Disentangling indirect effects through multiple 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

Draft version 1.0, September 6, 2019. This paper has not been peer reviewed. Please do not copy or cite without first author's permission.

Abstract

When there are multiple mediators on the causal pathway from a treatment to an outcome, path analysis is commonly used to disentangle the specific indirect effects transmitted along causal path(s) through each distinct mediator. However, fine-grained decompositions of specific indirect or mediated effects along separate paths are only valid under stringent assumptions, such as a correctly specified causal structure of the mediators, and no unobserved confounding of the mediators. In this article, we introduce a new type of direct and indirect effects for multiple mediators, called interventional effects, from the causal inference and epidemiology literature. While the framework for interventional direct and indirect effects can accommodate nonlinear models for the means of the mediators and/or the outcome in general, we will focus on a particular class of linear models widely used for multiple mediation analysis. We demonstrate how the interventional indirect effects through each distinct mediator can be unbiasedly estimated using prevalent path analysis methods within the linear structural equation modeling (SEM) framework. The estimators do not require specifying the directions of the causal effects between the mediators, and are unbiased (in large samples) even when the mediators share hidden or unobserved common causes. The estimation method is utilized to assess the effect of political inclusion on political prejudice that is possibly mediated by six distinct mediators.

Keywords: Direct and indirect effects; Effect decomposition; Interventional effects; Mediation analysis; Path analysis; Specific indirect effects

Disentangling indirect effects through multiple mediators whose causal structure is unknown

Introduction

Mediation analysis is widely used in the behavioral, psychological and social sciences to gain insight into the extent to which the causal effect of a treatment (A) on an outcome (Y) is transmitted through intermediate variables on the causal pathway from A to Y. For example, suppose researchers are investigating the causal effect of a political inclusion manipulation (A) on the level of prejudice toward a political outgroup (Y). Perceived worldview dissimilarity of the political outgroup (M_1) is considered a mediator if the manipulation affects how strongly an individual regards the political outgroup as holding political or social beliefs different from her/his own, which in turn causes a change in prejudice toward that outgroup. Similarly, perceived fairness of the political outgroup (M_2) is also considered a mediator if the manipulation affects how strongly an individual regards the outgroup as being open to different opinions, which in turn causes a change in prejudice toward that outgroup. Here and throughout subscripts in the notation for mediators are merely used to distinguish the different mediators, and not to indicate any assumed causal ordering of the mediators; e.g., M_1 is not necessarily assumed to causally precede M_2. Many mediation analyses involve multiple mediators, either because interventions are designed to affect outcome by changing multiple (repeated measures of) mediators, or because scientific interest is in trying to understand the various causal pathways through (simultaneous) competing candidate mediators. In the presence of multiple or competing mediators, *path analysis* (Wright, 1934) is commonly used to disentangle the *indirect* or *mediated* effects of A on Y along the causal path(s) through each distinct mediator.

Building on our example, the *causal diagram* of Figure 1(a) depicts the causal relations between the variables when worldview dissimilarity (M_1) and fairness (M_2) are independent conditional on A and baseline covariate(s), henceforth denoted by C, such as political ideology. In this article, a causal diagram is a causal directed acyclic graph (DAG) (Pearl, 1995; Hayduk et al., 2003; Pearl, 2012a) that, similar to path diagrams in the structural equation modeling (SEM) framework, represents causal relations among a set of variables. Vertices represent variables, and a directed edge e.g., from M_1 to Y, represents the causal effect M_1 may exert on Y. The absence of a directed edge between two variables, e.g., between M_1 and M_2 in Figure 1(a), implies that neither variable causally affects the other, conditional on the causal ancestors of both variables, e.g., A and C. The key concepts of causal DAGs are summarized in e.g., Moerkerke et al. (2015, Figure 2). It is assumed throughout that any specified causal structure of the variables is based on well-established scientific theoretical knowledge or empirical laws that satisfy logical and causal-temporal constraints such as the Hyman-Tate criterion (Tate, 2015; Fiedler et al., 2018). Unlike path diagrams, causal DAGs do not rely on (parametric) assumptions about the nature of the relationship between the variables; hence path coefficients and error terms are not displayed on causal diagrams in this article.

[Figure 1 about here.]

Mediation using path analysis within the SEM framework has predominantly employed linear regression models for the mediator(s) and the outcome; see e.g., Baron & Kenny (1986), MacKinnon (2008), and Hayes (2018) among many others. A linear path analysis model (or set of linear regression models) is first fitted to the outcome and the mediator(s) using SEM (or ordinary least squares). The *specific* effect along a particular path, as encoded by the (partial) regression coefficients of the variables on the path in question, is then calculated using the *product of coefficients* method (Alwin & Hauser, 1975; MacKinnon et al., 2002). Continuing our example above in the causal diagram of Figure 1(a), when M_1 and M_2 are causally unconnected (conditional on A and C), the specific indirect effect via M_1 in the corresponding (linear) path model is defined to be the product of the coefficient of A in the regression of M_1 on A and C, and the coefficient of M_1 in the regression of Y on A, M_1, M_2 and C. When the specified causal structure of the mediators allows for multiple ("compound") paths from treatment to outcome that intersect a particular mediator, different existing definitions of the indirect effect specific to that mediator may comprise different sets of paths

(Bollen, 1987). For example, Alwin & Hauser (1975) define the specific indirect effect via a mediator of interest to include all paths intersecting that mediator and any of its descendants, and to exclude all paths via mediators that causally precede the mediator of interest. Greene (1977) proposes a more restrictive definition that includes the path that intersects the particular mediator only and no other mediators, whereas Brown (1997), following Fox (1980), proposes a less restrictive definition that includes all paths that intersect the particular mediator. The combined specific effect for a set of paths can be obtained by adding the specific effects for each path. Using the motivating example, suppose that worldview dissimilarity (M_1) is assumed to causally affect fairness (M_2), as depicted in the causal diagram of Figure 1(b). The "three-path" mediated effect passing through both mediators along the path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ (Taylor et al., 2008) is defined to be the product of the coefficient of A in the regression of M_1 on A and C, the coefficient of M_1 in the regression of M_2 on A, M_1 and C, and the coefficient of M_2 in the regression of Y on A, M_1, M_2 and C. Following Alwin & Hauser's definition, the specific indirect effect via M_1 is then the combined effect along the separate paths $A \rightarrow M_1 \rightarrow Y$ and $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ (since M_2 is a causal descendant of M_1). But if fairness (M_2) is assumed to causally affect worldview dissimilarity (M_1) instead, as depicted in the causal diagram of Figure 1(d), then the three-path mediated effect passing through both mediators is along a different path $A \rightarrow M_2 \rightarrow M_1 \rightarrow Y$, and is defined differently. Furthermore, the specific indirect effect via M_1 (following Alwin & Hauser's definition) under this latter path model now consists of the specific effect along the path $A \rightarrow M_1 \rightarrow Y$ alone, since M_1 is now a causal descendant of M_2. Thus specific indirect effects and mediated effects along separate paths that traverse multiple mediators are well-defined exclusively in path models that correctly assume the (directions of the) causal effects among the mediators.

More recently, Hayes (2018) proposes separately assessing each specific indirect effect along a path that passes through at least one of the mediators. But estimates of such specific indirect effects can be biased when there is hidden or unobserved confounding of the mediators. For example, suppose that worldview dissimilarity (M_1)

and fairness (M_2) do not causally depend on each other, but instead share an unobserved common cause U (unaffected by treatment A and uncorrelated with other covariates C), such as prior adverse interactions with a political outgroup, as depicted in the causal diagram of Figure 1(c). Since neither mediator has a causal effect on the other, there is no indirect effect that passes through both mediators. But incorrectly assuming that M_1 causally affects M_2 in a fitted path model will result in biased estimates of the mediated effect along the path $A \to M_1 \to M_2 \to Y$. Specifically, the bias is due to biased estimates of the (partial) regression coefficient of M_1 (in the regression of M_2 on A, M_1 and C) since M_1 and M_2 are correlated (conditional on A and C) only due to unobserved U. Similarly, incorrectly assuming that M_2 causally affects M_1 in a (different) fitted path model will result in biased estimates of the mediated effect along the path $A \to M_2 \to M_1 \to Y$. In general, when multiple mediators are correlated, there may be several different causal mechanisms that are plausible explanations of the associations in the observed data, such as those depicted in the causal diagrams of Figures 1(b), 1(c), and 1(d) for M_1 and M_2. Distinguishing between various (conflicting) causal mechanisms requires scientifically-grounded theory and thoughtful experimentation (Fiedler et al., 2018). In most realistic scenarios the causal structure of the mediators either is unknown or cannot be correctly specified with absolute certainty. Furthermore, when the (repeatedly measured) mediators are manifestations of an underlying latent variable or process (see e.g., Derkach et al. (2019)), it may not be reasonable to assume that the mediators share no unobserved common causes.

A different definition of indirect effects via each mediator

In this article we propose a different definition of indirect effects via each mediator that does not require (correctly) specifying how the mediators causally depend on one another. We first provide an intuitive motivation, and defer its formal derivation to the next section. Suppose that the outcome obeys the following linear and additive mean model:

$$\mathrm{E}(Y|A, M_1, M_2, C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_C C. \tag{1}$$

The effect of each mediator M_s on outcome is encoded by the (partial) regression coefficient $\beta_s, s = 1, 2$ in (1). To avoid specifying any causal dependence among the mediators, consider the *marginal* mean model for each mediator $M_s, s = 1, 2$, that depends only on treatment A and baseline covariate(s) C, and does not depend on the other mediator. In particular, suppose that the linear and additive (marginal) mean model for M_s is:

$$\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, \quad s = 1, 2. \tag{2}$$

It is important to note that (2) does not imply the assumption that M_1 and M_2 are causally independent. Indeed, the intention is simply to leave the causal structure of the mediators unspecified, and to only consider the "overall" (i.e., marginal) effect of treatment A on each mediator $M_s, s = 1, 2$ (conditional on C). The overall effect of A on M_s captures all of the treatment effects that are transmitted through any causal ancestors of M_s (in the underlying causal diagram from which the observed data is generated), and is encapsulated by the (partial) regression coefficient $\delta_s, s = 1, 2$ in (2). Using the product of coefficients method, the *interventional* indirect effect via mediator M_s is encoded by $\beta_s \delta_s$; it is thus agnostic to the underlying causal structure of the mediators.

It may be seen that the mean models (1) and (2) are jointly parametrically equivalent to a (linear) path model corresponding to the causal diagram of Figure 1(a), where the mediators are assumed not to causally affect each other. Such path models (without baseline covariates C) are termed "single-step" or "parallel" multiple mediator models in the SEM literature (Hayes et al., 2011; Hayes, 2018). In fact, a prevalent multiple mediation analysis approach (MacKinnon, 2000; Preacher & Hayes, 2008) advocates fitting a parallel multiple mediator model to the observed data, so that there is only one path from A to Y that intersects each mediator M_s, i.e., $A \to M_s \to Y$. The specific indirect effect for that path, using the product of coefficients method, is $\beta_s \delta_s$. The specific indirect (and direct) effects under a fitted parallel multiple mediator model thus equal the interventional indirect (and direct) effects under the assumed mean models (1) and (2). However, the specific indirect effects are predicated on a parallel multiple mediator model where the mediators are a priori assumed to be causally independent; a different “multiple-step” or “serial” multiple mediator model (Hayes et al., 2011; Hayes, 2018) where the mediators are a priori assumed to be causally dependent yields different specific indirect effects along different paths. In contrast, the definition of each interventional indirect effect remains the same regardless of the underlying causal structure of the mediators. We will elaborate on how the interventional indirect effects and the specific indirect effects should be interpreted differently later in this article.

Nonetheless, we propose using the parallel multiple mediator model to obtain estimators of the interventional (in)direct effects under linear models (1) and (2) within the SEM framework. While SEM is not limited to linear models in general, the product of coefficients method is still used to combine effects along paths even in nonlinear models; see e.g., MacKinnon et al. (2007), Iacobucci (2012), MacKinnon & Cox (2012), and Rijnhart et al. (2019) for the single mediator setting. Notwithstanding such extensions, in this article we will exploit widely adopted estimation methods for linear models in SEM. The resulting estimators of the interventional (in)direct effects are unbiased (in large samples) even when the mediators share hidden common causes, or when the directions of the causal effects between the mediators are unknown. Thus the estimators, while obtained using conventional estimation methods in SEM, are now endowed with robustness against unobserved confounding of the mediators, and incorrect specification of the causal structure of the mediators, when employed for estimating interventional effects.

The remainder of this article is organised as follows. A brief introduction to mediation analysis using the potential outcomes framework is first given. Notation for the potential outcomes and the assumptions required to identify the potential outcomes from observed data are then introduced, using the minimal example with two mediators introduced above. The conceptual definitions of the interventional (in)direct effects and the exact decomposition of the total effect into the direct effect and indirect effects via each mediator are then presented. The interventional (in)direct effects are interpreted using causal paths in the causal diagrams of Figure 1 to give readers intuition into the proposed interventional effects. We then describe how to obtain estimators of the interventional (in)direct effects using a parallel multiple mediator model, henceforth termed a parallel path model for simplicity, using (linear) SEM. Simulation studies comparing the empirical biases of estimators of the proposed interventional indirect effects and of specific indirect and mediated effects along separate paths are conducted. Code for the open source statistical software `R` (R Core Team, 2019) is provided for each step of the procedure in Appendix B to demonstrate its straightforward implementation. The proposed estimation method is utilized to assess the effect of political inclusion on political prejudice that is possibly mediated by six distinct mediators. We conclude with recommendations for practitioners of multiple mediation analysis and discussion of possible areas for future work.

Interventional effects

Mediation analysis using a counterfactual-based framework

Notwithstanding the widespread use of parametric approaches to mediation analysis, a counterfactual-based framework for mediation analysis has been developed using model-free definitions of *natural direct and indirect effects* (Robins & Greenland, 1992; Pearl, 2001). This development enables extensions to nonlinear models and formalizes the "ignorability" assumptions needed to identify the natural (in)direct effects, without relying on a specific statistical model; see e.g., VanderWeele & Vansteelandt (2009), Imai, Keele, & Tingley (2010), and Pearl (2014) for the single mediator setting. Under these assumptions, the total effect can be decomposed into a direct and an indirect effect using the mediation formula (Pearl, 2012b). Using linear and additive (i.e., without interactions) models for the mediator and outcome in the mediation formula yields the same estimators as the path analysis approach using the product of coefficients method (Imai, Keele, & Yamamoto, 2010).

When there are multiple mediators, identifying the natural indirect effects via each mediator is challenging because one mediator that is affected by treatment can concurrently be a confounder of the mediator-outcome association for another mediator, also known as *post-treatment* or *treatment-induced confounding*; see e.g., Rosenbaum (1984) and Robins (2000). Continuing our example above, now suppose that worldview dissimilarity (M_1) affects fairness (M_2), as depicted in Figure 1(b), so that M_1 is a post-treatment confounder of the $M_2 - Y$ relation. Then the natural indirect effect via M_2 cannot be identified without further (parametric) assumptions (Avin et al., 2005; Shpitser & VanderWeele, 2011; Imai & Yamamoto, 2013). A *path-specific effect* for a path linking one variable to another variable considers causal effects along the set of edges on that path (Avin et al., 2005). For example, in Figure 1(a), the indirect effect via each mediator $M_s, s = 1, 2$, is the path-specific effect for the path $A \rightarrow M_s \rightarrow Y$. Using the counterfactual graphical model framework, Shpitser (2013) describes how path-specific effects for "bundles" (or sets) of paths can still be (nonparametrically) identified using the "recanting district" criteria, even in the presence of merely observed post-treatment confounding and some degree of hidden confounding. Given the difficulty in identifying natural indirect effects in the multiple mediator setting, recent proposals of counterfactual-based mediation analysis for multiple (repeated measures of) mediators have thus relied on stringent ("sequential ignorability") assumptions to carefully decompose the natural indirect effect into path-specific effects via each mediator; see e.g., Albert & Nelson (2011), Lange et al. (2013), VanderWeele & Vansteelandt (2014), Daniel et al. (2015), Steen et al. (2017) and Albert et al. (2019) among others. However, such fine-grained decompositions of natural indirect effects are only valid when the mediators are independent (conditional on treatment and baseline covariates) or can be causally ordered, and share no hidden confounders. In most realistic scenarios, the directions of the causal effects between the various mediators are unknown, thus either violating the assumptions needed to identify the path-specific effects, or demanding additional assumptions about the correct specification of the causal structure.

In contrast, *interventional (in)direct effects*, first introduced by Didelez et al. (2006) and VanderWeele et al. (2014) for a single mediator, then generalized by Vansteelandt & Daniel (2017) to the multiple mediator setting, can be identified under much weaker conditions than natural effects, and still achieve an exact decomposition of the total effect. Unlike natural effects that are defined in terms of individual-level (deterministic) interventions on the mediator, interventional effects consider population-level (stochastic) interventions that set the value of the mediator to a random draw from its counterfactual distribution. Interventional effects concern ideal (distinct) interventions on the treatment and the mediator distribution, without changing anything else in the causal structure (Quynh Nguyen et al., 2019), and thus remain meaningful even when the treatment cannot be manipulated at the individual level. For example, VanderWeele & Robinson (2014) and Jackson & VanderWeele (2018) describe interventional (in)direct effects using race as the treatment and socioeconomic status as the mediator, without having to define nested potential outcomes (for each individual) where race is set to one group but socioeconomic status is simultaneously set to its potential value under a different group, depending on the treatment effect of race. Quynh Nguyen et al. (2019) compare different types of direct and indirect effects used in causal mediation analysis that may be motivated by different research questions. Recent work in the causal inference and epidemiology literature discussing interventional effects include Lok (2016), Moreno-Betancur & Carlin (2018), Lok (2019), and Moreno-Betancur et al. (2019).

Definition of potential outcomes

For pedagogic purposes, we consider a setting with two mediators M_1 and M_2 throughout, and refer the reader to Loh et al. (2019) for more general results involving more than two mediators. Uppercase letters are used to denote (observed) random variables and (possibly unobserved) potential outcomes, and lowercase letters are used to denote specific values, for each individual. For $s = 1, 2$, let $M_{sa^{(s)}}$ denote the potential outcome for M_s if, possibly counter to fact, treatment A is set to $a^{(s)}$. Let

$\tilde{M}_{sa^{(s)}|C}$ denote a random draw from the counterfactual marginal distribution (given baseline covariates C) of $M_{sa^{(s)}}$ that does not depend on any other mediators. Continuing the example above, M_{10} denotes the potential outcome for worldview dissimilarity of the political outgroup (M_1) under the control condition ($a^{(1)} = 0$), and $\tilde{M}_{10|C}$ denotes a random draw from the distribution of M_{10} that depends on political ideology (C) but does not depend on perceived fairness (M_2). Let $Y_{am_1m_2}$ denote the potential outcome for Y if, possibly counter to fact, A is set to a, and each mediator M_s is set to the value $m_s, s = 1, 2$. Hence $Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\tilde{M}_{2a^{(2)}|C}}$ denotes the potential outcome for Y under treatment $A = a^{(0)}$, when the value of each mediator is set to a random draw from its (counterfactual) marginal distribution under treatment $A = a^{(s)}, s = 1, 2$.

Identification of average potential outcomes

Identification of the average potential outcomes defined above requires the following assumptions, respectively labelled (i′), (ii′) and (iii′) in Vansteelandt & Daniel (2017):

(A1) The effect of treatment A on outcome Y is unconfounded conditional on C.

Assumption (A1) states that there are no unobserved confounders between A and Y, or equivalently, that the observed covariate(s) C are sufficient to adjust for confounding of the effect of A on Y. This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of A and Y. Note that assumption (A1) is met in randomized trials when A is randomly assigned.

Since $Y_{am_1m_2}$ is unknown for each value of (a, m_1, m_2) except for the observed realization (A, M_1, M_2), it is further assumed that the following holds:

(A2) The effect of both mediators M_1, M_2 on outcome Y is unconfounded conditional on A and C.

Assumption (A2) states that there is available sufficient covariate information observed in C so that the association between any of (M_1, M_2) and Y is unconfounded within levels of the covariate(s) C. This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of any of (M_1, M_2) and Y.

Since $M_{1a^{(1)}}$ and $M_{2a^{(2)}}$ are unknown for each value of $\{a^{(1)}, a^{(2)}\}$ except when $a^{(1)} = a^{(2)} = A$, it is also assumed that the following holds:

(A3) The effect of treatment A on both mediators is unconfounded conditional on C.

Assumption (A3) states that there are no unobserved confounders between A and any of (M_1, M_2), or equivalently, that the observed covariate(s) C are sufficient to adjust for confounding of the effects of A on (M_1, M_2). This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of A and any of (M_1, M_2). Note that assumption (A3) is met in randomized trials when A is randomly assigned.

Under assumptions (A1)–(A3), the average potential outcomes are identified by:

$$\begin{aligned}
&\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\tilde{M}_{2a^{(2)}|C}}\right) \\
&= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y_{a^{(0)}m_1m_2}|C) \prod_{s=1}^{2} \Pr(M_{sa^{(s)}} = m_s|C)\right] \\
&= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y|A = a^{(0)}, M_1 = m_1, M_2 = m_2, C) \prod_{s=1}^{2} \Pr(M_s = m_s|A = a^{(s)}, C)\right]. \qquad (3)
\end{aligned}$$

Since $\tilde{M}_{1a^{(1)}}$ is drawn from a distribution that does not depend on M_2, and similarly for $\tilde{M}_{2a^{(2)}}$ and M_1, the result in (3) does not require any additional assumptions about the joint distribution of the mediator potential outcomes $M_{1a^{(1)}}$ and $M_{2a^{(2)}}$, when the hypothetical treatment levels $a^{(1)}$ and $a^{(2)}$ are either equal or unequal.

Definition of interventional indirect and direct effects

In this section, we define the interventional indirect and direct effects and describe the decomposition for a binary treatment A as stated in Vansteelandt & Daniel (2017). We also give interpretations of the interventional (in)direct effects in terms of the causal path(s) in the causal diagrams of Figure 1.

Define the interventional indirect effect of treatment on outcome via M_1 as:

$$\begin{aligned}
\mathrm{IE}_1 &= \mathrm{E}\left(Y_{1\tilde{M}_{11|C}\tilde{M}_{20|C}}\right) - \mathrm{E}\left(Y_{1\tilde{M}_{10|C}\tilde{M}_{20|C}}\right) \\
&= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y_{1m_1m_2}|C) \left\{\Pr(M_{11} = m_1|C) - \Pr(M_{10} = m_1|C)\right\} \Pr(M_{20} = m_2|C)\right].
\end{aligned} \qquad (4)$$

The interventional indirect effect via mediator M_1 is the treatment effect of changing the distribution of M_1 from its (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level $a^{(1)} = 1$ to its distribution under level $a^{(1)} = 0$, while fixing the individual values of treatment at $a^{(0)} = 1$ and the mediator M_2 to be a random draw $\tilde{M}_{2a^{(2)}}$ from its (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level $a^{(2)} = 0$.

The interventional indirect effect via M_1 (4) corresponds to the path-specific effect $A \to M_1 \to Y$ in the causal diagrams of Figures 1(a) – 1(c). By definition, the indirect effect via M_1 is a function of the difference $\Pr(M_{11} = m_1|C) - \Pr(M_{10} = m_1|C)$, which encodes the combination of the path-specific effects $A \to M_1 \to Y$ and $A \to M_2 \to M_1 \to Y$ in the causal diagram of Figure 1(d). The interventional indirect effect via M_1 thus captures all of the treatment effect that is mediated by M_1 and its causal ancestors, but not its causal descendants, in the underlying causal diagram.

Similarly, define the interventional indirect effect of treatment on outcome via M_2 as:

$$\begin{aligned}\mathrm{IE}_2 &= \mathrm{E}\left(Y_{1\tilde{M}_{11|C}\tilde{M}_{21|C}}\right) - \mathrm{E}\left(Y_{1\tilde{M}_{11|C}\tilde{M}_{20|C}}\right) \\ &= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y_{1m_1m_2}|C)\Pr(M_{11} = m_1|C)\left\{\Pr(M_{21} = m_2|C) - \Pr(M_{20} = m_2|C)\right\}\right]. \end{aligned} \tag{5}$$

The interventional indirect effect via mediator M_2 can be analogously interpreted as the indirect effect via M_1. In particular, the interventional indirect effect via M_2 (5) corresponds to the path-specific effect $A \to M_2 \to Y$ in the causal diagrams of Figures 1(a), 1(c), and 1(d). In the causal diagram of Figure 1(b), the indirect effect via M_2 is the combination of the path-specific effects $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$. As before, the interventional indirect effect via M_2 is interpreted as the effect of treatment that is mediated by M_2 and its causal ancestors, but not its causal descendants, in an underlying causal diagram. Note that the above definitions of the interventional indirect effects via each mediator are not contingent on an assumed causal ordering of the mediators. While switching the indices of the mediators, e.g., by denoting worldview dissimilarity and fairness by M_2 and M_1 respectively, may lead to different definitions of each indirect effect (due to fixing the other mediator at its distribution under a different treatment level), it does not change the interpretation of each interventional indirect effect in terms of the underlying causal dependence among the mediators.

The interventional joint indirect effect of treatment on outcome that is mediated by at least one mediator can thus be defined as the sum of the separate interventional indirect effects via each mediator, i.e., $\mathrm{IE}_1 + \mathrm{IE}_2$. The interventional direct effect of treatment on outcome that avoids both mediators is correspondingly defined as:

$$\begin{aligned}\mathrm{DE} &= \mathrm{E}\left(Y_{1\tilde{M}_{10|C}\tilde{M}_{20|C}}\right) - \mathrm{E}\left(Y_{0\tilde{M}_{10|C}\tilde{M}_{20|C}}\right) \\ &= \mathrm{E}\left[\sum_{m_1,m_2} \{\mathrm{E}(Y_{1m_1m_2}|C) - \mathrm{E}(Y_{0m_1m_2}|C)\} \prod_{s=1}^{2} \Pr(M_{s0} = m_s|C)\right]. \end{aligned} \tag{6}$$

The direct effect (6) is the treatment effect when controlling the individual values of the mediators M_1 and M_2 to be random draws $\tilde{M}_{1a^{(1)}}$ and $\tilde{M}_{2a^{(2)}}$ from their respective (counterfactual) marginal distributions (given covariates C) under hypothetical treatment levels $a^{(1)} = a^{(2)} = 0$. The direct effect (6) corresponds to the path-specific effect for the path $A \rightarrow Y$ in the causal diagrams of Figure 1. The sum of the separate indirect effects via each mediator (4) and (5), and the direct effect (6), may then be defined to be the total effect of treatment on outcome:

$$\begin{aligned}\mathrm{TE} &= \mathrm{E}\left(Y_{1\tilde{M}_{11|C}\tilde{M}_{21|C}}\right) - \mathrm{E}\left(Y_{0\tilde{M}_{10|C}\tilde{M}_{20|C}}\right) \\ &= \mathrm{E}\left[\sum_{m_1,m_2} \left\{\mathrm{E}(Y_{1m_1m_2}|C) \prod_{s=1}^{2} \Pr(M_{s1} = m_s|C) - \mathrm{E}(Y_{0m_1m_2}|C) \prod_{s=1}^{2} \Pr(M_{s0} = m_s|C)\right\}\right]. \end{aligned} \tag{7}$$

In other words, $\mathrm{TE} = \mathrm{DE} + \mathrm{IE}_1 + \mathrm{IE}_2$. In general when there are $t > 2$ distinct mediators, the interventional indirect effect via each mediator $M_s, s = 1, \ldots, t$, henceforth denoted by IE_s, is defined in Appendix A.

Estimators of interventional effects

In this section we describe how to obtain estimators of the interventional (in)direct effects defined in the previous section. Suppose that the outcome obeys the linear and additive mean model in (1), restated here as:

$$\mathrm{E}(Y|A, M_1, M_2, C) = \beta_0 + \beta_A A + \sum_{s=1}^{2} \beta_s M_s + \beta_C C.$$

The average potential outcome $\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\tilde{M}_{2a^{(2)}|C}}\right)$ is identified upon plugging the assumed outcome model (1) and mediator marginal distributions into (3); i.e.,

$$\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\tilde{M}_{2a^{(2)}|C}}\right) = \mathrm{E}\left[\left\{\beta_0 + \beta_A a^{(0)} + \beta_C C\right\} + \sum_{s=1}^{2} \beta_s \,\mathrm{E}(M_s|A = a^{(s)}, C)\right].$$

Suppose that the overall treatment effect on each mediator, given baseline covariate(s) C, is parametrized by the (partial) regression coefficient of treatment A in the linear and additive (marginal) mean model for each mediator in (2), restated here as:

$$\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, \quad s = 1, 2.$$

It follows that the interventional direct effect and indirect effects via each mediator M_s are identified by (functions of) the parameters in the assumed models; i.e., $\mathrm{DE} = \beta_A$ and $\mathrm{IE}_s = \beta_s\delta_s, s = 1, 2$ respectively. Again we note that the overall effect of A on M_s, as encoded by δ_s in (2), captures all of the underlying treatment effects that are transmitted from A to M_s through any causal ancestors of M_s. Unbiased estimation of the interventional (in)direct effects thus requires correctly specifying the outcome mean model (1) and mediator (marginal) mean models (2) under assumptions (A1)–(A3). Even though conventional mediation approaches typically do not adjust for baseline covariates C that are confounders of the mediator-outcome relation (Coffman, 2011), we include baseline covariates in (1) and (2) toward satisfying assumptions (A1)–(A3).

As previously noted in the introduction, the assumed models (1) and (2) are jointly parametrically equivalent to a parallel path model where the mediators are assumed not to causally affect each other. The interventional indirect effect via each mediator M_s thus equals the specific indirect effect using the product of coefficients method $\beta_s\delta_s$ for the path $A \to M_s \to Y$ in the parallel path model. Similarly, the interventional direct effect equals the specific effect β_A for the path $A \to Y$ that avoids both mediators in the parallel path model. Estimators of the interventional effects can thus be obtained by fitting the parallel path model to the observed data using SEM, then plugging in estimates of the (partial) regression coefficients in the respective specific effects. Standard errors can be estimated using a nonparametric percentile bootstrap procedure (Efron & Tibshirani, 1994) that randomly resamples observations with replacement. In general when there are $t > 2$ distinct mediators, the estimators of the interventional indirect effects via each mediator $M_s, s = 1, \ldots, t$, are described in Appendix A.

To see why fitting mean models (1) and (2) is sufficient to obtain unbiased estimators of the interventional (in)direct effects, consider the continuing example from the introduction corresponding to the causal diagram of Figure 1(b). Suppose that the observed data is generated from an underlying path model where M_1 has a causal effect on M_2, with the mediator and outcome models:

$$\begin{aligned}
\mathrm{E}(M_1|A,C) &= \alpha_{01}^* + \alpha_1^* A + \alpha_{C1}^* C,\\
\mathrm{E}(M_2|A,M_1,C) &= \alpha_{02}^* + \alpha_2^* A + \eta_{12}^* M_1 + \alpha_{C2}^* C,\\
\mathrm{E}(Y|A,M_1,M_2,C) &= \beta_0^* + \beta_A^* A + \beta_1^* M_1 + \beta_2^* M_2 + \beta_{CY}^* C.
\end{aligned}$$

(Asterisks denote parameters in the true data-generating model.) The mean of the implied marginal distribution of M_2, obtained by averaging over the distribution of M_1, is then:

$$\begin{aligned}
\mathrm{E}(M_2|A,C) &= \sum_{m_1} \mathrm{E}(M_2|A,M_1,C)\Pr(M_1 = m_1|A,C)\\
&= \alpha_{02}^* + \alpha_2^* A + \eta_{12}^* \,\mathrm{E}(M_1|A,C) + \alpha_{C2}^* C\\
&= \alpha_{02}^* + \alpha_2^* A + \eta_{12}^*(\alpha_{01}^* + \alpha_1^* A + \alpha_{C1}^* C) + \alpha_{C2}^* C\\
&= (\alpha_{02}^* + \eta_{12}^*\alpha_{01}^*) + (\alpha_2^* + \eta_{12}^*\alpha_1^*)A + (\alpha_{C2}^* + \eta_{12}^*\alpha_{C1}^*)C.
\end{aligned}$$

The interventional indirect effect via M_2 in the true model is thus identified by $\beta_2^*(\alpha_2^* + \eta_{12}^*\alpha_1^*)$. By fitting to the observed data a parallel path model with outcome mean model (1), so that $\beta_2 = \beta_2^*$, and mediator mean model (2), so that $\delta_2 = \alpha_2^* + \eta_{12}^*\alpha_1^*$, it follows that the interventional indirect effect can be unbiasedly estimated using the product of coefficients method since $\beta_2\delta_2 = \beta_2^*(\alpha_2^* + \eta_{12}^*\alpha_1^*)$ (assuming (A1)–(A3) hold). Hence the parallel path model is used merely to obtain estimators of the interventional indirect and direct effects within the linear SEM framework. Unbiased estimation does not require the mediators to be causally independent, as implied in the parallel path model; in fact, the (marginal) mean model (2) is used precisely so that the interventional indirect effects are agnostic to the underlying causal dependence among the mediators.

Simulation studies

Two simulation studies were conducted to assess the empirical biases and standard errors of estimators of the proposed interventional indirect effects, and of separate specific indirect and mediated effects along different paths from existing SEM approaches. In study 1, the setting with two causally dependent mediators that share an unobserved common cause, as motivated by the example in the introduction, was considered. This study was used to assess how unobserved (baseline) confounding of the mediators affected estimates of separate specific indirect effects under a correctly-specified path model. In study 2, the setting with four candidate mediators, one of which was causally unconnected to the other three, conditional on treatment and observed (baseline) confounding of the mediators and outcomes, was considered. However, the mediators were correlated due to unobserved (baseline) confounding of the mediators. This study was used to assess how incorrectly specifying the causal dependence among the mediators affected estimates of separate mediated effects along different paths.

Study 1

[Figure 2 about here.]

Each observed dataset was generated with the following linear and additive models based on the causal diagram in Figure 2:

$$\begin{aligned}
A &\sim \text{Bernoulli}(0.5)\\
U &\sim \mathcal{N}(1,1)\\
M_1 &= \alpha_{01} + \alpha_1 A + \alpha_{U1} U + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)\\
M_2 &= \alpha_{02} + \alpha_2 A + \alpha_{U2} U + \eta_{12} M_1 + \epsilon_2, \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2)\\
Y &= \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)
\end{aligned}$$

Continuing the example in the introduction, treatment A was a randomly assigned political inclusion manipulation, the mediators M_1 and M_2 were worldview dissimilarity and fairness respectively, with a common cause U being prior adverse interactions with a political outgroup, and the outcome Y was prejudice toward a political outgroup. The interventional indirect effect via M_1 corresponded to the path-specific effect for $A \to M_1 \to Y$, and was identified by $\beta_1\alpha_1$. It followed by definition that the interventional indirect effect via M_2 corresponded to the combination of the path-specific effects for $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$, and was identified by $\beta_2(\alpha_2 + \alpha_1\eta_{12})$. The variables and residual errors $A, U, \epsilon_1, \epsilon_2$, and ϵ_Y are mutually independent. For simplicity, the values of all parameters (including partial regression coefficients and residual variances) in the data-generating model, with the exception of $\alpha_{U1}, \alpha_{U2}, \eta_{12}$, were set to 1; α_{U1} and α_{U2} were set to 2 and -2 respectively so that M_1 and M_2 were negatively and strongly correlated, while η_{12} was set to -2 so that the values of the indirect effects via M_1 and via M_2 had the same magnitude but had different signs.

Two (linear) path models that correctly assumed M_1 (worldview dissimilarity) to causally precede M_2 (fairness) were fitted to each generated dataset. The first assumed path model was the true data-generating model (“correct”) and corresponded to the causal diagram depicted in Figure 2. To fit this model, the confounder U was assumed to be observed and hence adjusted for in the mediator models. The second assumed path model corresponded to the causal diagram of Figure 1(b), and consisted of the following mediator and outcome models:

$$M_1 = \alpha_{01} + \alpha_1 A + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$$

$$M_2 = \alpha_{02} + \alpha_2 A + \eta_{12} M_1 + \epsilon_2, \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$$

$$Y = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$$

There was unobserved confounding of M_1 and M_2 under this path model ("no U") as the common cause of the mediators U was not adjusted for in the mediator models. Under both these assumed path models, we considered estimates of specific indirect effects from existing SEM approaches. Let PE$_s$ denote the specific indirect effect mediated by M_s following Alwin & Hauser's definition. In the linear path model used to generate the observed data in this simulation study, PE$_1$ was the combined effect for the separate paths $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$, and PE$_2$ was the effect for the path $A \to M_2 \to Y$ alone. We also considered the three-path mediated effect for the path $A \to M_1 \to M_2 \to Y$ separately, which we denoted by PE$_{12}$. The specific indirect and three-path mediated effects were encoded by $\alpha_1(\beta_1 + \eta_{12}\beta_2)$, $\alpha_2\beta_2$, and $\alpha_1\eta_{12}\beta_2$ respectively using the product of coefficients method under the data-generating linear models. Note that the path $A \to M_1 \to M_2 \to Y$ contributed to the interventional indirect effect via M_2 but contributed to the specific indirect effect via M_1. Lastly, estimates of the interventional indirect effects were obtained by fitting to each generated dataset a parallel path model where M_1 and M_2 were incorrectly assumed to be causally independent (by assuming the arrow from M_1 to M_2 in the causal diagram of Figure 2 to be absent, thus corresponding to the causal diagram in Figure 1(c)).

The various indirect effects for each assumed path model were estimated by fitting the path model to each generated dataset using `lavaan` (Rosseel, 2012) in `R`. Within each fitted model, the (estimators of the) indirect effects were defined syntactically by combining the relevant (partial) regression coefficients using the product of coefficients method. We provide in Appendix B the model syntax in `lavaan` that describes each path model and the indirect effects to be estimated. 10000 observed datasets of size 400 were generated, and all three path models described above fitted to each dataset. Approximately 2% of generated datasets were discarded due to convergence errors for at least one of the fitted path models. Average estimates and empirical standard errors of the indirect effects under each fitted path model for the remaining datasets are displayed in Table 1. The estimated specific indirect effects PE_1 and PE_2, and three-path mediated effect PE_{12}, were empirically unbiased only when the common cause U of M_1 and M_2 was observed and adjusted for in the mediator models. However, when U was not adjusted for in the mediator models, so that there was unobserved confounding of the mediators, estimates of all three specific indirect and mediated effects were biased "away from zero," in the sense that the (absolute) magnitudes were greater than the true values on average. In contrast, when the causal structure of the mediators was left unspecified (by fitting a parallel path model), estimates of the interventional indirect effects using the product of coefficients method were empirically unbiased.

[Table 1 about here.]

Study 2

[Figure 3 about here.]

Each observed dataset was generated with the following linear and additive models corresponding to the top causal diagram of Figure 3:

$$\begin{aligned}
A &\sim \text{Bernoulli}(0.5)\\
C &\sim \mathcal{N}(0,1)\\
U &\sim \mathcal{N}(0,1)\\
M_1 &= \alpha_1 A + \alpha_{C1} C + \alpha_{U1} U + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0,\sigma_1^2)\\
M_2 &= \eta_{12} M_1 + \alpha_{C2} C + \alpha_{U2} U + \epsilon_2, \epsilon_2 \sim \mathcal{N}(0,\sigma_2^2)\\
M_3 &= \eta_{13} M_1 + \alpha_{C3} C + \alpha_{U3} U + \epsilon_3, \epsilon_3 \sim \mathcal{N}(0,\sigma_3^2)\\
M_4 &= \alpha_{C4} C + \alpha_{U4} U + \epsilon_4, \epsilon_4 \sim \mathcal{N}(0,\sigma_4^2)\\
Y &= \beta_2 M_2 + \beta_3 M_3 + \beta_4 M_4 + \beta_C C + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0,\sigma_Y^2)
\end{aligned}$$

Under the data-generating model, the unobserved variable U was a (baseline) confounder of all four mediators, and the observed variable C was a (baseline) confounder of all four mediators and the outcome. The interventional indirect effect via M_1 was zero since M_1 did not affect Y directly. The interventional indirect effects via M_2 and M_3 corresponded to the path-specific effects for $A \to M_1 \to M_2 \to Y$ and $A \to M_1 \to M_3 \to Y$ respectively, and were identified by $\beta_2\eta_{12}\alpha_1$ and $\beta_3\eta_{13}\alpha_1$. The interventional indirect effect via M_4 was zero since M_4 was not causally dependent on either M_1, M_2, M_3 or A. The variables and residual errors $A, C, U, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$, and ϵ_Y are mutually independent. For simplicity, the values of all parameters (including partial regression coefficients and residual variances) in the data-generating model, with the exception of η_{13} and β_3, were set to 1; η_{13} and β_3 were set to 0.5 and 2 respectively so that the values of the separate effects of $M_1 \to M_2$ and $M_1 \to M_3$ (as well as $M_2 \to Y$ and $M_3 \to Y$) were unequal but the values of the indirect effects via M_2 and M_3 both equalled one.

Suppose that interest was in mediated effects along specific paths that were predicated on assuming certain causal effects between the mediators, such as in the bottom causal diagram of Figure 3. This particular path model was posited by van der Linden et al. (2015) as a “gateway belief model” representing “causal associations” between perceptions of scientific consensus, key beliefs in climate change, and support for climate action. In van der Linden et al. (2015), treatment A was a randomly assigned consensus-message intervention, mediators M_1, M_2, M_3, and M_4 were perceived level of scientific consensus, belief that climate change is happening, belief in human causation (of climate change), and worry about climate change respectively, and outcome Y was support for public action. In particular, a causal structure among the variables was posited that assumed (i) the consensus-message intervention (A) affected only the level of perceived consensus (M_1), and no other variables, (ii) the level of perceived consensus (M_1) affected the key beliefs in climate change (M_2, M_3, M_4), (iii) belief that climate change is happening (M_2), and belief in human causation (M_3) subsequently affected worry about climate change (M_4), and (iv) support for public action (Y) was causally affected by the key beliefs in climate change (M_2, M_3, M_4), and neither level of perceived consensus (M_1) or the consensus-message intervention (A)

directly. A discussion of how to derive and justify such detailed posited causal effects based on established scientific knowledge and prior careful experimentation is beyond the scope of this paper. Instead, we will consider estimates of each separate mediated effect along different paths that traverse at least one mediator, by fitting the posited path model to each generated dataset. Denote the three-path mediated effect for the path $A \to M_1 \to M_s \to Y$ by PE_{1s} for $s = 2, 3, 4$, and denote the "four-path" mediated effect for the path $A \to M_1 \to M_s \to M_4 \to Y$ by PE_{1s4} for $s = 2, 3$. The three-path and four-path mediated effects were encoded by $\beta_s \eta_{1s} \alpha_1$ and $\beta_4 \eta_{s4} \eta_{1s} \alpha_1$ respectively using the product of coefficients method under the data-generating linear models.

In addition, estimates of the interventional indirect effects were obtained by fitting to each generated dataset a parallel path model where each mediator depended only on treatment A and observed (baseline) covariate C, and the outcome Y depended on all other variables. Estimators of the interventional indirect effects were then obtained by combining the relevant regression coefficients using the product of coefficients method. 10000 observed datasets of size 400 were generated, and both path models described above fitted to each dataset using `lavaan` (Rosseel, 2012) in `R`; the model syntax that describes each path model and the indirect effects to be estimated is provided in Appendix B. Both path models were fitted without error for all the generated datasets. Average estimates and empirical standard errors of the mediated or indirect effects under each fitted path model are displayed in Table 2. All the estimated three-path and four-path mediated effects for the separate paths under the (former) posited path model were empirically biased. In particular, the mediated effects that traversed M_4 were biased due to incorrectly specifying the causal effects entering M_4, and could be either positively or negatively biased. Even when the causal effects were correctly specified, such as in the three-path mediated effects for M_2 and M_3, the estimates were biased due to unobserved confounding between M_2, M_3 and M_4. In contrast, when the causal structure of the mediators was left unspecified (by fitting a parallel path model), estimates of the interventional indirect effects using the product of coefficients method were empirically unbiased.

[Table 2 about here.]

In conclusion, estimates of separate mediated effects or specific indirect effects along different paths that were predicated on certain causal ordering (or effects) among the mediators, were shown to be biased when the posited causal structure among the mediators in the fitted path model was incorrect, or when there was unobserved confounding between the mediators. In contrast, estimates of the interventional indirect effects using a joint model that left the causal structure among the mediators unspecified (by fitting a parallel path model) were empirically unbiased even when the mediators shared unobserved common causes.

Application

The proposed estimation procedure is illustrated using a publicly-available data set from a randomized study assessing the effect of (non-)political inclusion on political prejudice that is possibly mediated by six different mediators (Voelkel et al., 2019). The data set is available as part of a preregistered study via the Open Science Framework (`https://osf.io/jcmmp/?view_only=3af8cb6b1f2845b1ba3fd69cb0b89585`). The goal of the study was to assess the causal effect of either political inclusion or non-political inclusion versus control on momentary prejudice toward the political outgroup. The sample consisted of college freshmen from a large university in the Netherlands who received course credit in a psychology course for their participation. Participants were randomly assigned to one of three conditions: political inclusion, non-political inclusion, or control. For the purposes of illustrating the estimation procedure, we will only consider the 183 participants assigned to either political inclusion ($A = 1$) or control ($A = 0$). In the treatment group, participants' political inclusion experiences were manipulated using an online political discussion. In the control group, participants experienced a neutral scenario where no discussion (political or non-political) occurred, and they were only asked to fill in a questionnaire. The dependent variable Y (prejudice) was an average of three items: dislike of, social distance from, and perceived immorality of, the participant's political outgroup. Larger values indicated higher levels of prejudice. To understand how political inclusion affects prejudice, the authors of the study considered six possible mediators of the causal relationship between political inclusion and prejudice: satisfaction of the need to belong (M_1), satisfaction of the need for self-esteem (M_2), satisfaction of the need for control (M_3), satisfaction of the need for meaningful existence (M_4), perceived worldview dissimilarity of the political outgroup (M_5), and perceived fairness of the political outgroup (M_6). The first four mediators represent the satisfaction of basic human needs that are often used in social exclusion research, and were each measured by the average of three items specific to each assigned condition, with larger values indicating higher satisfaction. The fifth mediator (worldview dissimilarity) was measured by the average of eight items, where each item indicated the extent to which the participant saw a particular (political or social) group as holding political or social beliefs different from their own. The sixth mediator (fairness) was measured by the average of two items indicating how unfair or fair, and how disrespectful or respectful, they perceived their political outgroup. For the fifth and sixth mediators, larger values indicated stronger negative attitudes toward the political outgroup. All values of the mediators and the outcome were between zero and one. In addition, we considered political ideology (“`Ideology`”) indicated on a scale from -2.5 (very left) to 2.5 (very right) in increments of 0.5, age (“`Age`”), and gender (“`Gender`”) taking value 1 if the participant was a female or 0 otherwise, as (baseline) confounders of the mediator-outcome relation for all the mediators. Summaries of the variables for each treatment group are displayed in Table 3.

[Table 3 about here.]

To estimate the interventional (in)direct effects, the following parallel path model where each mediator depended only on treatment, political ideology, age, and gender, was fitted to the observed data in `lavaan`:

$$\begin{aligned}\mathrm{E}(M_s|A,C) &= \delta_{0s} + \delta_s A + \delta_{C1s}\mathrm{Ideology} + \delta_{\mathrm{C2s}}\mathrm{Age} + \delta_{\mathrm{C3s}}\mathrm{Gender}, \quad \mathrm{s} = 1,\ldots,6;\\ \mathrm{E}(Y|A,M_1,\ldots,M_6,C) &= \beta_0 + \beta_A A + \sum_{s=1}^{6}\beta_s M_s + \beta_{C1}\mathrm{Ideology} + \beta_{\mathrm{C2}}\mathrm{Age} + \beta_{\mathrm{C3}}\mathrm{Gender}.\end{aligned}$$

Under the assumed models, the interventional indirect effect via each mediator M_s was identified by $\beta_s \delta_s, s = 1, \ldots, 6$, and the interventional direct effect that avoids all the mediators was identified by β_A. We provide in Appendix B the model syntax in `lavaan` to fit the parallel path model and to estimate the (in)direct effects. Bootstrap standard errors and 95% bootstrap (percentile) confidence intervals were constructed using 10000 bootstrap samples. The results are shown in Table 4.

[Table 4 about here.]

The estimated total effect of the political inclusion manipulation (versus control) was an average change in prejudice by $-0.076\,(95\%$ confidence interval (CI) $= (-0.135, -0.016))$. The estimated interventional direct effect can be interpreted as politically included individuals having higher prejudice (than if assigned to the control condition) by $0.039\,(95\%\,\text{CI} = (-0.003, 0.083))$, holding the distributions of all mediators (given political ideology) fixed under those of the control condition. The estimated interventional indirect effect via worldview dissimilarity (M_5) can be interpreted as the estimated average change in prejudice among politically included individuals being $-0.012\,(95\%\,\text{CI} = (-0.027, -0.002))$ if the (counterfactual) distribution of worldview dissimilarity under the political inclusion manipulation is shifted to that under control, while treatment and the distributions of all other mediators were fixed. The estimated interventional indirect effect via fairness (M_6) can be similarly interpreted as the estimated average change in prejudice for politically included individuals being $-0.097\,(95\%\,\text{CI} = (-0.146, -0.054))$ if the (counterfactual) distribution of fairness under the political inclusion manipulation is shifted to that under control, while treatment and the distributions of all other mediators were fixed. The (absolute) magnitudes of the estimated interventional indirect effects via the other mediators were at most 0.005, and non-statistically significant at 5%. These results suggested that the total diminishing effect of political inclusion on prejudice was primarily explained by the mediating effects through perceived fairness and worldview dissimilarity of the political outgroup.

Discussion

When there are multiple or competing mediators on the causal pathway from treatment to outcome, path analysis is commonly used to disentangle the indirect effects transmitted along causal path(s) through each mediator. However, specific indirect or mediated effects along separate paths traversing distinct mediators may only be well-defined, and subsequently unbiasedly estimated, using the product of coefficients method when the causal dependence among the mediators is correctly specified, and there is no unobserved confounding of the mediators. In contrast, interventional (in)direct effects, first introduced by Didelez et al. (2006) and VanderWeele et al. (2014) for a single mediator, then generalized by Vansteelandt & Daniel (2017) to the multiple mediator setting, can be identified, and thus unbiasedly estimated under less stringent conditions.

In this article, we have introduced the interventional (in)direct effects using a motivating example from the psychology literature, and formalized the nonparametric assumptions (A1)–(A3) required to identify the effects. We have described how to obtain estimators of the interventional (in)direct effects using prevalent path analysis methods for multiple mediation analysis within the linear SEM framework. The estimation procedure was illustrated using simulation studies and an applied example; `lavaan` model syntax using the open source statistical software `R` is provided in Appendix B to show its straightforward implementation. Fitting a joint model for all variables using SEM exploits widely adopted estimation methods, and facilitates easy implementation of restrictions on the model parameters, such as fixing certain path coefficients to be equal for different mediators.

Recommendations for multiple mediation analysis

When the causal structure among the mediators is unknown or cannot be correctly specified with absolute certainty, or when unobserved confounding of the mediators is plausible, fitting a posited path model assuming certain causal effects among the mediators can yield biased estimates of specific indirect effects along separate paths. The biases were demonstrated empirically using simulation studies in this article.

Instead, the causal dependence among the mediators should be left unspecified, by fitting only the marginal mean model (1) for each mediator that captures the overall treatment effect (on that mediator). The interventional indirect effect via a mediator of interest is then identified by the product of (i) the (partial) regression coefficient of treatment in the mediator (marginal) mean model, and (ii) the partial regression coefficient of that mediator in the outcome model (2) (conditional on all other mediators, treatment and baseline confounders), under assumptions (A1)–(A3). The interventional indirect effect via a mediator of interest is thus interpreted as the combination of all underlying causal pathways from treatment to outcome that intersect that mediator and any other mediators causally preceding it, and avoid all of the mediators causally dependent on the mediator in question.

Since models (1) and (2) are jointly parametrically equivalent to a parallel path model where the mediators are (incorrectly) assumed to be causally independent of one another, the interventional effects can be estimated by fitting a parallel path model to the observed data. Estimators of the specific (in)direct effects using the product of coefficients method can then be used as estimators of the interventional (in)direct effects. However, unlike the specific indirect effects in a parallel path model that assume underlying causal independence among the mediators, the interventional indirect effects are agnostic to the underlying causal structure of the mediators. The resulting estimators of the interventional effects, while obtained using conventional estimation methods in SEM, are unbiased (in large samples) and are robust against both unobserved confounding of the mediators, and incorrect specification of the causal structure of the mediators.

Areas of future work

There are several avenues of possible future research related to mediation analyses with multiple mediators using interventional effects developed in this paper. The total effect in (7) is defined as the sum of the separate indirect effects via each mediator (4) and (5), and the direct effect (6). Vansteelandt & Daniel (2017) propose different definitions of the total and direct effects so that in addition to the indirect and direct effects, there is a separate indirect effect that exists when treatment changes the relationships between the mediators, which in turn affects the outcome. This latter indirect effect, termed the *indirect effect due to the mediators' mutual dependence* (Vansteelandt & Daniel, 2017), should therefore be considered separately from the indirect effects via each mediator, as it cannot be attributed to any one mediator alone. In the applied example, this indirect effect was estimated to be in the order of 10^{-8}. However, this was a consequence of fitting linear models without interactions, and cannot be viewed as evidence suggesting that the indirect effect equalled zero. Under linear models for the means of the mediators and the outcome, this indirect effect is non-zero if (i) there is non-zero mediator-mediator interaction in the outcome model; and (ii) the covariance of the mediators differs with treatment. Following Vansteelandt & Daniel (2017), we define the interventional effects that can be identified by functions of the model parameters in the assumed models for the mediators and outcome under the setting with two mediators in Appendix C. Note that the proposed method in Appendix C allows for treatment-mediator interactions in the mean model for the outcome.

One of the assumptions required to identify the interventional (in)direct effects presented in this paper is that there be no hidden common causes of the mediators and outcome, i.e., assumption (A2). Future research could include extending sensitivity analyses to unobserved confounding of the mediator-outcome relations for a single mediator (Fritz et al., 2016) to the multiple mediator setting. Similarly, sensitivity analyses investigating the (finite sample) biases of the direct and indirect effects due to incorrect specification of the outcome mean model (1) or the mediator mean models (2), or both, may also be considered. An alternative to causal mediation analysis using the potential outcomes framework is the stochastic theory of causal effects (Mayer et al., 2014) that does not require conceptual manipulation of variables to define and identify direct and indirect effects. Future work may extend the representation of causal effects based on probability theory for the single mediator setting to the multiple mediator setting, and contrast such an approach with existing interventional direct and indirect effects. Further investigation of valid confidence intervals in finite samples using parametric or robust (sandwich-based) standard errors in place of bootstrap standard errors may also be considered. For example, one may employ the general procedure of Bollen (1987) to first express the interventional effects in matrix form, then following Sobel (1982), apply the multivariate delta method to estimate the asymptotic variances of the interventional effects. The path analysis approach using SEM can be easily extended to accommodate latent mediators or outcome, or both, by including latent variable models in the fitted SEM; see e.g., Loeys et al. (2014), Loh et al. (2018) and Derkach et al. (2019). When there are nonlinear models for the means of the mediators, or the outcome, or both, estimation of interventional (in)direct effects in general can be carried out using a "mediation formula" approach as presented in Vansteelandt & Daniel (2017), where the average (potential) outcomes Y in (3) are averaged over random draws of the mediators from their respective (counterfactual) distributions. Alternatively, *interventional effect models* (Loh et al., 2019) that directly parameterize the direct and indirect effects through each distinct mediator may also be used. Under both aforementioned approaches, the mediators and outcome can be continuous or noncontinuous. Estimation proceeds via Monte Carlo integration and only requires specifying a joint distribution of the mediators and an outcome model.

References

Albert, J. M., Cho, J. I., Liu, Y., & Nelson, S. (2019). Generalized causal mediation and path analysis: Extensions and practical considerations. *Statistical Methods in Medical Research*, *28*(6), 1793–1807. doi: 10.1177/0962280218776483

Albert, J. M., & Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, *67*(3), 1028 – 1038. doi: 10.1111/j.1541-0420.2010.01547.x

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 37–47. doi: 10.2307/2094445

Avin, C., Shpitser, I., & Pearl, J. (2005, July). Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 357–363). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173. doi: 10.1037/0022-3514.51.6.1173

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 37–69. doi: 10.2307/271028

Brown, R. L. (1997). Assessing specific mediational effects in complex theoretical models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 142–156. doi: 10.1080/10705519709540067

Coffman, D. L. (2011). Estimating causal effects in mediation analysis using propensity scores. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 357–369. doi: 10.1080/10705511.2011.582001

Daniel, R., De Stavola, B., Cousens, S., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*(1), 1–14. doi: 10.1111/biom.12248

Derkach, A., Pfeiffer, R. M., Chen, T.-H., & Sampson, J. N. (2019). High dimensional mediation analysis with latent variables. *Biometrics*, 1–12. doi: 10.1111/biom.13053

Didelez, V., Dawid, A. P., & Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd conference on uncertainty in artificial intelligence* (pp. 138–146). Arlington, VA, USA: AUAI Press.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Chapman and Hall/CRC. doi: 10.1007/978-1-4899-4541-9

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–an analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. doi: 10.1016/j.jesp.2017.11.008

Fox, J. (1980). Effect analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Methods & Research*, *9*(1), 3-28. doi: 10.1177/004912418000900101

Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behavioral Research*, *51*(5), 681-697. doi: 10.1080/00273171.2016.1224154

Greene, V. L. (1977). An algorithm for total and indirect causal effects. *Political Methodology*, *4*(4), 369–381. Retrieved from https://www.jstor.org/stable/25791510

Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., . . . Boadu, K. (2003). Pearl's D-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 289–311. doi: 10.1207/s15328007sem1002_8

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis : a regression-based approach* (2nd ed.). New York, NY, USA: Guilford press.

Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. In E. P. Bucy & R. L. Holbert (Eds.), *Sourcebook for political communication research: Methods, measures, and analytical techniques* (1st ed., pp. 434–465). New York, NY, USA: Routledge.

Iacobucci, D. (2012). Mediation analysis and categorical variables: The final frontier. *Journal of Consumer Psychology*, *22*(4), 582 - 594. doi: 10.1016/j.jcps.2012.03.006

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309. doi: 10.1037/a0020761

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71. doi: 10.2307/41058997

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, *21*(2), 141–171. doi: 10.1093/pan/mps040

Jackson, J. W., & VanderWeele, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, *29*(6), 825–835. doi: 10.1097/EDE.0000000000000901

Lange, T., Rasmussen, M., & Thygesen, L. C. (2013). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, *179*(4), 513–518. doi: 10.1093/aje/kwt270

Loeys, T., Moerkerke, B., Raes, A., Rosseel, Y., & Vansteelandt, S. (2014). Estimation of controlled direct effects in the presence of exposure-induced confounding and latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 396–407. doi: 10.1080/10705511.2014.915372

Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G., & Vansteelandt, S. (2018). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomised studies. *Manuscript submitted for publication.*

Loh, W. W., Moerkerke, B., Loeys, T., & Vansteelandt, S. (2019). Interventional effect models for multiple mediators. *arXiv e-prints*, arXiv:1907.08415.

Lok, J. J. (2016). Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Statistics in Medicine*, *35*(22), 4008–4020. doi: 10.1002/sim.6990

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

MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis* (1st ed.). New York, NY, USA: Routledge. doi: 10.4324/9780203809556

MacKinnon, D. P., & Cox, M. G. (2012). Commentary on “mediation analysis and categorical variables: The final frontier” by dawn iacobucci. *Journal of Consumer Psychology*, *22*(4), 600 - 602. doi: 10.1016/j.jcps.2012.03.009

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, *4*(5), 499–513. doi: 10.1177/1740774507083434

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, *7*(1), 83 – 104. doi: 10.1037/1082-989X.7.1.83

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, *49*(5), 425–442. doi: 10.1080/00273171.2014.931797

Moerkerke, B., Loeys, T., & Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological Methods*, *20*(2), 204. doi: 10.1037/a0036368

Moreno-Betancur, M., Moran, P., Becker, D., Patton, G., & Carlin, J. B. (2019, Jul). Defining mediation effects for multiple mediators using the concept of the target randomized trial. *arXiv e-prints*, arXiv:1907.06734.

Moreno-Betancur, M., & Carlin, J. B. (2018). Understanding interventional effects: A more natural approach to mediation analysis? *Epidemiology*, *29*(5), 614–617. doi: 10.1097/EDE.0000000000000866

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. doi: 10.1093/biomet/82.4.702

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearl, J. (2012a). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York, NY, USA: Guilford Press. doi: 10.21236/ADA557445

Pearl, J. (2012b). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*(4), 426–436. doi: 10.1007/s11121-011-0270-1

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459. doi: 10.1037/a0036434

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. doi: 10.3758/BRM.40.3.879

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

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rijnhart, J. J., Twisk, J. W., Eekhout, I., & Heymans, M. W. (2019). Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. *BMC Medical Research Methodology*, *19*(1), 19. doi: 10.1186/s12874-018-0654-z

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–133). New York, NY, USA: Springer-Verlag. doi: 10.1007/978-1-4612-1284-3_2

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143–155. doi: 10.1097/00001648-199203000-00013

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79*(385), 41-48. doi: 10.1080/01621459.1984.10477060

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, *37*(6), 1011–1035. doi: 10.1111/cogs.12058

Shpitser, I., & VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *The International Journal of Biostatistics*, *7*(1), 1 – 24. doi: 10.2202/1557-4679.1297

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. doi: 10.2307/270723

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, *186*(2), 184–193. doi: 10.1093/aje/kwx051

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology*, *37*(4), 235–246. doi: 10.1080/01973533.2015.1062380

Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, *11*(2), 241–269. doi: 10.1177/1094428107300344

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015, 02). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLOS ONE*, *10*(2), 1-8. doi: 10.1371/journal.pone.0118489

VanderWeele, T. J., & Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, *25*(4), 473 – 484. doi: 10.1097/EDE.0000000000000105

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, *2*(4), 457–468. doi: 10.4310/SII.2009.v2.n4.a7

VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, *2*(1), 95–115. doi: 10.1515/em-2012-0010

VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, *25*(2), 300. doi: 10.1097/EDE.0000000000000034

Vansteelandt, S., & Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, *28*(2), 258-265. doi: 10.1097/EDE.0000000000000596

Voelkel, J. G., Ren, D., & Brandt, M. J. (2019). Political inclusion reduces political prejudice. *PsyArXiv*. doi: 10.31234/osf.io/dxwpu

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215. doi: 10.1214/aoms/1177732676

Table 1

Average estimates and empirical standard errors (in brackets) of the indirect effect estimators in simulation study 1 where the mediators were causally connected.

Assumed path model	Indirect effect	Truth	Estimate
Correct	PE_1	-1	-1 (0.12)
	PE_2	1	1 (0.12)
	PE_{12}	-2	-2 (0.22)
no U	PE_1	-1	-1.8 (0.41)
	PE_2	1	1.8 (0.15)
	PE_{12}	-2	-2.8 (0.64)
Parallel	IE_1	1	1 (0.25)
	IE_2	-1	-1.01 (0.64)

Table 2

Average estimates and empirical standard errors (in brackets) of the indirect effect estimators for each fitted path model in simulation study 2 with four mediators, one of which was causally unconnected to the other mediators and to treatment.

Assumed path model	Indirect effect	Truth	Estimate
Posited	PE_{12}	1	1.44 (0.21)
	PE_{13}	1	1.88 (0.27)
	PE_{14}	0	-0.18 (0.08)
	PE_{124}	0	0.38 (0.09)
	PE_{134}	0	0.25 (0.06)
Parallel	IE_{1}	0	0 (0.08)
	IE_{2}	1	1 (0.25)
	IE_{3}	1	1 (0.38)
	IE_{4}	0	0 (0.14)

Table 3

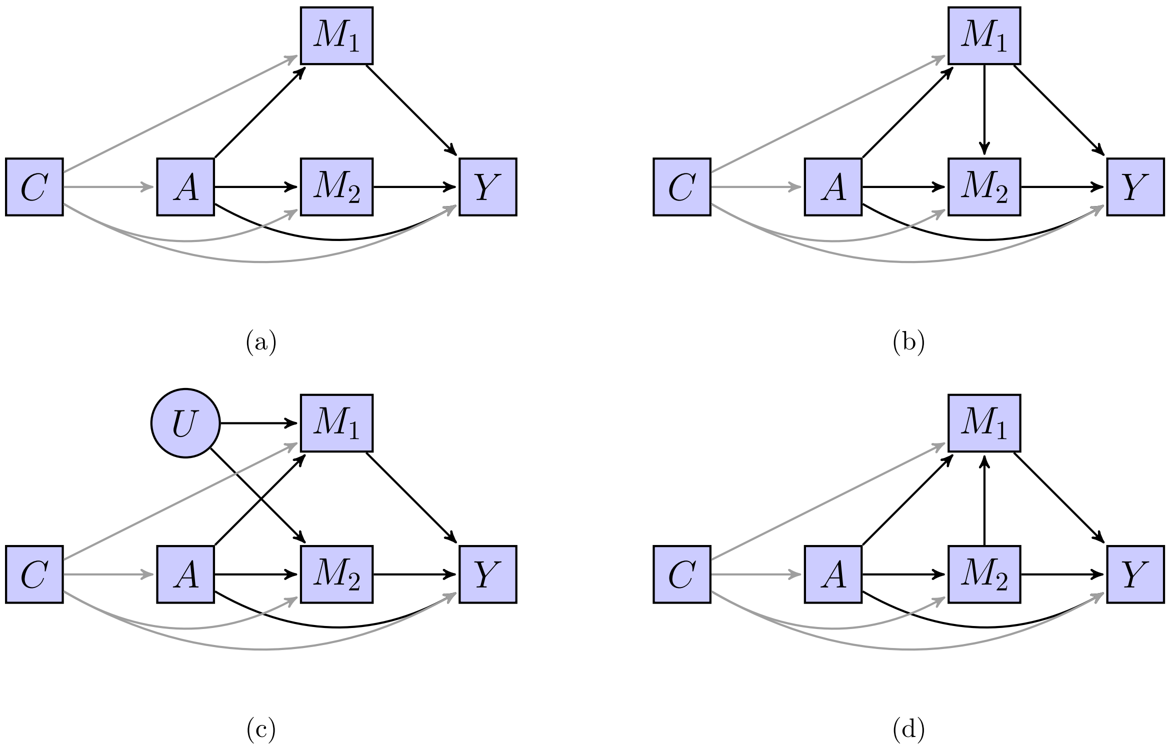
Sample means and standard deviations (in brackets) for the baseline confounders, mediators and outcome for each treatment group in the applied example.

Treatment group	$A = 0$	$A = 1$
Number of participants	95	88
Ideology	-0.51 (0.9)	-0.43 (0.7)
Gender	0.73 (0.4)	0.76 (0.4)
Age	20.0 (2.3)	20.1 (2.5)
M_1 (belong)	0.81 (0.2)	0.69 (0.2)
M_2 (self-esteem)	0.50 (0.2)	0.50 (0.2)
M_3 (control)	0.37 (0.2)	0.30 (0.2)
M_4 (meaningful existence)	0.83 (0.2)	0.81 (0.2)
M_5 (worldview dissimilarity)	0.65 (0.2)	0.56 (0.2)
M_6 (fairness)	0.48 (0.2)	0.62 (0.2)
Y	0.46 (0.2)	0.38 (0.2)

Table 4

Interventional effects estimates, standard errors ("SE") and 95% bootstrap (percentile) confidence intervals ("CI") for the applied example.

Interventional effect	Estimate	Bootstrap SE	95% CI
Direct effect	0.039	0.022	$(-0.003, 0.083)$
Joint indirect effect	-0.115	0.028	$(-0.171, -0.064)$
Indirect effect via M_1 (belong)	-0.003	0.007	$(-0.016, 0.013)$
Indirect effect via M_2 (self-esteem)	0.000	0.002	$(-0.005, 0.005)$
Indirect effect via M_3 (control)	-0.005	0.005	$(-0.017, 0.004)$
Indirect effect via M_4 (meaningful existence)	0.002	0.004	$(-0.006, 0.010)$
Indirect effect via M_5 (worldview dissimilarity)	-0.012	0.007	$(-0.027, -0.002)$
Indirect effect via M_6 (fairness)	-0.097	0.024	$(-0.146, -0.054)$
Total effect	-0.076	0.030	$(-0.135, -0.016)$



(a)

(b)

(c)

(d)

Figure 1. Causal diagrams with two mediators where either (a) M_1 and M_2 are independent conditional on A and C, or (b) M_1 causally precedes M_2, or (c) M_1 and M_2 do not affect each other but share an unobserved common cause U, or (d) M_2 causally precedes M_1. Rectangular nodes denote observed variables, while round nodes denote unobserved variables. For visual clarity, edges emanating from C are drawn in gray.

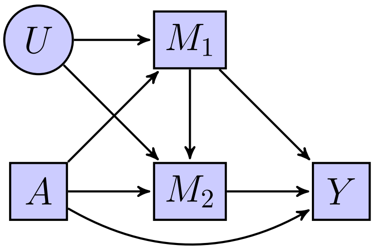


Figure 2. Causal diagram used to generate data for the simulation study with two mediators that share a common cause that does not depend on treatment. Rectangular nodes denote observed variables, while round nodes denote (possibly) unobserved variables.

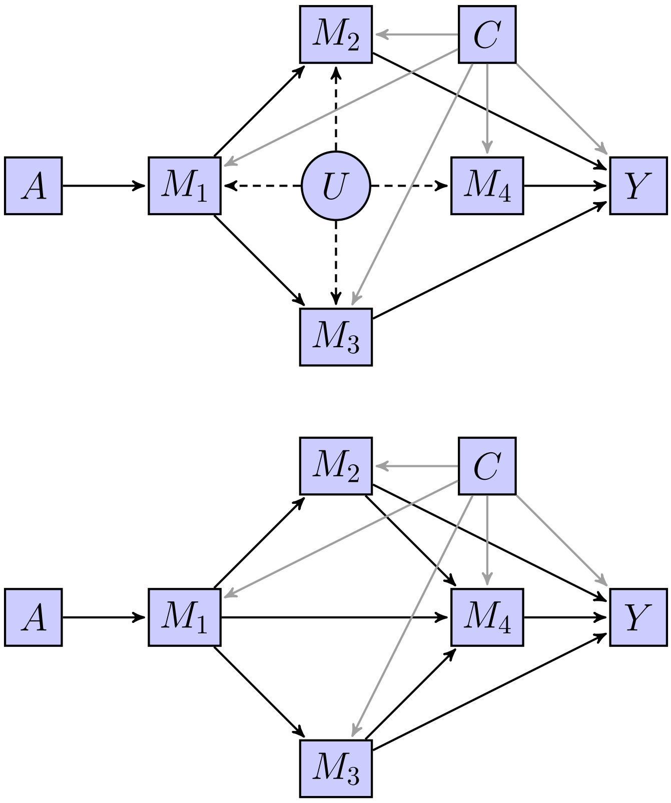


Figure 3. Causal diagrams used to generate data (top), and to posit specific indirect effects (bottom), in simulation study 2 with four candidate mediators. Rectangular nodes denote observed variables, while round nodes denote unobserved variables. For visual clarity, edges emanating from the baseline covariate C are drawn in gray, while edges emanating from the confounder U are drawn as broken lines.

Appendix A

Interventional indirect effects when there are more than two mediators

In general for $t > 2$ mediators, the indirect effect via each distinct mediator $M_s, s = 1, \ldots, t$, can be defined as:

$$\text{IE}_s = \text{E}\left(Y_{1\tilde{M}_{11|C}\cdots\tilde{M}_{s-1,1|C}\tilde{M}_{s1|C}\tilde{M}_{s+1,0|C}\cdots\tilde{M}_{t0|C}}\right) - \text{E}\left(Y_{1\tilde{M}_{11|C}\cdots\tilde{M}_{s-1,1|C}\tilde{M}_{s0|C}\tilde{M}_{s+1,0|C}\cdots\tilde{M}_{t0|C}}\right), \tag{A.1}$$

where the individual values of the first $s-1$ mediators are randomly drawn from their respective (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level(s) $a^{(1)} = \ldots = a^{(s-1)} = 1$, and the last $t-s$ mediators are randomly drawn from their respective (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level(s) $a^{(s+1)} = \ldots = a^{(t)} = 0$. As before, for a given causal diagram, the interventional indirect effect of treatment on outcome via the mediator M_s consists of all causal pathways from A to M_s (possibly intersecting any causal ancestors of M_s), then lead directly from M_s to Y, thus capturing all of the treatment effect that is mediated by M_s and its causal ancestors, but not its causal descendants.

Note that the definition in (A.1) differs from the definition in Vansteelandt & Daniel (2017, eAppendix B, Equation (1)): under the latter definition, the mediators are drawn from a mixture of joint, and possibly marginal, distributions. Under the definition in (A.1), each mediator is always drawn from its marginal distribution, thus making a clearer distinction between the joint (for all mediators) and separate (via each mediator) indirect effects, and providing a simpler expression and interpretation of the indirect effect via the mediators' mutual dependence, as defined later in (C.2) in Appendix C.

Estimators of the interventional (in)direct effects via each mediator can be obtained as follows. Suppose that the outcome obeys the linear and additive mean model:

$$\text{E}(Y|A, M_1, \ldots, M_t, C) = \beta_0 + \beta_A A + \sum_{s=1}^{t} \beta_s M_s + \beta_C C. \tag{A.2}$$

Let $Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\cdots\tilde{M}_{ta^{(t)}|C}}$ denote the potential outcome for Y under treatment $A = a^{(0)}$, when the value of each of the t mediators is set to a random draw from its (counterfactual) marginal distribution under treatment $A = a^{(s)}, s = 1, \ldots, t$. The average potential outcome $\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\cdots\tilde{M}_{ta^{(t)}|C}}\right)$ is identified upon plugging the assumed outcome model (A.2) and mediator marginal distributions into the analogous expression of (3) for t mediators:

$$\mathrm{E}\left(Y_{a^{(0)}\tilde{M}_{1a^{(1)}|C}\cdots\tilde{M}_{ta^{(t)}|C}}\right) = \mathrm{E}\left[\left\{\beta_0 + \beta_A a^{(0)} + \beta_C C\right\} + \sum_{s=1}^{t} \beta_s \,\mathrm{E}(M_s|A = a^{(s)}, C)\right].$$

Suppose that the overall treatment effect on each mediator, given baseline covariate(s) C, is parametrized by the (partial) regression coefficient of treatment A in the following linear and additive (marginal) mean model for each mediator:

$$\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, \quad s = 1, \ldots, t. \tag{A.3}$$

The outcome mean model and mediator (marginal) mean models for the setting with $t = 2$ mediators are stated as (1) and (2) respectively. It follows that the interventional direct effect and indirect effects via each mediator $M_s, s = 1, 2$, are identified by (functions of) the parameters in the assumed models; i.e., $\mathrm{DE} = \beta_A$ and $\mathrm{IE}_s = \beta_s\delta_s$ respectively. Again we note that the overall effect of A on M_s, as encoded by δ_s in (A.3), captures all of the underlying treatment effects that are transmitted through any causal ancestors of M_s. Unbiased estimation of the interventional (in)direct effects thus requires correctly specifying the outcome mean model (A.2) and mediator (marginal) mean models (A.3) under assumptions (A1)–(A3).

As previously noted, the assumed models (A.2) and (A.3) are jointly parametrically equivalent to a parallel path model where the mediators are assumed not to causally affect one another. The interventional indirect effect via each mediator M_s equals the specific indirect effect using the product of coefficients method $\beta_s\delta_s$ for the path $A \to M_s \to Y$ in the parallel path model. Similarly, the interventional direct effect equals the specific effect β_A for the path $A \to Y$ that avoids all the mediators in the parallel path model. Estimators of the interventional effects can thus be obtained by fitting the parallel path model to the observed data using SEM, then plugging in estimates of the (partial) regression coefficients in the respective specific effects.

Standard errors can be estimated using a nonparametric percentile bootstrap procedure (Efron & Tibshirani, 1994) that randomly resamples observations with replacement.

Appendix B

`lavaan` code for the simulation studies and the applied example

Simulation study 1

The following path models were considered for each generated dataset. Here and throughout example `R` code is displayed in gray text boxes.

1. The true data-generating path model.

```
models[["correct"]] <- '
    M1 ~ a1*A + U
    M2 ~ a2*A + U + e12*M1
    Y ~ bA*A + b1*M1 + b2*M2
    pe1 := b1*a1+(b2*e12*a1)
    pe2 := b2*a2
    pe12 := b2*e12*a1'
```

2. A path model that corresponded to the causal diagram of Figure 1(b), with U being unobserved so that M_1 and M_2 shared a hidden common cause (“no U”).

```
models[["noU"]] <- '
    M1 ~ a1*A
    M2 ~ a2*A + e12*M1
    Y ~ bA*A + b1*M1 + b2*M2
    pe1 := b1*a1+(b2*e12*a1)
    pe2 := b2*a2
    pe12 := b2*e12*a1'
```

3. The parallel path model for estimating the interventional indirect effects.

```
models[["parallel"]] <- '
    M1 ~ a1*A
    M2 ~ a2*A
    Y ~ bA*A + b1*M1 + b2*M2
    ie1 := b1*a1
    ie2 := b2*a2'
```

The specified path models were then fitted to each generated dataset using the `sem` function, and the parameter estimates obtained using the `parameterEstimates` function, in `lavaan` as follows:

```
lapply(models, function(mm) {
  fit <- sem(model=mm, data=data, estimator="ML", se="none")
  fit.est <- parameterEstimates(fit)
  return(fit.est)
})
```

Simulation study 2

The following path models were considered for each generated dataset.

1. The path model posited in van der Linden et al. (2015).

   ```
    models[["posited"]] <- ’
      M1 ~ a1*A + C
      M2 ~ e12*M1 + C
      M3 ~ e13*M1 + C
      M4 ~ e14*M1 + e24*M2 + e34*M3 + C
      Y ~ b2*M2 + b3*M3 + b4*M4 + C
      C ~~ 0*A # fix covariate and treatment to be uncorrelated
      pe12 := b2*e12*a1
      pe13 := b3*e13*a1
      pe14 := b4*e14*a1
      pe124 := b4*e24*e12*a1
      pe134 := b4*e34*e13*a1’
   ```

2. The parallel path model for estimating the interventional indirect effects.

   ```
   models[["parallel"]] <- ’
       M1 ~ a1*A + C
       M2 ~ a2*A + C
       M3 ~ a3*A + C
       M4 ~ a4*A + C
       Y ~ bA*A + b1*M1 + b2*M2 + b3*M3 + b4*M4 + C
       C ~~ 0*A # fix covariate and treatment to be uncorrelated
   ```

```
  ie1 := b1*a1
  ie2 := b2*a2
  ie3 := b3*a3
  ie4 := b4*a4’
```

The specified path models were then fitted to each generated dataset using the `sem` function, and the parameter estimates obtained using the `parameterEstimates` function, in `lavaan` as follows:

```
lapply(models, function(mm) {
  fit <- sem(model=mm, data=data, estimator="ML", se="none")
  fit.est <- parameterEstimates(fit)
  return(fit.est)
})
```

Applied example

The following parallel path model was fitted to the observed data, and estimates (and bootstrap standard errors) of the interventional direct and indirect effects obtained, as follows:

```
model <- ’
  Y ~ bA*A + b1*M1 + b2*M2 + b3*M3 + b4*M4 + b5*M5 + b6*M6 + Ideology + Age
      + Gender
  M1 ~ d1*A + Ideology + Age + Gender
  M2 ~ d2*A + Ideology + Age + Gender
  M3 ~ d3*A + Ideology + Age + Gender
  M4 ~ d4*A + Ideology + Age + Gender
  M5 ~ d5*A + Ideology + Age + Gender
  M6 ~ d6*A + Ideology + Age + Gender
  # fix covariates and treatment to be uncorrelated
  Ideology ~~ 0*Age + 0*Gender
  Age ~~ 0*Gender
  A ~~ 0*Ideology + 0*Age + 0*Gender
  # define interventional indirect effects via each mediator
  ie1 := b1*d1
  ie2 := b2*d2
  ie3 := b3*d3
  ie4 := b4*d4
  ie5 := b5*d5
  ie6 := b6*d6
  # define direct effect, sum of indirect effects, and total effect
  de := bA
  ie_jt := ie1 + ie2 + ie3 + ie4 + ie5 + ie6
  de_ie_sum := de + ie_jt
  '
nboots <- 10000
fit <- sem(model, data = data, estimator = "ML", se = "bootstrap",
   bootstrap = nboots, fixed.x=FALSE)
summary(fit)
parameterEstimates(fit)
```

Availability of R code

The R code used to implement the proposed methods and to carry out the simulation studies and the analysis of the applied example are available at the following web address:

https://github.com/wwloh/disentangle-multiple-mediators

Appendix C

Estimation assuming linear mean models

Following Vansteelandt & Daniel (2017), define the total effect as:

$$\begin{aligned}\mathrm{E}\Bigg[\sum_{m_1,m_2}\Big\{&\mathrm{E}(Y_{1m_1m_2}|C)\,\Pr(M_{11}=m_1,M_{21}=m_2|C)\\&-\mathrm{E}(Y_{0m_1m_2}|C)\,\Pr(M_{10}=m_1,M_{20}=m_2|C)\Big\}\Bigg],\end{aligned}\tag{C.1}$$

where the individual values of the mediators M_1 and M_2 are random draws from the (counterfactual) *joint* distribution, instead of the marginal distributions as defined in (7). The total effect in (C.1), henceforth termed the *overall* total effect, can be similarly decomposed into the sum of the separate indirect effects via each mediator (4) and (5), the direct effect (6), and another two additional components, respectively defined as the *indirect effect due to the mediators' mutual dependence* (Vansteelandt & Daniel, 2017):

$$\begin{aligned}\mathrm{E}\Bigg[\sum_{m_1,m_2}\mathrm{E}(Y_{1m_1m_2}|C)\Bigg\{&\Pr(M_{11}=m_1,M_{21}=m_2|C)-\prod_{s=1}^{2}\Pr(M_{s1}=m_s|C)\\&-\Pr(M_{10}=m_1,M_{20}=m_2|C)+\prod_{s=1}^{2}\Pr(M_{s0}=m_s|C)\Bigg\}\Bigg],\end{aligned}\tag{C.2}$$

and the *effect modification by the mediators' mutual dependence acting on the direct effect*:

$$\begin{aligned}\mathrm{E}\Bigg[\sum_{m_1,m_2}&\{\mathrm{E}(Y_{1m_1m_2}|C)-\mathrm{E}(Y_{0m_1m_2}|C)\}\\&\times\Bigg\{\Pr(M_{10}=m_1,M_{20}=m_2|C)-\prod_{s=1}^{2}\Pr(M_{s0}=m_s|C)\Bigg\}\Bigg].\end{aligned}\tag{C.3}$$

The indirect effect (C.2) exists when treatment changes the relationships between the mediators, which in turn affects the outcome. It should therefore be considered separately from the indirect effects via each mediator, as it cannot be attributed to any one mediator alone. The direct effect modification (C.3) exists when the direct effect that is defined in (6) using random draws from the mediators' (counterfactual) marginal distributions differs from the direct effect, to be defined later in (C.10), using the mediators' (counterfactual) *joint* distribution.

In the following, we derive closed form expressions of the estimators of the interventional (in)direct effects for the two mediator setting, assuming linear models for the means of the mediators and outcome under assumptions (A1)–(A3). Suppose that the outcome obeys the linear mean model:

$$\begin{aligned}\mathrm{E}(Y|A,M_1,M_2,C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_C C \\ + \beta_{12} M_1 M_2 + \beta_{A1} A M_1 + \beta_{A2} A M_2 + \beta_{A12} A M_1 M_2,\end{aligned} \tag{C.4}$$

and that the linear mean models for the mediators are as specified in (2), restated here as: $\mathrm{E}(M_s|A,C) = \delta_{0s} + \delta_s A + \delta_{Cs} C$, $s = 1, 2$. In addition, suppose that the covariance of the mediators, given covariates C, differs for different values of treatment A, which we denote by $\mathrm{cov}(M_1, M_2|A, C) = \Sigma(A)$ for notational simplicity. For a binary treatment, the interventional indirect effect via mediator M_1 (4) is identified by:

$$\begin{aligned}&\mathrm{E}\Bigg[\sum_{m_1,m_2} \mathrm{E}(Y|A=1,m_1,m_2,C)\Pr(M_2=m_2|A=0,C) \\ &\quad\times \{\Pr(M_1=m_1|A=1,C) - \Pr(M_1=m_1|A=0,C)\}\Bigg] \\ &= \{(\beta_1+\beta_{A1}) + (\beta_{12}+\beta_{A12})(\delta_{02}+\delta_{C2}\mu_C)\}\,\delta_1,\end{aligned} \tag{C.5}$$

and the interventional indirect effect via mediator M_2 (5) is identified by:

$$\begin{aligned}&\mathrm{E}\Bigg[\sum_{m_1,m_2} \mathrm{E}(Y|A=1,m_1,m_2,C)\Pr(M_1=m_1|A=1,C) \\ &\quad\times \{\Pr(M_2=m_2|A=1,C) - \Pr(M_2=m_2|A=0,C)\}\Bigg] \\ &= \{(\beta_2+\beta_{A2}) + (\beta_{12}+\beta_{A12})(\delta_{01}+\delta_{C1}\mu_C)\}\,\delta_2,\end{aligned} \tag{C.6}$$

where we denote $\mu_C = \mathrm{E}(C)$ for simplicity. Similarly, the interventional direct effect (6) is identified by:

$$\begin{aligned}&\mathrm{E}\left[\sum_{m_1,m_2} \{\mathrm{E}(Y|A=1,m_1,m_2,C) - \mathrm{E}(Y|A=0,m_1,m_2,C)\}\prod_{s=1}^{2}\Pr(M_s=m_s|A=0,C)\right] \\ &= \mathrm{E}[\beta_A + \beta_{A1}\,\mathrm{E}(M_1|A=0,C) + \beta_{A2}\,\mathrm{E}(M_2|A=0,C) + \beta_{A12}\,\mathrm{E}(M_1|A=0,C)\,\mathrm{E}(M_2|A=0,C)] \\ &= \beta_A + \beta_{A1}(\delta_{01}+\delta_{C1}\mu_C) + \beta_{A2}(\delta_{02}+\delta_{C2}\mu_C) + \beta_{A12}(\delta_{01}+\delta_{C1}\mu_C)(\delta_{02}+\delta_{C2}\mu_C).\end{aligned} \tag{C.7}$$

Recall that the overall total effect in (C.1) can be decomposed into the sum of the indirect and direct effects (4), (5), (6), the indirect effect due to the mediators' mutual dependence (C.2), and direct effect modification (C.3). The indirect effect due to the mediators' mutual dependence (C.2) is identified by:

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{m_1,m_2} \mathrm{E}(Y|A=1,m_1,m_2,C) \\
&\quad \times \Big\{ \Pr(M_1=m_1,M_2=m_2|A=1,C) - \Pr(M_1=m_1|A=1,C)\Pr(M_2=m_2|A=1,C) \\
&\qquad - \Pr(M_1=m_1,M_2=m_2|A=0,C) + \Pr(M_1=m_1|A=0,C)\Pr(M_2=m_2|A=0,C)\Big\}\Bigg] \\
&= (\beta_{12}+\beta_{A12})\,\mathrm{E}\{\mathrm{cov}(M_1,M_2|A=1,C) - \mathrm{cov}(M_1,M_2|A=0,C)\} \\
&= (\beta_{12}+\beta_{A12})\{\Sigma(1)-\Sigma(0)\}, \qquad \text{(C.8)}
\end{aligned}$$

and the direct effect modification (C.3) is identified by:

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{m_1,m_2} \{\mathrm{E}(Y|A=1,m_1,m_2,C) - \mathrm{E}(Y|A=0,m_1,m_2,C)\} \\
&\quad \times \{\Pr(M_1=m_1,M_2=m_2|A=0,C) - \Pr(M_1=m_1|A=0,C)\Pr(M_2=m_2|A=0,C)\}\Bigg] \\
&= \beta_{A12}\,\mathrm{E}\{\mathrm{cov}(M_1,M_2|A=0,C)\} \\
&= \beta_{A12}\Sigma(0). \qquad \text{(C.9)}
\end{aligned}$$

We make the perhaps obvious point that the indirect effect (C.8) is non-zero if (i) there is non-zero (treatment-)mediator-mediator interaction in the outcome model (C.4), i.e., $\beta_{12}+\beta_{A12} \neq 0$; and (ii) the covariance of the mediators differs depending on treatment, i.e., $\Sigma(1)-\Sigma(0) \neq 0$. Similarly, the direct effect modification (C.9) is non-zero if (i) there is non-zero treatment-mediator-mediator interaction in the outcome model (C.4), i.e., $\beta_{A12} \neq 0$; and (ii) the covariance of the mediators is non-zero in the absence of treatment, i.e., $\Sigma(0) \neq 0$. In the absence of interactions in the outcome model, (C.4) reduces to (1). It follows that the indirect effects via each mediator (C.5) and (C.6) simplify to $\beta_1\delta_1$ and $\beta_2\delta_2$ respectively, and the direct effect (C.7) simplifies to β_A. Furthermore, the indirect effect due to the mediators' mutual dependence (C.8), and direct effect modification (C.9), both equal zero.

Finally, note that Vansteelandt & Daniel (2017) define the direct effect as

$$\mathrm{E}\left[\sum_{m_1,m_2}\{\mathrm{E}(Y_{1m_1m_2}|C)-\mathrm{E}(Y_{0m_1m_2}|C)\}\Pr(M_{10}=m_1,M_{20}=m_2|C)\right], \tag{C.10}$$

where the individual values of the mediators M_1 and M_2 are random draws from the (counterfactual) *joint* distribution, instead of the marginal distributions as defined in (6), given covariates C under hypothetical treatment level $a^{(1)} = a^{(2)} = 0$. It follows that the direct effect as defined in (C.10) equals the sum of the direct effect (6), and the direct effect modification due to the mediators' mutual dependence (C.3). The interventional (in)direct effects defined in (C.10), (4), (5) and (C.2) are thus equivalent to the definitions in Vansteelandt & Daniel (2017, Equations (5)–(8)) for $a = 1, a^* = 0$.