Identification and robust estimation of swapped direct and indirect effects: Mediation analysis with unmeasured mediator–outcome confounding and intermediate confounding

An-Shun Tai[1] and 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 300
Cell: +886 (3) 5712121 ext.56822
E-mail: shenglin@stat.nctu.edu.tw

Abstract

Counterfactual-model-based mediation analysis can yield substantial insight into the causal mechanism through the assessment of natural direct effects (NDEs) and natural indirect effects (NIEs). However, the assumptions regarding unmeasured mediator–outcome confounding and intermediate mediator–outcome confounding that are required for the determination of NDEs and NIEs present practical challenges. To address this problem, we introduce an instrumental blocker, a novel quasi-instrumental variable, to relax both of these assumptions, and we define a swapped direct effect (SDE) and a swapped indirect effect (SIE) to assess the mediation. We show that the SDE and SIE are identical to the NDE and NIE, respectively, based on a causal interpretation. Moreover, the empirical expressions of the SDE and SIE are derived with and without an intermediate mediator–outcome confounder. Then, a bias formula is developed to examine the plausibility of the proposed instrumental blocker. Moreover, a multiply robust estimation method is derived to mitigate the model misspecification problem. We prove that the proposed estimator is consistent, asymptotically normal, and achieves the semiparametric efficiency bound. As an illustration, we apply the proposed method to genomic datasets of lung cancer to investigate the potential role of the epidermal growth factor receptor in the treatment of lung cancer.

Keywords: bias formula; mediation analysis; mediator–outcome confounding; multiply robust estimation; swapped direct and indirect effects

1. Introduction

1.1. Mediation analysis and mediator–outcome confounding

Causal mediation analysis is a technique used to investigate the mechanism of a confirmed causal effect. Methods have been proposed for various settings, including binary outcomes, time-varying covariates, and multiple mediators (Huang and Cai, 2015; Lin *et al.*, 2017; Lin *et al.*, 2017; VanderWeele and Tchetgen Tchetgen, 2017; VanderWeele and Vansteelandt, 2010; VanderWeele and Vansteelandt, 2014; Zheng and van der Laan, 2012). Although mediation analysis is popular and well-adapted to various applications, one concern that makes researchers hesitant to adapt causal mediation analysis is mediator–outcome confounding (Pearl, 2001; Robins and Greenland, 1992). Specifically, the aim of causal mediation is to decompose the causal effect of a treatment into its natural direct effect (NDE) and natural indirect effect (NIE) and thus quantify the importance of a particular mediator in the mechanism. To identify the NDE and NIE through causal mediation analysis based on empirical data, we assume that all mediator–outcome (M–Y) confounders are measured (i.e., the "no unmeasured M–-Y confounding" assumption is satisfied) and are not affected by the treatment (i.e., the "no treatment-induced M–Y confounding" or "no intermediate M–Y confounding" assumption is satisfied). However, both these assumptions present practical challenges. M–Y confounding is often not fully controllable even if the treatment is randomly assigned. For example, epidermal growth factor receptor (EGFR) and its cognate ligands are associated with numerous cancers, including lung cancer (Lynch *et al.*, 2004; Pao and Chmielecki, 2010), and they appear to promote solid tumor growth (Nicholson *et al.*, 2001). To explore the mediating role of EGFR in the effect of treatment on cancer mortality, mediation analysis was conducted with the expression of EGFR as the mediator. The assumption of no confounding due to unmeasured EGFR-related mortality is always violated because the

common causes of EGFR and mortality, such as genetic and epigenetic variants, are not fully understood; this makes it challenging to collect comprehensive data. Moreover, the assumption of no intermediate M–Y confounding is also a considerable limitation in practical applications. This is because the NDE and NIE results based on the no intermediate M–Y confounding assumption cannot be verified through randomized controlled trials (RCTs).

1.2. Related works

According to a recent literature review, several methods have been developed for mediation analysis when the assumptions of no unmeasured M–Y confounding and no intermediate M–Y confounding are infeasible. The first technique, called sensitivity analysis, derives the bounds of the bias that arises due to these assumptions not holding and assesses the possible influence on the observed direct and indirect effects accordingly (Ding and Vanderweele, 2016; Hafeman, 2011; Smith and VanderWeele, 2019; VanderWeele and Chiba, 2014). For example, Ding and Vanderweele (2016) proposed sharp bounds of the NDE and NIE to represent the strength of unmeasured M–Y confounding. Although sensitivity analysis is a valuable method for the quantification of robustness to confounding in the M–Y relationship, it cannot be used to identify and estimate the direct and indirect effects when these assumptions are violated.

Some other studies have emphasized the estimation of direct and indirect effects in the presence of an unmeasured M–Y confounder or an intermediate M–Y confounder (Lin and VanderWeele, 2017; Miles *et al.*, 2017; Miles *et al.*, 2020; Talloen *et al.*, 2016; Tchetgen and VanderWeele, 2014; VanderWeele, 2011; Vansteelandt and VanderWeele, 2012). For an intermediate M–Y confounder, Tchetgen and VanderWeele (2014) proposed a monotonicity assumption on the confounder and revealed that the NDE can be nonparametrically identified in the case of a binary confounder. Lin and VanderWeele (2017) applied the interventional approach to define a new measure for estimating the direct and indirect effects without

assuming the absence of intermediate M–Y confounders. In the interventional approach, which was initially proposed by Geneletti (2007), the counterfactual value of the mediator defined for the NIE and NDE is replaced with a stochastic intervention of the counterfactual distribution of the mediator in the absence of a treatment. Similarly, another causal definition for the indirect effect allows for an intermediate M–Y confounder and estimates the path-specific effect through the mediator alone (Miles *et al.*, 2017; Miles *et al.*, 2020). The aforementioned studies have focused on intermediate M–Y confounding; by contrast, Talloen *et al.* (2016) proposed an unbiased estimator of the indirect effect (i.e., the effect mediated through a lower-level mediator) based on linear models. This approach, which was developed for educational psychology, can be used to identify the indirect effect in the presence of unmeasured M–Y confounding at the upper level but not if the lower-level M–Y confounder is unmeasured.

1.3. Unsolved problems and contributions of the present study

Although existing methods have each addressed some set of issues in M–Y confounding, both of the aforementioned assumptions cannot be simultaneously relaxed. Moreover, in most of these methods, the estimated direct and indirect effects cannot be interpreted as the NDE and NIE, respectively. To address these weaknesses, the present study proposes a new model that allows the researcher to assess the direct and indirect effects without having to make either assumption. We introduce a novel quasi-instrumental variable called the instrumental blocker (IB), with behavior similar to that of a conventional instrumental variable (Angrist *et al.*, 1996), to model the intervention. The IB is primarily characterized by an antagonistic interaction with the treatment through the mediator's mechanism, and the intervention blocks the path from the treatment to the mediator. Specifically, when the IB is present, the status of the mediator with treatment is equivalent to the status of mediator without treatment. This equivalence in the presence of the IB facilitates the relaxation of the two M–Y confounding assumptions. The

assumptions and properties of the proposed IB are explicitly detailed in Section 2, and the swapped direct effect (SDE) and swapped indirect effect (SIE) are accordingly defined as alternatives for the NDE and NIE, respectively.

This study makes three substantial contributions. First, we explicitly establish the assumptions of the IB and detail its properties. Based on the IB, the SDE and SIE are interpreted as the NDE and NIE, respectively, and they rely on testable assumptions. Moreover, we conduct a sensitivity analysis by establishing bias formulas for the SDE and SIE to evaluate the plausibility of the assumptions required for the IB. Second, as mentioned, the SDE and SIE can be identified in the presence of unmeasured M–Y confounding and intermediate M–Y confounding. Two empirical expressions are separately derived with and without an intermediate M–Y confounder. By excluding both of the M–Y confounding assumptions from identification, the SDE and SIE have the advantage of being widely applicable. Third, we propose a robust estimator for the SDE and SIE based on the union of the three semiparametric model spaces; therefore, this estimator is less sensitive to model misspecification compared with previous methods. Furthermore, the proposed robust estimator is consistent and asymptotically normal. If all models can be correctly specified, then the robust estimator achieves the semiparametric efficiency bound.

The remainder of this paper is organized as follows: Section 2 introduces the definitions, assumptions, and identifications of the SDE and SIE with and without an intermediate M–Y confounder. Section 3 presents the sensitivity analysis and establishes a bias formula to examine the plausibility of the assumptions required for the IB. Section 4 introduces the robust estimators of the SDE and SIE and demonstrates its asymptotic properties. Section 5 details the results of a simulation study conducted to evaluate the performance of the proposed estimator. Section 6 applies the developed method to a genomic study of lung cancer. Finally, in Section 7, we conclude this study by discussing the contributions and limitations of the

proposed method.

2. Swapped direct and indirect effects

2.1. Notations, causal structures, and counterfactual models

Let A denote a binary treatment, M the mediator, C the baseline confounders, Y the outcome, and Z a binary IB. The causal relationships between these variables are illustrated in the directed acyclic graph shown in Figure 1(A). To conduct a causal mediation analysis, we further introduce counterfactual models to define all effects (Robins and Greenland, 1992). Let $Y(a)$ and $M(a)$ denote the counterfactual values of Y and M, respectively, where A is set to a. Similarly, let $Y(a,m)$ denote the counterfactual of Y when M is set to m and A is set to a. Additionally, let $Y(a,M(a^*))$ denote the counterfactual value of Y when the treatment is set to a and when the mediator is set to its value when treatment is set to a^*. Analogously to the counterfactual values for the natural direct and indirect effects, we focus on the counterfactual values when the treatment is set as $A=a$ and the IB is set as $Z=z$; then, the outcome and mediator are defined as $Y(a,z)$ and $M(a,z)$, respectively.

Next, we define consistency and composition assumptions (Gibbard and Harper, 1978; Robins and Greenland, 1992; VanderWeele and Vansteelandt, 2009). According to the consistency assumption for $Y(a,m)$, the outcome Y is observed as $Y(a,m)$ when the observed values of A and M are a and m, respectively. For $M(a)$, the consistency assumption states that the observed mediator M is equal to $M(a)$ when the observed treatment is given by $A=a$. Therefore, according to the composition assumption, $Y(a)=Y(a,M(a))$ and $Y(a,z)=Y(a,z,M(a,z))$.

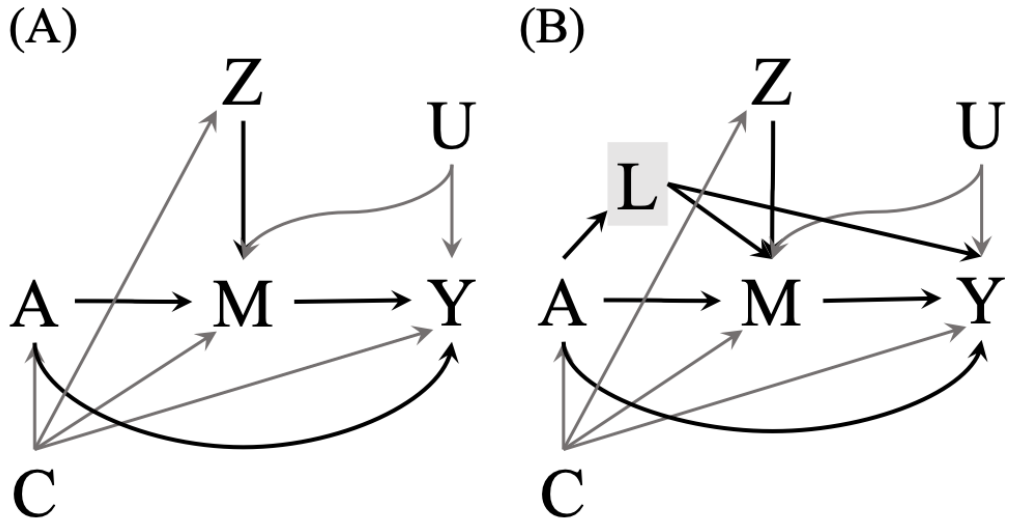

Figure 1. Direct acyclic graph of causal relationships between variables. (A) C, A, M, Z, and Y denote the confounder, treatment, mediator, IB, and outcome, respectively. U represents an unmeasured M–Y confounder. (B) L represents an intermediate M–Y confounder.

2.2. Review of natural direct and indirect effects

First, we consider the standard decomposition of the total causal effect or total effect (TE) in mediation analysis. The goal of mediation analysis is to evaluate the importance of a mediator within the mechanism of a confirmed TE. Technically, this method decomposes the TE into the part that is transmitted through mediator (the NIE) and the part that is not (the NDE). The TE, NDE, and NIE are defined as follows based on an additive scale (Pearl, 2001; Robins and Greenland, 1992): $\psi(1,1) - \psi(0,0)$, $\psi(1,0) - \psi(0,0)$, and $\psi(1,1) - \psi(1,0)$, respectively, where $\psi(a, a^*) \equiv E(Y(a, M(a^*)))$ is termed the conventional mediation parameter (Pearl, 2001; Robins and Greenland, 1992). To identity the NDE and NIE, four assumptions are required: (S1) no unmeasured treatment–mediator confounding; (S2) no unmeasured M–Y confounding; (S3) no unmeasured treatment–outcome confounding; (S4) no intermediate M–Y confounding. Typically, (S1), (S2), and (S3) are referred to as the exchangeability assumptions, and (S4) is termed the cross-world assumption (Robins and

Greenland, 1992). In practice, (S3) can be verified in an RCT, but this verification is difficult to achieve. In addition, (S4) is an untestable assumption, and (S3) and (S4) often restrict the utility of the NDE and NIE for assessing direct and indirect effects. In the following section, we propose a new definition for the direct and indirect effects without (S3) and (S4).

2.3. Definitions of the SDE and SIE

This section proposes alternatives for the NDE and NIE based on the IB. Z serves as an IB with respect to A, M, and Y if Z meets four conditions: (1) Z is associated with M; (2) the presence of Z blocks the path from A to M; (3) in the absence of A, Z has no causal effect on M; (4) Z affects the outcome Y only through M. These four conditions are formalized, respectively, in the following assumptions:

Assumption 1. $P(M|Z) \neq P(M)$.

Assumption 2. $M(a = 1, z = 1) = M(a = 0, z = 1)$.

Assumption 3. $M(a = 0, z) = M(a = 0, z^*) = M(a = 0)$ *for all* z *and* z^*.

Assumption 4. $Y(a, z, m) = Y(a, z^*, m) = Y(a, m)$ *for all* a, z, z^*, *and* m.

Assumptions 1 and 4 for the IB are similar to those for conventional instrumental variables (Angrist *et al.*, 1996). Specifically, Assumption 1 is similar to the relevance assumption for an instrumental variable, and Assumption 4 is similar to the exclusion restriction. In addition, Assumptions 2 and 3 state that the distribution of the mediator with the treatment (i.e., $A = 1$) in the presence of the IB (i.e., $Z = 1$) is identical to the distribution of the mediator without treatment (i.e., $A = 0$). From the perspective of mechanistic interaction, Assumptions 2 and 3 further imply that the IB and the treatment have an antagonistic or agonistic interaction effect on the mediator (Lin *et al.*, 2019). That is, if the treatment is present and the IB is absent, then the mediator is produced. Antagonistic treatment interactions can be observed empirically in many medical or biological studies. For example, in a lung cancer study, the amplification of *YES1*—the gene that encodes a protein that functions as a tyrosine kinase—is a mechanism of

acquired resistance to EGFR inhibition in EGFR-mutant lung cancer (Fan *et al.*, 2018). Therefore, *YES1* may have an antagonistic effect on the therapeutic effects of EGFR-tyrosine kinase inhibitors (TKIs).

Next, we define a new mediation parameter as $\phi(a,z) = E\big(Y(a,z)\big)$; this is the expectation of the counterfactual value of Y with the treatment set to a and the IB set to z. The definitions of the SDE and SIE provided based on this parameter as follows:

Definition 1. SDE and SIE

Given the IB Z, the SDE and SIE of treatment A on the outcome Y are separately defined as

$$SDE \equiv \phi(a=1,z=1) - \psi(a=0,a^*=0) \text{ and}$$

$$SIE \equiv \psi(a=1,a^*=1) - \phi(a=1,z=1),$$

where $\psi(a,a^*) \equiv E(Y(a,M(a^*)))$ *and* $\phi(a,z) = E\big(Y(a,z)\big)$.

The SDE and SIE provide an alternative approach to define the direct and indirect effects. Although the formulations of the SDE and SIE differ from those of the NDE and NIE, the following theorem demonstrates that the SDE and SIE can be strictly interpreted as the NDE and NIE, respectively, under Assumptions 1 to 4.

Theorem 1. (Equivalence)

If an IB satisfies Assumptions 1 to 4 and the composition assumption, then SDE = NDE and SIE = NIE.

To prove Theorem 1, we first prove that $\phi(1,1)$ is identical to $\psi(1,0)$; the proof is as follows:

$$\phi(1,1) \equiv E\big(Y(a=1,z=1)\big)$$
$$= E(Y(a=1,z=1,M(a=1,z=1))) \text{ (by the composition assumption)}$$
$$= E(Y(a=1,z=1,M(a=0,z=1))) \text{ (by Assumption 2)}$$
$$= E(Y(a=1,M(a=0,z=1))) \text{ (by Assumption 4)}$$
$$= E(Y(a=1,M(a=0))) \text{ (by Assumption 3)}$$
$$= \psi(1,0).$$

Accordingly, we can obtain

$$\text{SDE} = \phi(1,1) - \psi(0,0) = \psi(1,0) - \psi(0,0) = \text{NDE} \text{ and}$$
$$\text{SIE} = \psi(1,1) - \phi(1,1) = \psi(1,1) - \psi(1,0) = \text{NIE}.$$

Thus, Theorem 1 guarantees the equivalence of the SDE and SIE with the NDE and NIE. The result of Theorem 1 is valuable because it implies that the empirical expressions of the SDE and SIE (Sections 2.4 and 2.5) can also be used to assess the NDE and NIE. In Section 2.4, we identify the SDE and SIE in the absence of intermediate confounding regardless of M–Y confounding, that is (A4) holds but (A3) is relaxed. Furthermore, we identify the SDE and SIE in the presence of intermediate confounding (A3) but without requiring (A4) in Section 2.5.

2.4. Identification of the SDE and SIE in the absence of intermediate confounding

We assume no intermediate confounding between the mediator and outcome (S4), as illustrated in Figure 1(A). For observational data, a set of confounders C must be included in the analysis to avoid confounding:

Assumption 5. $M(a,z) \perp (A,Z)|C$ *for all* a *and* z.

Assumption 6. $\big(Y(a,z,m), M(a,z)\big) \perp (A,Z)|C$ *for all* a, z, *and* m.

Conditioned on the measured confounders C, Assumptions 5 and 6 state that there is no unmeasured confounding of the associations of the mediator with the treatment and the IB or of the associations of the outcome with the treatment and the IB. These assumptions correspond with (S1) and (S2). The assumption of no unmeasured M–Y confounding is no longer required for the identification of the SDE and SIE because the IB meets the three assumptions for the instrumental variable in the M–Y association. Assumptions 1 and 4 satisfy the relevance and exclusion assumptions, and Assumption 6 implies the exchangeability assumption for the IB and the outcome. Thus, the IB is a proper instrumental variable in the path from the mediator

to the outcome. Assumptions 1 to 6 can be formulated under a nonparametrical structural equation model (NPSEM; Pearl, 2009) to obtain a nonparametric algebraic interpretation of the diagram shown in Figure 1(A). This formulation is detailed in Appendix A.

According to Definition 1, the SDE and SIE are defined in terms of the differences between the conventional mediation parameter $\psi(a, a^*)$ and the proposed mediation parameter $\phi(a, z)$. $\psi(a, a^*)$ and $\phi(a, z)$ should both be identified from data. Given Assumption 6 and the consistency assumption, $\psi(1,1)$ and $\psi(0,0)$ can be simply identified as $\int_c E(Y|A = 1, c)Pr(c)dc$ and $\int_c E(Y|A = 0, c)Pr(c)dc$, respectively, where $Pr(\cdot)$ is a probability function. The identification of $\phi(1,1)$ is described in Theorem 2.

Theorem 2. (Identification of $\phi(1, 1)$ without intermediate confounding)
Under Assumptions 1 to 6 and the consistency assumption, $\phi(1,1) = E(Y(a = 1, z = 1))$ is identified from the data as the expression Q, where

$$Q = \int_{c,m} E(Y|A = 1, Z = 1, m, c)Pr(m|A = 0, c)Pr(c)dv(c, m).$$

In this empirical expression for Q, $v(\chi)$ denotes a probability measure of a combination of random variables χ. The proof of Theorem 2 is provided in Appendix A. The expression of Q does not coincide with the mediation formula proposed by Pearl (2009; 2010) for the empirical expressions of the NDE and NIE. However, if the IB is independent of the outcome conditional on the treatment, mediator, and confounder (i.e., $Y \perp Z|C, M, A$), which is a necessary condition for (S3), then the expression of Q can be reduced to Pearl's mediation formula. According to Theorem 2, the SDE and SIE can be directly identified under Assumptions 1 to 6 as follows:

$$\text{SDE} = \int_{c,m} [E(Y|A = 1, Z = 1, m, c)Pr(m|A = 0, c) - E(Y|A = 0, c)]Pr(c)dv(c, m) \text{ and}$$

$$\text{SIE} = \int_{c,m} [E(Y|A = 1, c) - E(Y|A = 1, Z = 1, m, c)Pr(m|A = 0, c)]Pr(c)dv(c, m)m.$$

Based on Theorem 1, the empirical expressions of the SDE and SIE can be used to quantify the NDE and NIE. Moreover, the assumptions required for the SDE and SIE are more plausible than those required for the NDE and NIE because the results of the SDE and SIE can be verified through RCTs in principle. This reveals that the development of the SDE and SIE provides considerable progress for mediation analysis. Notably, we assume no intermediate confounding in this section. However, this assumption is not necessary for identifying the SDE and SIE; we exclude this untestable assumption in Section 2.5.

2.5. Identification of the SDE and SIE in the presence of intermediate confounding

In this section, we assume that an intermediate confounder, L, is present in the causal diagram, as shown in Figure 1(B). In the presence of L, the assumptions for identification are modified to the following:

Assumption 2'. $M(a=1,z=1,l)=M(a=0,z=1,l)$ *for all* l.
Assumption 3'. $M(a=0,z,l)=M(a=0,z^*,l)=M(a=0,l)$ *for all* z, z^*, *and* l.
Assumption 4'. $Y(a,z,l,m)=Y(a,z^*,l,m)=Y(a,l,m)$ *for all* a, z, z^*, l, *and* m.
Assumption 5'. $\big(M(a,z,l),L(a)\big)\perp(A,Z)|C$ *for all* a, z, *and* l.
Assumption 6'. $(Y(a,z,l,m),M(a,z,l),L(a))\perp(A,Z)|C$ *for all* a, z, l, *and* m.
Assumption 7'. $M(a,z,l)\perp L|C$ *for all* a, z, *and* l.

These assumptions are verified using an NPSEM based on the diagram of Figure 1(B) in Appendix A. In contrast to Assumptions 2 and 3, Assumptions 2' and 3' indicate that the IB can block the path from A to M no matter what the value of the intermediate confounder is. Similarly, according to Assumption 4', the exclusion restriction is independent of L. Conditional on C, Assumptions 5' to 7' ensure no unmeasured confounding from the associations that L, M, and Y each have with A or Z and of the association that M has with L. Unmeasured confounding of the association between M and Y is permitted for the SDE and SIE. The identification of $\phi(1,1)$ in the presence of intermediate confounding is

described in Theorem 3.

Theorem 3. (Identification of $\phi(1,1)$ with intermediate confounding)
Given an intermediate confounder L, under Assumptions 2' to 7' and the consistency assumption, $\phi(1,1) = E(Y(a=1,z=1))$ is identified from data as Q_L, where

$$Q_L = \int_{c,m,l} E(Y|A=1,Z=1,l,m,c) \times Pr(m|A=0,l,c)Pr(l|A=1,c)Pr(c)\, dv(c,m,l).$$

The proof of Theorem 3 is provided in Appendix A. According to Theorem 3, the SDE and SIE can be identified under Assumptions 2' to 7' as follows:

$$\text{SDE} = \int_{c,m,l} [E(Y|A=1,Z=1,l,m,c) \times Pr(m|A=0,l,c)Pr(l|A=1,c) - E(Y|A=0,c)]Pr(c)dv(c,m,l) \text{ and}$$

$$\text{SIE} = \int_{c,m} [E(Y|A=1,c) - E(Y|A=1,Z=1,l,m,c) \times Pr(m|A=0,l,c)Pr(l|A=1,c)]Pr(c)dv(c,m,l).$$

In the following sections, we develop a sensitivity analysis and robust estimator for the SDE and SIE in the absence of intermediate confounding. Both techniques can be extended to the case with an intermediate confounder.

3. Sensitivity analysis and bias formulas

To assess the plausibility of assumptions required for the IB (i.e., Assumptions 1 to 4), this section establishes the bias formulas for the SDE and SIE and evaluates the sensitivity of the results due to the violation of these assumptions. We place particular emphasis on Assumptions 2 to 4. Assumption 1 is the statistical independence of the mediator and the IB, and the methods for testing independence can be applied to this assumption. After relaxing Assumptions 2 to 4, the SDE and SIE can be identified using alternative empirical expressions, which are referred to as wSDE and wSIE, respectively, and obtained as follows:

$$\text{wSDE} = \int_{c,m} [E(Y|A=1,Z=1,m,c)Pr(m|A=0,Z=1,c) - E(Y|A=0,c)]Pr(c)dv(c,m).$$

$$\text{wSIE} = \int_{c,m} [E(Y|A=1,c) - E(Y|A=1,Z=1,m,c)Pr(m|A=0,Z=1,c)]Pr(c)dv(c,m).$$

The detailed derivations of these are shown in Appendix B. Notably, wSDE and wSIE cannot be interpreted as the NDE and NIE because Assumptions 2 to 4 have been relaxed. Intuitively, the difference between the empirical expressions of SDE (or SIE) and wSDE (or wSIE) can be used to quantify the bias arising from the violation of Assumptions 2 to 4. Accordingly, we suggest the bias formulas for the IB in Theorem 4 as follows.

Theorem 4. (Bias formulas for SDE and SIE)

Suppose that the assumptions of no unmeasured confounding (Assumptions 5 and 6) hold and that the IB is a binary variable. Let $\Delta(m,c) \equiv Pr(m|A=0,c) - Pr(m|A=0,Z=1,c)$ *define the conditional correlation between the outcome and IB as* $\rho_{Y,Z}(m,c) \equiv corr(Y,Z|A=1,m,c)$*, and let the conditional probability of the IB be defined as* $p(m,c) \equiv Pr\,(Z=1|A=1,m,c)$*. Then, the bias formulas for SDE and SIE are given by* $B(\Delta,\rho_{Y,Z},p)$ *and* $-B(\Delta,\rho_{Y,Z},p)$*, respectively, where*

$$B\big(\Delta,\rho_{Y,Z},p\big) = \int_{c,m}[E(Y|A=1,m,c)\Delta(m,c) + \rho_{Y,Z}(m,c)\sigma_Y(m,c)\sqrt{(1-p(m,c))/p(m,c)}\Delta(m,c)]\,Pr(c)dv(c,m)$$

and $\sigma_Y(m,c) \equiv Var(Y|A=1,m,c)$.

The proof is given in Appendix B. In Theorem 4, the bias formulas rely on $\Delta(m,c)$, $\rho_{Y,Z}(m,c)$, and $p(m,c)$, which depend on the IB. $\Delta(m,c)$, $\rho_{Y,Z}(m,c)$, and $p(m,c)$ can be the predetermined functions of m and c that are suggested by experts. For example, the function $\rho_{Y,Z}(m,c)$ can be determined based on prior knowledge about the strength of the association between the outcome and IB. Alternatively, $\Delta(m,c)$, $\rho_{Y,Z}(m,c)$, and $p(m,c)$ can be empirically estimated from the data through a parametric approach. $B\big(\Delta,\rho_{Y,Z},p\big)$ can be used to assess how plausible the SDE and SIE are in an application. The nonparametric

bootstrapping method is suggested to estimate $B(\Delta, \rho_{Y,Z}, p)$ in practice.

4. Robust estimation

4.1. Three semiparametric estimators

In this section, we describe estimation methods for the SDE and SIE. We mainly focus on estimating Q because the remaining components of the SDE and SIE (i.e., $\psi(1,1)$ and $\psi(0,0)$) can easily be estimated by using marginal structure models or G-computation. To estimate Q, models must be specified for $E(Y|A,Z,M,C;\alpha)$, $Pr(M|A,Z,C;\beta)$, $Pr(Z|A,C;\gamma)$, and $Pr(A|C;\delta)$, which correspond to the outcome, mediator, treatment, and IB, respectively, and α, β, γ, and δ are the parameters in the corresponding models. We estimate α, β, γ, and δ by using the maximum likelihood approach, and the estimates of these parameters are denoted as $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\delta}$, respectively. We first introduce three semiparametric estimators for Q for the following sets of model assumptions:

(A) $\mathcal{M}_A$: the models for $E(Y|A,Z,M,C;\alpha)$, $Pr(M|A,Z,C;\beta)$, and $Pr(Z|A,C;\gamma)$ are correctly and separately specified.
(B) $\mathcal{M}_B$: the models for $Pr(M|A,Z,C;\beta)$, $Pr(Z|A,C;\gamma)$, and $Pr(A|C;\delta)$ are correctly and separately specified.
(C) $\mathcal{M}_C$: the models for $E(Y|A,Z,M,C;\alpha)$ and $Pr(A|C;\delta)$ are correctly and separately specified.

In $\mathcal{M}_A$, the model for the treatment is unrestricted, in $\mathcal{M}_B$, the model for the outcome is unrestricted, and, in $\mathcal{M}_C$, the models for the mediator and IB are unrestricted. Each semiparametric estimator is based on a specific set of model assumptions. In Section 4.2, we propose a multiply robust estimator of Q based on the three semiparametric estimators for the union of $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$.

The three semiparametric estimators of Q, denoted as $\hat{Q}^A$, $\hat{Q}^B$, and $\hat{Q}^C$, are given by

$$\hat{Q}^A = \mathbb{P}_n\left\{\int_m [E(Y|A=1,Z=1,m,C;\hat{\alpha})\left\{\int_z Pr(m|A=0,z,C;\hat{\beta})Pr(z|A=0,C;\hat{\gamma})dv(z)\right\}dv(m)\right\},$$

$$\hat{Q}^B = \mathbb{P}_n\left\{\frac{\int_z Pr\left(M|A=0,z,C;\hat{\beta}\right)Pr(z|A=0,C;\hat{\gamma})dv(z)\times I(A=1,Z=1)}{Pr\left(M|A,Z,C;\hat{\beta}\right)Pr(Z|A,C;\hat{\gamma})Pr\left(A|C;\hat{\delta}\right)}Y\right\}, \text{ and}$$

$$\hat{Q}^C = \mathbb{P}_n\left\{\frac{I(A=0)}{Pr\left(A|C;\hat{\delta}\right)}E(Y|A=1,Z=1,M,C;\hat{\alpha})\right\},$$

where $\mathbb{P}_n[\cdot] = n^{-1}\sum_i[\cdot]_i$ is the empirical average operator and $I(\cdot)$ is an indicator function. Under standard regularity conditions, $\hat{Q}^A$, $\hat{Q}^B$, and $\hat{Q}^C$ are consistent and asymptotically normal (CAN) for $\mathcal{M}_{\mathrm{A}}$, $\mathcal{M}_{\mathrm{B}}$, and $\mathcal{M}_{\mathrm{C}}$, respectively, based on the central limit theorem and Slutsky's theorem (see Appendix C for details). However, these estimators could be severely biased if their corresponding model assumptions are violated. For example, $\hat{Q}^A$ and $\hat{Q}^B$ are inconsistent if the model for the mediator is misspecified, even if the remaining models are correctly specified. To address this problem, we develop a novel estimator of Q, denoted as $\hat{Q}^R$, that is multiply robust to model misspecification. More specifically, $\hat{Q}^R$ remains CAN if the models are correctly specified for at least one of $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$. Moreover, knowing which model assumptions are correct is unnecessary.

4.2. Multiply robust estimators

To motivate the proposed multiply robust estimator of Q, the following theorem provides the efficient influence function (EIF) for Q in a nonparametric model $\mathcal{M}_{np}$, which does not rests on any assumption on the outcome, mediator, IB, or treatment.

Theorem 5. (EIF in $\mathcal{M}_{np}$)

Under the consistency assumption and on Assumptions 1 to 6, the EIF for Q in the nonparametric model $\mathcal{M}_{np}$ is given by

$$EIF(O;Q) = \frac{Pr(M|A=0,C)I(A=1,Z=1)}{Pr(M|A,Z,C)Pr(Z|A,C)Pr(A|C)}[Y - E(Y|A,Z,M,C)]$$
$$+\frac{I(A=0)}{Pr(A|C)}[E(Y|A=1,Z=1,M,C) - \xi(C)]$$
$$+[\xi(C) - Q],$$

where $\xi(C) = \int_m [E(Y|A=1,Z=1,m,C)Pr(m|A=0,C)dv(m)$ *and* $O=(Y,A,Z,M,C)$ *denotes the observed data. In addition, the semiparametric efficiency bound of* Q *in* $\mathcal{M}_{np}$ *is* $Var(EIF(O;Q))$.

The proof of Theorem 5 is provided in Appendix C. Theorem 5 indicates that any regular and asymptotically linear (RAL) estimators of Q have an identical influence function $EIF(O;\Delta)$ and satisfy

$$\sqrt{n}\left(\hat{Q}-Q\right) = \sqrt{n}\left(\sum\nolimits_i EIF(O_i;Q)\right) + o_p(1),$$

where the probability of $o_p(1)$ converges to 0. Theorem 5 motivates the establishment of a robust estimator of Q under the union model $\mathcal{M}_U = \mathcal{M}_A \cup \mathcal{M}_B \cup \mathcal{M}_C$ as follows:

$$\hat{Q}^R = \mathbb{P}_n\{\ \widehat{W}(M,Z,A,C)[Y - E(Y|A,Z,M,C;\hat{\alpha})] + \frac{I(A=0)}{Pr\left(A\middle|C;\hat{\delta}\right)}\left[E(Y|A=1,Z=1,M,C;\hat{\alpha}) - \hat{\xi}(C)\right] + \hat{\xi}(C)\ \},$$

where

$$\widehat{W}(M,Z,A,C) = \frac{\int_z Pr\left(m\middle|A=0,z,C;\hat{\beta}\right)Pr(z|A=0,C;\hat{\gamma})dv(z) \times I(A=1,Z=1)}{Pr\left(M\middle|A,Z,C;\hat{\beta}\right)Pr(Z|A,C;\hat{\gamma})Pr\left(A\middle|C;\hat{\delta}\right)} \text{ and}$$

$$\hat{\xi}(C) = \int_m [E(Y|A=1,Z=1,m,C;\hat{\alpha})\left\{\int_z Pr\left(m\middle|A=0,z,C;\hat{\beta}\right)Pr(z|A=0,C;\hat{\gamma})dv(z)\right\}dv(m).$$

The robust estimator $\hat{Q}^R$ solves the estimating equation defined as $\mathbb{P}_n\left\{EIF\left(O;Q,\widehat{\boldsymbol{\theta}}\right)\ \right\} = 0$, where $EIF\left(O;Q,\widehat{\boldsymbol{\theta}}\right)$ represents $EIF(O;Q)$ evaluated at $\widehat{\boldsymbol{\theta}} = (\hat{\alpha},\hat{\beta},\hat{\gamma},\hat{\delta})$. Theorem 6 summarizes the primary properties of $\hat{Q}^R$. The proof is provided in Appendix C.

Theorem 6. (Asymptotic property of $\widehat{Q}^R$)

Suppose that the assumptions of Theorem 3 hold and that the regularity conditions of Theorem A.1 in Robins et al. (1992) hold. Then, $\hat{Q}^R$ *is RAL for* $\mathcal{M}_U = \mathcal{M}_A \cup \mathcal{M}_B \cup \mathcal{M}_C$, *and its influence function is given by*

$$IF(O;Q,\boldsymbol{\theta}^*) = EIF(O;Q,\boldsymbol{\theta}^*) - \frac{\partial EIF(O;Q,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{T}}} E\left(\frac{\partial U(O;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{T}}}\right)^{-1} U(O;\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*},$$

where $\boldsymbol{\theta}^*$ *is the probability limit of* $\widehat{\boldsymbol{\theta}}$, *and* $U(O;\boldsymbol{\theta})$ *represents the collection of score*

functions for $Pr(Y|A,Z,M,C;\alpha)$, $Pr(M|A,Z,C;\beta)$, $Pr(Z|A,C;\gamma)$, *and* $Pr(A|C;\delta)$. *Thus,* $\hat{Q}^R$ *is a CAN estimator with asymptotic variance* $E(IF(O;Q,\boldsymbol{\theta}^*)^2)$. *Moreover,* $\hat{Q}^R$ *achieves the semiparametric efficiency bound of* Q *at the intersection submodel* $\mathcal{M}_U = \mathcal{M}_A \cap \mathcal{M}_B \cap \mathcal{M}_C$, *in which all models are correctly specified.*

The asymptotic variance formula of $\hat{Q}^R$ in Theorem 6 follows from the standard M-estimation method (Stefanski and Boos, 2002), which can be implemented in both simulation and application studies. Alternatively, the nonparametric bootstrapping method may be used to estimate the variance and confidence interval in practice (Cheng and Huang, 2010).

5. Simulation studies

In this section, we discuss simulation studies performed to evaluate the finite sample performance of the estimators of Q. For comparison, two conventional methods for mediation analysis, namely G-computation and inverse-probability-weighting (IPW) estimation, are also applied in simulation studies. Notably, the estimators of G-computation and IPW estimation correspond to $\hat{Q}^A$ and $\hat{Q}^B$; thus, G-computation and IPW estimation are expected to suffer from severe bias in the presence of model misspecification. To appropriately mimic the motivating example, we use a binary outcome and a continuous mediator for the simulation study. The data generation for the simulations is detailed as follows:

$C_1 \sim Ber(p = expit(0.5))$,

$C_2 \sim Normal(\mu = 0, \sigma^2 = 1)$,

$A|C_1, C_2 \sim Ber(p = expit(0.5 + C_1 - C_2 - \lambda_1 C_2^3))$,

$Z|C_1, C_2 \sim Ber(p = expit(0.5 - C_1 + C_2))$,

$M|C_1, C_2, A, Z \sim Normal(\mu = 0.5C_1 - 0.5C_2 + A - 0.5Z - AZ + \lambda_2 AC_2, \sigma^2 = 1)$,

$Y|C_1, C_2, A, Z, M \sim Ber(p = expit(0.3C_1 + 0.3C_2 - 0.5A - M + 0.4Z + \lambda_3 AM))$,

where Ber denotes the Bernoulli distribution function, $Normal$ is the normal distribution function, and $expit$ represents the expit function. In these data generating models, λ_1, λ_2,

and λ_3, which are arbitrary numbers, are used to control the degree of model misspecification. Specifically, we fitted $A|C_1, C_2$, $Z|C_1, C_2$, $M|C_1, C_2, A, Z$, and $Y|C_1, C_2, A, Z, M$ as follows:

$$A|C_1, C_2 \sim Ber(p = expit(\delta_{A,0} + \delta_{A,C_1}C_1 + \delta_{A,C_2}C_2)),$$

$$Z|C_1, C_2 \sim Ber(p = expit(\gamma_{Z,0} + \gamma_{Z,C_1}C_1 + \gamma_{Z,C_2}C_2)),$$

$$M|C_1, C_2, A, Z \sim Normal(\mu = \beta_{M,C_1}C_1 + \beta_{M,C_2}C_2 + \beta_{M,A}A + \beta_{M,Z}Z + \beta_{M,AZ}AZ, \sigma^2),$$

$$Y|C_1, C_2, A, Z, M \sim Ber(p = expit(\alpha_{Y,C_1}C_1 + \alpha_{Y,C_2}C_2 + \alpha_{Y,A}A + \alpha_{Y,M}M + \alpha_{Y,Z}Z)).$$

Thus, if λ_1, λ_2, and λ_3 are nonzero in the data generation process, which implies that the models specified in the estimation are inconsistent with the data generating models, then model misspecification occurs. Accordingly, we investigate the following three simulation scenarios:

Scenario (1): the model of the outcome can be misspecified, but the remaining models are correctly specified. That is, in the data generation, $\lambda_1 = 0$, $\lambda_2 = 0$, and λ_3 = 0, 0.5, 1, 1.5, or 2.

Scenario (2): the model of the mediator can be misspecified, but the remaining models are correctly specified. That is, in the data generation, $\lambda_1 = 0$, $\lambda_3 = 0$, and λ_2 = 0, 0.5, 1, 1.5, or 2.

Scenario (3): the model of the treatment can be misspecified, but the remaining models are correctly specified. That is, in the data generation, $\lambda_2 = 0$, $\lambda_3 = 0$, and λ_1 = 0, 0.5, 1, 1.5, or 2.

By using these scenarios, we assess the robustness of G-computation, IPW, and the proposed robust estimation methods when models were misspecified. Simulations were performed 10,000 times with sample sizes of 1,000. The results are summarized in Figure 1.

For Scenario 1, the top panels of Figure 1 indicate that the estimate produced through G-computation became increasingly biased as the degree of misspecification of the outcome model (λ_3) increased. By contrast, the IPW estimation and robust estimation precisely estimated Q regardless of the value of λ_3. This reveals the weakness of G-computation. Although the implementation of G-computation is more straightforward than that of the other methods (Snowden *et al.*, 2011), the outcome model must be correctly specified to ensure an unbiased estimation, which is generally more challenging than correctly specifying the

mediator or the treatment. The center panels of Figure 1 present the results of Scenario 2, in which the mediator model was incorrectly specified. In this scenario, the estimates of Q provided by G-computation and IPW estimation were biased. By contrast, the proposed robust estimation is theoretically consistent in Scenario 2, and the simulation study confirms that the proposed robust estimator was unbiased despite slight increases in the empirical variance of the robust estimator when the degree of mediator model misspecification (λ_2) was increased. In Scenario 3, we assessed the performance of three estimators when the treatment was not correctly specified. The bottom panels of Figure 1 show that the IPW estimator was sensitive to treatment model, whereas the G-computation approach and the proposed method were robust to misspecification of the treatment model. Although the IPW estimator is easily computed due to its straightforward formulation, the applicability of the IPW estimation may be limited if the treatment model is difficult to specify correctly. In summary, the robust estimator substantially outperformed the IPW estimator and G-computation in simulation studies.

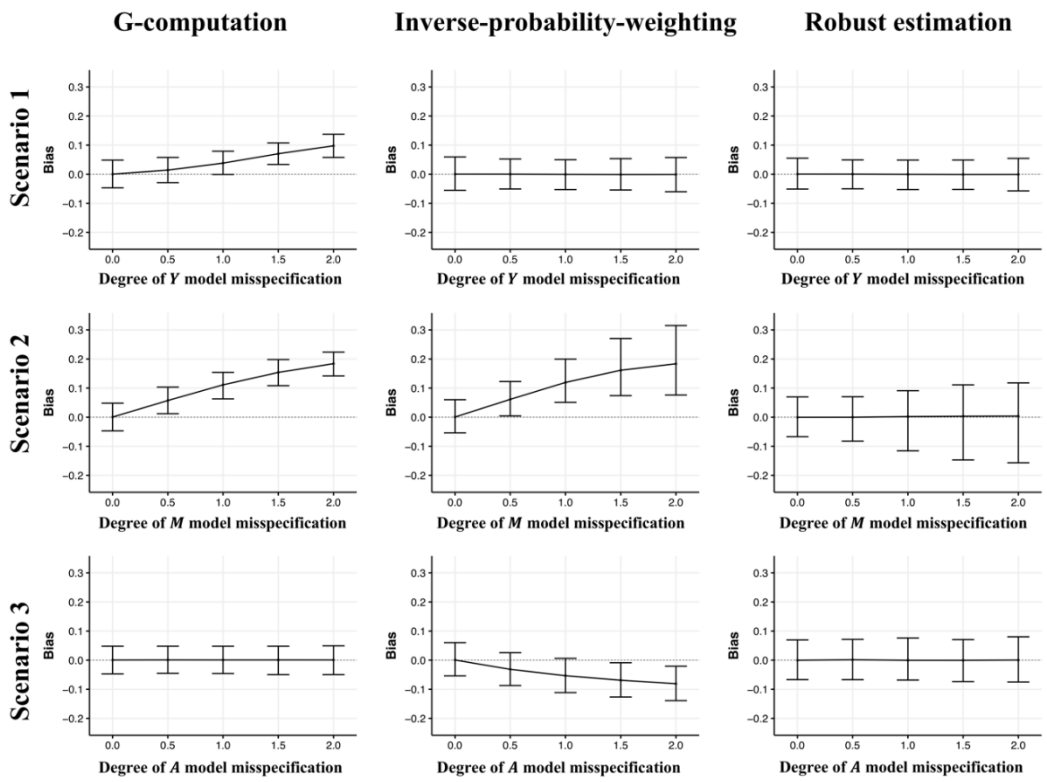

Figure 2. Bias and 95% confidence intervals for Q estimation. The *x* axis represents the degree of model misspecification. For scenarios 1, 2, and 3 the misspecification degrees are denoted by λ_1, λ_2, and λ_3, respectively. The *y* axis represents the bias. Bars represent 95% confidence intervals for the degrees of model misspecification. The dotted horizontal line represents zero bias.

6. Application to genomic datasets of lung cancer

To illustrate our method, we separately analyzed two genomic datasets of lung cancer from The Cancer Genome Atlas. The first dataset comprised data on 502 patients with lung squamous cell carcinoma; 9 of these 502 samples were excluded from the analysis due to incomplete data. The second dataset included 533 patients with lung adenocarcinoma; 19 of these 533 samples were removed after filtering missing data. The gene expression of primary tumor samples collected during surgery was measured using Agilent gene expression arrays. To reduce bias from the abundant transcript reads, the gene expression data were normalized across samples by using the unit of fragments per kilobase of transcript per million mapped reads (FPKM).

The elevated expression of EGFR and its cognate ligands are associated with numerous cancer types, including lung cancer (Lynch *et al.*, 2004; Pao and Chmielecki, 2010), and appear to promote solid tumor growth (Nicholson *et al.*, 2001). To investigate the mechanism of EGFR in the treatment (A) of lung cancer, we applied the proposed method to both datasets and assessed the mediating role of *EGFR* expression in the treatment of patients with lung cancer. Accordingly, we treated the *EGFR* expression as a continuous mediator (M) and the vital status (Y) as the primary outcome. Moreover, clinical studies have revealed the effect of *YES1* amplification on the mechanism of resistance to EGFR inhibitors in lung cancer (Fan *et al.*, 2018; Helena *et al.*, 2018; Ichihara *et al.*, 2017). Therefore, we considered *YES1* amplification (Z)–a dichotomous variable recording whether the gene expression level of *YES1* is abnormal– as the potential IB in the path from the treatment to EGFR in lung cancer. The bias formula for the IB proposed in Section 3 was applied to assess the plausibility of *YES1* amplification in this

study. In addition, demographic variables (age, gender, and ethnicity) and clinical variables (tumor, node, and metastasis staging) were adjusted for as baseline confounders ($\widetilde{\boldsymbol{C}}$). Figure 3 presents the causal diagram.

All variables were fitted according to the causal relationship shown in Figure 3 as follows:

$A|\widetilde{\boldsymbol{C}} \sim Ber(p = expit(\delta_0 + \delta_C\widetilde{\boldsymbol{C}}))$,

$Z|\widetilde{\boldsymbol{C}} \sim Ber(p = expit(\gamma_0 + \gamma_C\widetilde{\boldsymbol{C}}))$,

$M|\widetilde{\boldsymbol{C}}, A, Z \sim Normal(\mu = \beta_0 + \beta_C\widetilde{\boldsymbol{C}} + \beta_A A + \beta_Z Z + \beta_{AZ}AZ, \sigma^2)$,

$Y|\widetilde{\boldsymbol{C}}, A, Z, M \sim Ber(p = expit(\alpha_0 + \alpha_C\widetilde{\boldsymbol{C}} + \alpha_A A + \alpha_M M + \alpha_Z Z))$.

All the parameters in the preceding models were estimated using the regular maximum likelihood approach for lung squamous cell carcinoma and lung adenocarcinoma, separately. Accordingly, the TE, SDE, and SIE were estimated as shown in Table 1. In addition to the SDE and SIE, we further estimated the NDE and NIE for comparison, although the assumption of no unmeasured M–Y confounding for the NDE and NIE was violated. SDE and SIE were estimated using the proposed robust estimation, and the estimations of the TE, NDE, and NIE were obtained by using the conventional IPW approach (Table 1). The confidence intervals were determined by using the nonparametric bootstrapping method with 10,000 bootstraps for simplicity in calculation.

The estimated bias formulas for the IB in lung squamous cell carcinoma and lung adenocarcinoma (Table 1) both indicate that the biases arising from the IB assumptions being violated were slight. This suggested that *YES1* was an appropriate IB in this application. The conclusions regarding the direct and indirect effects from the analyses of lung squamous cell carcinoma and in lung adenocarcinoma were relatively consistent. Specifically, when mediated through *EGFR* expression, the treatment reduced mortality rates by 7.4% and 5.7% in the two datasets. The estimated SDEs were both positive, reflecting a negative therapeutic effect. The current treatment may have no significant effect on the patients without *EGFR* mutation. By contrast, the results obtained using the natural approaches (i.e., the NDE and NIE) were

inconsistent between lung squamous cell carcinoma and lung adenocarcinoma. The inconsistency in the natural approach was probably caused by the violation of the assumption of no unmeasured M–Y confounding.

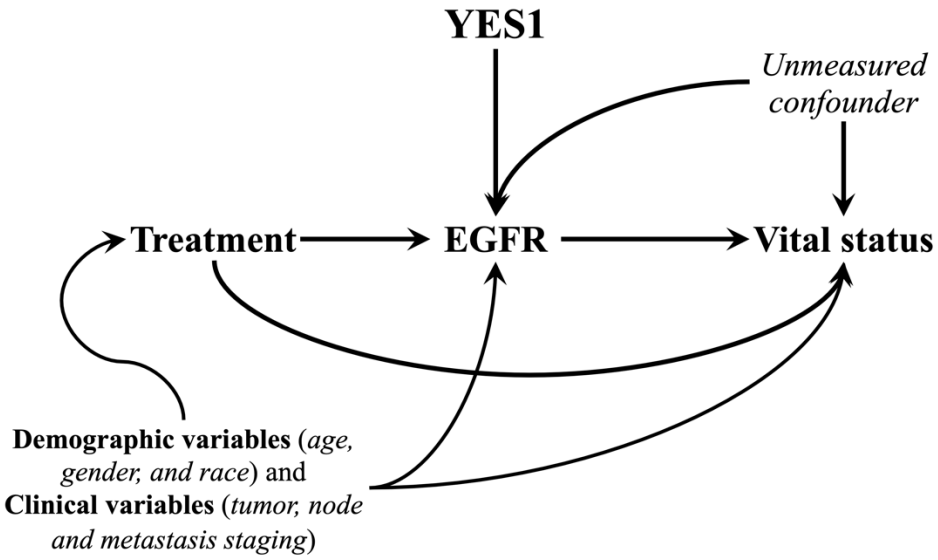

Figure 3. Causal diagram of the application to lung cancer.

Table 1. Results for lung squamous cell carcinoma and lung adenocarcinoma

	Lung squamous cell carcinoma		Lung adenocarcinoma	
	Estimate	95% CI	Estimate	95% CI
Proposed method				
SDE	0.027	(-0.071, 0.124)	0.118	(0.045, 0.189)
SIE	-0.074	(-0.150, -0.003)	-0.057	(-0.095, -0.019)
Natural approach				
NDE	-0.042	(-0.103, 0.018)	0.059	(0.001, 0.114)
NIE	-0.005	(-0.018, 0.005)	0.003	(-0.007, 0.015)
Bias formula	0.019	(-0.013, 0.051)	-0.010	(-0.026, 0.005)
TE	-0.047	(-0.107, 0.011)	0.062	(0.006, 0.116)

Abbreviations: SDE: swapped direct effect; SIE: swapped indirect effect; NDE: natural direct effect; NIE: natural indirect effect; TE: total effect; CI: confidence interval.

7. Discussion

This paper proposes a new method, namely the SDE and SIE, for causal mediation analysis based on the introduction of a novel quasi-instrumental variable, IB, which satisfies the relevance assumption and exclusion restriction of the conventional instrumental variable for the M–Y relationship. The proposed SDE and SIE can assess direct and indirect effects, respectively, in the presence of unmeasured M–Y confounding and intermediate M–Y confounding; this condition has been addressed in existing methods. Thus, the development of the SDE and SIE fills this research gap. Moreover, the causal interpretation of the SDE and SIE coincides with that of the NDE and NIE. This is a crucial theorem for the SDE and SIE because it implies that their empirical expressions are alternative approaches to inferring the NDE and NIE under verifiable assumptions. The key to the success of the SDE and SIE is to employ a variable that satisfies the assumptions for the IB in the analysis. To examine whether a variable meets the proposed assumptions for the IB, we conducted a sensitivity analysis by establishing a bias formula for the SDE and SIE. This bias formula enabled a determination of the plausibility of treating *YES1* as the IB in the pathway from treatment to mortality mediated through *EGFR* expression. From the perspective of statistical inference, we propose a robust estimation for the SDE and SIE. Moreover, Theorem 6 demonstrates that the robust estimation is CAN and achieves the semiparametric efficiency bound. In addition, simulation studies revealed that the proposed robust estimators mostly outperformed their counterparts in conventional methods, namely IPW estimation and G-computation, under various scenarios.

Reference

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. Commentaries. *Journal of the American statistical Association*; **91(434)**:444-472.

Cheng, G. and Huang, J.Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*; **38(5)**:2884-2915.

Ding, P. and Vanderweele, T.J. (2016). Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*; **103(2)**:483-490.

Fan, P.-D., Narzisi, G., Jayaprakash, A.D., Venturini, E., Robine, N., Smibert, P., Germer, S., Helena, A.Y., Jordan, E.J. and Paik, P.K. (2018). YES1 amplification is a mechanism of acquired resistance to EGFR inhibitors identified by transposon mutagenesis and clinical genomics. *Proceedings of the National Academy of Sciences*; **115(26)**:E6030-E6038.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society Series B*; **69(2)**:199-215.

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

Hafeman, D.M. (2011). Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *American journal of epidemiology*; **174(6)**:710-717.

Helena, A.Y., Suzawa, K., Jordan, E., Zehir, A., Ni, A., Kim, R., Kris, M.G., Hellmann, M.D., Li, B.T. and Somwar, R. (2018). Concurrent alterations in EGFR-mutant lung cancers associated with resistance to EGFR kinase inhibitors and characterization of MTOR as a mediator of resistance. *Clinical Cancer Research*; **24(13)**:3108-3118.

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

Ichihara, E., Westover, D., Meador, C.B., Yan, Y., Bauer, J.A., Lu, P., Ye, F., Kulick, A., De Stanchina, E. and Mcewen, R. (2017). SFK/FAK signaling attenuates osimertinib efficacy in both drug-sensitive and drug-resistant models of EGFR-mutant lung cancer. *Cancer research*; **77(11)**:2990-3000.

Lin, S.-H. and Vanderweele, T. (2017). Interventional Approach for Path-Specific Effects. *Journal of Causal Inference*; **5(1)**.

Lin, S.H., Huang, Y.T. and Yang, H.I. (2019). On identification of agonistic interaction: Hepatitis B and C interaction on hepatocellular carcinoma. *Statistics in medicine*; **38(13)**:2467-2476.

Lin, S.H., Young, J., Logan, R., Tchetgen Tchetgen, E.J. and Vanderweele, T.J. (2017). Parametric Mediational g-Formula Approach to Mediation Analysis with Time-varying Exposures, Mediators, and Confounders. *Epidemiology*; **28(2)**:266-274.

Lin, S.H., Young, J.G., Logan, R. and Vanderweele, T.J. (2017). Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med*; **36(26)**:4153-4166.

Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G. and Haluska, F.G. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*; **350(21)**:2129-2139.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. and Tchetgen Tchetgen, E.J. (2017). Quantifying an adherence path-specific effect of antiretroviral therapy in the nigeria pepfar program. *Journal of the American Statistical Association*; **112(520)**:1443-1452.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. and Tchetgen Tchetgen, E.J. (2020). On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*; **107(1)**:159-172.

Nicholson, R.I., Gee, J.M.W. and Harper, M.E. (2001). EGFR and cancer prognosis. *European journal of cancer*; **37**:9-15.

Pao, W. and Chmielecki, J. (2010). Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. *Nature Reviews Cancer*; **10(11)**:760-774.

Pearl, J. Direct and indirect effects. In, *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. San Francisco, CA, USA: Morgan kaufmann publishers Inc.; 2001. p. 411-420.

Pearl, J. Causality: models, reasoning, and inference. New York: Cambridge University Press; 2009.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*; **6(2)**.

Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3(2)**:143-155.

Robins, J.M., Mark, S.D. and Newey, W.K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*:479-495.

Smith, L.H. and Vanderweele, T.J. (2019). Mediational E-values: approximate sensitivity analysis for unmeasured mediator–outcome confounding. *Epidemiology*; **30(6)**:835-837.

Snowden, J.M., Rose, S. and Mortimer, K.M. (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*; **173(7)**:731-738.

Stefanski, L.A. and Boos, D.D. (2002). The calculus of M-estimation. *The American Statistician*; **56(1)**:29-38.

Talloen, W., Moerkerke, B., Loeys, T., De Naeghel, J., Van Keer, H. and Vansteelandt, S. (2016). Estimation of indirect effects in the presence of unmeasured confounding for the mediator–Outcome relationship in a multilevel 2-1-1 mediation model. *Journal of Educational and*

Behavioral Statistics; **41(4)**:359-391.

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

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

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

Vanderweele, T.J. and Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, biostatistics, and public health*; **11(2)**.

Vanderweele, T.J. and Tchetgen Tchetgen, E.J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*; **79(3)**:917-938.

Vanderweele, T.J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*; **172(12)**:1339-1348.

Vanderweele, T.J. and Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiol Method*; **2(1)**:95-115.

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

Zheng, W. and Van Der Laan, M.J. (2012). Causal mediation in a survival setting with time-dependent mediators.