediator density functions, conditional on their causally antecedent mediators. The counterfactual mediators can then be randomly sampled using these conditional distributions in steps A2 and B2 of the proposed estimation procedure. We emphasize that the proposed estimation strategy remains valid even when the (counterfactual) mediators' joint density may be factorized according to an incorrectly specified causal ordering. Nonetheless, a potential shortcoming is the imposition of parametric models for each individual mediator distribution conditional on other mediators. Using parametric models may conceptually resemble existing methods (Vansteelandt and Daniel, 2017; Lin and VanderWeele, 2017), albeit targeting different definitions of interventional indirect effects as described earlier in this article. But as we will describe later in Section 6, the interventional effects proposed in this article are equivalent to existing definitions by Vansteelandt and Daniel (2017) under most common settings. Finally, we reiterate that the proposed estimators of interventional indirect effects put forth in this article

draw the (counterfactual) mediators from their marginal distributions, as in VanderWeele and Tchetgen Tchetgen (2017). This permits the proposed estimation strategy using non-parametric estimates of the mediator distributions, thus avoiding (i) assumptions on the causal structure among the mediators, and (ii) modeling the mediators' joint distribution.

3.2 *High-dimensional mediators*

In this section, we consider high-dimensional mediation settings with fewer observations than mediators (and covariates). Separate mediation analyses can lead to biased estimates of indirect effects because merely omitted mediators can act as unobserved (post-treatment) confounders of the mediator(s) in question and the outcome, thus violating assumption (9). Recall the example in Figure 1, where M_1 and M_2 are confounders of the $M_3 - Y$ relation. A single mediation analysis for M_3 that ignores either M_1 or M_2, or both, will yield biased estimates of the indirect effect via M_3. In this section, we propose a strategy to estimate the separate indirect effects via each mediator, by focusing on each distinct mediator $M_s, s = 1, \ldots, p$, in turn. For the mediator M_s in question, possible confounding of the mediator-outcome relation (by other mediators or possibly high-dimensional covariates) is carefully adjusted for to ensure unbiased estimation of the interventional indirect effect.

3.2.1 *Outcome model.* First, fit a penalized regression model for the outcome to the observed data. Here and throughout, we will consider an elastic net penalty (Zou and Hastie, 2005), which is a compromise between the ridge regression penalty and the lasso penalty. The elastic net penalty is especially useful when there are many correlated predictor variables, and has been applied to different gene expression datasets with highly correlated genes (Friedman et al., 2010). Because evaluating the indirect effect via M_s requires assessing the associations between M_s and Y, (prematurely) shrinking the coefficient of M_s to zero can lead to biased inference. The coefficient for M_s is therefore better left unpenalized, e.g., by setting its penalty factor to zero, to retain M_s in the outcome model. Denote the subset

of the remaining mediators and covariates with non-zero (penalized) coefficients by $\mathcal{M}_s$, where the subscript emphasizes the focus on mediator M_s. Denote the resulting (penalized) outcome model by $h(A, M_s, \mathcal{M}_s)$, where the dependence on only the selected variables in $\mathcal{M}_s$ is explicit.

3.2.2 *Indirect effect via each distinct mediator.* To estimate the indirect effect via M_s, carry out steps B1 – B5 using only rows $s+1$ and 1 of Table 2. Predictions from the (penalized) outcome model $h(A, M_s, \mathcal{M}_s)$ fitted to the observed data are used to impute potential outcomes in step B4.

3.3 *Observational studies with non-randomly assigned treatments*

When treatment is not randomly assigned, the counterfactual mediator values used to construct the duplicated data in the estimation procedure cannot be sampled by merely selecting mediator values at random within each observed treatment group. The observed baseline confounders L of the treatment-mediator(s) and treatment-outcome relations have to be adjusted for toward ensuring that the identifying assumptions (8) and (10) hold. For example, the counterfactual density for mediator M_s when setting A to a is:

$$\begin{aligned} f(M_{sa}) &= \int f(M_s|A=a, L=l) f(L=l) f(L=l|A=a)^{-1}\, dF_{L|A=a}(l) \\ &= \Pr(A=a) \int f(M_s|A=a, L=l) \Pr(A=a|L=l)^{-1}\, dF_{L|A=a}(l), \quad a=0,1. \end{aligned}$$

The above result motivates a modified estimation procedure for non-randomly assigned treatments, by sampling the (observed) mediator values with probability proportional to the inverse of the conditional probability of receiving the observed treatment given the confounders. For example, suppose that $a^{(s)} = 0$. Consider an individual in the treatment group $A = 0$ with covariate values $L = l$ and observed mediator value M_s. The sampling probability of this particular individual's value of M_s is $\Pr(A=0|L=l)^{-1} / \{\sum_i \Pr(A=0|L=l^i)^{-1} \mathbb{1}(A^i=0)\}$, where the sum in the denominator is over all individuals i (as indexed

by the i superscripts without brackets) with covariate values $L = l^i$ in the (same) treatment group $A^i = 0$. The (counterfactual) mediator value(s) can be readily sampled by randomly selecting an individual from the treatment group $A = a^{(s)}$ according to the aforementioned sampling probabilities, then setting the counterfactual mediator to the selected individual's observed mediator value(s). In practice, the probabilities of the observed treatments $\Pr(A = a|L), a = 0, 1$, may be estimated using predictions from a (saturated) logistic regression model fitted to the observed data, with the treatment as the dependent variable, and main effects (and when feasible, higher order and interaction effects) for the covariates as the predictors.

4. Simulation studies

Two simulation studies were conducted to empirically assess the operating characteristics of the proposed estimation strategy in finite samples across different settings. Details of the procedures and results of these simulation studies are deferred to Web Appendix B. To provide an overview, in study 1, we considered a setting with two (continuous) mediators where one was causally affected by the other. The estimators of the interventional effects proposed in this article were compared with existing estimators of natural effects (Steen et al., 2017) that required stricter identification assumptions. We considered an extensive range of settings where the causal ordering of the mediators was either correctly or incorrectly assumed, and unobserved confounding of the mediators was either present or absent. In study 2, settings with high-dimensional (non-continuous) mediators were considered for assessing the indirect effect estimators proposed in Section 3.2. We considered binary outcomes in study 1, and both continuous and binary outcomes in study 2.

The results of the simulation studies showed that estimators of the existing natural indirect effects were unbiased only when the identifying assumptions were met; i.e., the mediators' causal ordering was correctly assumed, and there was no unobserved confounding of the

mediators. As expected, when at least one of the assumptions was violated, estimates were biased even at large sample sizes. In contrast, estimators of the interventional effects proposed in this article were unbiased under all the considered settings. The proposed estimation strategy for high-dimensional mediators was able to detect the mediators through which the (non-zero) indirect effects were transmitted with high probability (90% or above) empirically. In general, the interventional indirect effects were unbiasedly estimated, with empirical coverage of the confidence intervals at approximately their nominal levels. However, under extreme scenarios when (i) the mediators were strongly associated with each other, and (ii) true mediators were only indirectly affected by treatment (via other mediators), and (iii) the outcome was binary, the estimated indirect effects were empirically biased in finite samples. The biases were due to the penalized logistic regression model under-selecting the relevant predictors which were strongly correlated with irrelevant predictors, resulting in inadequate adjustment for common causes (and intermediate variables) of a mediator's effect on the outcome. Nonetheless, the biases shrunk to zero empirically when the unpenalized MLE of the outcome model (which did not require variable selection) was used to estimate the indirect effects. More complex working solutions to reduce such finite sample biases are described in the Discussion section and deferred to future work.

5. Illustration with an example dataset

We illustrate the proposed estimation strategy using publicly available data from a previous high-dimensional mediation analysis by Huang and Pan (2016). Huang and Pan assessed whether the causal effect of the microRNA miR-223 expression (the treatment of interest) on mortality within three months due to glioblastoma multiforme (GBM), a malignant brain tumor, was mediated by expression values of different genes in the tumor genome. Expression values for $p = 1220$ genes (with no missing data) from 490 patients suffering from GBM were included in the online supplemental materials of Huang and Pan (2016). Interventional effects

are suitable for such high-dimensional mediator settings because the mediators' internal causal structure, such as gene regulation networks or protein signaling networks, are often unknown in practice. For the purposes of illustrating the proposed estimation strategy, we made the following simplifying assumptions. We dichotomized the miR-223 expression at the empirical median so that $A = 1$ if the expression was above the median, or 0 otherwise. We used a dichotomous indicator of whether death from GBM had occurred in the first three months as the outcome (Y). Among the 490 patients, 44 died within the first three months ($Y = 1$). We assumed that the patients lost to follow-up prior to three months (four in the $A = 0$ group, and five in the $A = 1$ group) were alive ($Y = 0$). We assumed that the baseline demographic variables in the publicly available data (age at diagnosis, gender, and ethnicity), which we jointly denoted by L, were sufficient for the identifying assumptions (8)–(10) to hold. In practice, a richer set of (baseline) covariates should be adjusted for in substantive analyses to reduce the possibility of biases due to unobserved confounding.

The proposed estimation strategy for high-dimensional mediators described in Section 3.2 was carried out. Estimating each indirect effect required fitting only an outcome model to the observed data, and did not require assuming any model for the mediators. For each mediator $M_s, s = 1, \ldots, p$, in turn, a penalized logistic regression model for the binary outcome using an elastic net penalty was fitted using the `glmnet` package (Friedman et al., 2010) in `R`. The value of the tuning parameter that controlled the overall strength of the penalty in each candidate model was selected using n-fold (or "leave-one-out") cross-validation on the misclassification error. The default sequence of models in the `cv.glmnet` function employs $K = 100$ (equally-spaced) values of the penalty between $\lambda_{min} = \epsilon\lambda_{max}, \epsilon = 10^{-4}$, and λ_{max} (on the log scale), with λ_{max} computed using the empirical data (Friedman et al., 2010). To ensure a wider spread of models and to improve the convergence properties for a nonlinear model, we instead set $K = 200$ and $\lambda_{min} = \epsilon p/n, \epsilon = 10^{-6}$. We further forced treatment (A),

the baseline covariates (L), and the mediator in question (M_s) into the outcome model by setting their penalty factors to zero. All other arguments were set to their default values. Because the resulting penalized coefficient estimates may be biased (Hastie et al., 2009, page 91), we refitted the (un)penalized regression model to the selected set of predictors (with non-zero coefficient estimates in the previous step) following the "relaxed LASSO" approach (Meinshausen, 2007) to obtain the estimated outcome model.

Because treatment was not randomly assigned, counterfactual mediator values were sampled using the proposed weights in Section 3.3 under a saturated logistic regression model for treatment conditional on the baseline covariates. For each mediator, we plotted the estimated coefficient of treatment on the mediator in question in a simple linear regression, and the estimated coefficient of the mediator in the (penalized) logistic outcome model, in Figure 2. Each panel corresponded to one of the nine (overlapping) sets of mediators defined in Table 2 of Huang and Pan (2016). We refer readers to Huang and Pan for details of the biological functions of each gene set.

[Figure 2 about here.]

Non-parametric 95% percentile bootstrap confidence intervals (CIs) were constructed using 2000 bootstrap samples that randomly resampled observations with replacement and repeated the estimation procedure for each sample. Genes whose 95% CIs excluded zero are displayed in Table 3. The interventional indirect effect via an individual gene can be interpreted using the following example. The odds of death from GBM within three months due to shifting the marginal distribution of the CLEC5A gene expression from those with observed miR-223 expressions above the median to those below the median, while holding the distributions of all other gene expressions fixed among those with miR-223 expressions below the median, was estimated to increase by $\exp(0.34) \approx 1.40$ times (95% CI $= (1.13, 1.97)$). To compare our results with those in Huang and Pan, we indicated in Table 3 which

gene set(s) each distinct gene belonged to. Huang and Pan found that all nine gene sets significantly mediated the causal effect of miR-223 expression on GBM mortality within three months (with p-values less than 0.05). We (similarly) found that each gene set contained at least one gene with a non-zero interventional indirect effect whose 95% CI excluded zero. Because several simplifying assumptions were made for the sole purpose of illustrating the proposed estimation strategy, substantive conclusions about the biological significance of the statistically significant mediators are beyond the scope of this article.

[Table 3 about here.]

6. Comparison

In this section, we argue that in common settings, the interventional effects proposed in this article are equal to existing definitions by Vansteelandt and Daniel (2017). In particular, the existing definitions rely on (counterfactual) mediator distributions that depend on treatment and all baseline covariates, including confounders of the mediator(s)-outcome relation(s). Suppose that all mediators are continuous, so that linear mean models for the mediators (conditional on L) may be assumed. Suppose that the outcome is dichotomous, and sufficiently rare so that the logit link in a mean model for the outcome can be approximated by the log link. When (i) the outcome model has main effects only, and (ii) the effect of treatment on the mediators is not moderated by the confounders (i.e., there are no interaction terms involving treatment in the mediator mean models), the proposed interventional effects are equal to existing definitions. Detailed results are provided in Web Appendix C.

7. Conclusion

In this article, we have developed interventional effects for high-dimensional mediators that permit a novel estimation strategy allowing for any mean model for the outcome, and requiring no models for the mediators. There are several avenues of possible future related research. The proposed estimation strategy for high-dimensional mediators is designed to deliver

reasonable approximations of the indirect effects through each distinct mediator, by selecting other variables (including mediators) that may be confounders of the mediator-outcome relation in question. To increase the chances of detecting such confounders, a separate (penalized regression) model for the mediator, conditional on all other mediators, treatment, and covariates, can be employed. The selected predictors (with non-zero coefficients) in either the penalized outcome model, or the penalized mediator model, can then be included in a new (unpenalized) outcome model to impute potential outcomes for estimating the interventional indirect effects. The use of separate outcome and mediator models is inspired by *double selection* principles (Belloni et al., 2014), where in settings without mediators, the (partial) associations between the outcome and the covariates, and between the treatment and the covariates, are evaluated. We acknowledge that specifying mediator models may appear contradictory to the key feature of our proposal that avoids modeling the relations between the mediators. But we emphasize that the (penalized) regression model for each mediator in question is used merely to select possible confounders of the mediator-outcome relation, by essentially viewing the other mediators as possible confounders. These models do not assume (or empirically imply) any causal structure among the mediators, and do not require specifying a model for the joint distribution of the mediators.

When selecting a subset of the mediators for further investigation, a threshold that either is pre-determined, such as in Huang and Pan (2016), or controls the familywise error rate under multiple testing scenarios, such as in Derkach et al. (2020), can be used. When further interest is in estimating the direct, joint, or mutual indirect effects, a separate penalized regression model for the outcome (conditional on all the mediators, covariates, and treatment) must be fitted to the observed data for predicting the potential outcomes in Table 1, and in the first and last rows of Table 2, of the estimation strategy in Section 3.1. The estimated interventional indirect effects via each distinct mediator can be used to determine

the penalties; e.g., the indirect effects can be used in place of the product of coefficients in the penalties for the pathway lasso (Zhao and Luo, 2016). High-dimensional mediation methods that employ principal components analysis, such as Huang and Pan (2016) and Zhao et al. (2020), can be used to select mediators with high loadings in the principal components. Substantive comparisons of the indirect effects through specific selected (sets of) mediators using such approaches, with the interventional indirect effects developed in this article, may be made in future work. Establishing theoretical properties to ensure valid inference of the proposed Monte Carlo-based estimators, especially following mediator selection, is an area of future development. Unbiased estimation of the interventional effects defined in this article depends on correctly specifying a mean model for the outcome. Vansteelandt et al. (2012) recommend using sufficiently rich outcome models that e.g., include higher-order or interaction terms between treatment and mediators, or between mediators. Assessing robustness to misspecification of the outcome model, e.g., due to omitting such interaction terms, is a direction for future research. Other more general non-parametric prediction methods for the potential outcomes that leverage data-adaptive techniques (Díaz et al., 2019; Benkeser, 2020) can also be considered. Extending such machine learning methods to high-dimensional settings requires further work to avoid the need for inverse weighting by the joint mediator density.

Acknowledgements

The authors would like to thank the Editor, Associate Editor, and two reviewers for their comments on prior versions of this manuscript. This research was supported by the Research Foundation – Flanders (FWO) under Grant G019317N. Computational resources and services were provided by the VSC (Flemish Supercomputer Center), funded by the FWO and the Flemish Government – department EWI. The content is solely the responsibility of the authors and does not represent the official views of the authors' institutions or FWO.

Data Availability Statement

The data used in the illustration in Section 5 is openly available as part of the web supplemental materials (filename: "biom12421-sup-0002-SuppData-Code.zip") of Huang and Pan (2016) at the Biometrics website on Wiley Online Library. The `R` code used to implement the proposed estimation procedure in carrying out the simulation studies in Section 4, and the illustration in Section 5, are available at: `https://github.com/wwloh/interventional-hdmed`

References

Albert, J. M., Cho, J. I., Liu, Y., and Nelson, S. (2019). Generalized causal mediation and path analysis: Extensions and practical considerations. *Statistical Methods in Medical Research* **28,** 1793–1807.

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

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81,** 608–650.

Benkeser, D. (2020). Nonparametric inference for interventional effects with multiple mediators. *arXiv preprint arXiv:2001.06027* .

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

Daniel, R., De Stavola, B., Cousens, S., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71,** 1–14.

Derkach, A., Moore, S. C., Boca, S. M., and Sampson, J. N. (2020). Group testing in mediation analysis. *Statistics in Medicine* .

Díaz, I., Hejazi, N. S., Rudolph, K. E., and van der Laan, M. J. (2019). Non-parametric efficient causal mediation with intermediate confounders. *arXiv preprint arXiv:1912.09936* .

Didelez, V., Dawid, A. P., and Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 138–146, Arlington, VA, USA. AUAI Press.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1–22.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72,** 402–413.

Lange, T., Rasmussen, M., and Thygesen, L. C. (2013). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology* **179,** 513–518.

Lin, S.-H. and VanderWeele, T. (2017). Interventional approach for path-specific effects. *Journal of Causal Inference* **5,**.

Loh, W. W., Moerkerke, B., Loeys, T., and Vansteelandt, S. (2019). Disentangling indirect effects through multiple mediators whose causal structure is unknown. *PsyArXiv preprint doi:10.31234/osf.io/3q4hg* .

Lok, J. J. (2019). Causal organic direct and indirect effects: closer to Baron and Kenny. *arXiv preprint arXiv:1903.04697* .

MacKinnon, D. P. (2000). *Contrasts in multiple mediator models*, pages 141–160. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA.

MacKinnon, D. P., Valente, M. J., and Gonzalez, O. (2018). The correspondence between causal and traditional mediation analysis: the link is the mediator by treatment interaction. *Prevention Science* pages 1–11.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52,** 374–393.

Moreno-Betancur, M. and Carlin, J. B. (2018). Understanding interventional effects: A more natural approach to mediation analysis? *Epidemiology* **29,** 614–617.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Preacher, K. J. and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods* **40,** 879–891.

Quynh Nguyen, T., Schmid, I., and Stuart, E. A. (2019). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *arXiv preprint arXiv:1904.08515* .

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3,** 143–155.

Smith, L. H. and VanderWeele, T. J. (2019). Mediational e-values: approximate sensitivity analysis for unmeasured mediator–outcome confounding. *Epidemiology* **30,** 835–837.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2017). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology* **186,** 184–193.

Taguri, M., Featherstone, J., and Cheng, J. (2018). Causal mediation analysis with multiple

causally non-ordered mediators. *Statistical Methods in Medical Research* **27,** 3–19.

VanderWeele, T. J. (2013). Policy-relevant proportions for direct effects. *Epidemiology* **24,** 175–176.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 917–938.

VanderWeele, T. J. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods* **2,** 95–115.

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

Vansteelandt, S., Bekaert, M., and Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods* **1,** 131–158.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology* **28,** 258–265.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32,** 3150–3154.

Zhao, Y., Li, L., and Caffo, B. S. (2019). Multimodal neuroimaging data integration and pathway analysis. *arXiv preprint arXiv:1908.10925* .

Zhao, Y., Lindquist, M. A., and Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Computational Statistics & Data Analysis* **142,** 106835.

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749* .

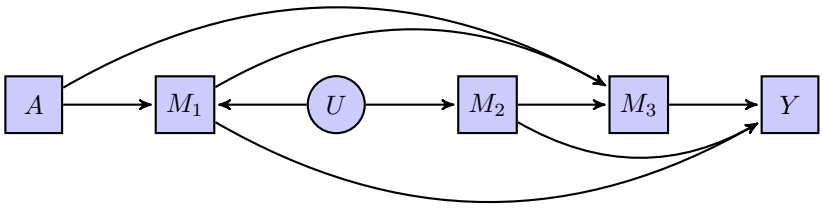
Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net.

Journal of the royal statistical society: series B (statistical methodology) **67,** 301–320.

Supporting Information

Web Appendices A to C referenced in Sections 2 to 6 are available with this paper at the Biometrics website on Wiley Online Library.

Figure 1. Causal diagram for an example with three mediators, two of which share an unobserved confounder U. Observed variables are drawn as rectangular nodes and unobserved variables are drawn as round nodes.



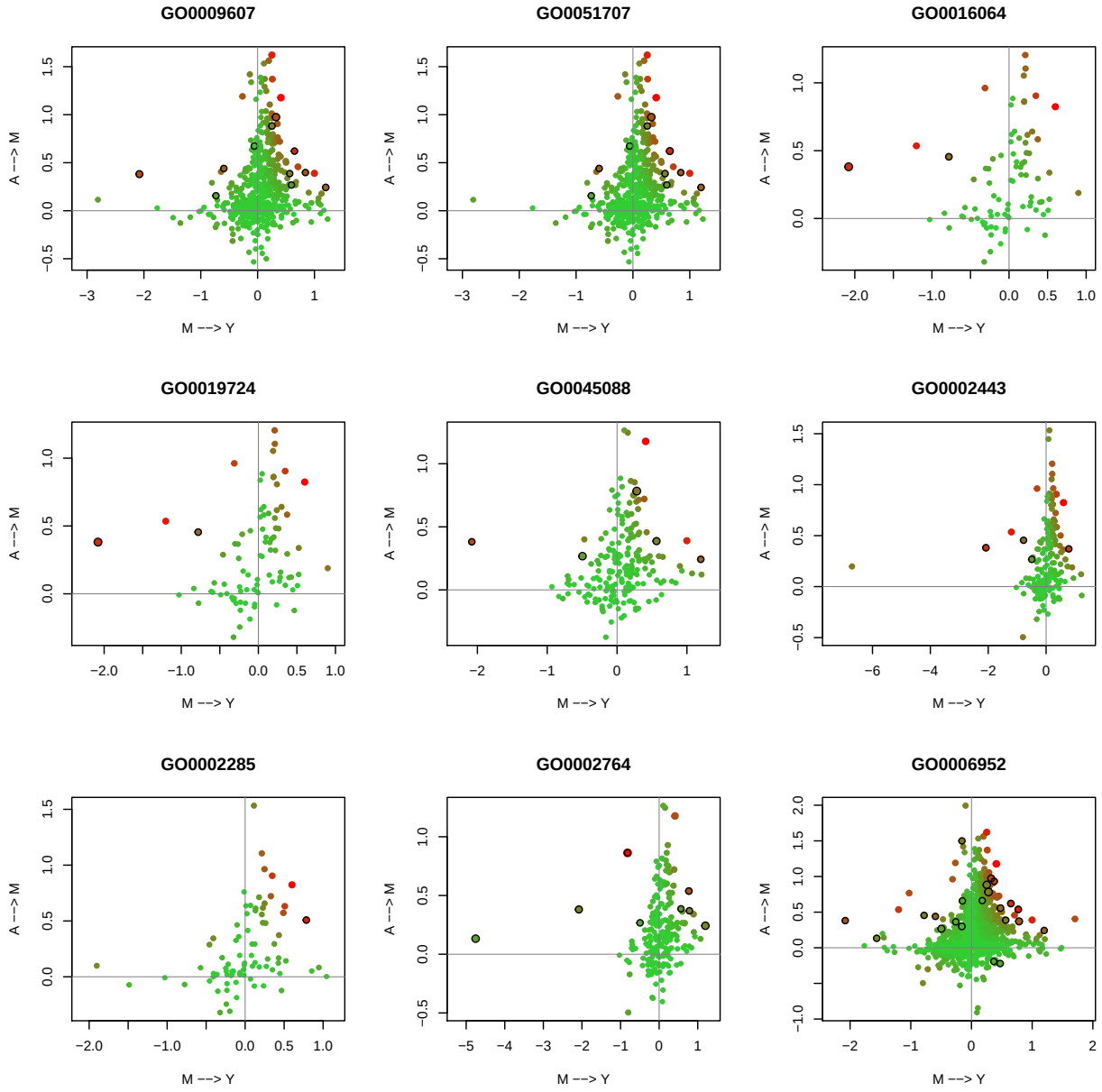


Figure 2. Scatterplots of the (penalized) regression coefficients in the linear mediator and logistic outcome models for the GBM data set. Each point corresponds to a distinct mediator, with the coefficient of treatment on the mediator in question on the vertical axis ("$A \to M$"), and the coefficient of the mediator on the outcome on the horizontal axis ("$M \to Y$"). The size and color of the points are proportional to the (absolute) magnitudes of the indirect effects, with larger points and darker (red) hues indicating indirect effect estimates further from zero. Mediators whose 95% CIs excluded zero are circled in black. Each panel corresponds to one of the nine sets of mediators described in Huang and Pan. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1

Duplicated data for each individual when estimating the direct and joint indirect effects. The counterfactual mediators $\{\tilde{M}_{1a^{(1)}} \cdots \tilde{M}_{pa^{(1)}}\}$ *are randomly drawn (jointly) from the observed treatment group* $A = a^{(1)}$*. The column* L *is omitted for simplicity.*

$a^{(0)}$	$a^{(1)}$	$\{\tilde{M}_{1a^{(1)}} \quad \cdots \quad \tilde{M}_{pa^{(1)}}\}$	$\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(1)}}}\vert L\right)$
0	0	$\{\tilde{M}_{10} \quad \cdots \quad \tilde{M}_{p0}\}$	$\hat{h}(0, \tilde{M}_{10}, \ldots, \tilde{M}_{p0}, L)$
0	1	$\{\tilde{M}_{11} \quad \cdots \quad \tilde{M}_{p1}\}$	$\hat{h}(0, \tilde{M}_{11}, \ldots, \tilde{M}_{p1}, L)$
1	1	$\{\tilde{M}_{11} \quad \cdots \quad \tilde{M}_{p1}\}$	$\hat{h}(1, \tilde{M}_{11}, \ldots, \tilde{M}_{p1}, L)$

Table 2

Duplicated data for each individual when estimating the separate indirect effects via each mediator. There are $t+2$ rows for a binary treatment A. The counterfactual mediators $\tilde{M}_{sa^{(s)}}$ are randomly drawn from the observed treatment group $A = a^{(s)}, s = 1, \ldots, p$. The column L is omitted for simplicity.

$a^{(0)}$	$a^{(1)}$	$a^{(2)}$	$\cdots$	$a^{(p-1)}$	$a^{(p)}$	$\tilde{M}_{1a^{(1)}}$	$\tilde{M}_{2a^{(2)}}$	$\cdots$	$\tilde{M}_{p-1,a^{(p-1)}}$	$\tilde{M}_{pa^{(p)}}$	$\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}}\cdots\tilde{M}_{pa^{(p)}}}\|L\right)$
0	0	0	$\cdots$	0	0	$\tilde{M}_{10}$	$\tilde{M}_{20}$	$\cdots$	$\tilde{M}_{p-1,0}$	$\tilde{M}_{p0}$	$\hat{h}(0, \tilde{M}_{10}, \ldots, \tilde{M}_{p0}, L)$
0	1	0	$\cdots$	0	0	$\tilde{M}_{11}$	$\tilde{M}_{20}$	$\cdots$	$\tilde{M}_{p-1,0}$	$\tilde{M}_{p0}$	$\hat{h}(0, \tilde{M}_{11}, \ldots, \tilde{M}_{p0}, L)$
0	0	1	$\cdots$	0	0	$\tilde{M}_{10}$	$\tilde{M}_{21}$	$\cdots$	$\tilde{M}_{p-1,0}$	$\tilde{M}_{p0}$	$\hat{h}(0, \tilde{M}_{10}, \ldots, \tilde{M}_{p0}, L)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	0	0	$\cdots$	1	0	$\tilde{M}_{10}$	$\tilde{M}_{20}$	$\cdots$	$\tilde{M}_{p-1,1}$	$\tilde{M}_{p0}$	$\hat{h}(0, \tilde{M}_{10}, \ldots, \tilde{M}_{p0}, L)$
0	0	0	$\cdots$	0	1	$\tilde{M}_{10}$	$\tilde{M}_{20}$	$\cdots$	$\tilde{M}_{p-1,0}$	$\tilde{M}_{p1}$	$\hat{h}(0, \tilde{M}_{10}, \ldots, \tilde{M}_{p1}, L)$
0	1	1	$\cdots$	1	1	$\tilde{M}_{11}$	$\tilde{M}_{21}$	$\cdots$	$\tilde{M}_{p-1,1}$	$\tilde{M}_{p1}$	$\hat{h}(0, \tilde{M}_{11}, \ldots, \tilde{M}_{p1}, L)$

Table 3

Estimates ("Est.") on the log-odds scale, and non-parametric bias-corrected and accelerated bootstrap 95% confidence intervals ("CIs"), for the interventional indirect effects via each distinct gene in the GBM data set. Only genes whose 95% CIs excluded zero are displayed. The genes are ordered by increasing lower bound of their 95% CI. Each gene either belongs ("1") or does not belong ("0") to a gene set described in Huang and Pan. The gene sets in the column headings are labelled as (i) GO0009607, (ii) GO0051707, (iii) GO0016064, (iv) GO0019724, (v) GO0045088, (vi) GO0002443, (vii) GO0002285, (viii) GO0002764, and (ix) GO0006952. All results are rounded to two decimal places.

Gene	Est.	95% CI		(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
FYB	-0.61	-1.18	-0.27	0	0	0	0	0	0	0	1	0
SAA2	-0.22	-0.68	-0.06	0	0	0	0	0	0	0	0	1
ADM	-0.04	-0.56	-0.01	1	1	0	0	0	0	0	0	0
NUPR1	-0.09	-0.53	-0.06	0	0	0	0	0	0	0	0	1
FCN2	-0.17	-0.50	-0.12	0	0	0	0	0	0	0	0	1
IRF5	-0.25	-0.49	-0.07	1	1	0	0	0	0	0	0	1
ARRB2	-0.12	-0.47	-0.04	0	0	0	0	1	1	0	1	1
SPON2	-0.10	-0.45	-0.03	0	0	0	0	0	0	0	0	1
CD226	-0.29	-0.44	-0.10	1	0	1	1	1	1	0	1	1
NCAM1	-0.07	-0.39	-0.02	0	0	0	0	0	0	0	0	1
MAPK8IP2	-0.10	-0.39	-0.02	0	0	0	0	0	0	0	0	1
BTN3A1	-0.18	-0.35	-0.01	0	0	0	0	0	0	0	1	0
ECM1	-0.04	-0.29	-0.00	0	0	0	0	0	0	0	0	1
CYP11A1	-0.10	-0.29	-0.01	1	1	0	0	0	0	0	0	0
C6	-0.18	-0.29	-0.03	0	0	1	1	0	1	0	0	1
CFH	0.11	0.02	0.40	0	0	0	0	0	0	0	0	1
LGALS8	0.32	0.02	0.57	0	0	0	0	0	0	1	0	0
SPHK1	0.23	0.03	0.50	0	0	0	0	0	0	0	0	1
SPP1	0.29	0.03	0.62	0	0	0	0	0	0	0	0	1
SERPINE1	0.29	0.04	0.69	1	1	0	0	0	0	0	0	1
TSPO	0.27	0.05	0.56	1	1	0	0	0	0	0	0	0
PIK3CD	0.27	0.05	0.77	0	0	0	0	0	1	0	1	1
FZD5	0.13	0.06	0.36	1	1	0	0	0	0	0	0	0
IRF1	0.19	0.06	0.49	1	1	0	0	1	0	0	1	1
ITK	0.39	0.06	0.59	0	0	0	0	0	0	0	1	1
MAPK1	0.25	0.06	0.46	1	1	0	0	1	0	0	1	1
A2M	0.21	0.09	0.64	0	0	0	0	1	0	0	0	1
GBP3	0.19	0.12	0.49	1	1	0	0	0	0	0	0	1
CLEC5A	0.34	0.12	0.68	1	1	0	0	0	0	0	0	1