When Do Natural Mediation Effects Differ from Their Randomized Interventional Analogues: Test and Theory

Ang Yu[*], Li Ge[†], and Felix Elwert[‡]

June 28, 2024

[**This paper is currently under active development.**]

Abstract

In causal mediation analysis, the natural direct and indirect effects (natural effects) are nonparametrically unidentifiable in the presence of treatment-induced confounding, which motivated the development of randomized interventions analogues (RIAs) of the natural effects. The RIAs are easier to identify and widely used in practice. Applied researchers often interpret RIA estimates as if they were the natural effects, despite reasons to suspect that the RIAs could be poor proxies of the natural effects. This calls for practical and theoretical guidance on when the RIAs differ from or coincide with the natural effects, which is the goal of this paper. We develop a novel empirical test for the divergence between the RIAs and the natural effects under the weak assumptions sufficient for identifying the RIAs. We illustrate the test using the Moving to Opportunity Study. We also provide new theoretical insights on the relationship between the RIAs and the natural effects from a covariance perspective and a structural equation perspective. We also discuss previously undocumented connections between the natural effects, the RIAs, and estimands in instrumental variable analysis and Wilcoxon-Mann-Whitney tests.

Keywords— causal mediation analysis, falsification test, nonparametric structural equation, randomized interventional analogue, Wilcoxon-Mann-Whitney test

[*]Department of Sociology, University of Wisconsin-Madison. Email: ayu33@wisc.edu

[†]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

[‡]Department of Sociology and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

1 Introduction

1.1 Background

Causal mediation analysis explains the mechanisms of a total causal effect by decomposing it into direct and indirect effects in terms of some mediators. The direct effect is the part of the total effect that does not go through the researcher-specified mediators, and the indirect effect is the part that does. As a central task in the social and health sciences, causal mediation analysis is widely used in applied research.

We adopt the conventional notation in causal mediation analysis. Y is the observed outcome, A is a binary treatment with support $\{0, 1\}$,[1] and M is a vector of mediators. Y_a and M_a are respectively the potential values of Y and M under the assignment of treatment value a. We further define two groups of confounders that may be empty, C is a vector of baseline confounders, and L is a vector of post-treatment confounders. Figure 1 illustrates the relationship between variables, when any variable may affect any temporally subsequent variables.

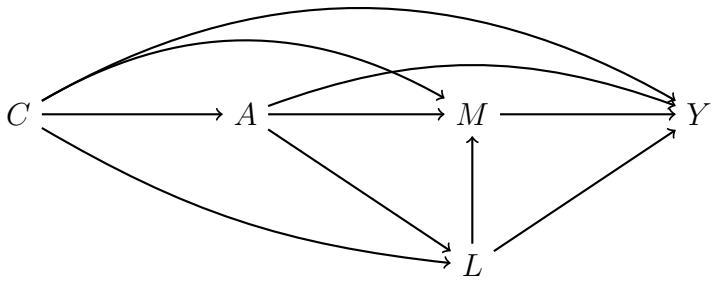

Figure 1: Variable Set-up in Causal Mediation Analysis

The most classic approach of causal mediation analysis decomposes the total effect (TE) into the natural indirect effect (NIE) and the natural direct effect (NDE) (Robins and Greenland, 1992; Pearl, 2001).

$$\underbrace{\mathrm{E}(Y_1 - Y_0)}_{\text{TE}} = \underbrace{\mathrm{E}(Y_{1,M_1} - Y_{0,M_0})}_{\text{TE}} = \underbrace{\mathrm{E}(Y_{1,M_1} - Y_{1,M_0})}_{\text{NIE}} + \underbrace{\mathrm{E}(Y_{1,M_0} - Y_{0,M_0})}_{\text{NDE}},$$

where $Y_{a,M_{a'}}$ denotes the potential outcome of Y under the assignment of treatment a and the mediator value that would be realized under the assignment of treatment a'. The NIE is defined by varying the mediator assignment from M_1 to M_0 but otherwise fixing treatment assignment at 1, capturing the part of the total effect that only goes through M. The NDE is defined by varying the treatment assignment from 1 to 0 but holding mediator assignment at the baseline

[1] This is generalizable to any pair of two values for a multivalued treatment.

mediator value, capturing the part of the total effect that does not go through M. Since the natural effects (NIE and NDE) are defined in terms of individual-level potential mediators (M_1 and M_0), they capture causal mechanisms at the individual level, even though they are ultimately summarized as population-average effects through the expectation operator.

The natural effects are notoriously difficult to identify. Without parametric assumptions, they are unidentifiable when there is any treatment-induced confounders L, regardless of whether L is observed (Robins, 2003; Avin et al., 2005). This severely limits the application of the natural effects in practice, as ruling out L is impossible in most empirical settings.

Motivated by the difficulty of identifying the natural effects, an alternative decomposition has been proposed (VanderWeele et al., 2014). Nonparametrically, its identification does not require the absence of treatment-induced confounders. This alternative decomposition is based on the randomized interventional analogues (RIA) of the TE, the NIE and the NDE, namely the TE^R, the NIE^R and the NDE^R:

$$\underbrace{\text{E}(Y_{1,G_1} - Y_{0,G_0})}_{\text{TE}^R} = \underbrace{\text{E}(Y_{1,G_1} - Y_{1,G_0})}_{\text{NIE}^R} + \underbrace{\text{E}(Y_{1,G_0} - Y_{0,G_0})}_{\text{NDE}^R},$$

where $G_{a'}$ is a value randomly drawn from the mediator distribution that would realize under the assignment of treatment value a' given C, and $Y_{a,G_{a'}}$ is the potential outcome of Y under the assignment of the treatment value a and the mediator value $G_{a'}$. Clearly, the RIAs differ from the natural effects in mediator assignments. Instead of M_1 and M_0, the mediator assignments for the RIAs are G_1 and G_0. As G_1 and G_0 are random draws from population distributions, the RIAs are not aggregations of individual-level causal contrasts like the natural effects.

Seen as much less demanding and more widely applicable than the natural effects, the RIAs are popular in empirical research.[2] In practice, applied researchers frequently estimate the RIAs as proxies of the natural effects. In fact, the RIA estimates are often interpreted as if they were estimates of the natural effects (after all, the RIAs are named as "analogues"!). Sarvet et al. (2023) reviewed 16 applied studies that estimate RIAs, all of which contain interpretive statements that elide the difference between the RIAs and the natural effects. Indeed, the methodological literature has encouraged this ambiguity. For example, VanderWeele and Tchetgen Tchetgen (2017) write that "it will only be in extremely unusual settings that the interventional analogue is non-zero, with there being no natural indirect effects".

[2]Methodological work that further develops the RIA decomposition includes (VanderWeele and Tchetgen Tchetgen, 2017; Vansteelandt and Daniel, 2017; Zheng and Van Der Laan, 2017; Rudolph et al., 2018; Wodtke and Zhou, 2020; Loh et al., 2020; Díaz et al., 2021; Benkeser and Ran, 2021; Hejazi et al., 2022; Nguyen et al., 2022; Díaz et al., 2023; Rudolph et al., 2024).

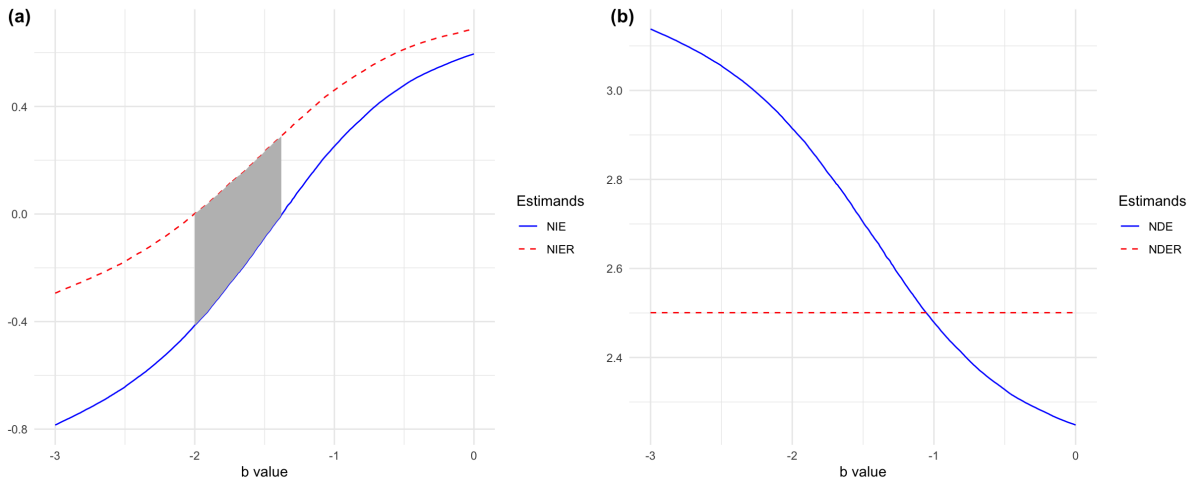

Figure 2: Illustration of Possible Divergence Between the Natural Mediation Estimands and Their RIAs. Panel (a) depicts NIE and NIE^R, and Panel (b) depicts NDE and NDE^R. The DGP is as follows. $L \sim \mathcal{N}(A, 1)$; $M \sim \text{expit}(A{+}L{+}bAL)$; $Y \sim \mathcal{N}(A{+}L{+}M{+}LM, 1)$. Thus, the x axis is the b value that we vary, which is the coefficient of the A-L interaction in the generative model for M. In Panel (a), the shaded area indicates sign reversal between NIE and NIE^R.

However, there are reasons to suspect that the RIAs can be poor proxies of the natural effects. Unlike the natural effects, they are not interpretable as explanatory mechanisms at the individual as opposed to the population level. Formalizing this intuition, Miles (2023) proposes a set of null criteria that valid causal indirect effect measures should satisfy and shows that the NIE is valid by these criteria while the NIE^R is not. In particular, the NIE^R can be nonzero even if the mediator does not "mediate" the treatment effect for any individual. In addition, it has been frequently noted in the methodological literature that the NIE^R and the NDE^R do not generally sum to the TE, which is problematic because the canonical task of causal mediation analysis is to understand the TE (Vansteelandt and Daniel, 2017; Nguyen et al., 2021).

Beyond the violation of null criteria, which focuses on a knife-edge scenario, we draw attention to possible quantitative differences between the natural effects and the RIAs in a wide range of data generating processes (DGPs). These quantitative differences may be large and even involve sign reversal. In the illustration of Figure 2, data are simulated according to a set of very simple and seemingly innocuous DGPs. By varying one parameter of the DGP, we observe areas of significant divergence and sign reversal, where the RIAs can hardly be used to draw conclusions about the natural effects.

Therefore, it is natural to ask when the natural effects differ from their RIAs. If they are identical or at least close to each other, then it might be warranted to interpret estimates of the RIAs as the natural effects. Conversely, if they substantially differ, then more caution and precision in interpretation is called for. In this paper, we answer this question with one practical

test and two theoretical perspectives.

1.2 Contributions

We make a practical contribution by proposing a novel test for the differences between the NIE, the NDE, and their respective RIAs. The empirical testability of these differences may be surprising, because under the standard assumptions for identifying the NIE and the NDE, the natural effects necessarily coincide with their RIAs (VanderWeele and Tchetgen Tchetgen, 2017, p.921). And under the standard assumptions identifying the NIE^R and the NDE^R, the NIE and the NDE are unidentified. Thus, it may appear that under no set of common assumptions can one test the differences. However, our test is made possible by leveraging two simple facts. First, the TE and the TE^R are identified under the standard assumptions for the NIE^R and the NDE^R. Second, when $\text{TE} - \text{TE}^R \neq 0$, it is necessarily the case that either $\text{NIE} \neq \text{NIE}^R$ or $\text{NDE} \neq \text{NDE}^R$. Hence, instead of hoping that "the natural and interventional effects may coincide empirically" (Loh et al., 2020), we can actually test their divergence by testing $\text{TE} - \text{TE}^R = 0$ under weak identifying assumptions that are sufficient for the RIAs but not the natural effects.

We make a theoretical contribution by clarifying and illustrating the substantive conditions under which the natural effects differ from or coincide with their RIAs. We do so from a nonparametric covariance perspective and a structural equation perspective. First, we derive a covariance-based representation of the differences between the natural effects and their RIAs. Second, we derive parametric constraints on the structural equations generating the data under which the the natural effects will coincide with the RIAs. These two novel perspectives provide intuitive insights on the substantive mechanisms underpinning the relationship between the natural effects and the RIAs. In Miles' (2023) discussion of the relationship between NIE and NIE^R, he proves the null criteria violation using one specific numerical counterexample. With two new analytic perspectives that are general and intuitive, we thus demystify and expand on Miles' (2023) results.

The remaining of this paper is organized as follows. In Section 2, we review some standard assumptions in causal mediation analysis that were referred to above. In Section 3, we present our empirical test for the differences between the natural effects and the RIAs and apply it to the Moving to Opportunity (MTO) study. Section 4 and 5, respectively, introduce the covariance perspective and the structural equation perspective. Section 6 discusses related estimands, including those in the instrumental variable (IV) settings and those underlying the Wilcoxon-Mann-Whitney tests. We present novel results that unify causal mediation analysis with these other fields of causal inference. Technical proofs are collected in the appendix.

R code for simulating Figure 2 and empirical data analysis in Section 3 can be found at https://github.com/ang-yu/diff_naturals_rias.

2 Review of Conventional Mediation Assumptions

We review conventional assumptions included in the literature of causal mediation analysis.

Assumption 1 (Consistency)**.** $f(M_a \mid a, C) = f(M \mid a, C)$ and $\mathrm{E}(Y_{a,m} \mid a, m, C) = \mathrm{E}(Y \mid a, m, C)$, for all a and m.

Assumption 2 (Ignorability of A conditional on C)**.** $Y_{a,m} \perp\!\!\!\perp A \mid C$ for all a and m; $M_a \perp\!\!\!\perp A \mid C$ for all a.

Assumption 3 (Ignorability of M conditional on C, A, L)**.** $Y_{a,m} \perp\!\!\!\perp M \mid C, A = a, L$ for all a and m.

Assumption 4 (Ignorability of M conditional on C, A)**.** $Y_{a,m} \perp\!\!\!\perp M \mid C, A = a$ for all a and m.

Assumption 5 (Cross-world Independence)**.** $Y_{a,m} \perp\!\!\!\perp M_{a'} \mid C$ for all a, a', and m.

Assumption 1 is a standard statement linking the potential values to the observed values. Assumption 2 requires the treatment A be ignorable conditional on baseline confounders C. Assumption 3 states that the mediator M is conditionally ignorable given both baseline confounders C and post-treatment confounders L, as well as the treatment. Assumption 4 imposes conditional ignorability of the mediator given only baseline confounders and the treatment, which is stronger than Assumption 3. Finally, Assumption 5 requires the conditional independence between the potential outcomes $Y_{a,m}$ and potential mediators $M_{a'}$ under two possibly different treatment assignments a and a', hence its name (cross-world independence).

In the literature, Assumptions 1, 2, and 3, are the standard identifying assumptions for the RIAs (VanderWeele et al., 2014), while Assumptions 1, 2, 4, and 5 are the standard assumptions for identifying the NIE and the NDE (Pearl, 2001; VanderWeele, 2015, p.463-4; Imai, Keele, and Yamamoto, 2010 for a slightly stronger version). Notably, the cross-world independence assumption requires the absence of any post-treatment confounding of the mediator-outcome relationship ($L = \{\}$) (Robins, 2003; Avin et al., 2005; Andrews and Didelez, 2021). Hence, it is clear that the standard assumptions for the RIAs are weaker, as they allow for the existence of post-treatment confounders. Furthermore, when the cross-world independence assumption holds, the natural effects are necessarily equivalent to their RIAs.

3 Empirical Test

We propose to use the empirical estimate of $\text{TE} - \text{TE}^R$ as a test statistic for the divergence between the NIE and the NIE^R and the divergence between the NDE and the NDE^R. This test relies on the fact that if $\text{TE} - \text{TE}^R \neq 0$, it is necessarily the case that either $\text{NIE} \neq \text{NIE}^R$ or $\text{NDE} \neq \text{NDE}^R$, or both. Thus, if we reject the null hypothesis that $\text{TE} - \text{TE}^R = 0$, we also reject the null hypothesis that $\text{NIE} = \text{NIE}^R$ and $\text{NDE} = \text{NDE}^R$.[3] In addition, since $|\text{TE} - \text{TE}^R| \leq |\text{NIE} - \text{NIE}^R| + |\text{NDE} - \text{NDE}^R|$ by the triangle inequality, $|\text{TE} - \text{TE}^R|$ also provides a lower bound for the sum of the absolute differences between the NIE and the NIE^R and between the NDE and the NDE^R.

Under assumptions 1, 2 and 3, $\text{TE} - \text{TE}^R = \text{E}(Y_1) - \text{E}(Y_0) - \text{E}(Y_{1,G_1}) + \text{E}(Y_{0,G_0})$ is identified by the functionals below (VanderWeele et al., 2014).

$$\text{E}(Y_a) = \iint y f(y \mid c, a) f(c) \mathrm{d}y \mathrm{d}c$$
$$\text{E}(Y_{a,G_a}) = \iiiint y f(y \mid c, a, l, m) f(m \mid c, a) f(l \mid c, a) f(c) \mathrm{d}y \mathrm{d}m \mathrm{d}l \mathrm{d}c.$$

Hence, importantly, our test is nonparametrically identifiable when there are treatment-induced confounders and Assumption 5 is invalid. This is because although the NIE and the NDE are not nonparametrically identifiable under treatment-induced confounding, their sum is.

The task now is to estimate $\text{TE} - \text{TE}^R$. This can be done using various estimators of TE and TE^R. Below, we discuss a nonparametric approach based on the efficient influence functions (EIF) (Bickel et al., 1998; Hines et al., 2022).

The EIF for $\text{E}(Y_a)$ is well-known (Hahn, 1998):

$$\phi_a := \frac{\mathbb{1}(A = a)}{\Pr(A = a \mid C)}[Y - \text{E}(Y \mid C, a)] + \text{E}(Y \mid C, a) - \text{E}[\text{E}(Y \mid C, a) \mid a].$$

The EIF for $\text{E}(Y_{a,G_a})$ is derived by Díaz et al. (2021). Let $\text{E}(Y \mid C, A, L, M) = \mu(C, A, L, M)$ and $\xi(C, A) = \iint \mu(C, A, l, m) f(m \mid C, A) f(l \mid C, A) \mathrm{d}m \mathrm{d}l$, this EIF is

$$\psi_a := \frac{\mathbb{1}(A = a)}{\Pr(A = a \mid C)} \left\{ \frac{\Pr(M \mid C, a)}{\Pr(M \mid C, a, L)} [Y - \mu(C, a, L, M)] + \int \mu(C, a, L, m) f(m \mid C, a) \, \mathrm{d}m - \xi(C, a) + \int \mu(C, a, l, M) f(l \mid C, a) \, \mathrm{d}l - \xi(C, a) \right\} - \text{E}[\xi(C, a)].$$

[3] In a recent work, Vo et al. (2024) constructs a test using a similar premise. In the context of path-specific effects, they propose using the difference between the TE and the sum of a set of RIA-type path-specific estimands to test the absence of intermediate confounding.

Then the EIF for $\text{TE} - \text{TE}^R$ is $\phi_1 - \phi_0 + \psi_1 - \psi_0$. Using these EIFs, we can construct either double machine learning (Chernozhukov et al., 2018) or targeted maximum likelihood (Van der Laan and Rose, 2011) estimators that allow nonparametric estimation with desirable theoretical properties (Díaz et al., 2021; Rudolph et al., 2024). In particular, these estimators are data-adaptive and can handle high-dimensional covariates. They are also multiply robust to misspecification of components of the EIF. Furthermore, using cross-fitting, they attain semiparametric efficiency and asymptotic normality under relatively weak conditions. In practice, researchers may take advantage of the output of *medoutcon* package (Hejazi et al., 2022) in R to estimate $\text{TE} - \text{TE}^R$ and the associated confidence interval.[4]

3.1 Empirical Illustration

We apply our test to mediation analysis of the Moving to Opportunity (MTO) study, a large-scale longitudinal randomized control trial conducted by the Department of Housing and Urban Development of the United States (Ludwig et al., 2013; Kling et al., 2005). We follow the conceptual set-up of Rudolph et al. (2021) and Rudolph et al. (2024), who estimated the RIAs.[5] The treatment is a binary indicator of whether or not a family living in a high-poverty neighborhood was randomized to receive a Section 8 housing voucher that allowed them to move to a less poor neighborhood. We consider two mediators measured between 10-15 years of follow up, neighborhood poverty and the number of residential moves. The outcome is a composite score for mental health. For causal identification, we account for a post-treatment confounder which is whether the family used the voucher to move within the 90 days allotted. We also account for 12 baseline confounders, which capture baseline household socioeconomic and demographic characteristics, as well as neighborhood-related perceptions and aspirations.

We implement our test using double machine learning estimators with two-fold cross-fitting. The nuisance functions are estimated using random forests (Wright and Ziegler, 2017). For confidence intervals, we leverage the asymptotic normality of the estimates and calculate the variances using the mean squared estimated EIFs. We present our estimates in Table 1. Our estimate of $\text{TE} - \text{TE}^R$ is significantly different from 0 (95% Confidence Interval= $(0.081, 0.087)$). Therefore, we reject the null hypothesis that $\text{NIE} = \text{NIE}^R$ and $\text{NDE} = \text{NDE}^R$. In this empirical example, one should not interpret the RIA estimates as the natural effects. Furthermore, the

[4]Another R package, *HDmediation* (Williams et al., 2024), may also be useful. *medoutcon* only accommodates a single binary L variable but directly outputs individual-level EIF estimates, which allows the calculation of p value and confidence interval. On the other hand, *HDmediation* accommodates vector-valued and non-binary L but does not directly outputs EIF estimates.

[5]Due to lack of access to the restricted-use dataset, we follow their variable and sample choices only conceptually, not precisely. Hence, our estimates should be regarded as purely illustrative.

Estimand	Estimate	95% Confidence Interval
TE	0.084	(0.081, 0.087)
TE^R	0.002	(-0.002, 0.005)
$\text{TE} - \text{TE}^R$	0.083	(0.069, 0.072)

Table 1: Empirical Estimates from the MTO study. The treatment is the receipt of a randomized housing voucher. The mediators are neighborhood poverty and the number of residential moves. The outcome is mental health. Estimation is implemented using double machine learning estimators with cross-fitting. Confidence intervals are Wald-type and calculated using the estimated EIF of the estimands.

sum of the absolute differences between the NIE and the NIE^R and between the NDE and the NDE^R is greater than $|\text{TE} - \text{TE}^R|$, which is estimated to be 0.083.

4 Covariance Perspective

We present a covariance-based representation of the differences between the natural effects and their RIAs. We first focus on the simple case with a scalar binary mediator and a randomized treatment such that C is empty. This simple case most easily captures the core intuition. Then we generalize the covariance representation to vector mediators with arbitrary distributions and a non-randomized treatment. The expressions are derived just using the definitions of the estimands, without any identifying assumptions or functional form restrictions.

4.1 Single Binary Mediator, Randomized Treatment

Proposition 1. When the treatment is randomized and the support of M is $\{0, 1\}$,

$$\begin{aligned}
\text{TE} - \text{TE}^R &= \text{Cov}(M_1, Y_{1,1} - Y_{1,0}) - \text{Cov}(M_0, Y_{0,1} - Y_{0,0}) \\
\text{NIE} - \text{NIE}^R &= \text{Cov}(M_1 - M_0, Y_{1,1} - Y_{1,0}) \\
\text{NDE} - \text{NDE}^R &= \text{Cov}(M_0, Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}).
\end{aligned}$$

First, the difference between the TE and the TE^R reflects how the treatment changes the covariance between an individual's mediator value (M_a) and their mediator effect on the outcome ($Y_{a,1} - Y_{a,0}$).[6] The TE will be greater than the TE^R if the treatment (rather than the control) induces more accurate ex ante expectations of the individual-level mediator effect such that

[6] In the causal decomposition of group disparities proposed by Yu and Elwert (2024), the "selection" component captures the contribution of group-differential selection into treatment to an outcome disparity. Relabelling the group and the treatment in the framework of Yu and Elwert (2024) as the treatment and mediator, the selection component can be written as $\text{Cov}(M, Y_{M=1} - Y_{M=0} \mid A = 1) - \text{Cov}(M, Y_{M=1} - Y_{M=0} \mid A = 0)$, which coincides with $\text{TE} - \text{TE}^R$ when treatment is randomized.

individuals with a higher mediator effect are more likely to select into the mediator value 1.

Second, the difference between the NIE and the NIE^R equals the covariance between the treatment effect on the mediator ($M_1 - M_0$) and a net mediator effect on the outcome ($Y_{1,1} - Y_{1,0}$). Thus, if there are common determinants these two effects, the NIE will differ from the NIE^R. These determinants could be either pre-treatment modifiers of both effects or post-treatment mediators of the first effect which also modify the second effect. In the MTO example, those who are better able to take advantage of the housing voucher (the treatment) to move to a lower-poverty neighborhood (the mediator) may, in turn, be better able to leverage the resources in their new lower-poverty neighborhood to improve mental health outcomes. In that case, the covariance between the treatment effect on the mediator and the net mediator effect on the outcome will be positive.

Third, the difference between the NDE and the NDE^R is the covariance between the mediator value under control (M_0) and the interaction effect between treatment and mediator on outcome ($Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}$).

Generally, the natural effects and the RIAs differ to the extent that the potential mediators (M_a) and the potential outcomes ($Y_{a',m}$) are correlated with one another. This makes sense as the RIAs are defined using random draws of potential mediators, G_a, that are independent of everything else, whereas the natural effects do not remove the naturally occurring dependency between the potential mediators and the potential outcomes.

Miles (2023) proposes a set of mediation null criteria. In particular, the definition of the "sharper mediation null" is: For each individual in the population of interest, either $M_1 = M_0$ or $Y_{a,m} = Y_{a,m'}$ for a, m, and m'. And a valid measure of indirect effect should be zero when the sharper mediation null is true. We note that NIE $= \mathrm{E}[(M_1 - M_0)(Y_{1,1} - Y_{1,0})]$. Thus, the NIE clearly satisfies this criterion, while, by Proposition 1, NIE^R does not. For example, if half of the population has $M_1 - M_0 = 1$ and $Y_{1,1} - Y_{1,0} = 0$ while the other half has $M_1 - M_0 = 0$ and $Y_{1,1} - Y_{1,0} = 1$, the NIE will be zero, but the NIE^R will be 1/4. This is consistent with Miles (2023)'s results. However, Miles (2023) proves that the NIE^R does not satisfy the null criterion using a specific counterexample, which might be viewed as a contrived example. In contrast, Proposition 1 analytically reveals why and when the NIE^R deviates from the null criterion: it is because the NIE^R omits the natural dependency between the treatment effect on the mediator and the mediator effect on the outcome, which is a part of the NIE. To the extent that the correlation between these effects is pervasive in practice, there is nothing "contrived" in the deviation of the NIE^R from the null criterion.

4.2 General Case

In last subsection, we focused on the case of a binary M and a randomized treatment. Now we generalize our results to a continuous or multivalued discrete vector of mediators and a non-randomized treatment. Again, we do not make any identifying assumptions or parametric restrictions.

Proposition 2.

$$\begin{aligned}
\text{TE} - \text{TE}^R &= \sum_{m \in \mathcal{M}} \text{E}\{\text{Cov}[\mathbb{1}(M_1 = m), Y_{1,m} \mid C]\} - \text{E}\{\text{Cov}[\mathbb{1}(M_0 = m), Y_{0,m} \mid C]\} \\
\text{NIE} - \text{NIE}^R &= \sum_{m \in \mathcal{M}} \text{E}\{\text{Cov}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m), Y_{1,m} \mid C]\} \\
\text{NDE} - \text{NDE}^R &= \sum_{m \in \mathcal{M}} \text{E}\{\text{Cov}[\mathbb{1}(M_0 = m), Y_{1,m} - Y_{0,m} \mid C]\},
\end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $\mathcal{M}$ is the support of M. The relationships above directly hold for discrete mediators, but they also hold for continuous mediators if summations are replaced with integrals and the indicator function is replaced with the Dirac delta function.

We thus obtain a covariance-based representation analogous to Proposition 1. Here, the building blocks are conditional covariances between the potential mediators (M_a) and the potential outcomes ($Y_{a',m}$) given baseline confounders C. We further summarize the c- and m-specific covariances by taking expectation over the distribution of C and uniformly taking sum over the support of M. The natural effects and the RIAs generally differ due to the dependency between the mediator and outcome potential values conditional on baseline confounders. Clearly, the natural effects and the RIAs coincide when the cross-world independence assumption (Assumption 5) is satisfied.

In particular, $\text{TE} - \text{TE}^R$ still has a highly interpretable form. It reflects the treatment effect on selection into mediator values based on the corresponding potential outcomes. Consider a treatment assignment a, a mediator value m, and a baseline confounder value c. If those who, when assigned a, would take the mediator value m tend to be those whose corresponding potential outcome is higher (among individuals with $C = c$), then $\text{Cov}[\mathbb{1}(M_a = m), Y_{1,m} \mid C = c]$ will be positive. This may happen if treatment a induces somewhat accurate ex-ante anticipation of what outcome m would bring about, and individuals choose M based on this anticipation. And $\text{TE} - \text{TE}^R$ will generally be non-zero if this induction differs by treatment status for some m and c.

An alternative RIA-based decomposition is developed by Lok (2016) and Lok and Bosch

(2021)[7]. In this decomposition, the TE is decomposed to what are called the organic indirect and direct effects ($\text{NIE}^{\text{organic}}$ and $\text{NDE}^{\text{organic}}$).

$$\underbrace{\text{E}(Y_1 - Y_0)}_{\text{TE}} = \underbrace{\text{E}(Y_1 - Y_{1,G_0})}_{\text{NIE}^{\text{organic}}} + \underbrace{\text{E}(Y_{1,G_0} - Y_0)}_{\text{NDE}^{\text{organic}}}.$$

We again show a corresponding covariance representation in the general case.

Proposition 3.

$$\text{NIE} - \text{NIE}^{\text{organic}} = -\sum_{m \in \mathcal{M}} \text{E}\{\text{Cov}[\mathbb{1}(M_0 = m), Y_{1,m} \mid C]\}$$
$$\text{NDE} - \text{NDE}^{\text{organic}} = \sum_{m \in \mathcal{M}} \text{E}\{\text{Cov}[\mathbb{1}(M_0 = m), Y_{1,m} \mid C]\}.$$

Finally, Zheng and Van Der Laan (2017) propose a related decomposition (also see Nguyen et al., 2022, p.264). The intervention underlying this decomposition involves assigning to people with $C = c, L_a = l$ values of mediator randomly drawn from the distribution of $M_{a'}$ conditional on $C = c, L_{a'} = l$. Denoting these random draws by $G_{a'|L_a}$, their decomposition is

$$\underbrace{\text{E}(Y_{G_{1|C,L_1}} - Y_{G_{0|L_0}})}_{\text{TE}^{RL}} = \underbrace{\text{E}(Y_{1,G_{1|L_1}} - Y_{1,G_{0|L_1}})}_{\text{NIE}^{RL}} + \underbrace{\text{E}(Y_{1,G_{0|L_1}} - Y_{0,G_{0|L_0}})}_{\text{NDE}^{RL}}.$$

The differences between the natural effects and the estimands above do not have a covariance representation. This is because the way L enters into the estimands makes the NIE^{RL} the path-specific effect through M but not L (see Appendix S2 in Miles, 2023). Thus, these estimands are conceptually further removed from the natural effects.

5 Structural Equation Perspective

To further facilitate substantive reasoning on the differences between the RIAs and the natural effects, we illustrate some specific data generating processes (DGPs) that would make the NDE coincide with the NDE^R or the NIE with the NIE^R. We express these DGPs using structural equations (generative models) with parametric constraints. Throughout this section, we allow for the existence of L, such that the equivalence between the natural estimands and the RIAs is not guaranteed by cross-world independence. We also do not restrict the number or the distribution of mediators.

We first present results with assumed linearity and a randomized treatment, which provides

[7]With relabelled variables, this decomposition also coincides with a disparity decomposition in Jackson (2021)

the easiest intuition. Then we extend the results to structural equations without the linearity conditions and treatment randomization. For comparison with parametric constraints below, we note that the nonparametric structural equations with no constraints are as follows:

$$\begin{aligned} C &= \epsilon_C \\ A &= g_A(C, \epsilon_A) \\ L &= g_L(C, A, \epsilon_L) \\ M &= g_M(C, A, L, \epsilon_M) \\ Y &= g_Y(C, A, L, M, \epsilon_Y), \end{aligned}$$

where g_C, g_A, g_L, and g_M are arbitrary functions of their arguments. And $\epsilon_C, \epsilon_A, \epsilon_L, \epsilon_M$ and ϵ_Y are unspecified inputs for each variable. Importantly, throughout this section, we allow these unspecified inputs to be arbitrarily dependent on one another and all other variables.

5.1 Linear Structural Equations, Randomized Treatment

Since the treatment is assumed to be randomized, C is empty. We consider the structural equations for A, L, M, and Y. In this subsection, the notation technically only applies to one L and one M variables, but our expressions can be easily extended to accommodate multiple L and M variables without compromising intuition.

Proposition 4. Under the following linear structural equations with constant coefficients (i.e., all α, β, γ terms are constants),

$$\begin{aligned} A &= \epsilon_A \\ L &= \alpha_0 + \alpha_1 A + \epsilon_L \\ M &= \beta_0 + \beta_1 A + \beta_2 L + \beta_3 AL + \epsilon_M \\ Y &= \gamma_0 + \gamma_1 A + \gamma_2 L + \gamma_3 M + \gamma_4 AL + \gamma_5 AM + \gamma_6 LM + \gamma_7 ALM + \epsilon_Y, \end{aligned}$$

we have $\text{NIE} - \text{NIE}^R = (\gamma_6 + \gamma_7)\beta_3 \operatorname{Var}(\epsilon_L)$, and $\text{NDE} - \text{NDE}^R = \gamma_7 \beta_2 \operatorname{Var}(\epsilon_L) + \gamma_7 \operatorname{Cov}(\epsilon_L, \epsilon_M)$.[8]

Hence, under the linear structural equations, there are multiple sufficient conditions for either the NIE or the NDE to coincide with their respective RIAs. For example, the NIE and the NIE^R are equivalent if $\beta_3 = 0$, i.e., there is no AL interaction in the equation for M. And the

[8] Clearly, Proposition 4 is a special case of Proposition 2. It is also easy to show that when M is binary, Proposition 4 recovers Proposition 1.

NDE and the NDE^R are equivalent if $\gamma_7 = 0$, i.e., there is no three-way interaction ALM in the equation for Y. In summary, equivalences can be established by ruling out certain interaction effects.

It is possible to have only one of the NIE and the NDE coincide with their RIA. When only one of the natural effects equal its RIA, our test statistic in Section 3, TE $-$ TE^R, will capture the deviaion of the other natural effect from its RIA. The next subsection shows that the intuitions from the linear analysis can be extended to the settings where the structural equations are much more unrestricted.

5.2 Nonlinear Structural Equations, Nonrandomized Treatment

Throughout this subsection, we focus on constraints on the structural equations for Y. Thus, we maintain completely unconstrained structural equations for C, A, L, and M. Again, the structural equations we consider allow L to affect M and Y in some manner. Below, we let g_{Y1} and g_{Y2} denote arbitrary functions of their arguments. Thus, within these functions, the effects of the variables are left completely unconstrained.

Proposition 5. If $Y = (1 - A)g_{Y1}(C, L, M, \epsilon_{Y1}) + Ag_{Y2}(C, L, \epsilon_{Y2})$, NIE=NIE^R; If $Y = g_{Y1}(C, A, L, \epsilon_{Y1}) + g_{Y2}(C, M, \epsilon_{Y2})$, NDE=NDE^R.

The first structural equation rules out any effect of M when $A = 1$. The second structural equation rules out A-M and L-M interactions in the equation for Y, in the sense that the nonparametric function containing M is additively separable from the nonparametric function containing A and L.

In summary, in the presence of treatment-induced confounders, it is still possible to make NIE = NIE^R or NDE = NDE^R. However, these equivalences require imposing constrains on relevant structural equations by ruling out interaction effects or two-sided effects. The structural equation constraints we present are sufficient but not necessary to establish equivalences between the natural effects and the RIA. Nevertheless, they are derived with the goal of being maximally flexible, in the sense that they allow as much complexity in functional form as possible without incurring other strong constraints.

5.3 Summary

In this paper, we answer the question of when natural mediation estimands differ from their randomized interventional analogues. In order to do so, we provide tools for both empirical testing and theoretical reasoning to researchers who wish to estimate and interpret the RIAs.

Our test and theories are complementary to one another: when the researcher rejects the test, they would know for sure that the natural effects and the RIAs differ (up to the type one error of the statistical test); when the researcher believe some structural equations are true, they would know for sure that a natural effect and its RIA are the same. With respect to the two theoretical perspectives, the covariance perspective is complete, in the sense that it provides necessary and sufficient conditions for the equivalence between the natural effects and the RIAs; while the structural equation perspective provides simple and intuitive conditions of equivalence even when M is vector-valued with arbitrary distributions.

6 Related Estimands

In causal inference, it is not unusual that a pair of competing estimands is present, where one has a more natural interpretation and the other is easier to identify. Apart from the natural mediation effects and their RIAs, we discuss two other such pairs of estimands: the average treatment effect (ATE) versus the local average treatment effect (LATE) in the IV context; and what we call the natural Mann-Whitney estimand and its RIA. The theory we developed for causal mediation analysis proves to be useful for unifying these three long-standing literatures in causal inference. In particular, we establish a formal equivalence result between estimands in the IV literature and the mediation literature. And we reveal a striking resemblance in structure between the Mann-Whitney estimands and the mediation estimands.

6.1 ATE and LATE

We first define the ATE and LATE estimands. In line with the notation we use for causal mediation analysis, we consider three temporally ordered variables, A, M, and Y. In the IV context, A is the IV, M is the treatment, and Y is the outcome. Here, we focus on the case where A and M are both binary, and A is randomized, which is a classic setting considered in the modern IV literature (Angrist et al., 1996; Balke and Pearl, 1997). Then, the ATE is defined to be $\mathrm{E}(Y_{A=1} - Y_{A=0})$, and the LATE is defined as $\mathrm{E}(Y_{M=1} - Y_{M=0} \mid M_{A=1} = 1, M_{A=0} = 0)$, i.e., the average effect of M on Y among those whose M value is induced to increase by an increase in A (those who are the "compliers"). In this subsection, we explicitly write the assignment variables in the potential outcomes to avoid ambiguity. Also note that the labelling of the ATE and the LATE involves a slight abuse of terminology in juxtaposition to mediation estimands, as the "treatment" refers to A in the IV context, while it refers to M in the mediation context.

In the IV context, the estimand with a more natural interpretation is ATE, while LATE

requires weaker identifying assumptions (Robins and Greenland, 1996; Aronow and Carnegie, 2013; Wang and Tchetgen Tchetgen, 2018). Just like in the mediation context, applied researchers often interpret a LATE estimate as if it was the ATE (Aronow and Carnegie, 2013; Sarvet et al., 2023). We show that there exists a direct equivalence between ATE−LATE and NIE−NIE^R under four standard identifying assumptions for LATE: 1) Exclusion: $Y_{A=a,M=m} = Y_{M=m}, \forall\{a, m\}$; 2) Independence: $A \perp\!\!\!\perp \{M_{A=1}, M_{A=0}, Y_{A=1}, Y_{A=0}\}$; 3) Relevance: $\mathrm{E}(M \mid A = 1) - \mathrm{E}(M \mid A = 0) > 0$; and 4) Monotonicity: $M_{A=1} \geq M_{A=0}$. We also denote the identified functional called the Wald estimand as $\text{Wald} := \frac{\mathrm{E}(Y|A=1)-\mathrm{E}(Y|A=0)}{\mathrm{E}(M|A=1)-\mathrm{E}(M|A=0)}$.

Proposition 6. Under assumptions of exclusion, independence, and relevance,

$$\text{Wald} - \text{ATE} = \frac{\mathrm{Cov}(M_{A=1} - M_{A=0}, Y_{M=1} - Y_{M=0})}{\mathrm{E}(M_{A=1} - M_{A=0})} = \frac{\text{NIE} - \text{NIE}^R}{\mathrm{E}(M_1 - M_0)},$$

which, further under monotonicity, also equals LATE − ATE.[9]

Thus, under the four assumptions identifying the LATE, the difference between the LATE and the ATE is simply the difference between the NIE and the NIE^R scaled by the effect of A on M. This means that, under these assumptions, the LATE differs from the ATE if and only if the NIE differs the NIE^R. Intuitively, $\mathrm{Cov}(M_{A=1} - M_{A=0}, Y_{M=1} - Y_{M=0}) = \mathrm{Cov}[\mathbb{1}(M_{A=1} = 1, M_{A=0} = 0), Y_{M=1} - Y_{M=0}]$ captures selection into the subpopulation of compliers based on the effect of M on Y. If there is strong selection, then the local average effect of M on Y among compliers must differ substantially from the corresponding global average effect.

There is a long-standing literature on using the Wald estimand to estimate the ATE based on exclusion, independence, relevance, and another additional assumption (Heckman, 1997; Hernán and Robins, 2006; Wang and Tchetgen Tchetgen, 2018). A weak form of the additional assumption has recently appeared in Hernán and Robins (2020, Section 16.3) and Hartwig et al. (2023), which can be written as $\mathrm{Cov}(M_{A=1} - M_{A=0}, Y_{M=1} - Y_{M=0}) = 0$. Proposition 6 shows that this is, in fact, the weakest possible among such assumptions.

6.2 Natural Mann-Whitney estimand and its RIA

We define the natural Mann-Whitney estimand as $\mathrm{E}[\mathbb{1}(Y_1 \geq Y_0)]$, i.e., the probability of the potential outcome under treatment being greater or equal to the potential outcome under control. It is often referred to as the probability of no harm (the probability of the treatment not worsening the outcome), given that a larger value of Y is desired. This estimand is broadly useful for

[9] Also, by Proposition 1 and the exclusion assumption, $\text{NIE} - \text{NIE}^R = \text{TE} - \text{TE}^R$.

rank-based evaluation of treatment effects, especially for noncontinuous ordinal outcomes.[10] We call this estimand a "natural" estimand, because it is an aggregation of an individual-level contrast of potential outcomes.

The natural Mann-Whitney estimand is difficult to identify for the same reason that the NIE and the NDE are difficult to identify: just as $\mathrm{E}(Y_{1,M_0})$, the natural Mann-Whitney estimand involves the assignment of two different treatment values to the same individual. Due to the fundamental problem of causal inference (Holland, 1986), the joint distribution of two potential outcomes is impossible to identify even with a randomized treatment. Hence, an assumption analogous to cross-world independence (Assumption 5) can also be used to identify the natural Mann-Whitney estimand: $Y_1 \perp\!\!\!\perp Y_0$ (Greenland et al., 2020). However, this assumption is clearly unlikely to hold.

Consequently, an alternative estimand has been used in practice: $\mathrm{E}[\mathbb{1}(H_1 \geq H_0)]$, where H_a is a value randomly drawn from the marginal distribution of Y_a. Clearly, this alternative estimand has the interpretation of a RIA. In contrast to the natural Mann-Whitney estimand, the Mann-Whitney RIA does not aggregate an individual-level contrast. On the other hand, randomization of treatment does enable the identification of the Mann-Whitney RIA. The Mann-Whitney RIA has a long history in statistics, dating back to the Mann-Whitney U test (Mann and Whitney, 1947) and the Wilcoxon rank-sum test (Wilcoxon, 1945). Recent methodological development based on the Mann-Whitney RIA includes the probability index model (Thas et al., 2012), the win ratio (Pocock et al., 2012; Mao, 2018), and the rank average treatment effect (Lei, 2024).

Similar to the mediation literature, conflation of the natural Mann-Whitney estimand and its RIA is pervasive even in methodological work. For example, in a textbook discussion on the Mann-Whitney RIA, Thas (2010) claims that "If this conclusion is statistically significant, it is very relevant evidence to a physician that most of his patients will be better off with the treatment." Wu et al. (2014) states "This allows us to make inference about the potential outcome-based δ through the estimable quantity ξ...", where δ and ξ are respectively the natural Mann-Whitney estimand and its RIA. And Demidenko (2016) names the Mann-Whitney RIA the "D-value" and argues that "The D-value has a clear interpretation as the proportion of patients who get worse after the treatment", in the context where a smaller value of a continuous Y is desirable.

Interestingly, despite recurrent confusion, the literature on Mann-Whitney estimands has been clarifying the important differences between the natural Mann-Whitney estimand and

[10] A related estimand, $\Pr(Y_1 > Y_0 \mid A = 1)/\Pr(Y_1 = 1 \mid A = 1)$, for a binary Y, is called the probability of necessity (Tian and Pearl, 2000).

its RIA since decades before Miles (2023) pioneered an analogous inquiry in causal mediation analysis. The early work of Hand (1992) already notes the possibility of sign reversal in the relationship between the natural Mann-Whitney estimand and its RIA (when 1/2 is subtracted from both), which has been known as the Hand's paradox. Multiple work since has considered various DGPs under which the Hand's paradox is present or absent (Hand, 1992; Fay et al., 2018; Greenland et al., 2020). This line of work is in the same spirit as our theoretical analysis on the relationship between the natural mediation estimands and their RIAs.

Lastly, we note that there is also a covariance representation for the difference between the natural Mann-Whitney estimand and its RIA.

Proposition 7.

$$\mathrm{E}[\mathbb{1}(Y_1 \geq Y_0)] - \mathrm{E}[\mathbb{1}(H_1 \geq H_0)] = \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \mathbb{1}(t \geq s)\, \mathrm{Cov}[\mathbb{1}(Y_1 = t), \mathbb{1}(Y_0 = s)],$$

where $\mathcal{T}$ and $\mathcal{S}$ are respectively the supports of Y_1 and Y_0. When Y is binary with the support of $\{0, 1\}$, the expression simplifies to $\mathrm{Cov}(Y_1, Y_0)$.

Clearly, the natural Mann-Whitney estimand differs from its RIA to the extent that Y_1 and Y_0 are dependent on each other. This is in parallel to the natural mediation effects differing from their RIAs to the extent that M_a and $Y_{a',m}$ are dependent. By redefining the estimands using random draws, RIAs in both cases miss a naturally occurring dependency. The thorny issue created by cross-world treatment assignments for identification cannot be magically waved away by redefining the estimand.

6.3 Summary

The dilemma facing researchers in all these three fields of causal inference (causal mediation analysis, instrumental variable, and Mann-Whitney estimands) is that a natural estimand is more interpretable but hard to identify while an alternative estimand is less interpretable but easier to identify. Going forward, we recommend four strategies to applied researchers. First, we join Sarvet et al. (2023) to call for more clarity in interpreting estimates of the alternative estimands in all three areas. Second, with the addition of our two theoretical perspectives in this paper, now researchers in all three areas are able to reason about when the natural estimand coincides with at least does not have the opposite sign to the alternative estimand. Third, in all three areas, bounding methods have been developed to provide partial identification for the natural estimands (e.g., Miles et al., 2017; Swanson et al., 2018; Lu et al., 2020). Fourth, in causal mediation analysis, we uniquely also provide a falsification test for interpreting the RIAs

as the natural mediation effects, which goes beyond theoretical reasoning and provides empirical guidance.

Acknowledgement

We are grateful for a comment from a reviewer at the Annals of Applied Statistics for Yu and Elwert (2024), which inspired us to start this project. We also thank Sameer Deshpande, Hyunseung Kang, Xinran Miao, Chan Park, and Michael Sobel for helpful suggestions. An earlier version of this paper was presented at the American Causal Inference Conference in 2024. We thank the audience for an engaging discussion.

Appendices

A1. Proof of Proposition 1

The NIE and NDE are defined in terms of $\mathrm{E}(Y_{a,M_{a'}})$ for two treatment values (a, a'). When M is binary and its support is $\{0, 1\}$, we rewrite this quantity just using its definition:

$$\begin{aligned}
&\mathrm{E}(Y_{a,M_{a'}}) \\
&= \mathrm{E}[Y_{a,1}M_{a'} + Y_{a,0}(1 - M_{a'})] \\
&= \mathrm{E}(Y_{a,0}) + \mathrm{E}[M_{a'}(Y_{a,1} - Y_{a,0})] \\
&= \mathrm{E}(Y_{a,0}) + \mathrm{E}\{\mathrm{E}[M_{a'}(Y_{a,1} - Y_{a,0}) \mid C]\}.
\end{aligned}$$

The NIE^R and NDE^R are defined in terms of $\mathrm{E}(Y_{a,G_{a'}})$ for two treatment values (a, a'). When M is binary, we again rewrite this quantity using its definition:

$$\begin{aligned}
&\mathrm{E}(Y_{a,G_{a'}}) \\
&= \mathrm{E}[\mathrm{E}(Y_{a,G_{a'}} \mid C)] \\
&= \mathrm{E}[\mathrm{E}(Y_{a,1} \mid G_{a'} = 1, C)\Pr(G_{a'} = 1 \mid C) + \mathrm{E}(Y_{a,0} \mid G_{a'} = 0, C)\Pr(G_{a'} = 0 \mid C)] \\
&= \mathrm{E}\{\mathrm{E}(Y_{a,1} \mid C)\mathrm{E}(M_{a'} \mid C) + \mathrm{E}(Y_{a,0} \mid C)[1 - \mathrm{E}(M_{a'} \mid C)]\} \\
&= \mathrm{E}(Y_{a,0}) + \mathrm{E}\{\mathrm{E}(M_{a'} \mid C)[\mathrm{E}(Y_{a,1} - Y_{a,0} \mid C)]\} \\
&= \mathrm{E}(Y_{a,M_{a'}}) - \mathrm{E}[\mathrm{Cov}(M_{a'}, Y_{a,1} - Y_{a,0} \mid C)].
\end{aligned}$$

Then using the results above, we have the following representations:

$$\begin{aligned}
\text{NIE} &= \text{E}(Y_{1,M_1} - Y_{1,M_0}) = \text{E}[(M_1 - M_0)(Y_{1,1} - Y_{1,0})] \\
\text{NIE}^R &= \text{E}(Y_{1,G_1} - Y_{1,G_0}) = \text{E}[\text{E}(M_1 - M_0 \mid C)\,\text{E}(Y_{1,1} - Y_{1,0} \mid C)] \\
\text{NDE} &= \text{E}(Y_{1,M_0} - Y_{0,M_0}) = \text{E}(Y_{1,0} - Y_{0,0}) + \text{E}\{M_0[Y_{1,1} - Y_{1,0} - (Y_{0,1} - Y_{0,0})]\} \\
\text{NDE}^R &= \text{E}(Y_{1,G_0} - Y_{0,G_0}) = \text{E}(Y_{1,0} - Y_{0,0}) + \text{E}\{\text{E}(M_0 \mid C)\,\text{E}[Y_{1,1} - Y_{1,0} - (Y_{0,1} - Y_{0,0}) \mid C]\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\text{NIE} &= \text{NIE}^R + \text{E}[\text{Cov}(M_1 - M_0, Y_{1,1} - Y_{1,0} \mid C)] \\
\text{NDE} &= \text{NDE}^R + \text{E}\{\text{Cov}[M_0, Y_{1,1} - Y_{1,0} - (Y_{0,1} - Y_{0,0}) \mid C]\} \\
\text{TE} &= \text{TE}^R + \text{E}[\text{Cov}(M_1, Y_{1,1} - Y_{1,0} \mid C) - \text{Cov}(M_0, Y_{0,1} - Y_{0,0} \mid C)].
\end{aligned}$$

When the treatment is randomized, C becomes an empty set, and we obtain the results shown in Proposition 1.

A2. Proof of Propositions 2 and 3

The NIE and NDE are still defined in terms of $\text{E}(Y_{a,M_{a'}})$ for two treatment values (a, a'). Treating M as a vector of continuous variable, we rewrite this quantity using its definition:

$$\begin{aligned}
&\text{E}(Y_{a,M_{a'}}) \\
&= \text{E}\left[\int Y_{a,m}\mathbb{1}(M_{a'} = m)\text{d}m\right] \\
&= \int \text{E}[Y_{a,m}\mathbb{1}(M_{a'} = m)]\text{d}m \\
&= \int \text{E}\{\text{E}[Y_{a,m}\mathbb{1}(M_{a'} = m) \mid C]\}\text{d}m,
\end{aligned}$$

where the first equality holds by treating the Dirac delta function $\mathbb{1}(M_{a'} = m)$ as a limiting case of a probability density function concentrated at $M_{a'} = m$. This allows us to express a function of $M_{a'}$ as an integral over the support of $M_{a'}$.

The NIE^R and NDE^R are defined in terms of $\text{E}(Y_{a,G_{a'}})$ for two treatment values (a, a'). We rewrite this quantity as follows:

$$\begin{aligned}
&\text{E}(Y_{a,G_{a'}}) \\
&= \text{E}[\text{E}(Y_{a,G_{a'}} \mid C)]
\end{aligned}$$

$$
\begin{aligned}
&= \iint \mathrm{E}(Y_{a,m} \mid G_{a'} = m, C = c) f_{G_{a'}|c}(m) f_C(c) \mathrm{d}m \mathrm{d}c \\
&= \iint \mathrm{E}(Y_{a,m} \mid C = c) f_{M_{a'}|c}(m) f_C(c) \mathrm{d}m \mathrm{d}c \\
&= \iint \mathrm{E}(Y_{a,m} \mid C = c) \, \mathrm{E}[\mathbb{1}(M_{a'} = m) \mid C = c] f_C(c) \mathrm{d}m \mathrm{d}c,
\end{aligned}
$$

where the last equality is by the property of the Dirac delta function $\mathbb{1}(M_{a'} = m)$.

Therefore,

$$
\begin{aligned}
\mathrm{NIE} &= \mathrm{E}(Y_{1,M_1} - Y_{1,M_0}) = \int \mathrm{E}\{\mathrm{E}\{[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m)] Y_{1,m}\} \mid C\} \mathrm{d}m \\
\mathrm{NIE}^R &= \mathrm{E}(Y_{1,G_1} - Y_{1,G_0}) = \int \mathrm{E}\{\mathrm{E}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m) \mid C] \, \mathrm{E}(Y_{1,m} \mid C)\} \mathrm{d}m \\
\mathrm{NDE} &= \mathrm{E}(Y_{1,M_0} - Y_{0,M_0}) = \int \mathrm{E}\{\mathrm{E}[(Y_{1,m} - Y_{0,m}) \mathbb{1}(M_0 = m) \mid C]\} \mathrm{d}m \\
\mathrm{NDE}^R &= \mathrm{E}(Y_{1,G_0} - Y_{0,G_0}) = \int \mathrm{E}\{\mathrm{E}[(Y_{1,m} - Y_{0,m}) \mid C] \, \mathrm{E}[\mathbb{1}(M_0 = m) \mid C]\} \mathrm{d}m.
\end{aligned}
$$

And

$$
\begin{aligned}
\mathrm{NIE} &= \mathrm{NIE}^R + \int \mathrm{E}\{\mathrm{Cov}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m), Y_{1,m} \mid C]\} \mathrm{d}m \\
\mathrm{NDE} &= \mathrm{NDE}^R + \int \mathrm{E}\{\mathrm{Cov}[Y_{1,m} - Y_{0,m}, \mathbb{1}(M_0 = m) \mid C]\} \mathrm{d}m \\
\mathrm{TE} &= \mathrm{TE}^R + \int \mathrm{E}\{\mathrm{Cov}[Y_{1,m}, \mathbb{1}(M_1 = m) \mid C]\} - \mathrm{E}\{\mathrm{Cov}[Y_{0,m}, \mathbb{1}(M_0 = m) \mid C]\} \mathrm{d}m.
\end{aligned}
$$

When M is a vector of discrete variables, we replace the integrals with summations to obtain the results in Proposition 2.

Proposition 3 similarly follows from the expressions of $\mathrm{E}(Y_{a,M_{a'}})$ and $\mathrm{E}(Y_{a,G_{a'}})$ derived above.

A3. Proof of Proposition 4

We let L_a denote the potential values of L under treatment assignment a. Under the structural equations of Proposition 4,

$$
\begin{aligned}
Y_{1M_1} &= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) L_1 + (\gamma_3 + \gamma_5) M_1 + (\gamma_6 + \gamma_7) L_1 M_1 + \epsilon_Y \\
Y_{1M_0} &= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) L_1 + (\gamma_3 + \gamma_5) M_0 + (\gamma_6 + \gamma_7) L_1 M_0 + \epsilon_Y \\
Y_{0M_0} &= \gamma_0 + \gamma_2 L_0 + \gamma_3 M_0 + \gamma_6 L_0 M_0 + \epsilon_Y \\
Y_{1G_1} &= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) L_1 + (\gamma_3 + \gamma_5) G_1 + (\gamma_6 + \gamma_7) L_1 G_1 + \epsilon_Y \\
Y_{1G_0} &= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_4) L_1 + (\gamma_3 + \gamma_5) G_0 + (\gamma_6 + \gamma_7) L_1 G_0 + \epsilon_Y
\end{aligned}
$$

$$Y_{0G_0} = \gamma_0 + \gamma_2 L_0 + \gamma_3 G_0 + \gamma_6 L_0 G_0 + \epsilon_Y.$$

Hence,

$$\begin{aligned}
\text{NDE} &= \gamma_1 + (\gamma_2 + \gamma_4)\,\text{E}(L_1) - \gamma_2\,\text{E}(L_0) + \gamma_5\,\text{E}(M_0) + (\gamma_6 + \gamma_7)\,\text{E}(L_1 M_0) - \gamma_6\,\text{E}(L_0 M_0) \\
\text{NDE}^R &= \gamma_1 + (\gamma_2 + \gamma_4)\,\text{E}(L_1) - \gamma_2\,\text{E}(L_0) + \gamma_5\,\text{E}(G_0) + (\gamma_6 + \gamma_7)\,\text{E}(L_1 G_0) - \gamma_6\,\text{E}(L_0 G_0) \\
\text{NIE} &= (\gamma_3 + \gamma_5)\,\text{E}(M_1 - M_0) + (\gamma_6 + \gamma_7)\,\text{E}(L_1 M_1 - L_1 M_0) \\
\text{NIE}^R &= (\gamma_3 + \gamma_5)\,\text{E}(G_1 - G_0) + (\gamma_6 + \gamma_7)\,\text{E}(L_1 G_1 - L_1 G_0).
\end{aligned}$$

Noting that $\text{E}(M_a) = \text{E}(G_a)$, and

$$\begin{aligned}
&\text{E}(L_a M_{a'}) - \text{E}(L_a G_{a'}) \\
&= \text{E}(L_a M_{a'}) - \text{E}(L_a)\,\text{E}(G_{a'}) \\
&= \text{Cov}(L_a, M_{a'}) \\
&= \text{Cov}[\alpha_0 + \alpha_1 a + \epsilon_L, \beta_0 + \beta_1 a' + \beta_2(\alpha_0 + \alpha_1 a' + \epsilon_L) + \beta_3 a'(\alpha_0 + \alpha_1 a' + \epsilon_L) + \epsilon_M] \\
&= (\beta_2 + \beta_3 a')\,\text{Var}(\epsilon_L) + \text{Cov}(\epsilon_L, \epsilon_M).
\end{aligned}$$

we have

$$\begin{aligned}
\text{NDE} - \text{NDE}^R &= (\gamma_6 + \gamma_7)\,\text{Cov}(L_1, M_0) - \gamma_6\,\text{Cov}(L_0, M_0) \\
&= \gamma_7 \beta_2\,\text{Var}(\epsilon_L) + \gamma_7\,\text{Cov}(\epsilon_L, \epsilon_M) \\
\text{NIE} - \text{NIE}^R &= (\gamma_6 + \gamma_7)\{\text{Cov}(L_1, M_1) - \text{Cov}(L_1, M_0)\} \\
&= (\gamma_6 + \gamma_7)\beta_3\,\text{Var}(\epsilon_L).
\end{aligned}$$

A4. Proof of Proposition 5

For the NDE part, our proof leverages an assumption in Robins (2003): $Y_{1,m} - Y_{0,m}$ is a random variable not dependent on m. Originally, this assumption was proposed to identify NDE in the presence of treatment-induced confounding. We first prove that this assumption is sufficient for $\text{NDE} = \text{NDE}^R$. Then we prove that the structural equation in Proposition 5 is, in turn, sufficient for this assumption to hold.

According to our Proposition 2, we just need to show that under the assumption of Robins

(2003), $\int \mathrm{E}\{\mathrm{Cov}[Y_{1,m} - Y_{0,m}, \mathbb{1}(M_0 = m) \mid C]\}\mathrm{d}m = 0$. Let $Y_{1,m} - Y_{0,m} = B$, then,

$$
\begin{aligned}
&\int \mathrm{E}\{\mathrm{Cov}[\mathbb{1}(M_0 = m), Y_{1,m} - Y_{0,m} \mid C]\}\mathrm{d}m \\
&= \int \mathrm{E}\{\mathrm{E}[\mathbb{1}(M_0 = m)B \mid C] - \mathrm{E}[\mathbb{1}(M_0 = m) \mid C]\,\mathrm{E}(B \mid C)\}\mathrm{d}m \\
&= \mathrm{E}\left\{\mathrm{E}\left[\int \mathbb{1}(M_0 = m)\mathrm{d}m B \mid C\right] - \int f_{M_0}(m \mid C)\mathrm{d}m\,\mathrm{E}(B \mid C)\right\} \\
&= \mathrm{E}[\mathrm{E}(B \mid C) - \mathrm{E}(B \mid C)] = 0.
\end{aligned}
$$

Next, we show that, if $Y = g_{Y1}(C, A, L, \epsilon_{Y1}) + g_{Y2}(C, M, \epsilon_{Y2})$, the assumption of Robins (2003) is satisfied. Under this structural equation for Y,

$$
\begin{aligned}
&Y_{1,m} - Y_{0,m} \\
&= g_{Y1}(C, 1, g_L(C, 1, \epsilon_L), \epsilon_{Y1}) + g_{Y2}(C, m, \epsilon_{Y2}) - g_{Y1}(C, 0, g_L(C, 0, \epsilon_L), \epsilon_{Y1}) - g_{Y2}(C, m, \epsilon_{Y2}) \\
&= g_{Y1}(C, 1, g_L(C, 1, \epsilon_L), \epsilon_{Y1}) - g_{Y1}(C, 0, g_L(C, 0, \epsilon_L), \epsilon_{Y1}),
\end{aligned}
$$

which is not dependent on m.

For the NIE part, we propose a novel condition that is analogous to the assumption of Robins (2003) used above: $Y_{1,m}$ is a random variable not dependent on m. We refer to this condition as the analogous assumption. We first show that the analogous assumption is sufficient for NIE to be equal to NIE^R. According to Proposition 2, it suffices to show $\int \mathrm{E}\{\mathrm{Cov}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m), Y_{1,m} \mid C]\}\mathrm{d}m = 0$. Let $Y_{1,m} = B$, then under the analogous assumption,

$$
\begin{aligned}
&\int \mathrm{E}\{\mathrm{Cov}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m), Y_{1,m} \mid C]\}\mathrm{d}m \\
&= \int \mathrm{E}\{\mathrm{Cov}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m), B \mid C]\}\mathrm{d}m \\
&= \int \mathrm{E}\{\mathrm{E}[\mathbb{1}(M_1 = m)B \mid C] - \mathrm{E}[\mathbb{1}(M_0 = m)B \mid C] \\
&\quad - \mathrm{E}[\mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m) \mid C]\,\mathrm{E}(B \mid C)\}\mathrm{d}m \\
&= \mathrm{E}\{\mathrm{E}[\int \mathbb{1}(M_1 = m)\mathrm{d}m B \mid C] - \mathrm{E}[\int \mathbb{1}(M_0 = m)\mathrm{d}m B \mid C] \\
&\quad - \mathrm{E}[\int \mathbb{1}(M_1 = m) - \mathbb{1}(M_0 = m)\mathrm{d}m \mid C]\,\mathrm{E}(B \mid C)\} \\
&= \mathrm{E}\{\mathrm{E}[B \mid C] - \mathrm{E}[B \mid C]\} \\
&= 0.
\end{aligned}
$$

Then, we show that if $Y = (1 - A)g_{Y1}(C, L, M, \epsilon_{Y1}) + Ag_{Y2}(C, L, \epsilon_{Y2})$, the analogous

assumption is satisfied. Under this structural equation, $Y_{1,m} = g_{Y2}(C, g_L(C, 1, \epsilon_L), \epsilon_{Y2})$, which clearly does not depend on m.

A5. Proof of Proposition 6

$$
\begin{aligned}
&\text{Wald} \\
&= \frac{\mathrm{E}(Y_{A=1} - Y_{A=0})}{\mathrm{E}(M_{A=1} - M_{A=0})} \\
&= \frac{\mathrm{E}[(M_{A=1} - M_{A=0})(Y_{M=1} - Y_{M=0})]}{\mathrm{E}(M_{A=1} - M_{A=0})} \\
&= \frac{\mathrm{E}(M_{A=1} - M_{A=0})\,\mathrm{E}(Y_{M=1} - Y_{M=0}) + \mathrm{Cov}(M_{A=1} - M_{A=0}, Y_{M=1} - Y_{M=0})}{\mathrm{E}(M_{A=1} - M_{A=0})} \\
&= \mathrm{ATE} + \frac{\mathrm{NIE} - \mathrm{NIE}^R}{\mathrm{E}(M_{A=1} - M_{A=0})}.
\end{aligned}
$$

The first equality is by the independence assumption, the second is by the exclusion assumption (equation 9 in Angrist et al. (1996)), the third is by the definition of covariance, the fourth is by Proposition 1 and the exclusion assumption. The relevance assumption makes sure the denominator is nonzero. Finally, under assumptions of exclusion, independence, relevance, and monotonicity, the classic result of Angrist et al. (1996) equates Wald with LATE.

A6. Proof of Proposition 7

$$
\begin{aligned}
&\mathrm{E}[\mathbb{1}(Y_1 \geq Y_0)] - \mathrm{E}[\mathbb{1}(H_1 \geq H_0)] \\
&= \iint \mathbb{1}(t \geq s) f_{Y_1,Y_0}(t,s)\mathrm{d}t\mathrm{d}s - \iint \mathbb{1}(t \geq s) f_{H_1,H_0}(t,s)\mathrm{d}t\mathrm{d}s \\
&= \iint \mathbb{1}(t \geq s) f_{Y_1,Y_0}(t,s)\mathrm{d}t\mathrm{d}s - \iint \mathbb{1}(t \geq s) f_{H_1}(t) f_{H_0}(s)\mathrm{d}t\mathrm{d}s \\
&= \iint \mathbb{1}(t \geq s) f_{Y_1,Y_0}(t,s)\mathrm{d}t\mathrm{d}s - \iint \mathbb{1}(t \geq s) f_{Y_1}(t) f_{Y_0}(s)\mathrm{d}t\mathrm{d}s \\
&= \iint \mathbb{1}(t \geq s)\,\mathrm{E}[\mathbb{1}(Y_1 = t)\mathbb{1}(Y_0 = s)]\mathrm{d}t\mathrm{d}s - \iint \mathbb{1}(t \geq s)\,\mathrm{E}[\mathbb{1}(Y_1 = t)]\,\mathrm{E}[\mathbb{1}(Y_0 = s)]\mathrm{d}t\mathrm{d}s \\
&= \iint \mathbb{1}(t \geq s)\,\mathrm{Cov}[\mathbb{1}(Y_1 = t), \mathbb{1}(Y_0 = s)]\mathrm{d}t\mathrm{d}s.
\end{aligned}
$$

When Y is discrete, this becomes the expression in Proposition 7. Furthermore, when the support of Y is $\{0, 1\}$,

$$
\begin{aligned}
&\sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \mathbb{1}(t \geq s)\,\mathrm{Cov}[\mathbb{1}(Y_1 = t), \mathbb{1}(Y_0 = s)] \\
&= \mathrm{Cov}[\mathbb{1}(Y_1 = 1), \mathbb{1}(Y_0 = 1)] + \mathrm{Cov}[\mathbb{1}(Y_1 = 1), \mathbb{1}(Y_0 = 0)] + \mathrm{Cov}[\mathbb{1}(Y_1 = 0), \mathbb{1}(Y_0 = 0)] \\
&= \mathrm{E}[\mathbb{1}(Y_1 = 1)\mathbb{1}(Y_0 = 1)] - \mathrm{E}[\mathbb{1}(Y_1 = 1)]\,\mathrm{E}[\mathbb{1}(Y_0 = 1)]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathrm{E}[\mathbb{1}(Y_1 = 1)\mathbb{1}(Y_0 = 0)] - \mathrm{E}[\mathbb{1}(Y_1 = 1)]\,\mathrm{E}[\mathbb{1}(Y_0 = 0)] \\
&\quad + \mathrm{E}[\mathbb{1}(Y_1 = 0)\mathbb{1}(Y_0 = 0)] - \mathrm{E}[\mathbb{1}(Y_1 = 0)]\,\mathrm{E}[\mathbb{1}(Y_0 = 0)] \\
&= \mathrm{E}(Y_1 Y_0) - \mathrm{E}(Y_1)\,\mathrm{E}(Y_0) \\
&\quad + \mathrm{E}[Y_1(1 - Y_0)] - \mathrm{E}(Y_1)[1 - \mathrm{E}(Y_0)] + \mathrm{E}[(1 - Y_1)(1 - Y_0)] - \mathrm{E}[(1 - Y_1)]\,\mathrm{E}[(1 - Y_0)] \\
&= \mathrm{E}(Y_1 Y_0) - \mathrm{E}(Y_1)\,\mathrm{E}(Y_0) = \mathrm{Cov}(Y_1, Y_0).
\end{aligned}
$$

References

Andrews, R. M. and V. Didelez (2021, March). Insights into the Cross-world Independence Assumption of Causal Mediation Analysis. *Epidemiology 32*(2), 209–219.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Aronow, P. M. and A. Carnegie (2013). Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable. *Political Analysis 21*(4), 492–506.

Avin, C., I. Shpitser, and J. Pearl (2005). Identifiability of Path-Specific Effects. In *Proceedings of International Joint Conference on Artificial Intelligence*, Edinburgh, Schotland, pp. 357–363.

Balke, A. and J. Pearl (1997, September). Bounds on Treatment Effects From Studies With Imperfect Compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Benkeser, D. and J. Ran (2021, July). Nonparametric inference for interventional effects with multiple mediators. *Journal of Causal Inference 9*(1), 172–189.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York, NY: Springer.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, February). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Demidenko, E. (2016, January). The p -Value You Can't Buy. *The American Statistician 70*(1), 33–38.

Díaz, I., N. S. Hejazi, K. E. Rudolph, and M. J. Van Der Laan (2021, August). Nonparametric efficient causal mediation with intermediate confounders. *Biometrika 108*(3), 627–641.

Díaz, I., N. Williams, and K. E. Rudolph (2023, July). Efficient and flexible mediation analysis with time-varying mediators, treatments, and confounders. *Journal of Causal Inference 11*(1), 20220077.

Fay, M. P., E. H. Brittain, J. H. Shih, D. A. Follmann, and E. E. Gabriel (2018, September). Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Statistics in Medicine 37*(20), 2923–2937.

Greenland, S., M. P. Fay, E. H. Brittain, J. H. Shih, D. A. Follmann, E. E. Gabriel, and J. M. Robins (2020, July). On Causal Inferences for Personalized Medicine: How Hidden Causal Assumptions Led to Erroneous Causal Claims About the *D* -Value. *The American Statistician 74*(3), 243–248.

Hahn, J. (1998, March). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica 66*(2), 315.

Hand, D. J. (1992, August). On Comparing Two Treatments. *The American Statistician 46*(3), 190–192.

Hartwig, F. P., L. Wang, G. D. Smith, and N. M. Davies (2023). Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. *Epidemiology 34*(3), 325–332.

Heckman, J. (1997). Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *The Journal of Human Resources 32*(3), 441.

Hejazi, N., K. Rudolph, and I. Díaz (2022, January). medoutcon: Nonparametric efficient causal mediation analysis with machine learning in R. *Journal of Open Source Software 7*(69), 3979.

Hernán, M. A. and J. M. Robins (2006, July). Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology 17*(4), 360–372.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hines, O., O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician 76*(3), 292–304.

Holland, P. (1986, 12). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imai, K., L. Keele, and T. Yamamoto (2010, February). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science 25*(1), 51 – 71.

Jackson, J. W. (2021, March). Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology 32*(2), 282–290.

Kling, J. R., J. Ludwig, and L. F. Katz (2005, February). Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment. *The Quarterly Journal of Economics 120*(1), 87–130.

Lei, L. (2024, June). Causal Interpretation of Regressions With Ranks. arXiv:2406.05548 [econ, math, stat].

Loh, W. W., B. Moerkerke, T. Loeys, and S. Vansteelandt (2020). Heterogeneous indirect effects for multiple mediators using interventional effect models. *Epidemiologic Methods 9*(1), 20200023.

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

Lok, J. J. and R. J. Bosch (2021, May). Causal Organic Indirect and Direct Effects: Closer to the Original Approach to Mediation Analysis, with a Product Method for Binary Mediators. *Epidemiology 32*(3), 412–420.

Lu, J., Y. Zhang, and P. Ding (2020). Sharp bounds on the relative treatment effect for ordinal outcomes. *Biometrics 76*(2), 664–669.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. R. Katz, R. Kessler, J. R. Kling, and L. Sanbonmatsu (2013, March). Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults From All Five Sites of the Moving to Opportunity Experiment, 2008-2010 [Public Use Data].

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics 18*(1), 50–60.

Mao, L. (2018, March). On causal estimation using u-statistics. *Biometrika 105*(1), 215–220.

Miles, C., P. Kanki, S. Meloni, and E. Tchetgen Tchetgen (2017, February). On Partial Identification of the Natural Indirect Effect. *Journal of Causal Inference 5*(2), 20160004.

Miles, C. H. (2023, September). On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(4), 1154–1172.

Nguyen, T. Q., I. Schmid, E. L. Ogburn, and E. A. Stuart (2022, September). Clarifying causal mediation analysis: Effect identification via three assumptions and five potential outcomes. *Journal of Causal Inference 10*(1), 246–279.

Nguyen, T. Q., I. Schmid, and E. A. Stuart (2021, April). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods 26*(2), 255–271.

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intel ligence*, San Francisco, CA, pp. 411–20. Morgan Kaufmann.

Pocock, S. J., C. A. Ariti, T. J. Collier, and D. Wang (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European heart journal 33*(2), 176–182.

Robins, J. M. (2003, May). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 70–82. Oxford: Oxford University Press.

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

Robins, J. M. and S. Greenland (1996). Identification of causal effects using instrumental variables: comment. *Journal of the American Statistical Association 91*(434), 456–458.

Rudolph, K. E., C. Gimbrone, and I. Díaz (2021, May). Helped into Harm: Mediation of a Housing Voucher Intervention on Mental Health and Substance Use in Boys. *Epidemiology 32*(3), 336–346.

Rudolph, K. E., O. Sofrygin, W. Zheng, and M. J. Van Der Laan (2018, November). Robust and Flexible Estimation of Stochastic Mediation Effects: A Proposed Method and Example in a Randomized Trial Setting. *Epidemiologic Methods 7*(1), 20170007.

Rudolph, K. E., N. T. Williams, and I. Diaz (2024, April). Practical causal mediation analysis: extending nonparametric estimators to accommodate multiple mediators and multiple intermediate confounders. *Biostatistics*, kxae012.

Sarvet, A. L., M. J. Stensrud, and L. Wen (2023, December). Interpretational errors in statistical causal inference. arXiv:2312.07610 [stat].

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018, April). Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes. *Journal of the American Statistical Association 113*(522), 933–947.

Thas, O. (2010). *Comparing distributions*. Springer Series in Statistics. New York: Springer.

Thas, O., J. D. Neve, L. Clement, and J.-P. Ottoy (2012, September). Probabilistic Index Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 74*(4), 623–671.

Tian, J. and J. Pearl (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence 28*(1), 287–313.

Van der Laan, M. J. and S. Rose (2011). *Targeted learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. New York: Springer.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and E. J. Tchetgen Tchetgen (2017, June). 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., S. Vansteelandt, and J. M. Robins (2014, March). Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology 25*(2), 300–306.

Vansteelandt, S. and R. M. Daniel (2017, March). Interventional Effects for Mediation Analysis with Multiple Mediators:. *Epidemiology 28*(2), 258–265.

Vo, T.-T., N. Williams, R. Liu, K. E. Rudolph, and I. Dıaz (2024, January). Recanting twins: addressing intermediate confounding in mediation analysis. arXiv:2401.04450 [stat].

Wang, L. and E. Tchetgen Tchetgen (2018, June). Bounded, Efficient and Multiply Robust Estimation of Average Treatment Effects Using Instrumental Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 531–550.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*(6), 80–83.

Williams, N., K. Rudolph, and I. Díaz (2024). *HDmediation: Interventional (In)Direct Effects*. R package version 0.1.0.9000.

Wodtke, G. T. and X. Zhou (2020, May). Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-residuals Approach. *Epidemiology 31*(3), 369–375.

Wright, M. N. and A. Ziegler (2017). **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software 77*(1), 1–17.

Wu, P., Y. Han, T. Chen, and X. Tu (2014, April). Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. *Statistics in Medicine 33*(8), 1261–1271.

Yu, A. and F. Elwert (2024). Nonparametric causal decomposition of group disparities. arXiv:2306.16591 [stat].

Zheng, W. and M. Van Der Laan (2017, June). Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes. *Journal of Causal Inference 5*(2), 20160006.