Integrated multiple mediation analysis: A robustness–specificity trade-off in causal structure

An-Shun Tai[1], Sheng-Hsuan Lin[1]*

1. Institute of Statistics, National Chiao Tung University, Hsin-Chu, Taiwan. 1001 University Road, Hsinchu, Taiwan 300

***Corresponding author**

Sheng-Hsuan Lin, MD, ScD
Institute of Statistics, National Chiao Tung University, Hsin-Chu, Taiwan
1001 University Road,
Hsinchu, Taiwan 30010
Cell: +886-3-5712121 ext.56822
E-mail: shenglin@stat.nctu.edu.tw

Abstract

Recent methodological developments in causal mediation analysis have addressed several issues regarding multiple mediators. However, these developed methods differ in their definitions of causal parameters, assumptions for identification, and interpretations of causal effects, making it unclear which method ought to be selected when investigating a given causal effect. Thus, in this study, we construct an integrated framework, which unifies all existing methodologies, as a standard for mediation analysis with multiple mediators. To clarify the relationship between existing methods, we propose four strategies for effect decomposition: two-way, partially forward, partially backward, and complete decompositions. This study reveals how the direct and indirect effects of each strategy are explicitly and correctly interpreted as path-specific effects under different causal mediation structures. In the integrated framework, we further verify the utility of the interventional analogues of direct and indirect effects, especially when natural direct and indirect effects cannot be identified or when cross-world exchangeability is invalid. Consequently, this study yields a robustness–specificity trade-off in the choice of strategies. Inverse probability weighting is considered for estimation. The four strategies are further applied to a simulation study for performance evaluation and for analyzing the Risk Evaluation of Viral Load Elevation and Associated Liver Disease/Cancer data set from Taiwan to investigate the causal effect of hepatitis C virus infection on mortality.

1. Introduction

1.1 Existing methods

Mediation analysis quantifies the role of a mediator or set of mediators in the total causal effect of a known exposure on an outcome; this is crucial for investigating causal mechanisms (MacKinnon, 2008). Because most existing methods are applicable to only one mediator, they do not allow all mechanisms to be captured. Thus, several methods have been proposed for multiple mediators. In particular, path analysis, which is also integrated as part of structural equation modelling, is a standard method for conducting mediation analysis when all variables are continuous. Avin, Shpitser and Pearl (2005) proposed a method for multiple mediators based on the causal inference framework, under which all paths are quantitatively defined based on a counterfactual model; this extended path analysis to discrete variables. Avin et al. (2005) noted that empirical data cannot lead to the identification of all paths. As an alternative, VanderWeele and Vansteelandt (2014) extended the method with a single mediation analysis by treating all multiple mediators as one multivariate mediator and by decomposing the total effect (TE) of the exposure on the outcome into the natural direct effect (NDE) and natural indirect effect (NIE). This method furnishes information regarding the importance of the mediators, but it does not provide detailed information about each mediator. As a trade-off, the order of causal relations among all mediators and the confounders of all mediators are not required.

To determine the importance of each mediator, mediation analysis for path-specific effects (PSEs) can be used. PSEs are derived from the decomposition of TE according to mediation paths. The PSE with no mediator is the direct effect, and the remaining PSEs are the so-called indirect effects. Albert and Nelson (2011) and Daniel et al. (2015) have decomposed TE completely and derived four PSEs by using two causally ordered mediators. However, to identify a PSE, two counterfactuals of the mediator must be independent. Sensitivity analysis was performed to verify this stronger assumption. To avoid this unrealistic assumption, Steen et al. (2017) considered an alternative definition of the multimediation parameter—the expectation of the counterfactual of the outcome for multiple mediators—to partially decompose the TE. Although this decomposition did not yield the full PSEs, it was the finest natural TE decomposition under regular causal assumptions. Recently, the concept of partial decomposition has been implemented for survival outcomes (Huang and Yang, 2017; Huang and Cai, 2015; Tai et al., 2019). Moreover, Lin and VanderWeele (2017) and Lin (2019) applied an interventional approach (Didelez, Dawid and Geneletti, 2012; Geneletti, 2007) to decompose the interventional analogue of TE (iTE) for complete decomposition. The strong assumption of cross-world exchangeability was not required for this approach.

For causally nonordered mediators, Wang, Nelson and Albert (2013) and Taguri, Featherstone and Cheng (2018) have defined the parallel multimediation parameter by extending the mediation formula of one mediator (Avin et al., 2005), and they have then decomposed TE into NDE and mediator-specific NIEs. Because mediators are assumed to be causally independent, their natural causal effects, including NDE and NIEs, can be identified without the strong assumption adopted by Albert and Nelson (2011) and Daniel et al. (2015). In contrast to the previous approaches for a particular causal structure, Vansteelandt and Daniel (2017) proposed a decomposition method to derive interventional causal effects when the causal structure is unknown. Their method was defined in terms of causal effects instead of the mediation parameter, but their interventional causal effects were essentially the intermediate product obtained during the identification of the parallel multimediation parameter in the interventional approach.

1.2 Open questions and contributions of this study

Although the methods outlined above address several issues regarding mediation analysis with multiple mediators, it remains unclear which method ought to be selected when investigating a given causal effect. This difficulty lies in the differences between the definitions, assumptions, and interpretations of these methods. For example, Lin and VanderWeele (2017) and Vansteelandt and Daniel (2017) have both relied on the interventional approach, but they have performed different decomposition strategies, relied on different assumptions, and provided different interpretations of causal effects.

Therefore, to unify these various methods, we construct an integrated framework as a standard for causal mediation analysis with multiple mediators. This framework makes three contributions. First, the proposed framework clarifies the relationships between the assumptions, identification, and interpretation of causal effects in all existing methods. Moreover, four decomposition strategies are proposed: two-way, partially forward (PF), partially backward (PB), and complete decompositions. Existing methods for mediation analysis with multiple mediators (Albert and Nelson, 2011; Daniel et al., 2015; Fasanelli et al., 2019; Huang and Yang, 2017; Lin, 2019; Steen et al., 2017; Taguri et al., 2018; Tai et al., 2019; VanderWeele and Vansteelandt, 2014; VanderWeele, Vansteelandt and Robins, 2014; Vansteelandt and Daniel, 2017; Wang et al., 2013) can be classified into one of these four strategies. The unification of formulations in this article facilitates the comparability of existing methods of mediation analysis. We comprehensively characterize the features of the four strategies and provide a comparison between them; in doing so, we help researchers select the decomposition strategy for mediation analysis that (particularly in its assumptions) is most appropriate to their object of study.

Second, we propose four multimediation formulas corresponding to the four decomposition strategies; these formulas are a generalized version of mediation formula provided by (Pearl, 2009, 2010). Multimediation formulas have been restricted to particular causal mediation structures. For example, the multimediation formula under the PB decomposition strategy is applicable only when the mediators are mutually independent (Taguri et al., 2018; Wang et al., 2013). However, in this study, we demonstrate that the proposed multimediation formulas are adaptable to different mediation structures. Moreover, we demonstrate that the multimediation formula for PB decomposition is structure-free. This implies that the PB decomposition strategy can be implemented to investigate causal effects without considering structure; this allows the causal effects to be interpreted according to the causal structure of interest. The characteristic of structure-free PB decomposition has also been studied by Vansteelandt and Daniel (2017).

Third, we verify the utility of the interventional analogues of direct and indirect effects, which are termed interventional causal effects. In previous studies, the interventional approach has been primarily used when natural causal effects cannot be identified, meaning that the cross-world exchangeability assumptions are invalid (Lin and VanderWeele, 2017; Vansteelandt and Daniel, 2017). However, interventional causal effects can necessarily be derived regardless of mediation conditions. Under the proposed framework, we show that when the natural causal effects and interventional causal effects can be identified, they are derived using an identical multimediation formula for the various strategies. Accordingly, statistical inferences for causal effects that are based on a multimediation formula can be always interpreted as interventional analogues. If the cross-world exchangeability assumptions hold, the results can be further interpreted as natural causal effects based on the cross-world counterfactuals.

The remainder of this article is organized as follows: In Section 2, we introduce the symbolism and assumptions for the integrated framework. Section 3 reviews single mediator analysis and presents four decomposition strategies for mediation analysis with multiple mediators. Section 4 provides the estimation of each strategy through inverse probability weighting. Section 5 describes a simulation study to evaluate the performance of the four strategies. In Section 6, all strategies are illustrated based on the dataset of the Risk Evaluation of Viral Load Elevation and Associated Liver Disease/Cancer (REVEAL) study from Taiwan. Finally, we conclude with a discussion in Section 7.

2. Symbolism and assumptions of the integrated framework

2.1. Symbolism

In Sections 2 and 3, we focus on two mediators in our demonstration. Let A and Y denote the exposure and outcome of interest; $\widetilde{\boldsymbol{M}} = (M_1, M_2)$ denote the two mediators of interest; and

C denote the baseline covariate preceding A.

To define all causal effects, we introduce a counterfactual model (also called the potential outcome model), as follows (Little and Rubin, 2000). Let $X(a)$ be the hypothetical value of X given that A is intervened as a for all a, where X is M_1, M_2, $\widetilde{\boldsymbol{M}}$, or Y. We also define the cross-world counterfactual $Y(a_1, \widetilde{\boldsymbol{M}}(a_2))$ as the counterfactual of Y given that A is a and $\widetilde{\boldsymbol{M}}$ is $\widetilde{\boldsymbol{M}}(a)$, as previously defined.

We now define the interventional counterfactuals. Let $\widetilde{\mathbf{G}}(a) = \{G_1(a), G_2(a)\}$ be the joint random draw of $\widetilde{\boldsymbol{M}}(a) = \{M_1(a), M_2(a)\}$. In contrast to the curly brackets used in $\{G_1(a), G_2(a)\}$, the round-bracket notation in $\big(G_1(a), G_2(a)\big)$ represents $G_i(a)$ as being the separate random draw of $M_i(a)$ for $i = 1$ and 2; $\big(G_1(a), G_2(a)\big)$ are thus mutually independent. If A is a, then $Y(a, \widetilde{\mathbf{G}}(a))$ and $Y(a, G_1(a), G_2(a))$ are the hypothetical values of Y when $\widetilde{\boldsymbol{M}}$ is set to $\widetilde{\mathbf{G}}(a')$ and $\big(G_1(a), G_2(a)\big)$, respectively.

2.2. Causal structure

A causal structure is generally regarded as a necessary assumption for mediation analysis. Precisely, in mediation analysis, the prespecification of a causal structure among mediators is necessary for interpreting the causal relationship but not necessary for identifying and deriving causal effects. For instance, Vansteelandt and Daniel (2017) proposed a novel decomposition strategy for mediation analysis to derive causal effects when the mediation structure is unknown. In this article, we comprehensively reveal the relationship between all effect decomposition strategies and causal structures.

We now list all conditions of the causal structures for the two mediators. The causal effect of A on Y is the effect for the mechanism of interest. M_1 and M_2 are the mediators whose mediated effects in this mechanism must be quantified. Therefore, the causal structure of the mediators (M_1, M_2) fall under one of the following three conditions:

Mediation structure 1 (MS1): *M_1 and M_2 are causally independent.*

Mediation structure 2 (MS2): *M_1 is the cause of M_2.*

Mediation structure 3 (MS3): *M_2 is the cause of M_1.*

The conditions for a causal interpretation of causal effects can be explicitly characterized using causality diagrams; the causality diagrams corresponding to (*MS1*), (*MS2*), and (*MS3*) are shown in Figure 1(a) to (c), respectively. In previous studies, (*MS2*) and (*MS3*) have also been termed as the sequential or ordered mediation structure (Huang and Yang, 2017; Lin, 2019; Steen et al., 2017; Tai et al., 2019), and (*MS1*) has been termed as the parallel or nonordered mediation structure (Taguri et al., 2018; Wang et al., 2013).

To causally interpret the effects of each strategy, we specify PSEs for the three structures. For (*MS1*), three PSEs (PSE_0, PSE_1, and PSE_2) are present. PSE_0 is equal to the direct effect. PSE_1 and PSE_2 are the indirect effects of the exposure on the outcomes mediated solely through M_1 and M_2, respectively. For (*MS2*), the causal mechanism includes four PSEs (PSE_0, PSE_1, PSE_2, and PSE_{12}), where PSE_{12} represents the indirect effect sequentially mediated through M_1 and M_2. Similarly, PSE_0, PSE_1, PSE_2, and PSE_{21} are included in the mechanism for (*MS3*), where PSE_{21} is the indirect effect mediated sequentially through M_2 and M_1.

2.3. Assumptions for identification

In this article, we assume the following consistency and composition assumptions (Gibbard and Harper, 1978; Robins and Greenland, 1992; VanderWeele and Vansteelandt, 2009):

Consistency assumption: The observed value of *Y is equal to the counterfactual value of $Y(a)$ given that A is a.*

The consistency assumption is also called the well-defined assumption (Hernán and Robins, 2020) or the stable unit treatment value assumption (Rubin, 1980). It is also applied to other counterfactual models, including $Y(a, m)$, $M_1(a)$, and $M_2(a, m)$.

Composition assumption: $Y(a) = Y(a, \widetilde{\boldsymbol{M}}(a))$.

For (*MS2*), the composition assumption for M_2 is as follows: $M_2(a) = M_2\big(a, M_1(a)\big)$. Similarly, for (*MS3*), the additional composition assumption for M_1 is stated as $M_1(a) = M_1\big(a, M_2(a)\big)$.

In addition to the consistency and composition assumptions, several types of exchangeability assumptions and cross-world exchangeability assumptions are required for identification in all strategies.

Assumption of Exchangeability between A and Y (Ax1): *No unmeasured confounders are present between A and Y; that is,* $Y(a, \widetilde{\boldsymbol{m}}) \perp A|C$.

Assumption of Exchangeability between $\widetilde{\boldsymbol{M}}$ ***and Y (Ax2):*** *No unmeasured confounders are present between* $\widetilde{\boldsymbol{M}}$ *and* Y*; that is,* $Y(a, \widetilde{\boldsymbol{m}}) \perp \widetilde{\boldsymbol{M}}|C, A$. *Based on the fundamental properties of probability, (Ax2) implies* $Y(a, \widetilde{\boldsymbol{m}}) \perp M_1|C, A$, $Y(a, \widetilde{\boldsymbol{m}}) \perp M_2|C, A$, *and* $Y(a, \widetilde{\boldsymbol{m}}) \perp M_2|C, A, M_1$.

Assumption of Exchangeability between $\widetilde{\boldsymbol{M}}$ ***and A (Ax3):*** *No unmeasured confounders are present between* $\widetilde{\boldsymbol{M}}$ *and A. This assumption comprises four subtypes:*

(Ax3.1) $\widetilde{\boldsymbol{M}}(a) \perp A|C$

(Ax3.2) $M_1(a) \perp A|C$

(Ax3.3) $M_2(a) \perp A|C$

(Ax3.4) $M_2(a, m_1) \perp A|C$ *for any* m_1

Assumption of Exchangeability between $\boldsymbol{M_1}$ ***and*** $\boldsymbol{M_2}$ ***(Ax4):*** *No unmeasured confounders are present between* M_1 *and* M_2*; that is,* $M_2(a, m_1) \perp M_1|A, C$.

Additionally, five cross-world assumptions are required for all strategies. We defined these assumptions in terms of cross-world counterfactuals as follows:

Assumption of cross-world exchangeability 1 (Acx1): $Y(a, \widetilde{\boldsymbol{m}}) \perp \widetilde{\boldsymbol{M}}(a^*)$

Assumption of cross-world exchangeability 2 (Acx2): $Y(a, \widetilde{\boldsymbol{m}}) \perp (M_1(e_1), M_2(e_2))$

Assumption of cross-world exchangeability 3 (Acx3): $M_1(e_1) \perp M_2(e_2)$

Assumption of cross-world exchangeability 4 (Acx4): $M_1(e_1) \perp M_2(e_2, m_1)$

Assumption of cross-world exchangeability 5 (Acx5):

$Y(a, \widetilde{\boldsymbol{m}}) \perp \big(M_1(e_1), M_2(e_2, m_1)\big)$

The absence of time-varying confounders affected by the exposure, including mediator–mediator and mediator–outcome confounders, is necessary (but not sufficient) for the cross-world exchangeability assumptions. In this section, we assumed that all time-varying confounders can be captured by C.

3. Causal estimand, interventional analogue, and multimediation formula for various decomposition strategies

3.1. Review of causal mediation analysis with a single mediator

The average TE of A on Y when $A = 1$ versus $A = 0$ can be defined as $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$. Without loss of generality, we can replace $(1, 0)$ with any two level (a_1, a_0). Moreover, we can replace the difference with any comparative function, such as the risk ratio or odds ratio if Y is a disease status (VanderWeele and Vansteelandt, 2010). We can further replace the expectation with a hazard function if Y is a time-to-event variable (Lange and Hansen, 2011; VanderWeele, 2011a).

When the mechanism includes a single mediator, only one strategy is available for decomposing TE, namely decomposition into a part with the mediator (i.e., NIE) and another part without the mediator (i.e., NDE). These are defined as $\mathrm{NIE} \equiv \Phi(1,1) - \Phi(1,0)$ and $\mathrm{NDE} \equiv \Phi(1,0) - \Phi(0,0)$, where $\mathrm{TE} = \mathrm{NIE} + \mathrm{NDE}$. Here, $\Phi(a,e) \equiv E[Y(a, M(e))]$ is the conventional mediation parameter with respect to a single mediator. Definitions other than NDE and NIE are possible for the direct and indirect effects, such as either the total direct effect and pure indirect effect or controlled direct effect and controlled mediated effect (Hafeman and VanderWeele, 2011; VanderWeele, 2011b). However, these still represent a decomposition of TE into a part with the mediator and a part without the mediator. Additionally, decomposition for both mediation and interaction (VanderWeele, 2014; VanderWeele and Shrier, 2016) is not considered in this study.

3.2. Effect decomposition strategies for causal mediation analysis with multiple mediators

For multiple mediators, several options are available for effect decomposition depending on practical identifiability conditions and the substantive characteristics of the object the researcher is interested in. To classify all existing methods, we propose four strategies for mediation analysis with multiple mediators, namely two-way decomposition, PF decomposition, PB decomposition, and complete decomposition. Interpretations of the causal mechanism differ between these four strategies. Two-way decomposition is primarily used to interpret the indirect effect mediated through all mediators. PF decomposition and PB decomposition can further decompose mediator-specific (M-specific) indirect effects from the indirect effect determined using two-way decomposition, but the causal interpretations of the

M-specific indirect effects for PF and PB decomposition differ. The M-specific indirect effect of PF decomposition is termed the M-leading indirect effect because it indicates the effect of exposure on the outcome through the mediation paths led by mediator M. By contrast, in the PB decomposition strategy, the M-specific indirect effect is termed the M-inducing indirect effect; this is because the M-inducing indirect effect represents the sum of the effects in which M directly induces the outcome. The complete decomposition strategy enables the extraction of PSEs for all possible mediation paths. The strengths and weaknesses of each strategy are discussed as follows.

Under each decomposition strategy, we propose unified definitions of causal effects in terms of the natural multimediation parameter (Φ) and interventional multimediation parameter (Ψ). Additionally, we unify the multimediation formula (Q) corresponding to the mediation parameter for statistical inference. We then specify the formulations of Φ, Ψ, and Q under the four decomposition strategies. To simplify the notation, we omit the confounders from the following formulations.

3.2.1. Two-way decomposition strategy

In the two-way decomposition strategy, all mediators are treated as one multivariate mediator ($\widetilde{\boldsymbol{M}}$). TE is decomposed into the part passing through $\widetilde{\boldsymbol{M}}$ and the part not passing through $\widetilde{\boldsymbol{M}}$; following the definition for a single mediator, these parts are defined as $\mathrm{NIE_{TW}} \equiv \Phi_{TW}(1,1) - \Phi_{TW}(1,0)$ and $\mathrm{NDE_{TW}} \equiv \Phi_{TW}(1,0) - \Phi_{TW}(0,0)$ (Fasanelli et al., 2019; VanderWeele and Vansteelandt, 2014; VanderWeele et al., 2014), where

$$\Phi_{TW}(a,e) \equiv \mathrm{E}\big[Y(a, \widetilde{\boldsymbol{M}}(e))\big].$$

Herein, Φ_{TW} is the natural multimediation parameter for the two-way decomposition strategy. According to (*Acx1*), (*Ax1*), (*Ax2*), and (*Ax3.1*), we have

$$\Phi_{TW}(a,e) = \mathrm{Q}_{TW}(a,e)\ \textit{a.s.}, \qquad (1)$$

where $\mathrm{Q}_{TW}(a,e) \equiv \int \mathrm{E}[Y|a,\widetilde{\boldsymbol{m}}]\, f(\widetilde{\boldsymbol{m}}|e)\, d\widetilde{\boldsymbol{m}}$. $\mathrm{Q}_{TW}(a,e)$ is the multimediation formula for two-way decomposition. A detailed description of (1) was provided by VanderWeele and Vansteelandt (2014), and it is presented in Appendix A.

Instead of using Φ_{TW}, the causal effects can be alternatively defined for the interventional multimediation parameter, as follows:

$$\Psi_{TW}(a,e) \equiv \mathrm{E}\big[Y(a,\widetilde{\boldsymbol{G}}(e))\big].$$

The causal effects based on Ψ_{TW} for the two-way decomposition strategy are defined as $\mathrm{IIE_{TW}} \equiv \Psi_{TW}(1,1) - \Psi_{TW}(1,0)$ and $\mathrm{IDE_{TW}} \equiv \Psi_{TW}(1,0) - \Psi_{TW}(0,0)$, where IIE and IDE refer to the interventional indirect effect and interventional direct effect, respectively. According to (*Ax1*), (*Ax2*), and (*Ax3.1*), we have

$$\Psi_{TW}(a,e) = \mathrm{Q}_{TW}(a,e) \; a.s., \qquad (2)$$

The equality in (2) is proven in Appendix A. By comparing (1) and (2), two features can be recognized. First, $(\mathrm{NIE_{TW}}, \mathrm{NDE_{TW}})$ and $(\mathrm{IIE_{TW}}, \mathrm{IDE_{TW}})$ are defined in terms of $\Phi_{TW}(a,e)$ and $\Psi_{TW}(a,e)$, which are identified by the identical multimediation formula $\mathrm{Q}_{TW}(a,e)$. Thus, the inference for two-way decomposition relies only on $\mathrm{Q}_{TW}(a,e)$ for the natural or interventional multimediation parameter. Second, identifying $\Phi_{TW}(a,e)$ requires the additional assumption (*Acx1*) compared with the identification of $\Psi_{TW}(a,e)$. Table 1 lists the required assumptions for each strategy. Therefore, based on these two features, we conclude that the causal effects of the two-way decomposition strategy necessarily have interventional causal interpretations. If a study satisfies the cross-world exchangeability assumption (*Acx1*), then the corresponding quantity differences of $\mathrm{Q}_{TW}(a,e)$ can be interpreted as representing natural causal effects. This provides the guidelines for the two-way decomposition strategy.

Notably, the two-way decomposition strategy requires minimal assumptions (Table 1). For example, (*Ax4*) is not required for two-way decomposition. Moreover, a causal mediation structure is not required for two-way decomposition. However, although two-way decomposition furnishes the causal effect mediated by a given set of mediators, it cannot furnish the detailed causal mechanism concerning a particular path of mediators. Thus, if a study is primarily focused on PSEs, then the following three decomposition strategies can provide a finer decomposition of TE under relatively stronger assumptions.

3.2.2. PF decomposition strategy

The PF decomposition strategy has recently been developed for mediation analysis with causally ordered mediators (Huang and Yang, 2017; Steen et al., 2017). For two mediators, this strategy decomposes TE into three parts: via $\mathrm{M_1}$, via $\mathrm{M_2}$, and via either $\mathrm{M_1}$ or $\mathrm{M_2}$, which are defined as $\mathrm{NIE}_{F1} \equiv \Phi_F(1,1,0) - \Phi_F(1,0,0)$, $\mathrm{NIE}_{F2} \equiv \Phi_F(1,1,1) - \Phi_F(1,1,0)$, and $\mathrm{NDE}_F \equiv \Phi_F(1,0,0) - \Phi_F(0,0,0)$, respectively. The natural multimediation parameter under PF decomposition is defined as

$$\Phi_F(a, e_1, e_2) \equiv \mathrm{E}[Y(a, M_1(e_1), M_2(e_2, M_1(e_1)))].$$

As shown in Table 1, based on assumptions (*Acx4*), (*Acx5*), (*Ax1*), (*Ax2*), (*Ax3.2*), (*Ax3.4*), and (*Ax4*), we identify $\Phi_F(a, e_1, e_2)$ as follows:

$$\Phi_F(a, e_1, e_2) = \mathrm{Q}_F(a, e_1, e_2) \ a.s., \qquad (3)$$

where $\mathrm{Q}_F(a, e_1, e_2) \equiv \int E[Y|a, \widetilde{\boldsymbol{m}}] f(m_1|e_1) f(m_2|e_2, m_1) d\widetilde{\boldsymbol{m}}$, which is the multimediation formula under PF decomposition. The proof of (3) was provided by Steen et al. (2017), and it is presented in Appendix A.

We further introduced the interventional multimediation parameter under PF decomposition as

$$\Psi_F(a, e_1, e_2) \equiv E[Y(a, G_1(e_1), G_2(e_2, G_1(e_1)))],$$

where the two instances of $G_1(e_1)$ represent the same random draw. Based on Ψ_F, the interventional analogues of causal effects in PF decomposition are defined as $\mathrm{IIE}_{F1} \equiv \Psi_F(1,1,0) - \Psi_F(1,0,0)$, $\mathrm{IIE}_{F2} \equiv \Psi_F(1,1,1) - \Psi_F(1,1,0)$, and $\mathrm{IDE}_F \equiv \Psi_F(1,0,0) - \Psi_F(0,0,0)$. According to (*Ax1*), (*Ax2*), (*Ax3.2*), (*Ax3.4*), and (*Ax4*), we have

$$\Psi_F(a, e_1, e_2) = \mathrm{Q}_F(a, e_1, e_2) \ a.s., \qquad (4)$$

Similar to two-way decomposition, (3) and (4) reveal that the PF decomposition strategy provides a unique multimediation formula for inference. Thus, if the assumptions (*Acx4*) and (*Acx5*) hold, the effects obtained by $\mathrm{Q}_F(a, e_1, e_2)$ have a natural causal interpretation; otherwise, the causal effects should be interpreted through the interventional analogues.

For (*MS2*), NIE_{F2} represents the causal effect mediated solely through M_2. Because the change of exposure status in NIE_{F2} only relates to M_2. NIE_{F1} can be rewritten as the sum of

$$\mathrm{E}[Y(1, M_1(1), M_2(0, M_1(1)))] - \mathrm{E}[Y(1, M_1(1), M_2(0, M_1(0)))]$$

and

$$\mathrm{E}[Y(1, M_1(1), M_2(0, M_1(0)))] - \mathrm{E}[Y(1, M_1(0), M_2(0, M_1(0)))],$$

where the first is interpreted as PSE_{12} and the second as PSE_1. Notably, PSE_1 and PSE_{12} are unidentifiable because of the cross-world exchangeability assumptions (Avin et al., 2005). Therefore, NIE_{F1} includes all the effects first mediated through M_1 (i.e., PSE_1 and PSE_{12}). For some arbitrary number of mediators, we conclude that a particular mediator led the mediation paths corresponding to the M-specific indirect effect of PF decomposition. We refer to this type of indirect effect as an M-leading indirect effect.

3.2.3. PB decomposition strategy

In this section, we propose the PB decomposition strategy, which is an alternative approach to the partial decomposition of TE. Similarly, for two mediators, this strategy decomposes TE into three parts: via M_1, via M_2, and neither via M_1 nor via M_2, which are defined as $\mathrm{NIE}_{B1} \equiv \Phi_B(1,1,0) - \Phi_B(1,0,0)$, $\mathrm{NIE}_{B2} \equiv \Phi_B(1,1,1) - \Phi_B(1,1,0)$, and $\mathrm{NDE}_B \equiv \Phi_B(1,0,0) - \Phi_B(0,0,0)$, respectively. The natural multimediation parameter under PB decomposition is defined as

$$\Phi_B(a, e_1, e_2) \equiv \mathrm{E}[Y(a, M_1(e_1), M_2(e_2))].$$

As shown in Table 1, based on assumptions (*Acx2*), (*Acx3*), (*Ax1*), (*Ax2*), (*Ax3.2*), and (*Ax3.3*), we identify $\Phi_B(a, e_1, e_2)$ as follows:

$$\Phi_B(a, e_1, e_2) = \mathrm{Q}_B(a, e_1, e_2) \ a.s., \qquad (5)$$

where $\mathrm{Q}_B(a, e_1, e_2) \equiv \int E[Y|a, \widetilde{\boldsymbol{m}}] f(m_1|e_1) f(m_2|e_2) d\widetilde{\boldsymbol{m}}$, which is the multimediation formula under PB decomposition. The proof of (5) is provided in Appendix A. Notably, (*Acx3*) is valid only when the mediators are mutually independent, implying that the identification of Φ_B is restricted to (*MS1*). Recently, several mediation analysis methodologies have been proposed using the PB decomposition strategy to address specific conditions. For example, Wang et al. (2013) and Taguri et al. (2018) have developed methodologies for mediation analysis specifically for the independent mediation structure (*MS1*) based on $\Phi_B(a, e_1, e_2)$.

In contrast to $\Phi_B(a, e_1, e_2)$, the interventional multimediation parameter for PB decomposition is as follows:

$$\Psi_B(a, e_1, e_2) \equiv E[Y(a, G_1(e_1), G_2(e_2))],$$

where $G_1(e_1)$ and $G_2(e_2)$ are separate random draws. This can be identified under three structures because the cross-world exchangeability is not required. More specifically, assuming (*Ax1*), (*Ax2*), (*Ax3.2*), and (*Ax3.3*), we have

$$\Psi_B(a, e_1, e_2) = \mathrm{Q}_B(a, e_1, e_2)\ a.s., \tag{6}$$

The details are provided in Appendix A. The corresponding interventional causal effects are $\mathrm{IIE}_{B1} \equiv \Psi_B(1,1,0) - \Psi_B(1,0,0)$, $\mathrm{IIE}_{B2} \equiv \Psi_B(1,1,1) - \Psi_B(1,1,0)$, and $\mathrm{IDE}_B \equiv \Psi_B(1,0,0) - \Psi_B(0,0,0)$. If (*Acx2*) and (*Acx3*) hold, then (5) and (6) support the interpretation of these interventional causal effects as natural causal effects under (*MS1*). By contrast, under (*MS2*) and (*MS3*), IIE_{B1}, IIE_{B2}, and IDE_B lack natural interpretations of these assumptions because the conventional causal effects of PB decomposition cannot be identified. Thus, the causal effects obtained through the PB decomposition strategy are always treated as interventional analogues of direct and indirect effects regardless of mediation structures, but they are natural only under (*MS1*).

Although the PF and PB decomposition strategies both decompose M-specific indirect effects from TE, as mentioned in Section 3.1, the interpretations of the derived indirect effects are distinct. For (*MS1*), NIE_{B1} and NIE_{B2} (or IIE_{B1} and IIE_{B2}) are the causal effects mediated solely through M_1 and M_2, respectively. For sequential structures, such as (*MS2*) and (*MS3*), IIE_{Bk} represents the sum of PSEs mediated through M_k for $k = 1, 2$. To prove this, we consider (*MS2*); the proof for (*MS3*) follows the same procedure. First, IIE_{B1} can be rewritten as $E[Y(1, G_1(1), G_2(0, G_1(0)))] - E[Y(1, G_1(0), G_2(0, G_1(0)))]$ based on the composition assumption. Clearly, IIE_{B1} is identical to IIE_{F1}, and they represent the causal effect mediated solely through M_1 . Second, based on the composition assumption, $\mathrm{IIE}_{B2} = E[Y(1, G_1(1), G_2(1, G_1(1)))] - E[Y(1, G_1(1), G_2(0, G_1(0)))]$ can be rewritten as the sum of

$$\mathrm{E}[Y(1, G_1(1), G_2(1, G_1(1)))] - \mathrm{E}[Y(1, G_1(1), G_2(1, G_1(0)))]$$

and

$$\mathrm{E}[Y(1, G_1(1), G_2(1, G_1(0)))] - \mathrm{E}[Y(1, G_1(1), G_2(0, G_1(0)))],$$

where the first is interpreted as PSE_{12} and the second as PSE_2. Therefore, IIE_{B2} includes all the effects finally mediated through M_2 (i.e., PSE_2 and PSE_{12}). In general, the M-specific indirect effect of PB decomposition passes through all the mediation paths in which a mediator directly induces the outcome. Therefore, we named the indirect effects of PB decomposition as M-inducing indirect effects.

As shown in Table 1, PB decomposition is the only strategy that allows structure-free decomposition. Structure-free mediation analysis is more useful because prespecifying an appropriate mediation structure is challenging. Vansteelandt and Daniel (2017) also proposed a structure-free decomposition strategy. They defined the direct and indirect effects based on $\Psi_B(a, e_1, e_2)$ by using the following random draw approaches for $G_1(e_1)$ and $G_2(e_2)$: if $e_1 \neq e_2$, then $G_1(e_1)$ and $G_2(e_2)$ are drawn separately, and if $e_1 = e_2$, then $G_1(e_1)$ and $G_2(e_2)$ are drawn jointly. Therefore, this decomposition essentially mixes the proposed PB decomposition with two-way decomposition through interventional analogues of causal effects.

3.2.4. Complete decomposition strategy

In the complete decomposition strategy, TE is decomposed into four parts: solely via M_1, solely via M_2, via the dependence of M_1 and M_2, and neither via M_1 nor via M_2, which can be defined as $\mathrm{NIE}_{C1} \equiv \Phi_C(1,1,0,0) - \Phi_C(1,0,0,0)$, $\mathrm{NIE}_{C2} \equiv \Phi_C(1,1,1,0) - \Phi_C(1,1,0,0)$, $\mathrm{NIE}_{C3} \equiv \Phi_C(1,1,1,1) - \Phi_C(1,1,1,0)$, and $\mathrm{NDE}_C \equiv \Phi_C(1,0,0,0) - \Phi_C(0,0,0,0)$, respectively. The natural multimediation parameter for complete decomposition is defined as

$$\Phi_C(a, e_1, e_2, e_3) \equiv \mathrm{E}[Y(a, M_1(e_1), M_2(e_2, M_1(e_3)))].$$

Although Φ_C can define each PSE, it is generally unidentifiable if no stronger assumptions can be used (Daniel et al., 2015). Therefore, we consider the following interventional analogues of direct and indirect effects: $\mathrm{IIE}_{C1} \equiv \Psi_C(1,1,0,0) - \Psi_C(1,0,0,0)$, $\mathrm{IIE}_{C2} \equiv \Psi_C(1,1,1,0) - \Psi_C(1,1,0,0)$, $\mathrm{IIE}_{C3} \equiv \Psi_C(1,1,1,1) - \Psi_C(1,1,1,0)$, and $\mathrm{IDE}_C \equiv \Psi_C(1,0,0,0) - \Psi_C(0,0,0,0)$. In these expressions, Ψ_C is the interventional multimediation parameter for complete decomposition defined as

$$\Psi_C(a, e_1, e_2, e_3) \equiv \mathrm{E}[Y(a, G_1(e_1), G_2(e_2, G_1(e_3)))],$$

where $G_1(e_1)$ and $G_1(e_3)$ are distinct random draws even when $e_1 = e_3$. Assuming (*Ax1*), (*Ax2*), (*Ax3.2*), (*Ax3.4*), and (*Ax4*), we can prove

$$\Psi_C(a, e_1, e_2, e_3) = \mathrm{Q}_C(a, e_1, e_2, e_3) \textit{ a.s.}, \qquad (7)$$

where

$$\mathrm{Q}_C(a, e_1, e_2, e_3) \equiv \int E[Y|a, \widetilde{\boldsymbol{m}}]\, f(m_1|e_1) \{\int f(m_2|e_2, m_1^*)\, f(m_1^*|e_3)\, dm_1^*\}\, d\widetilde{\boldsymbol{m}}.$$

The details of (7) are presented in Appendix A. In the literature, a generalized form of (7) for an arbitrary number of mediators has been provided by Lin (2019) and Tai et al. (2019). In contrast to the preceding three strategies, the direct and indirect effects obtained using the complete decomposition strategy typically have only interventional interpretations, even when cross-world exchangeability is assumed. However, this strategy can furnish the most detailed mechanism for the causal effect of the exposure on the outcome.

3.3. Robustness–specificity trade-off for the mediation structure based on comparison of PF and PB decompositions

Conventionally, when using PF decomposition strategies, a specific mediation structure must be specified. For example, if (*MS2*) is assumed by virtue of background knowledge, $\Phi_F(a, e_1, e_2)$ or its interventional analogue $\Psi_F(a, e_1, e_2)$ are adapted to define the direct and M-specific indirect effects. They can be identified as $\text{Q}_F(a, e_1, e_2)$ under the aforementioned set of assumptions. By contrast, if M_2 is the cause of M_1 (i.e., (*MS3*) is assumed), then we can swap M$_1$ and M$_2$ and use $\Phi_{F'}(a, e_1, e_2) \equiv \text{E}\big[Y(a, M_1(e_1, M_2(e_2)), M_2(e_2))\big]$ or its interventional analogue $\Psi_{F'}(a, e_1, e_2) \equiv E[Y(a, G_1(e_1, G_2(e_2)), G_2(e_2))]$ to define the direct effect and M-specific indirect effect, which is identified as

$$\text{Q}_{F'}(a, e_1, e_2) \equiv \int E[Y|a, \widetilde{\boldsymbol{m}}] f(m_1|e_1, m_2) f(m_2|e_2) d\widetilde{\boldsymbol{m}}.$$

In this subsection, we demonstrate the interpretation of $\Phi_F(a, e_1, e_2)$, $\Psi_F(a, e_1, e_2)$, and $\text{Q}_F(a, e_1, e_2)$ when the mediation structure is (*MS1*) or (*MS3*). The performance of $\Phi_{F'}(a, e_1, e_2)$, $\Psi_{F'}(a, e_1, e_2)$, and $\text{Q}_{F'}(a, e_1, e_2)$ under (*MS1*) and (*MS2*) is also used for demonstration through an approach similar to that where M_1 and M_2 are swapped. We shall now demonstrate a deep relationship between PF and PB decomposition.

For (*MS1*) and (*MS3*), $\Phi_F(a, e_1, e_2)$ reduces to $\Phi_B(a, e_1, e_2)$ and $\Psi_F(a, e_1, e_2)$ reduces to $\Psi_B(a, e_1, e_2)$ because M_1 does not affect M_2. Therefore, both $\Phi_F(a, e_1, e_2)$ and $\Psi_F(a, e_1, e_2)$ are interpreted as $\Phi_B(a, e_1, e_2)$ and $\Psi_B(a, e_1, e_2)$ (i.e., the corresponding parallel IE$_1$ and M-inducing IE$_2$) under (*MS1*) and (*MS3*), respectively. Under the same identification assumptions, $\Phi_F(a, e_1, e_2)$ and $\Psi_F(a, e_1, e_2)$ can be identified as $\text{Q}_B(a, e_1, e_2)$. Notably, $\text{Q}_F(a, e_1, e_2)$ reduces to and has the same interpretation as $\text{Q}_B(a, e_1, e_2)$ for (*MS1*), but it does not have the corresponding interventional or natural causal interpretation for (*MS3*).

Following a similar logic, we also show that for (*MS1*) and (*MS2*), $\Phi_{F'}(a, e_1, e_2)$ reduces to $\Phi_B(a, e_1, e_2)$ and $\Psi_{F'}(a, e_1, e_2)$ reduces to $\Psi_B(a, e_1, e_2)$. Both $\Phi_{F'}(a, e_1, e_2)$ and $\Psi_{F'}(a, e_1, e_2)$ have the same interpretations of $\Phi_B(a, e_1, e_2)$ and $\Psi_B(a, e_1, e_2)$ if the underlying mediation structure is not correctly specified (i.e., it is *MS1* or *MS2*). Then, $\Phi_{F'}(a, e_1, e_2)$ and $\Psi_{F'}(a, e_1, e_2)$ can be identified as $\mathrm{Q}_B(a, e_1, e_2)$, and $\mathrm{Q}_{F'}(a, e_1, e_2)$ reduces to and has the same interpretation as $\mathrm{Q}_B(a, e_1, e_2)$ for (*MS1*). $\mathrm{Q}_{F'}(a, e_1, e_2)$ has no corresponding causal interpretation for (*MS2*).

Figure 2 summarizes the relation between the PF (in the directions of M_1 and M_2) and PB decompositions. All counterfactual definitions (natural and interventional) of PB and PF decompositions have causal interpretations for (*MS1*), (*MS2*), and (*MS3*). However, the indirect effects defined based on the PB decomposition are always M-inducing for (*MS2*) and (*MS3*), whereas the indirect effects of the PF decomposition are M-leading when the mediation structure is appropriately specified (i.e., *MS2*) and M-inducing when the mediation structure is in the opposite direction (i.e., *MS3*). For (*MS1*), both PB and PF decompositions are reduced to the parallel multiple mediators formula (Taguri et al., 2018). Although the PB decomposition strategy is considerably more robust to different mediation structures than is the PF decomposition strategy, it can only be interpreted as an interventional effect for (*MS2*) and (*MS3*). By contrast, PF decomposition is relatively specific to a certain mediation structure at two levels. In terms of the mediation formula, $\mathrm{Q}_F(a, e_1, e_2)$ has no causal interpretation for (*MS3*), and $\mathrm{Q}_{F'}(a, e_1, e_2)$ has no causal interpretation for (*MS2*). In terms of the mediation parameter, $\Psi_F(a, e_1, e_2)$ has the same interpretation as $\Psi_B(a, e_1, e_2)$ and is identified as Q_B for (*MS1*) and (*MS3*). However, it can be interpreted as both a natural and an interventional indirect effect for (*MS2*). In conclusion, if the mediation structure is assured, the corresponding PF decomposition is recommended because both interventional and natural effects can be derived; however, if the mediation structure is not assured, the PB decomposition is recommended for a more flexible interpretation.

4. Inverse probability of weighting (IPW)

In this study, we adopt IPW to calculate direct and indirect effects for two mediators.

Suppose that $f_{A|C}(a|C)$, $f_{M_1|A,C}(m_1|a,C)$, $f_{M_2|A,C}(m_2|a,C)$, and $f_{M_2|A,M_1,}(m_2|a,m_1,C)$ are the density functions of A, M_1, M_2, and $M_2|M_1$, respectively. The joint density function $\widetilde{\boldsymbol{M}} = (M_1, M_2)$ is referred to as $f_{\widetilde{\boldsymbol{M}}|A,C}(m_1, m_2|a, C)$. Assume that the outcome model is $\mathrm{E}[Y|A = a, \widetilde{\boldsymbol{m}}, C]$. Then, the multimediation parameters of the four strategies are rewritten, and the IPW estimators of each strategy are defined as follows:

<u>Two-way decomposition</u>

$\mathrm{Q}_{TW}(a,e) = \int \mathrm{E}[Y|a, \widetilde{\boldsymbol{m}}, C]\, f_{\widetilde{\boldsymbol{M}}|A,C}(\widetilde{\boldsymbol{m}}|e,C)\, d\widetilde{\boldsymbol{m}} = \mathrm{E}(W_{TW}(a,e;M_1,M_2) \times Y)$,

where $W_{TW}(a,e;M_1,M_2) = [f_{M_1|A,C}(M_1|e,C) f_{M_2|A,M_1,C}(M_2|e,M_1,C)\mathrm{I}(A=a)]/$

$$[f_{A|C}(A|C) f_{M_1|A,C}(M_1|A,C) f_{M_2|A,M_1,C}(M_2|A,M_1,C)].$$

Thus, the IPW estimator for $\mathrm{Q}_{TW}(a,e)$ is

$$\widehat{\Delta}_{TW}^{IPW}(a,e) = \mathbb{P}_n(\widehat{W}_{TW}(a,e;M_1,M_2) \times Y),$$

where $\mathbb{P}_n(X_i) = 1/n \sum_i X_i$ is the empirical average operator, and

$$\widehat{W}_{TW}(a,e;M_1,M_2) = [\hat{f}_{M_1|A,C}(M_1|e,C)\hat{f}_{M_2|A,M_1,C}(M_2|e,M_1,C)\mathrm{I}(A=a)]/$$

$$[\hat{f}_{A|C}(A|C)\hat{f}_{M_1|A,C}(M_1|A,C)\hat{f}_{M_2|A,M_1,C}(M_2|A,M_1,C)].$$

<u>PF decomposition</u>

$$\mathrm{Q}_F(a,e_1,e_2) = \int E[Y|a,\widetilde{\boldsymbol{m}},C] f_{M_1|A,C}(m_1|e_1,C) f_{M_2|A,M_1,C}(m_2|e_2,m_1,C) d\widetilde{\boldsymbol{m}}$$

$$= \mathrm{E}(W_F(a,e_1,e_2;M_1,M_2) \times Y)$$

where $W_F(a,e_1,e_2;M_1,M_2) = [f_{M_1|A,C}(M_1|e_1,C) f_{M_2|A,M_1,C}(M_2|e_2,M_1,C)\mathrm{I}(A=a)]/$

$$[f_{A|C}(A|C) f_{M_1|A,C}(M_1|A,C) f_{M_2|A,M_1,C}(M_2|A,M_1,C)].$$

The IPW estimator for $\mathrm{Q}_F(a,e_1,e_2)$ is

$$\widehat{\Delta}_F^{IPW}(a,e_1,e_2) = \mathbb{P}_n(\widehat{W}_F(a,e_1,e_2;M_1,M_2) \times Y),$$

where $\widehat{W}_F(a,e_1,e_2;M_1,M_2)$ is the weight estimated by substituting $\hat{f}_{A|C}$, $\hat{f}_{M_1|A,C}$, and $\hat{f}_{M_2|A,M_1,C}$.

<u>PB decomposition</u>

$\mathrm{Q}_B(a,e_1,e_2) = \int E[Y|a,\widetilde{\boldsymbol{m}},C] f_{M_1|A,C}(m_1|e_1,C) f_{M_2|A,C}(m_2|e_2,C) d\widetilde{\boldsymbol{m}} = \mathrm{E}(W_B(a,e_1,e_2;M_1,M_2) \times Y)$,

where $W_B(a,e_1,e_2;M_1,M_2) = [f_{M_1|A,C}(M_1|e_1,C) f_{M_2|A,C}(M_2|e_2,C)\mathrm{I}(A=a)]/$

$$[f_{A|C}(A|C) f_{M_1|A,C}(M_1|A,C) f_{M_2|A,M_1,C}(M_2|A,M_1,C)].$$

The IPW estimator for $\mathrm{Q}_B(a,e_1,e_2)$ is

$$\widehat{\Delta}_B^{IPW}(a, e_1, e_2) = \mathbb{P}_n(\widehat{W}_B(a, e_1, e_2; M_1, M_2) \times Y),$$

where $\widehat{W}_B(a, e_1, e_2; M_1, M_2)$ is the weight estimated by substituting $\hat{f}_{A|C}$, $\hat{f}_{M_1|A,C}$, $\hat{f}_{M_2|A,C}$, and $\hat{f}_{M_2|A,M_1,C}$.

<u>Complete decomposition</u>

$$\mathrm{Q}_C(a, e_1, e_2, e_3)$$

$$= \int E[Y|a, \tilde{\boldsymbol{m}}, C]\, f_{M_1|A,C}(m_1|e_1, C)\, \{\int f_{M_2|A,M_1,C}(m_2|e_2, m_1^*, C)\, f_{M_1|A,C}(m_1^*|e_3, C)\, dm_1^*\}\, d\tilde{\boldsymbol{m}}$$

$$= \mathrm{E}(W_C(a, e_1, e_2, e_3; M_1, M_2) \times Y),$$

where $W_C(a, e_1, e_2, e_3; M_1, M_2) = [f_{M_1|A,C}(M_1|e_1, C)$

$$\times \int f_{M_2|A,M_1,C}(M_2|e_2, m_1^*, C) f_{M_1|A,C}(m_1^*|e_3, C) dm_1^*$$

$$\times \mathrm{I}(A = a)]/[f_{A|C}(A|C) f_{M_1|A,C}(M_1|A, C) f_{M_2|A,M_1,C}(M_2|A, M_1, C)].$$

The IPW estimator for $\mathrm{Q}_C(a, e_1, e_2, e_3)$ is

$$\widehat{\Delta}_C^{IPW}(a, e_1, e_2, e_3) = \mathbb{P}_n(\widehat{W}_C(a, e_1, e_2, e_3; M_1, M_2) \times Y),$$

where $\widehat{W}_C(a, e_1, e_2, e_3; M_1, M_2)$ is the weight estimated by substituting $\hat{f}_{A|C}$, $\hat{f}_{M_1|A,C}$, and $\hat{f}_{M_2|A,M_1,C}$.

The aforementioned derivations are detailed in Appendix B.

To determine the IPW, the only remaining step is to estimate the conditional density functions of A, M_1, M_2, and $M_2|M_1$ (i.e., $f_{A|C}$, $f_{M_1|A,C}$, $f_{M_2|A,C}$, and $f_{M_2|A,M_1,C}$). These conditional density functions can be estimated using parametric methods, such as the maximum likelihood (ML) approach, or using nonparametric methods, such as kernel density estimation. In the following analysis, we adopt the ML approach by assuming conditional models to infer direct and indirect effects. As a consequence, $\widehat{W}_{TW}(a, e; M_1, M_2)$, $\widehat{W}_F(a, e_1, e_2; M_1, M_2)$, and $\widehat{W}_B(a, e_1, e_2; M_1, M_2)$ can be directly derived by substituting the estimated density functions into these weights. For $W_C(a, e_1, e_2, e_3)$, the importance sampling and Monte Carlo integration techniques are further incorporated into the estimation procedure because recursive integrations are required to calculate $\widehat{W}_C(a, e_1, e_2, e_3; M_1, M_2)$.

5. Simulation study

5.1. Data generation

To evaluate the finite sample performance of the proposed estimators, we conducted a simulation study using two mediators in the (*MS2*) mediation structure. In the simulations, the baseline confounder C was generated from a Bernoulli distribution with a success probability of 0.5. Conditional on C, the exposure A, mediators (M_1, M_2), and outcome Y were generated as follows:

$A|C \sim Ber\big(p = expit(0.5 + C)\big)$,

$M_1|C, A \sim Norm(\mu = 0.1C + 0.3A, \sigma^2 = 1)$,

$M_2|C, A, M_1 \sim Norm(\mu = 0.3C + 0.5A + 0.1M_1, \sigma^2 = 1)$, and

$Y|C, A, M_1, M_2 \sim Ber\big(p = expit(-0.5 - C + 0.5A + 0.1M_1 + 0.5M_2 + \theta_{int}M_1M_2)\big)$,

where $expit$ denotes the expit function, $Norm$ denotes the normal distribution, and Ber denotes the Bernoulli distribution. In the outcome model, θ_{int} is the interaction parameter, which was separately set as 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, and 7. Simulations were performed 1000 times with a sample size of 10,000 for each value of the interaction parameter.

We subsequently applied the IPW approach for four multimediation formulas to the simulated dataset, and we used the conventional regression-based approach to analyze the simulated dataset for comparison. The regression-based approach is a substitution method for estimation based on fitting the models of the outcome and mediators through the ML approach. In this simulation study, we considered a scenario in which the exposure and mediator models were correctly specified, but the outcome was regressed on M_1 and M_2 only. The model of the outcome was misspecified when θ_{int} was nonzero.

5.2. Results

In the simulation, the direct and indirect effects corresponding to each decomposition strategy were produced separately through regression-based and IPW approaches, and the results are summarized in Figure 3 and Appendix C. In Figure 3, the mediator-specific indirect effects were summed as a single indirect effect, and the biases and 95% confidence intervals were calculated for the different values of the interaction parameter. The results of the mediator-specific indirect effects are detailed in Appendix C.

As expected, the biases of the indirect effects of the regression-based approach in the complete and PB decompositions significantly increased as the model misspecification of the outcome became more severe—that is, the effect of interaction on the outcome model increased (Figure 3). However, for two-way and PF decompositions, the indirect effect estimation using the regression-based approach was unbiased regardless of the increase in the interaction parameter. The regression-based approach is theoretically biased in indirect effect estimation if the outcome is misspecified, but it can tolerate misspecifications of the outcome under the two-way and PF decomposition strategies. By contrast, the IPW approach is robust to the outcome model, regardless of the strategy used.

6. Causal mechanism of hepatitis C virus (HCV) infection on mortality

To apply our framework, we considered the REVEAL-HBV study—a community-based cohort study conducted in Taiwan that assessed the effect of viral hepatitis on the development of hepatocellular carcinoma (HCC) (Chen et al., 2006). In the REVEAL-HBV study, 23,820 residents aged 30–65 years from seven townships of Taiwan were recruited from 1991 to 1992 and followed up until 2008. A total of 477 cases of HCC were reported. HCV and HBV infection status and clinical data, such as alanine aminotransferase (ALT) level and ultrasound images, were measured at baseline. Mortality was confirmed every few years based on Taiwan's death certification system.

We applied the proposed method to the REVEAL-HBV study to investigate the mechanism through which HCV infection affects mortality in patients with HBV. We considered the following two mediators: elevated viral load of HBV—which was defined as viral load > 10,000 copies/mL (Chen et al., 2009)—was regarded as M1, and abnormal ALT was regarded as M2. In the diagnosis of HBV infection, an elevated ALT level indicates immune-mediated inflammation, which eliminates HBV-infected hepatocytes. In particular, high HBV viral load is the cause of abnormal ALT in the mechanism of HBV infection, and (*MS2*) is the potential mediation structure. Although the proposed decision rule suggests a particular strategy for this application in terms of the mediation structure and assumptions, we still analyzed the REVEAL-HBV data by using four strategies separately. Age, sex, and smoking status were included as baseline confounders.

In this study, we adopted the IPW approach for estimating the effects of binary survival status. The estimates of direct and indirect effects on the risk scales for (MS2) are summarized in Table 2. The standard deviations and P values were calculated using bootstrap resampling with 1000 replicates. The complete and PB decompositions both indicate that the indirect effect was mediated solely through the high HBV viral load among patients with HBV-positive status (Table 2). The negative value of this indirect effect reflects the inhibition of HBV replication by HCV. Furthermore, the positive indirect effect mediated solely through abnormal ALT level in the complete and PF decompositions reveals the mechanism of liver damage induced by HCV infection. Comparing the results of the four strategies revealed that the incomplete decomposition strategies, namely the PF, PB, and two-way decompositions, failed to provide meaningful estimates of the indirect effects when the directions of the underlying PSEs were inconsistent. For example, in the two-way decomposition, the indirect effect mediated through all mediators was nonsignificant, whereas the M1- and M2-specific indirect effects were observed in this population through the other deconvolution strategies.

7. Discussion

The investigation of causal mechanisms is crucial in many fields. Using different assumptions and definitions, many researchers have developed methodologies for causal mediation analysis with multiple mediators. Direct and indirect effects can be derived by decomposing TE into several components. In this article, we integrate (with a unified symbolism and set of definitions and assumptions) existing mediation analysis methods by proposing the four decomposition strategies of two-way, PF, PB, and complete decompositions. Based on this integrated framework, we develop the multimediation parameters and multimediation formulas for causal interpretations and statistical inferences, respectively. Moreover, we clarify the correct interpretation of the decomposed indirect effects. Two-way decomposition indicates the entire indirect effect mediated by all mediators; PF decomposition indicates the M-leading indirect effects; PF decomposition indicates the M-inducing indirect effects; and complete decomposition indicates all PSEs. The required assumptions for natural interpretation and interventional interpretation are explicitly specified.

Moreover, we illustrate the robustness–specificity trade-off to reveal the applicability of the four strategies to different mediation structures. The robustness–specificity trade-off permits considerable flexibility for mediation analysis. If researchers have empirical warrant for the mediation structure, a structure-specific strategy such as PF decomposition is suggested for investigating the causal mechanism. By contrast, the PB decomposition strategy is a suitable option to avoid misinterpreting causality when there is uncertainty surrounding the mediation structure.

In the assessment of assumptions, bias formulas for the sensitive analysis of direct and indirect effects under different conditions have recently been proposed (Arah, Chiba and Greenland, 2008; VanderWeele, 2010; VanderWeele and Arah, 2011). As indicated in the proposed decision rule, mediation analysis requires three assumptions: exchangeability between the outcome and exposure, exchangeability between the outcome and mediators, and exchangeability between the mediators and exposure. Thus, the bias formula can facilitate empirical quantification of the effect of bias when an assumption is invalid. We reveal that the remaining assumptions of exchangeability between mediators and cross-world exchangeability are optional for mediation analysis. The assumption of exchangeability between mediators is relative to the choice of PF and PB decomposition. The cross-world exchangeability assumption is related to natural interpretation. Thus, the integrated framework developed in this study aids mediation analysis with multiple mediators.

Acknowledgments

We thank Professor Hwai-I Yang (Genomics Research Center, Academia Sinica, Taipei, Taiwan) for suggesting revisions to refine the model and providing the example data set. We also thank Yi-Juan Du for the simulation study. This study was supported by a grant from the Ministry of Science and Technology in Taiwan (No. 109-2636-B-009 -001). This manuscript was edited by Wallace Academic Editing.

References

Albert, J. M., and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics* **67**, 1028-1038.

Arah, O. A., Chiba, Y., and Greenland, S. (2008). Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Annals of epidemiology* **18**, 637-646.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on Artificial intelligence*, 357-363: Morgan Kaufmann Publishers Inc.

Chen, C.-J., Yang, H.-I., Su, J*., et al.* (2006). Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *Jama* **295**, 65-73.

Chen, C. J., Yang, H. I., Iloeje, U. H., and Group, R. H. S. (2009). Hepatitis B virus DNA levels and outcomes in chronic hepatitis B. *Hepatology* **49**, S72-S84.

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

Didelez, V., Dawid, P., and Geneletti, S. (2012). Direct and indirect effects of sequential treatments. *arXiv preprint arXiv* **arXiv:1206.6840**.

Fasanelli, F., Giraudo, M. T., Ricceri, F., Valeri, L., and Zugna, D. (2019). Marginal Time-Dependent Causal Effects in Mediation Analysis With Survival Data. *American journal of epidemiology* **188**, 967-974.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society Series B* **69**, 199-215.

Gibbard, A., and Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In *Ifs*, 153-190: Springer.

Hafeman, D. M., and VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* **22**, 753-764.

Hernán, M., and Robins, J. (2020). Causal inference: What if. *Boca Raton: Chapman & Hill/CRC.*

Huang, Y.-T., and Yang, H.-I. (2017). Causal Mediation Analysis of Survival Outcome with Multiple Mediators. *Epidemiology* **28**, 370-378.

Huang, Y. T., and Cai, T. (2015). Mediation analysis for survival data using semiparametric probit models. *Biometrics*.

Lange, T., and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology* **22**, 575-581.

Lin, S.-H. (2019). Generalized interventional approach for causal mediation analysis with causally ordered multiple mediators.

Lin, S.-H., and VanderWeele, T. (2017). Interventional Approach for Path-Specific Effects. *Journal of Causal Inference* **5**.

Little, R. J., and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health* **21**, 121-145.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*: Routledge.

Pearl, J. (2009). *Causality: models, reasoning, and inference*, 2nd edition. New York: Cambridge University Press.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics* **6**.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association* **75**, 591-593.

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

Taguri, M., Featherstone, J., and Cheng, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical methods in medical research* **27**, 3-19.

Tai, A.-S., Lin, P.-H., Huang, Y.-T., and Lin, S.-H. (2019). General approach of causal mediation analysis with causally ordered multiple mediators and survival outcome. *Harvard University Biostatistics Working Paper Series*.

VanderWeele, T., and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface* **2**, 457-468.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21**, 540-551.

VanderWeele, T. J. (2011a). Causal interactions in the proportional hazards model. *Epidemiology*

22, 713-717.

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

VanderWeele, T. J. (2014). A unification of mediation and interaction: a four-way decomposition. *Epidemiology* **25**, 749-761.

VanderWeele, T. J., and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22**, 42-52.

VanderWeele, T. J., and Shrier, I. (2016). Sufficient cause representation of the four-way decomposition for mediation and interaction. *Epidemiology* **27**, e32.

VanderWeele, T. J., and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology* **172**, 1339-1348.

VanderWeele, T. J., and Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiol Method* **2**, 95-115.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* **25**, 300-306.

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

Wang, W., Nelson, S., and Albert, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in medicine* **32**, 4211-4228.

Table 1. Assumptions of the four decomposition strategies

	Decomposition strategy.							
	Two-way decomposition		**PF decomposition**		**PB decomposition**		**Complete decomposition**	
Assumptions	*Nature*	*Intervention*	*Nature*	*Intervention*	*Nature*	*Intervention*	*Nature**	*Intervention*
Exchangeability among A and Y								
Ax1: $Y(a,\widetilde{m}) \perp A \mid C$	V	V	V	V	V	V		V
Exchangeability among $\widetilde{M}$ *and Y*								
Ax2.1: $Y(a,\widetilde{m}) \perp \widetilde{M} \mid C, A$	V	V	V	V	V	V		V
Exchangeability among $\widetilde{M}$ *and A*								
Ax3.1: $\widetilde{M}(a) \perp A \mid C$	V	V						
Ax3.2: $M_1(a) \perp A \mid C$			V	V	V	V		V
Ax3.3: $M_2(a) \perp A \mid C$					V	V		
Ax3.4: $M_2(a, m_1) \perp A \mid C$			V	V				V
Exchangeability among M_1 *and* M_2								
Ax4: $M_2(a, m_1) \perp M_1 \mid A, C$			V	V				V
Cross-world Exchangeability								
Acx1: $Y(a,\widetilde{m}) \perp \widetilde{M}(a^*)$	V							
Acx2: $Y(a,\widetilde{m}) \perp \big(M_1(e_1), M_2(e_2)\big)$					V			
Acx3: $M_1(e_1) \perp M_2(e_2)$					V			
Acx4: $M_1(e_1) \perp M_2(e_2, m_1)$			V					
Acx5:$Y(a,\widetilde{m}) \perp \big(M_1(e_1), M_2(e_2, m_1)\big)$			V					

* complete decomposition only identifies interventional causal effects.

Table 2. Effect decomposition of HCV (A) on mortality (Y) through HBV (M1) and abnormal ALT (M2) under the four decomposition strategies.

Path	Strategy							
	Complete decomposition		PF decomposition		PB decomposition		Two-way decomposition	
	effect (SD)	P value	effect (SD)	P value	effect (SD)	P value	effect (SD)	P value
A→Y	0.080 (0.026)	0.002*	0.080 (0.027)	0.003*	0.080 (0.027)	0.003*	0.080 (0.027)	0.003*
A→M_1→Y	-0.015 (0.005)	0.003*	-0.016 (0.006)	0.004*	-0.015 (0.005)	0.003*	-0.004 (0.007)	0.543
A→M_1→M_2→Y	-0.001 (0.002)	0.399			0.011 (0.004)	0.004*		
A→M_2→Y	0.012 (0.004)	0.002*	0.012 (0.004)	0.002*				
Total effect	0.076 (0.026)	0.004*	0.076 (0.026)	0.004*	0.076 (0.027)	0.004*	0.076 (0.027)	0.004*

Abbreviations: HCV: hepatitis C virus; HBV: hepatitis B virus; ALT: alanine aminotransferase; PF: partially forward; PB: partially backward; SD: standard deviation

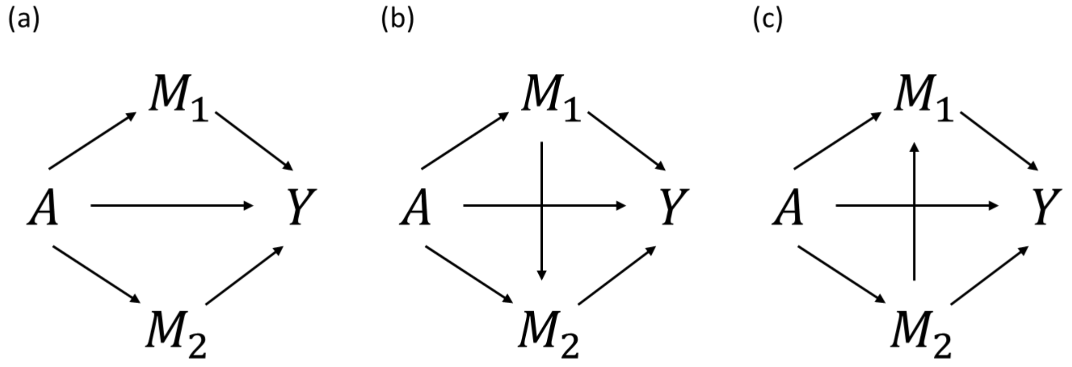

Figure 1. Causality diagram of A, M_1, M_2 and Y where (a) M_1 and M_2 are causally independent; (b) M_1 is the cause of M_2; and (c) M_2 is the cause of M_1.

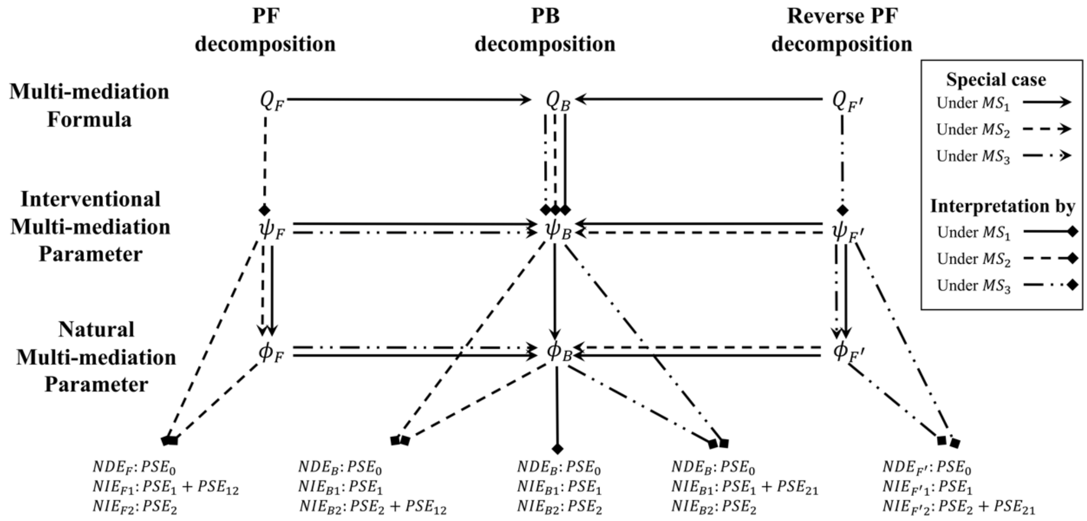

Figure 2. Relationship between PF and PB decompositions.
Abbreviations: NDE: natural direct effect; NIE: natural indirect effect; PF: partially forward; PB: partially backward; **(MS1):** M_1 and M_2 are causally independent; **(MS2):** M_1 is the cause of M_2; **(MS3):** M_2 is the cause of M_1; PSE: path-specific effect.

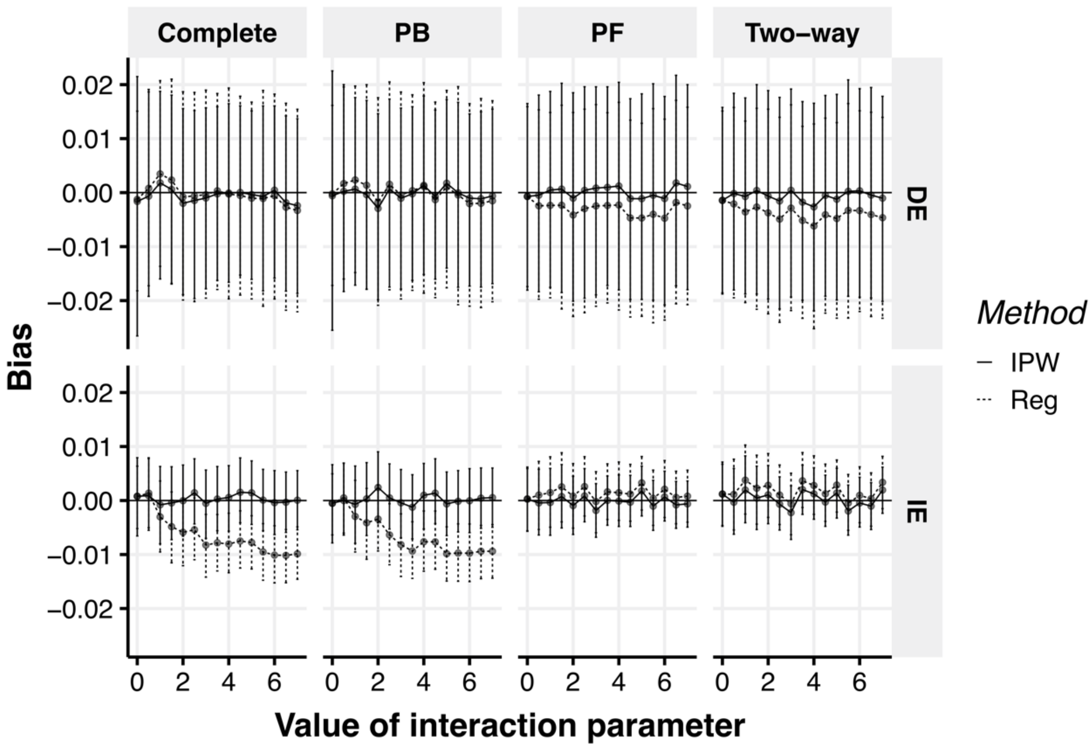

Figure 3. Bias and 95% confidence intervals for direct and indirect effects. The *x* axis represents the value of the interaction parameter of the outcome model. The interaction parameter was set at 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, and 7. The *y* axis represents the bias. Points indicate mean bias, and intervals represent 95% confidence intervals for the different interaction parameters.

Abbreviations: IPW: inverse probability weighting; Reg: regression-based approach; PF: partially forward; PB: partially backward; DE: direct effect; ID: indirect effect.