Received: 1 June 2021 | Revised: 5 May 2022 | Accepted: 18 May 2022

DOI: 10.1111/biom.13705

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Posttreatment confounding in causal mediation studies: A cutting-edge problem and a novel solution via sensitivity analysis

Guanglei Hong[1] | **Fan Yang**[2] | **Xu Qin**[3]

[1]University of Chicago, Chicago, Illinois, USA

[2]University of Colorado Denver, Aurora, Colorado, USA
(Email: fan.3.yang@cuanschutz.edu)

[3]University of Pittsburgh, Pittsburgh, Pennsylvania, USA
(Email: xuqin@pitt.edu)

Correspondence
Guanglei Hong, University of Chicago, 1126 E 59th Street, Chicago, IL 60637, USA.
Email: ghong@uchicago.edu

Abstract

In causal mediation studies that decompose an average treatment effect into indirect and direct effects, examples of posttreatment confounding are abundant. In the presence of treatment-by-mediator interactions, past research has generally considered it infeasible to adjust for a posttreatment confounder of the mediator–outcome relationship due to incomplete information: for any given individual, a posttreatment confounder is observed under the actual treatment condition while missing under the counterfactual treatment condition. This paper proposes a new sensitivity analysis strategy for handling posttreatment confounding and incorporates it into weighting-based causal mediation analysis. The key is to obtain the conditional distribution of the posttreatment confounder under the counterfactual treatment as a function of not only pretreatment covariates but also its counterpart under the actual treatment. The sensitivity analysis then generates a bound for the natural indirect effect and that for the natural direct effect over a plausible range of the conditional correlation between the posttreatment confounder under the actual and that under the counterfactual conditions. Implemented through either imputation or integration, the strategy is suitable for binary as well as continuous measures of posttreatment confounders. Simulation results demonstrate major strengths and potential limitations of this new solution. A reanalysis of the National Evaluation of Welfare-to-Work Strategies (NEWWS) Riverside data reveals that the initial analytic results are sensitive to omitted posttreatment confounding.

KEYWORDS

causal inference, direct effect, indirect effect, posttreatment confounding, potential outcomes, RMPW

Causal mediation analysis decomposes a total average treatment effect (ATE) on an outcome into indirect and direct effects; the most common example is the decomposition of the ATE into a natural indirect effect (NIE) and a natural direct effect (NDE). Past research has generally considered it infeasible to adjust for a posttreatment confounder when there exists treatment-by-mediator interaction (Avin et al., 2005; Robins, 2003). A posttreatment

confounder may be viewed as an additional mediator that precedes the focal mediator on causal pathways. Handling posttreatment confounding is a cutting-edge problem in causal mediation analysis. This is because a posttreatment confounder is only *partially observed*. Specifically, if an individual has been assigned to the experimental condition, the individual's potential posttreatment confounder value associated with the counterfactual control condition is unobserved. In this sense, the problem with statistical adjustment for a posttreatment confounder is *a problem of missing data*.

This paper proposes a new sensitivity analysis strategy for handling posttreatment confounding and incorporates it into weighting-based causal mediation studies. The key is to obtain, for individuals in the experimental group, the conditional distribution of the posttreatment confounder under the counterfactual control condition, which is then adjusted for along with the distribution of the same confounder under the experimental condition. The analyst can make the adjustment flexibly through ratio-of-mediator-probability weighting (RMPW) that allows for a treatment-by-mediator interaction (Hong, 2010, 2015; Hong & Nomi, 2012; Hong et al., 2015; Lange et al., 2012; Tchetgen Tchetgen & Shpister, 2012). Corresponding to a plausible range of the conditional correlation between the posttreatment confounder under the experimental condition and that under the control condition, our analytic procedure generates a bound for the NIE and that for the NDE, and thereby enabling the analyst to assess the sensitivity of the initial results to the omitted confounding.

Researchers have proposed several alternative strategies for handling posttreatment confounding in the presence of treatment-by-mediator interaction. The most relevant alternatives invoke an extra assumption about the ignorability of the posttreatment covariate that is viewed as the first among several sequentially ordered mediators, along with a series of model-based assumptions within the linear or generalized linear structural model framework (Albert & Nelson, 2011; Albert et al., 2019; Daniel et al., 2015; Imai & Yamamoto, 2013). Past research that has proposed to use weighting to adjust for posttreatment confounding similarly invokes additional strong assumptions involving the posttreatment confounder (Hong, 2015; Hong et al., 2018; Huber, 2014). Other researchers have opted to change the causal estimands such that the interest is no longer in decomposing the ATE into NIE and NDE (Geneletti, 2007; Rudolph et al., 2018; VanderWeele et al., 2014; Miles et al., 2017; Wodtke & Zhou, 2020) except under special conditions (Vansteelandt & Daniel, 2017).

Our solution focuses on decomposing the ATE into an NIE and an NDE. There are important distinctions between our key identification assumptions and the extended sequential ignorability assumptions proposed in the past research. For example, unlike Daniel et al. (2015), our strategy does not require the strong assumption that the posttreatment covariate be conditionally independent of the potential outcomes given the observed pretreatment covariates. Rather, we allow for unmeasured confounding of the relationship between the posttreatment covariate and the outcome. This is because instead of attempting to identify the treatment effect mediated via the posttreatment covariate, this paper has a much less ambitious goal, that is, to assess the potential bias associated with the omission of a posttreatment covariate that precedes the focal mediator. A unique feature of our method is that we obtain the conditional distribution of the posttreatment covariate under the counterfactual treatment condition as a function of not only pretreatment covariates but also its counterpart under the actual treatment condition. Similar to Albert and Nelson (2011), we consider the conditional correlation between the posttreatment covariate under the actual treatment condition and that under the counterfactual treatment condition as a sensitivity parameter and estimate confidence bands for NIE and NDE within the range of plausible values of this sensitivity parameter.

This article is organized as follows. Section 1 introduces the application study and illustrates the need for handling posttreatment confounding. Section 2 presents the theoretical rationale for our new solution. Section 3 lays out a sensitivity analysis strategy for assessing the potential consequence of omitting a posttreatment confounder. Section 4 investigates the performance of the new method across a range of realistic scenarios through a series of simulations. Section 5 demonstrates the implementation in the real-data application. Section 6 concludes and discusses further extensions.

1 | APPLICATION CONTEXT

The welfare-to-work reform bill in the mid-1990s was intended to reduce welfare applicants' dependence on the cash assistance system by providing incentives for participation in the labor force. Shortly before the new legislation, a randomized evaluation assessed the potential impacts of this radical overhaul of the welfare system. In Riverside, California, 694 welfare applicants with young children between ages 3 and 5 were assigned at random to either an experimental condition ($T = 1$)—a labor force attachment (LFA) program—or a control condition ($T = 0$). The LFA program offered job search services and incentives including a threat of sanctions should one fail to meet the program requirements for actively seeking and securing employment; in contrast, the control group members were guaranteed cash assistance without the requirement for employment. The psychological well-being of welfare applicants with young children was of concern because many were single mothers already disproportionately depressed at the baseline. Under the experimental

condition, the prospect of potentially losing the public safety net would likely trigger or aggravate depression.

1.1 | Research questions and causal estimands

Hong et al. (2015) investigated whether a treatment-induced change in employment played a mediating role in transmitting the program impact on maternal depression (Y) 2 years later. A self-administered 12-item questionnaire (Center for Epidemiologic Studies–Depression Scale; Radloff, 1977) measured depressive symptoms during the past week. The score ranged from 0 to 34 with a mean value of 7.49 and a standard deviation of 7.74. The mediator (M) indicated one's posttreatment employment record prior to the measure of depression. For simplicity, we reexamine the results based on a binary mediator that takes values $M = 1$ if an individual was ever employed and $M = 0$ if the individual was never employed during the 2 years after randomization. The following research questions correspond to the ATE, the NIE, and the NDE:

ATE: What is the average impact of the LFA program on maternal depression 2 years later?

NIE: How much of this impact is attributable to program-induced change in employment?

NDE: What would be the average program impact on maternal depression if every individual's employment status counterfactually took the value it would have in the absence of the program?

Each causal effect is defined in terms of potential outcomes:

$$ATE = E\{Y(1) - Y(0)\} \equiv E\{Y(1, M(1)) - Y(0, M(0))\};$$

$$NIE = E\{Y(1, M(1)) - Y(1, M(0))\};$$

$$NDE = E\{Y(1, M(0)) - Y(0, M(0))\}.$$

The sum of NIE and NDE is equal to ATE (Pearl, 2001; Robins & Greenland, 1992). Here $M(1)$ and $M(0)$ denote an individual's potential employment status under the experimental condition and the control condition, respectively; $Y(1, M(1))$ and $Y(0, M(0))$ denote the individual's potential depression level under the respective treatment conditions; the third potential outcome $Y(1, M(0))$ denotes the same individual's potential depression level under the experimental condition should the treatment counterfactually fail to change the individual's employment status from its value associated with the control condition $M(0)$. Each potential mediator or potential outcome is a random variable that "naturally" takes different values. The NIE is the average impact of the program on maternal depression associated with a change from $M(0)$ to $M(1)$ should all individuals be subjected to the new policy requirements; the NDE is the program impact on maternal depression should all individuals' employment status remain unchanged by the program.

1.2 | Initial results and potential bias

Although individuals were assigned at random to one of the two treatment conditions, they were not assigned at random to different mediator values under each treatment condition. Researchers of the original study (Hong, 2015) made adjustment for a set of pretreatment covariates including demographics, family structure, education, baseline depressive symptoms, prior employment history, and prior history of welfare dependence through an RMPW analysis. This adjustment strategy simply transforms through weighting the mediator distribution of the experimental group to resemble that of the control group within levels of the pretreatment covariates. The average weighted outcome of the experimental group identifies the average counterfactual outcome $E\{Y(1, M(0))\}$ under the assumption of no omitted confounders. The RMPW strategy allows for treatment-by-mediator interaction. This is relevant because the mediator–outcome relationship differed between the two treatment conditions as reported in Hong et al. (2015): as anticipated, an individual's failure in seeking employment heightened depressive symptoms under the experimental condition but not under the control condition. More generally, RMPW is flexible for handling any types of nonlinearity because it does not require an explicit specification of the outcome model; it also has the flexibility of accommodating multicategory and multivalued mediators.

The researchers reported tentative evidence that, on the one hand, assignment to LFA indeed increased employment rate from 40% to 65%, which would lead to a considerable reduction in depression on average as a result; on the other hand, should the employment rate have failed to improve, assignment to LFA would have increased depression on average. The estimated effect size of NIE was −0.11 with a 95% confidence interval [−0.24, 0.01]; and that of NDE was 0.13 with a 95% confidence interval [−0.09, 0.35].

The identification required the *sequential ignorability* assumptions (Imai et al., 2010):

(1) *Ignorable treatment assignment* given the observed pretreatment covariates $\mathbf{X} = \mathbf{x}$:

$$Y(t, m), M(t), M(t') \perp\!\!\!\perp T \mid \mathbf{X} = \mathbf{x}.$$

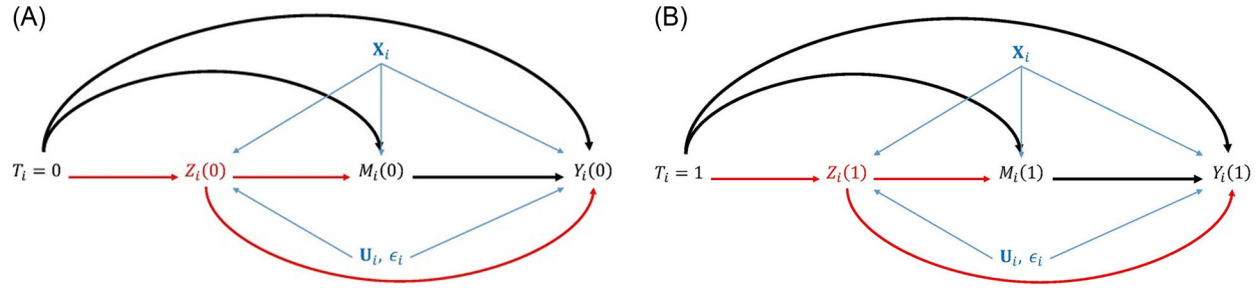

FIGURE 1 (A) Causal diagram for individual i under the control condition. (B) Causal diagram for individual i under the experimental condition (this figure appears in color in the electronic version of this article, and any mention of color refers to that version)

(2) *Ignorable mediator value assignment* under each treatment condition given $\mathbf{X} = \mathbf{x}$:

$$Y(t, m) \perp\!\!\!\perp M(t), M\left(t'\right) \mid T = t, \mathbf{X} = \mathbf{x}.$$

While Assumption (1) was guaranteed by the randomized treatment assignment, Assumption (2) would be violated in the presence of omitted pretreatment confounders. Hong et al. (2018) proposed a weighting-based approach to sensitivity analysis that is integrated with the RMPW strategy for causal mediation analysis. Conducting a sensitivity analysis with the same data, they reported that the initial estimate of NIE was sensitive to bias associated with some of the omitted pretreatment covariates.

Assumption (2) would also be violated in the presence of a posttreatment covariate that confounds the mediator–outcome relationship under either or both treatment conditions. One example is posttreatment welfare amount. On average, individuals in the experimental group received significantly less welfare amount than did their counterparts in the control group during the first year after randomization. Within a treatment group, a higher amount of posttreatment welfare appears to be associated with a lower propensity of employment. Another example is whether an individual was continuously on welfare during the first year after randomization. The proportion of such individuals was significantly higher in the control group than in the experimental group. Those with continuous welfare were also less likely to become employed.

We use Z to denote a post-treatment covariate. To clarify the unique challenge to causal mediation analysis in the presence of posttreatment confounding despite the randomization of treatment assignment, we illustrate with a pair of graphs each representing one of the two treatment conditions. As shown in Figure 1(A), individual i under the control condition would display potential posttreatment covariate $Z_i(0)$, potential mediator $M_i(0)$, and potential outcome $Y_i(0)$. Figure 1(B) shows that the same individual, under the alternative experimental condition, would display $Z_i(1)$, $M_i(1)$, and $Y_i(1)$. Under treatment condition t for $t = 0, 1$, a vector of pretreatment covariates $\mathbf{X}_i$ may predict $Z_i(t)$, $M_i(t)$, and $Y_i(t)$. Moreover, a vector of unobserved pretreatment covariates $\mathbf{U}_i$ along with a set of individual-specific random events denoted by ϵ_i may predict $Z_i(t)$ and $Y_i(t)$. In the current application, examples of $\mathbf{U}_i$ might include whether an individual was financially dependent on relatives at the baseline; examples of ϵ_i might include whether important communications from the social services administration were accidentally lost in mail, a random incident that could cause an unintended lapse of welfare. Here the value of $\mathbf{U}_i$ is fixed for individual i; in contrast, ϵ_i is a random variable that generates uncertainty in the individual's value of $Z_i(t)$. The conditional correlation between $Z_i(0)$ and $Z_i(1)$ depends on the variance of $\mathbf{U}_i$ relative to the variance of ϵ_i. In our new strategy for sensitivity analysis, we will relax the second component of the sequential ignorability assumption by instead assuming ignorable mediator value assignment given $\mathbf{X}_i$ and $Z_i(t)$. Under this assumption, $\mathbf{U}_i$ and ϵ_i are conditionally independent of $M_i(1)$ and $M_i(0)$.

However, it is well known that statistical adjustment for a posttreatment covariate would introduce bias in identifying the ATE (Rosenbaum, 1984). The observed value of Z_i is related to the potential values of the posttreatment covariate as follows: $Z_i = T_i Z_i(1) + (1 - T_i)Z_i(0)$. Because $Z(1)$ and $Z(0)$ have different distributions in the population, comparing individuals in the experimental group whose $Z(1) = z$ with those in the control group whose $Z(0) = z$ is tantamount to comparing "apples" and "oranges" in many cases. To remove bias associated with the posttreatment confounder in identifying the NIE and NDE, it is necessary to make statistical adjustment for both $Z(0)$ and $Z(1)$ along with $\mathbf{X}$. The fundamental difficulty is that the analyst could observe either $Z(0)$ or $Z(1)$ but not both.

2 | THEORETICAL RATIONALE FOR THE NEW SOLUTION

In the presence of a posttreatment confounder that precedes the focal mediator, $M(t)$ now represents $M(t, Z(t))$,

whereas $Y(t, M(t))$ now represents $Y(t, Z(t), M(t, Z(t)))$, for $t = 0, 1$. In a causal mediation analysis, the major challenge is to use the observed information to identify the counterfactual quantity $E\{Y(1, M(0))\}$, which now represents $E\{Y(1, Z(1), M(0, Z(0)))\}$. We modify the sequential ignorability assumptions as follows for $t, t' = 0, 1$ and $t \neq t'$:

(1*) *Ignorable treatment assignment* given the observed pretreatment covariates:

$$Y(t, m), M(t), M(t'), Z(t), Z(t') \perp\!\!\!\perp T | \mathbf{X} = \mathbf{x}.$$

Given that $Z(t)$ and $Z(t')$ are each specified under a respective fixed treatment condition, the following result can be easily derived from assumption (1*). This variant of assumption (1*) states that the treatment assignment is ignorable given the potential posttreatment covariate values $Z(t) = z$ and $Z(t') = z'$ in addition to $\mathbf{X} = \mathbf{x}$. Web Appendix 1 shows the derivation.

$$Y(t, m), M(t), M(t') \perp\!\!\!\perp T | Z(t) = z, Z(t') = z', \mathbf{X} = \mathbf{x}.$$

Assumption (1*) and its variant are guaranteed when the treatment is randomized, which is indeed the case in the current application. However, like assumption (1), assumption (1*) could be overly strong in a quasi-experimental study in which treatment selection might be associated with unobserved pretreatment covariates within levels of $\mathbf{X}$.

(2*) *Ignorable mediator value assignment* under each treatment condition given the observed pretreatment and posttreatment covariates:

$$Y(t, m) \perp\!\!\!\perp M(t), M(t') | T = t, Z(t) = z, \mathbf{X} = \mathbf{x}.$$

This assumption rules out cross-world connections between the mediator and the outcome. Assumption (2*) is considerably more plausible than the standard ignorability assumption (2) that conditions on pretreatment covariates only. This is simply because posttreatment confounding is often inevitable. As we discussed, in the NEWWS data, welfare receipt under the experimental condition during the year after the randomization is regarded as a posttreatment covariate $Z(1)$ that could confound the mediator–outcome relationship. In general, individuals eligible for a greater amount of welfare tended to face more barriers to employment, which could also make them more depressed. Under assumption (2*), among individuals with the same baseline covariate values $\mathbf{X} = \mathbf{x}$ and additionally with the same posttreatment welfare receipt $Z(1) = z$, the potential level of depression 2 years after the randomization $Y(1, m)$ is assumed to be independent of potential employment status associated with each of the two treatment conditions $M(1)$ and $M(0)$. Nonetheless, assumption (2*) might not hold should the mediator-outcome relationship be confounded by additional pretreatment or posttreatment covariates. For example, a family member's health problems before or after the randomization might predict employment and depression. Omitting a potential confounder as such would likely violate assumption (2*).

(3*) *Conditional cross-world independence between the* post*treatment covariate and the mediator*:

$$M(t') \perp\!\!\!\perp Z(t) | T = t', Z(t') = z', \mathbf{X} = \mathbf{x}.$$

Assumption (3*) simply states that under treatment condition t', when $Z(t')$ and $\mathbf{X}$ are already given, $Z(t)$ does not supply additional information about $M(t')$. Potential violations of assumption (3*) are conceivable if additional pretreatment or posttreatment covariates predict $M(t')$ and $Z(t)$ without affecting $Z(t')$. However, in the current application, it seems unlikely that among individuals in the control group who received the same amount of posttreatment welfare and shared the same baseline characteristics, their employment status under the control condition would be additionally predicted by their counterfactual posttreatment welfare amount associated with the experimental condition.

The following theoretical results are key to our new solution. Given the focus on decomposing the ATE into the NIE and the NDE, we identify $E\{Y(1, M(0))\}$ by applying the RMPW method to individuals in the experimental group for whom $Z(1)$ is observed. We explain in Section 6 how the results may apply to an alternative decomposition of ATE.

Theorem.

Under assumptions (1*) and (3*),

$$P(M(0) = m | T = 1, Z(1) = z, \mathbf{X} = \mathbf{x})$$

$$= \int P(M(0) = m | T = 0, Z(0) = z', \mathbf{X} = \mathbf{x}) h_{z'} dz',$$

where $h_{z'} = h_{z'}(\mathbf{x}) = P(Z(0) = z' | T = 1, Z(1) = z, \mathbf{X} = \mathbf{x})$.

See Web Appendix 2 for the proof. This theorem states that, for an individual who has been assigned to $T = 1$ and displays covariate values $\mathbf{X} = \mathbf{x}$ and $Z(1) = z$, the individual's conditional probability of displaying a certain mediator value under the counterfactual control condition $M(0) = m$ can be obtained when we relate it to the conditional probability of the mediator value for the individual's counterparts who have actually been assigned to $T = 0$ with covariate values $\mathbf{X} = \mathbf{x}$ and $Z(0) = z'$. Importantly, for the

focal individual who has been assigned to $T = 1$, the conditional probability of $M(0)$ under the counterfactual condition needs to be averaged over the individual's conditional distribution of $Z(0)$ denoted by $h_{z'}$.

Let $W_t(\mathbf{x}) = \frac{P(T=t)}{P(T=t|\mathbf{X}=\mathbf{x})}$ for individuals assigned to $T = t$, for $t = 0, 1$, with covariate values $\mathbf{X} = \mathbf{x}$. When the treatment is completely randomized, $W_t(\mathbf{x}) = 1$ for all individuals. Let $W(\mathbf{x}, z, z') = \frac{P(M(0)=m|T=0, Z(0)=z', \mathbf{X}=\mathbf{x})}{P(M(1)=m|T=1, Z(1)=z, \mathbf{X}=\mathbf{x})}$ for individuals assigned to $T = 1$ with covariate values $\mathbf{X} = \mathbf{x}$ and $Z(1) = z$. Henceforth we use $W(z')$ as a shorthand for $W(\mathbf{x}, z, z')$.

Lemma.

Under assumptions (1*), (2*), and (3*), NIE and NDE can be identified through weighting:

$$NIE = E\{W_1(\mathbf{x})Y|T = 1\} - E\left\{\int W_1(\mathbf{x})W(z')Yh_{z'}dz'|T = 1\right\};$$

$$NDE = E\left\{\int W_1(\mathbf{x})W(z')Yh_{z'}dz'|T = 1\right\} - E\{W_0(\mathbf{x})Y|T = 0\}.$$

Web Appendix 3 provides a proof of this lemma.

The theoretical result enables sensitivity analysis for assessing the potential consequence of omitting a posttreatment confounder. In the subsequent sections, we propose and evaluate analytic strategies for empirically estimating the parameters that define the distribution of $Z(0)$ when $\mathbf{X} = \mathbf{x}$ and $Z(1) = z$ are given. The conditional distribution of $Z(0)$ depends on the unknown conditional correlation between $Z(0)$ and $Z(1)$ regarded as a sensitivity parameter. Although we restrict the discussion to the case of a completely randomized experiment, it is straightforward to extend these strategies to evaluations of treatments randomized within levels of pretreatment covariates.

3 | SENSITIVITY ANALYSIS FOR A POSTTREATMENT CONFOUNDER

This section describes the theoretical models and the analytic steps that we use to obtain the conditional distribution of $Z(0)$ for individuals assigned to the experimental condition in a completely randomized trial. Assuming a bivariate normal distribution of $Z(0)$ and $Z(1)$ conditioning on the observed pretreatment covariates $\mathbf{X} = \mathbf{x}$, we obtain the conditional distribution of $Z(0)$ for individuals assigned to $T = 1$. For a binary confounder, we instead assume a bivariate probit model for the joint relationship between $Z(0)$ and $Z(1)$. We then impute $Z(0)$ by taking multiple random draws from its conditional distribution; an alternative strategy is to take the integral of $Z(0)$ over its conditional distribution. For the simplicity of presentation, below we describe the analytic strategy for the case of a continuous posttreatment confounder and leave to Web Appendix 4 additional details about the case of a binary posttreatment confounder.

3.1 | Theoretical models

As illustrated in Figures 1(A) and 1(B), $Z_i(0)$ and $Z_i(1)$ are each a function of observed pretreatment covariates $\mathbf{X}_i$, unobserved pretreatment covariates $\mathbf{U}_i$, and the impact of random events ϵ_i. In general, the causal effect of T on Z may depend on $\mathbf{X}$, $\mathbf{U}$, and ϵ. Because the impacts of $\mathbf{U}_i$ and ϵ_i on $Z_i(t)$ are not empirically distinguishable, we use r_{ti} to denote their joint impact: $Z_i(t) = \mu(t, \mathbf{x}) + r_{ti}$, for $t = 0, 1$.

Let $\mu(t, \mathbf{x})$ denote a function of $\mathbf{x}$ under treatment condition t; $\mu(t, \mathbf{x})$ may take any flexible parametric or nonparametric functional form. For individual i whose $T_i = 1$, $Z_i(1) = z$, and $\mathbf{X}_i = \mathbf{x}$, due to the randomness of r_{1i} and r_{0i}, the counterfactual $Z_i(0)$ is a random variable. We assume that $\begin{pmatrix} r_{0i} \\ r_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{10}\sigma_1\sigma_0 & \sigma_1^2 \end{pmatrix}\right)$. Here r_{1i} and r_{0i} are assumed to be bivariate normal; $\sigma_t^2 = \text{var}(r_{ti}|T_i = t, \mathbf{X}_i = \mathbf{x})$ for $t = 0, 1$; and $\rho_{10} = \text{corr}(Z_i(1), Z_i(0)|T_i = t, \mathbf{X}_i = \mathbf{x}) = \text{corr}(Z_i(1), Z_i(0)|\mathbf{X}_i = \mathbf{x})$ under assumption (1*). By convention, σ_1, σ_0, and ρ_{10} are assumed to be invariant across different values of $\mathbf{x}$. We will investigate potential implications of these distributional assumptions through simulations presented in Section 4. We derive the conditional distribution of $Z_i(0)$ for individual i whose $T_i = 1, Z_i(1) = z$, and $\mathbf{X}_i = \mathbf{x}$:

$$E\{Z_i(0)|T_i = 1, Z_i(1) = z, \mathbf{X}_i = \mathbf{x}\} = \mu(0, \mathbf{x}) + \rho_{10}\frac{\sigma_0}{\sigma_1}(z - \mu(1, \mathbf{x}));$$

$$var\{Z_i(0)|T_i = 1, Z_i(1) = z, \mathbf{X}_i = \mathbf{x}\} = \left(1 - \rho_{10}^2\right)\sigma_0^2. \quad (1)$$

In its essence, ρ_{10} determines the relative contribution of the observed value of $Z_i(1)$ in predicting the counterfactual $Z_i(0)$ for individual i. The value of ρ_{10}, which is bounded between -1 and 1, cannot be empirically obtained. To further narrow the bounds, the analyst may utilize an additional baseline covariate C. Let $\rho_{tC} = corr(Z, C|T = t, \mathbf{X} = \mathbf{x})$ denote the partial correlation under treatment condition t. As shown in past research

(Olkin, 1981; Stanley & Wang, 1969; also see Yang et al., 2017), ρ_{10} is restricted by the inequalities:

$$\rho_{1C}\rho_{0C} - \sqrt{\left(1-\rho_{1C}^2\right)\left(1-\rho_{0C}^2\right)} \le \rho_{10} \le \rho_{1C}\rho_{0C} + \sqrt{\left(1-\rho_{1C}^2\right)\left(1-\rho_{0C}^2\right)}. \quad (2)$$

When J additional covariates are available, the analyst may compute J bounded sets, the intersection of which defines a conservative bounded set of values for ρ_{10}. Alternatively, viewing C as a vector of dimension J, one may explicitly derive the bounds on ρ_{10} (Olkin, 1981).

3.2 | Analytic steps

There are four major steps in conducting the sensitivity analysis. Step 3 can be carried out through either imputation or integration.

Step 1. Predict Z as a function of observed baseline covariates and obtain the residuals. For individuals with $Z = z$ and $\mathbf{X} = \mathbf{x}$, the residuals are $\hat{r}_0 = z - \widehat{\mu(0,\mathbf{x})}$ if $T = 0$ and $\hat{r}_1 = z - \widehat{\mu(1,\mathbf{x})}$ if $T = 1$; the respective sample variances are $\widehat{\sigma_0^2}$ and $\widehat{\sigma_1^2}$. The set of covariates that predict Z should contain but should not be limited to the confounders of the mediator–outcome relationship; and the prediction model for Z may take any flexible form. This is because, as we will show through the simulations, including strong predictors of Z may reduce the range of the bounds in a sensitivity analysis.

Step 2. Obtain the conditional distribution of $Z(0)$. The analyst may choose a set of evenly spaced hypothetical values of ρ_{10} within its bounds including the maximum and the minimum. Corresponding to each hypothetical value of ρ_{10}, the parameters of the conditional distribution of $Z_i(0)$ are estimated for individuals assigned to $T = 1$ when $\mu(0,\mathbf{x})$, $\mu(1,\mathbf{x})$, σ_0, and σ_1 in Equation (1) are replaced by their sample analogues.

Step 3. Use imputed values of $Z(0)$ in an RMPW analysis. When implementing the lemma through multiple imputation (Little & Rubin, 2019), the analyst may take K random draws from the conditional distribution of $Z_i(0)$ for individual i whose $T_i = 1, Z_i(1) = z$, $\mathbf{X}_i = \mathbf{x}$, and $M_i(1) = m$. Let z'_{ik} be the kth random draw. The imputed values are viewed as given in each set of imputed data. To estimate the weight $\widehat{W_{ik}} = \frac{\hat{P}(M_i(0)=m|T_i=0,Z_i(0)=z'_{ik},\mathbf{X}_i=\mathbf{x})}{\hat{P}(M_i(1)=m|T_i=1,Z_i(1)=z,\mathbf{X}_i=\mathbf{x})}$, a separate propensity score model for the mediator is fitted to the data in each of the two treatment groups. The denominator of the weight is individual i's estimated propensity of displaying the observed mediator value m under the experimental condition; the numerator is the same individual's propensity of displaying the same mediator value under the counterfactual control condition as a function of $\mathbf{x}$ and the imputed value z'_{ik}. The model fitted to the control group data is used for predicting the numerator. The sample estimator of NIE and that of NDE are averaged over the K random draws of $Z(0)$, which will generate $\widehat{NIE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\sum_i T_i Y_i}{\sum_i T_i} - \frac{\sum_i T_i \widehat{W_{ik}} Y_i}{\sum_i T_i \widehat{W_{ik}}}\right)$ and $\widehat{NDE} = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{\sum_i T_i \widehat{W_{ik}} Y_i}{\sum_i T_i \widehat{W_{ik}}} - \frac{\sum_i (1-T_i) Y_i}{\sum_i (1-T_i)}\right\}$ in a completely randomized study. In accordance with Rubin's rules (Little & Rubin, 2019), the standard error for $\widehat{NIE}$ and that for $\widehat{NDE}$ are each pooled over the K estimates.

Alternative Step 3. Integrate over the predicted distribution of $Z(0)$ in an RMPW analysis. Alternatively, we may implement the lemma through taking the integral with respect to z' over the conditional distribution of $Z_i(0)$ for individual i whose $T_i = 1, Z_i(1) = z$, $\mathbf{X}_i = \mathbf{x}$, and $M_i(1) = m$. We estimate $W_i(z')$ as a function of z', as specified in the lemma, and apply the normal density function to $h_{iz'} = f(Z_i(0) = z' \mid T_i = 1, Z_i(1) = z, \mathbf{X}_i = \mathbf{x})$ in the integration. Subsequently, the sample estimator of NIE and that of NDE can be obtained as $\widehat{NIE} = \frac{\sum_i T_i Y_i}{\sum_i T_i} - \frac{\sum_i T_i Y_i \int \widehat{W_i(z')}\widehat{h_{iz'}}dz'}{\sum_i T_i \int \widehat{W_i(z')}\widehat{h_{iz'}}dz'}$; $\widehat{NDE} = \frac{\sum_i T_i Y_i \int \widehat{W_i(z')}\widehat{h_{iz'}}dz'}{\sum_i T_i \int \widehat{W_i(z')}\widehat{h_{iz'}}dz'} - \frac{\sum_i (1-T_i) Y_i}{\sum_i (1-T_i)}$ in a completely randomized study. Extending the previous results for RMPW-based causal mediation analysis that take into consideration the estimation uncertainty in the propensity score-based weight (Bein et al., 2018), we derive the asymptotic standard error for $\widehat{NIE}$ and that for $\widehat{NDE}$ when the integration method is employed and obtain estimates of these standard errors accordingly. The derivations are outlined in Web Appendix 5.

Step 4. Sensitivity analysis. Steps 2 and 3 are repeated at each hypothetical value of ρ_{10} to obtain the bounds for NIE and NDE estimates. These are to be contrasted with the initial estimates of NIE and NDE, respectively. In addition, to assess whether the results of hypothesis testing are sensitive to the omission of posttreatment confounding, the

analyst may estimate the confidence bands for NIE and NDE across the range of ρ_{10}. These are to be contrasted with the respective confidence intervals for NIE and NDE obtained from the initial analysis.

4 | SIMULATIONS

To assess the feasibility of this novel sensitivity analysis procedure and to evaluate its performance, the simulation study addresses two sets of research questions. The first set of questions concerns the performance of the imputation-based RMPW approach and the integration-based RMPW approach when the distributional assumptions about the posttreatment confounder are valid; the second set of questions is about their performance when such assumptions are invalid. We use the oracle estimator of NIE and that of NDE as the benchmark because they are consistent estimators. These oracle estimators are obtained by using the true value of $Z_i(0)$ in the RMPW analysis. We also present the results for the naïve estimators of NIE and NDE when the RMPW analysis makes no adjustment for the posttreatment confounder.

In the imputation-based RMPW analysis, the number of imputations is set to be 25. In the integration-based RMPW analysis, we apply the method of Gauss–Hermite quadrature with 10 quadrature points to approximate the values of integrals when Z is continuous. The sample size is set to be $n = 200$ in contrast with $n = 2000$. Additionally, in simulation scenario 10, we consider a sample size similar to the application study with $n = 700$. The number of replications in each simulation scenario is 1000. We generate data under assumptions (1*), (2*), and (3*), where we allow for unmeasured confounding of the relationship between Z and Y. In addition, the treatment-by-mediator interaction is present in all simulation scenarios. Web Appendix 6 provides details of the data generation. The simulation results for all 12 different scenarios of data generation are summarized in Tables 1 and 2 for NIE and NDE, respectively.

4.1 | Performance when the distributional assumptions are valid

Research Question 1: When the model for Z is correctly specified, are the imputation-based RMPW estimator and the integration-based RMPW estimator consistent with their benchmark values when ρ_{10} is equal to its true value? There is clear evidence that, for both a continuous Z and a binary Z, the results are consistent when ρ_{10} is equal to its true value. This conclusion holds across a wide range of scenarios. Specifically, simulation scenarios 1–6 evaluate the performance of the proposed estimators for a continuous Z; and scenario 11 for a binary Z. For a continuous Z, scenarios 1–3 consider a true value of ρ_{10} being 0.5, 0, and −0.5, respectively, when the sample size is 2000; in parallel, scenarios 4–6 consider the same set of true values of ρ_{10} when the sample size is reduced to 200.

Research Question 2: For a correctly specified model for Z, does the result deviate from the benchmark value when the hypothetical value of ρ_{10} deviates from its true value? As anticipated, the deviation from the benchmark value increases when the hypothetical value of ρ_{10} is farther away from the true value. This is true for both a continuous Z and a binary Z. Figure 2 provides a graphical illustration based on scenario 1 where Z is continuous, $\rho_{10} = 0.5$, and $n = 2000$. For both the NIE and the NDE, the imputation-based estimate and the integration-based estimate increasingly deviate from the benchmark values as the hypothetical value of ρ_{10} shifts farther away from its true value. The patterns are similar for different true values of ρ_{10} (scenarios 2 and 3), for a binary Z (scenario 11), and for different sample sizes (scenarios 4–6).

Research Question 3: Does the integration method outperform the imputation method in terms of the efficiency of estimation? Do the confidence bands accurately reflect the estimation error when the imputation method or the integration method is employed? When comparing the true standard error (i.e., the standard deviation of the NIE estimator or the NDE estimator over the 1000 replications) between the imputation-based strategy and the integration-based strategy, we find no evidence that one is more efficient than the other. This is true for both a continuous Z and a binary Z. For both the imputation-based estimator and the integration-based estimator, there is a general tendency that the estimated standard error converges to the true standard error when the sample size increases. The estimated standard error for the imputation-based estimator tends to be slightly greater than the true standard error regardless of sample size. Comparing between the two different sample sizes (2000 vs. 200), it appears that the imputation-based estimator is preferred when the sample size is relatively small as it tends to err on the conservative side; when the sample size is relatively large, the integration-based estimator is preferred as it tends to closely approximate the true standard error.

Research Question 4: When the prediction model for Z does not match the data generation model, is the result still robust at its true value of ρ_{10}? Is there an increase in the width of the bounds for the estimated values of NIE and NDE? Is there an increase in the width of the confidence bands for NIE and NDE? A comparison between scenarios 7a and 7b reveals the implications. The prediction model for Z specified by the analyst is different from the data generation model in scenario 7a and is the same as the data generation model in scenario 7b. The

TABLE 1 Simulation results for natural indirect effects

					Imputation-based RMPW estimator				Integration-based RMPW estimator			
Scenario	ρ_{10}	**True** NIE	$\widehat{NIE}_{w/oZ}$	$\widehat{NIE}_{ora}$	$\widehat{NIE}_{imp}$ **bounds**	$\widehat{NIE}_{imp}$ **at** ρ_{10}	**SE of** $\widehat{NIE}_{imp}$ **at** ρ_{10}	**SD of** $\widehat{NIE}_{imp}$ **at** ρ_{10}	$\widehat{NIE}_{int}$ **bounds**	$\widehat{NIE}_{int}$ **at** ρ_{10}	**SE of** $\widehat{NIE}_{int}$ **at** ρ_{10}	**SD of** $\widehat{NIE}_{int}$ **at** ρ_{10}
1	0.5	0.352	0.310	0.351	[0.314, 0.465]	0.352	0.051	0.048	[0.314, 0.465]	0.352	0.047	0.048
2	0.0	0.387	0.306	0.385	[0.311, 0.458]	0.384	0.056	0.051	[0.311, 0.458]	0.385	0.049	0.050
3	−0.5	0.425	0.311	0.425	[0.316, 0.463]	0.426	0.059	0.057	[0.316, 0.463]	0.426	0.054	0.056
4	0.5	0.352	0.208	0.348	[0.313, 0.449]	0.348	0.167	0.161	[0.313, 0.449]	0.348	0.152	0.161
5	0.0	0.387	0.308	0.383	[0.316, 0.452]	0.384	0.178	0.165	[0.316, 0.452]	0.385	0.160	0.165
6	−0.5	0.425	0.311	0.418	[0.316, 0.452]	0.417	0.186	0.186	[0.316, 0.452]	0.419	0.172	0.188
7a	0.8	0.333	0.296	0.336	[0.320, 0.489]	0.337	0.049	0.049	[0.320, 0.489]	0.337	0.047	0.049
7b	0.5	0.333	0.293	0.333	[0.318, 0.378]	0.333	0.049	0.046	[0.318, 0.378]	0.333	0.047	0.046
8	0.4	0.090	0.032	0.087	[0.071, 0.141]	0.090	0.058	0.056	[0.071, 0.141]	0.090	0.055	0.056
9	0.2	0.375	0.364	0.374	[0.345, 0.374]	0.354	0.047	0.036	[0.345, 0.374]	0.354	0.046	0.047
10	0.0	0.260	0.248	0.258	[0.248, 0.263]	0.256	0.088	0.082	[0.248, 0.263]	0.256	0.082	0.082
11	0.5	−0.423	−0.352	−0.427	[−0.366, −0.450]	−0.427	0.048	0.047	[−0.366, −0.450]	−0.427	0.046	0.047
12	0.0	−0.365	−0.272	−0.360	[−0.416, −0.305]	−0.361	0.053	0.048	[−0.416, −0.305]	−0.361	0.048	0.047

Note. Each scenario was simulated 1000 times. See the Supporting Information for details of data generation. This table lists the underlying true value of ρ_{10}. The true NIE is calculated numerically by generating a dataset with a sample size of 5,000,000. $\widehat{NIE}_{w/oZ}$ is the mean of NIE estimates across 1000 simulations without adjusting for Z in the analysis; and $\widehat{NIE}_{ora}$ is the mean of NIE estimates using the oracle estimator. For the imputation-based RMPW estimator (with 25 imputations), $\widehat{NIE}_{imp}$ bounds provide the means of the lower and upper bounds of NIE estimates by varying ρ_{10} in the range of $[-1, 1]$. At the true value of ρ_{10}, $\widehat{NIE}_{imp}$ is the mean of NIE estimates; SE of $\widehat{NIE}_{imp}$ at ρ_{10} is the mean of standard errors of NIE estimates; and SD of $\widehat{NIE}_{imp}$ at ρ_{10} is the standard deviation of NIE estimates across 1000 simulations. $\widehat{NIE}_{int}$ *bounds*, $\widehat{NIE}_{int}$ at ρ_{10}, SE of $\widehat{NIE}_{int}$ at ρ_{10}, and SD of $\widehat{NIE}_{int}$ at ρ_{10} have analogous interpretations for the integration-based RMPW estimator.

TABLE 2 Simulation results for natural direct effects

					Imputation-based RMPW estimator				Integration-based RMPW estimator			
Scenario	ρ_{10}	**True** NDE	$\widehat{NDE}_{w/oZ}$	$\widehat{NDE}_{ora}$	$\widehat{NDE}_{imp}$ **bounds**	$\widehat{NDE}_{imp}$ **at** ρ_{10}	**SE of** $\widehat{NDE}_{imp}$ **at** ρ_{10}	**SD of** $\widehat{NDE}_{imp}$ **at** ρ_{10}	$\widehat{NDE}_{int}$ **bounds**	$\widehat{NDE}_{int}$ **at** ρ_{10}	**SE of** $\widehat{NDE}_{int}$ **at** ρ_{10}	**SD of** $\widehat{NDE}_{int}$ **at** ρ_{10}
1	0.5	0.962	1.002	0.962	[0.848, 0.998]	0.961	0.039	0.030	[0.848, 0.998]	0.961	0.032	0.030
2	0.0	0.926	1.001	0.928	[0.855, 1.002]	0.928	0.044	0.033	[0.855, 1.002]	0.928	0.036	0.032
3	−0.5	0.888	1.004	0.890	[0.852, 0.999]	0.889	0.048	0.039	[0.852, 0.999]	0.889	0.042	0.039
4	0.5	0.962	1.004	0.964	[0.862, 1.000]	0.965	0.127	0.104	[0.862, 1.000]	0.964	0.107	0.104
5	0.0	0.926	1.009	0.935	[0.866, 1.002]	0.933	0.140	0.119	[0.866, 1.002]	0.932	0.115	0.119
6	−0.5	0.888	1.004	0.897	[0.863, 0.999]	0.898	0.148	0.138	[0.865, 1.001]	0.896	0.130	0.139
7*a*	0.8	0.988	1.028	0.988	[0.835, 1.004]	0.987	0.035	0.031	[0.835, 1.004]	0.987	0.032	0.031
7*b*	0.5	0.988	1.028	0.988	[0.943, 1.003]	0.988	0.035	0.031	[0.943, 1.003]	0.988	0.032	0.031
8	0.4	0.576	0.632	0.576	[0.522, 0.593]	0.574	0.042	0.039	[0.522, 0.593]	0.574	0.040	0.039
9	0.2	0.912	0.924	0.914	[0.914, 0.943]	0.934	0.036	0.027	[0.914, 0.943]	0.934	0.029	0.027
10	0.0	0.678	0.692	0.682	[0.677, 0.691]	0.683	0.058	0.051	[0.677, 0.691]	0.683	0.049	0.051
11	0.5	0.506	0.431	0.506	[0.446, 0.530]	0.506	0.059	0.055	[0.446, 0.530]	0.506	0.057	0.055
12	0.0	0.398	0.301	0.390	[0.335, 0.446]	0.391	0.081	0.073	[0.335, 0.446]	0.391	0.078	0.073

Note. Each scenario was simulated 1000 times. See the Supporting Information for details of data generation. This table lists the underlying true value of ρ_{10}. The true NDE is calculated numerically by generating a dataset of sample size of 5,000,000. $\widehat{NDE}_{w/oZ}$ is the mean of NDE estimates across 1000 simulations without adjusting for Z in the analysis; and $\widehat{NDE}_{ora}$ is the mean of NDE estimates using the oracle estimator. For the imputation-based RMPW estimator (with 25 imputations), $\widehat{NDE}_{imp}$ bounds provide the means of the lower and upper bounds of NDE estimates by varying ρ_{10} in the range of $[-1, 1]$; at the true value of ρ_{10}, $\widehat{NDE}_{imp}$ is the mean of NDE estimates; SE of $\widehat{NDE}_{imp}$ at ρ_{10} is the mean of standard errors of NDE estimates; and SD of $\widehat{NDE}_{imp}$ at ρ_{10} is the standard deviation of NDE estimates across 1000 simulations. $\widehat{NDE}_{int}$ *bounds*, $\widehat{NDE}_{int}$ at ρ_{10}, SE of $\widehat{NDE}_{int}$ at ρ_{10}, and SD of $\widehat{NDE}_{int}$ at ρ_{10} have analogous interpretations for the proposed integration-based RMPW estimator.

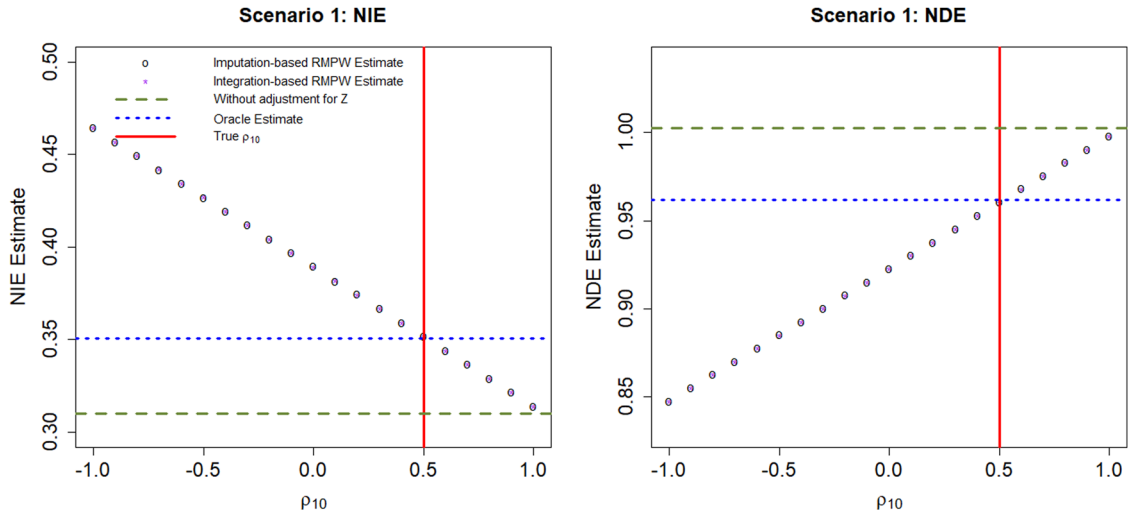

FIGURE 2 Simulation results for scenario 1 (this figure appears in color in the electronic version of this article, and any mention of color refers to that version)

former explains less variation in Z when compared with the latter. Figure 3 displays the estimation results. Despite the differences in how the structural part of the prediction model for Z is specified, the point estimates of NIE and those of NDE are always consistent at the true values of ρ_{10} corresponding to each model. However, the bounds for the NIE and NDE estimates become wider in scenario 7a. The impact on the width of the confidence bands appears to be negligible. We conclude that even though the structural part of the prediction model specified by the analyst tends to deviate from the data generation model in practice, the estimation results are robust. However, an increase in the predictive power of the model for Z will effectively reduce the width of the bounds for the NIE and NDE estimates.

4.2 | Performance when the distributional assumptions are violated

Research Question 5: Is there a consequence when homoscedasticity is violated in the prediction model for a continuous Z? In scenario 8, we specify σ_0, σ_1, and ρ_{10} each to be a function of a pretreatment covariate; the analysis, however, assumes homoscedasticity. As shown in Web Figure A1, violations of the homoscedasticity assumption do not appear to be consequential. At the value of the conditional correlation ρ_{10} for the overall population, the NIE and NDE estimates are consistent with their respective benchmark values.

Research Question 6: Is there a consequence when multivariate normality is violated in the prediction model for Z? We find that when violations of the multivariate normality assumption are not severe, such as in the case of a continuous Z taking non-negative values with a small degree of zero inflation (scenario 10) or in the case of a binary Z generated with logistic random errors (scenario 12), the NIE and NDE estimates deviate only slightly from their respective benchmark values at the true value of ρ_{10}. However, when violations are severe, as in the case of a continuous Z that follows a gamma distribution rather than a normal distribution (scenario 9), the estimation results are no longer robust. Web Figure A2 provides a graphical illustration of the above results.

Conclusion. For a continuous or a binary posttreatment confounder Z, when the distributional assumptions are valid, the imputation-based strategy and the integration-based strategy both produce NIE and NDE estimates that are consistent with their respective benchmark values at the true value of ρ_{10}. This is true even when the structural part of the prediction model for Z deviates from its data generation model. The results also remain robust when homoscedasticity is violated or when multivariate normality is violated to a minor degree. However, these strategies do not produce robust results when multivariate normality is severely violated. The simulation results provide important implications for data analysis. As the true value of ρ_{10} is unknown to the analyst, a sensitivity analysis must compare the initial estimates of NIE and NDE with the bounds for the new estimates obtained over the plausible range of ρ_{10}. By including strong predictors of Z that help to increase the explained proportion of variation, the analyst may obtain relatively narrow bounds for the NIE and NDE estimates and thereby increasing the chance of arriving at a definitive conclusion.

5 | NEWWS APPLICATION

The initial analysis of the NEWWS-Riverside data did not adjust for potential posttreatment confounders. We assess

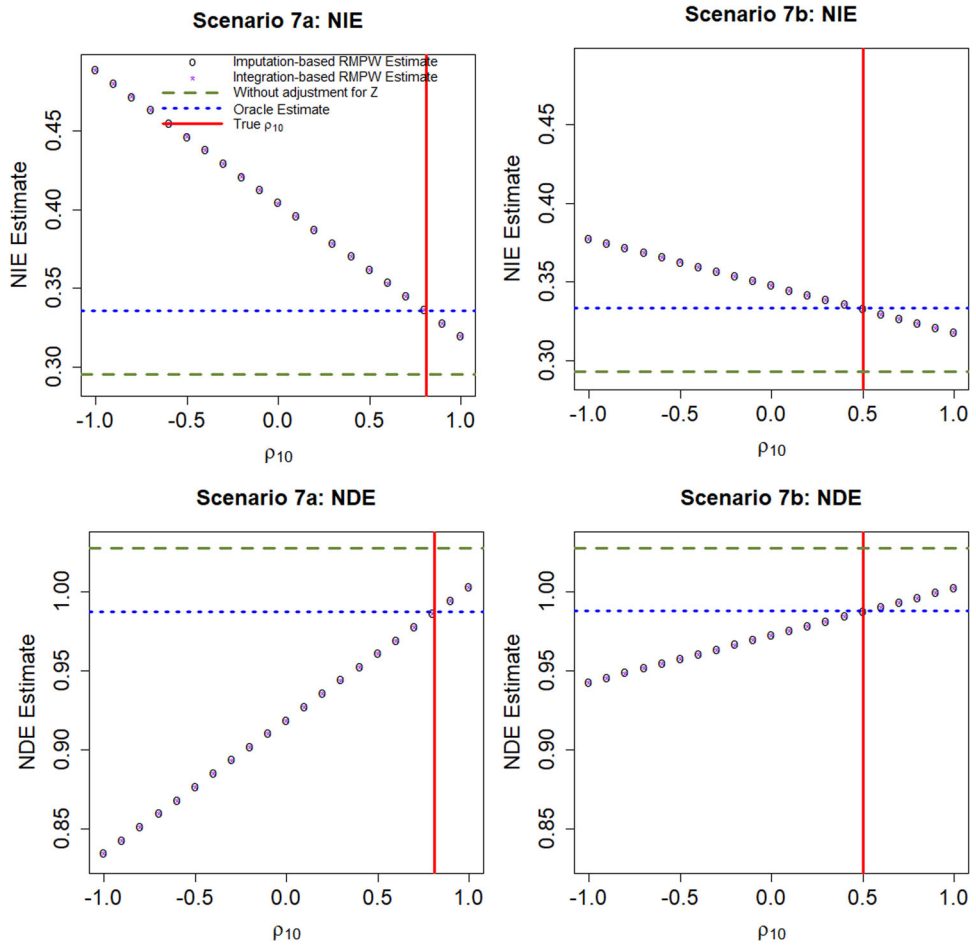

FIGURE 3 Simulation results comparing scenarios 7a and 7b (this figure appears in color in the electronic version of this article, and any mention of color refers to that version)

the sensitivity of the initial results to such omissions by applying the proposed strategies. We first evaluate the influence of a continuous posttreatment confounder—the amount of welfare received in the first year after the randomization. The prediction model for this continuous Z is specified as a function of not only the pretreatment covariates that predicted the mediator but also two additional strong predictors of Z (namely, welfare amount received in the pretreatment year and number of children in the household at baseline). To constrain the range of ρ_{10}, we further utilize a measure of duration of welfare dependence prior to the randomization. Applying Equation (2), the range of ρ_{10} becomes $[-0.86, 0.98]$, within which we choose 20 evenly spaced values. To enhance the smoothness of the results, we conduct 200 imputations when implementing the imputation-based procedure. Figures 4(A) and 4(B) display the estimation results obtained from the imputation-based analysis and the integration-based analysis, respectively. In each figure, the horizontal dashed line in the middle indicates the initial estimate of the NIE or the NDE in effect size; the upper and lower dashed lines correspond to the initial 95% confidence interval of the effect size; each circle represents the adjusted estimate at a given value of ρ_{10}; and the corresponding vertical line represents the adjusted estimate of the 95% confidence interval. Over the range of the plausible values of ρ_{10}, we find that with adjustment for posttreatment welfare amount, the estimated effect size of NIE is bounded between -0.15 and -0.12 and that of NDE is bounded between 0.14 and 0.16. These are distinctly different from the initial estimate of the effect size of NIE (-0.11) and that of NDE (0.13). Apparently, omitting posttreatment welfare amount led to a positive bias in the initial NIE estimate and a negative bias in the initial NDE estimate.

We then evaluate the influence of a binary posttreatment confounder indicating whether an individual was on welfare constantly during the first year after the randomization. From both the imputation-based analysis and the integration-based analysis with adjustment for this binary Z, the bounds for the NIE estimates are $(-0.15, -0.12)$ and those for the NDE estimates are (0.13, 0.16). Unsurprisingly, as this binary Z (indicator for constant

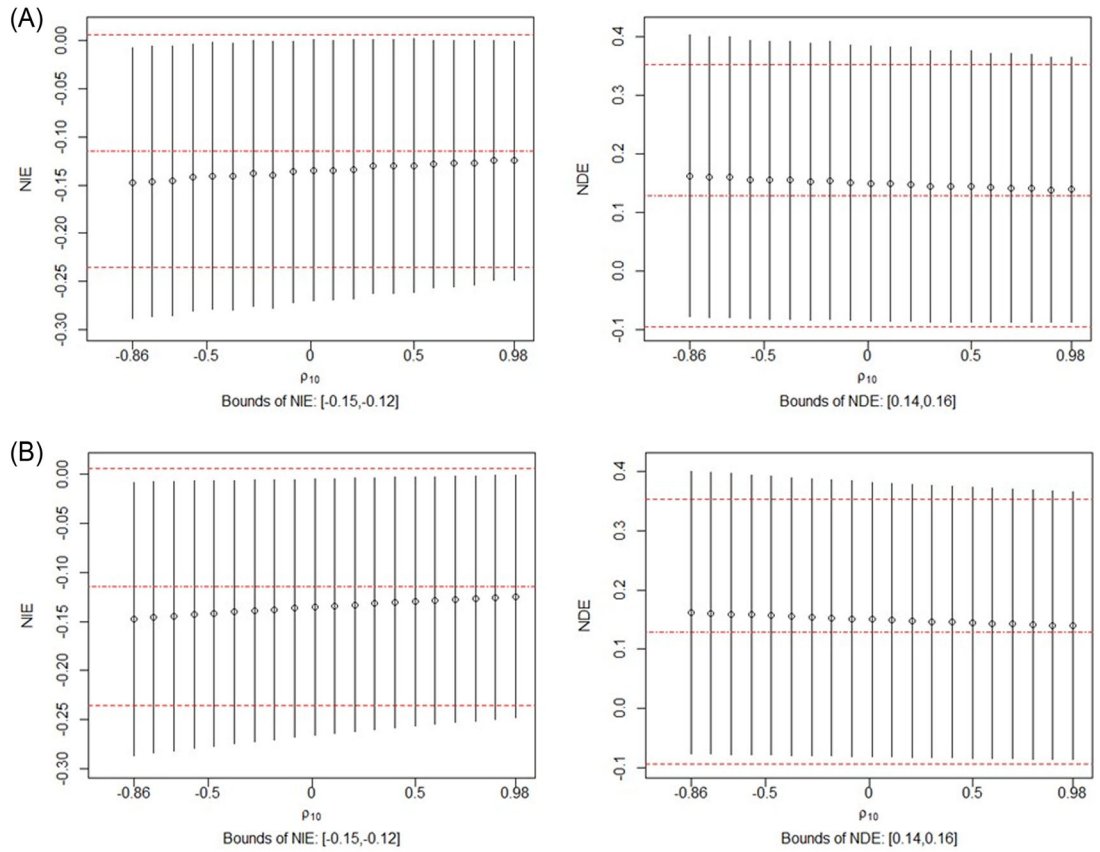

FIGURE 4 (A) Sensitivity of the initial estimate of the effect size of NIE and that of NDE to the omission of posttreatment welfare amount (imputation-based). (B) Sensitivity of the initial estimate of the effect size of NIE and that of NDE to the omission of posttreatment welfare amount (integration-based). (this figure appears in color in the electronic version of this article, and any mention of color refers to that version)

welfare dependence) is closely associated with the continuous Z (welfare amount), the sensitivity analysis arrives at a very similar conclusion.

6 | DISCUSSION

Adjusting for posttreatment confounders that precede the focal mediator has been a major challenge in causal mediation analysis. This paper proposes a new sensitivity analysis for assessing the consequences of omitting observed posttreatment confounders. The key is to obtain predicted values of the posttreatment covariate under the counterfactual control condition given an individual's observed posttreatment covariate value under the experimental condition as well as the observed pretreatment covariates. The analyst will then adjust for the posttreatment covariate through an RMPW analysis and obtain bounds for the NIE and NDE estimates. The analysis can be implemented through either imputation or integration over the conditional distribution of the posttreatment covariate. Unlike the linear or generalized linear structural modeling approach that requires correct specifications of not only the mediator model but also the outcome model, the weighting approach does not require the analyst to specify the response surface and therefore prevents bias induced by outcome model misspecification.

This paper makes a few important contributions to the literature on handling posttreatment confounding in causal mediation analysis. First, our new solution invokes a set of identification assumptions that are different from the standard and the extended sequential ignorability assumptions, which distinguishes our approach from the alternative methods. Conditioning on not only the observed pretreatment covariates but also a potentially important posttreatment confounder increases the plausibility of mediator ignorability. Moreover, instead of assuming the ignorability of the posttreatment covariate given the observed pretreatment covariates, we assume cross-world independence between the observed mediator and the counterfactual posttreatment covariate conditioning on the observed posttreatment covariate as well as the pretreatment covariates, which is arguably plausible in many cases including the current application.

Second, this new solution applies to both continuous and binary posttreatment covariates. Third, this approach allows for a treatment-by-mediator interaction and can be extended naturally to the alternative decomposition of the ATE into a pure indirect effect and a total direct effect (Robins & Greenland, 1992), which involves identifying $E\{Y(0,M(1))\}$. Specifically, under assumptions (1*), (2*), and (3*), $E\{Y(0,M(1))\}$ can be identified by $E\{\int W_0(\mathbf{x})W(z')Yh_{z'}dz'|T=0\}$ where $W_0(\mathbf{x}) = \frac{P(T=0)}{P(T=0|\mathbf{X}=\mathbf{x})}$, $W(z') = \frac{P(M(1)=m|T=1,Z(1)=z',\mathbf{X}=\mathbf{x})}{P(M(0)=m|T=0,Z(0)=z,\mathbf{X}=\mathbf{x})}$, and $h_{z'} = P(Z(1)=z'\,|\,T=0, Z(0)=z, \mathbf{X}=\mathbf{x})$ for individuals whose $T=0$, $Z(0)=z$, and $\mathbf{X}=\mathbf{x}$. Fourth, as we have shown through simulations, the estimation results are robust even when the structural part of the prediction model for the posttreatment confounder deviates from its data generation model, when homoscedasticity is violated, or when multivariate normality is violated to a minor degree. Fifth, by obtaining the bounds for the adjusted NIE and NDE estimates and additionally by obtaining the confidence bands for these causal effects, we can assess whether adjustment for the posttreatment confounder would alter the initial conclusion in terms of not only the practical significance but also the statistical significance. And lastly, we underscore the practical value of including strong predictors of the posttreatment covariate in the prediction models as it will effectively make the bounds more informative for sensitivity analysis.

Several limitations of our approach are to be addressed in future research. First, while assumption (1*) was guaranteed by the treatment randomization in the current application, violations of assumptions (2*) and (3*), if exist, could potentially invalidate the sensitivity analysis results. Second, the estimation results are not robust when multivariate normality is severely violated. In some cases, the analyst may overcome this limitation through appropriately transforming the posttreatment covariate. For example, if Z follows a lognormal distribution, a log transformation will be a suitable solution. Future research may explore other strategies such as deriving the conditional distribution of Z that follows a nonnormal distribution or applying copula to the residuals obtained from the prediction models for Z. Third, the proposed weighting approach is semiparametric and avoids bias induced by outcome model misspecification, however, possibly at the price of a decrease in efficiency when compared to the competing linear or generalized linear structural modeling approaches. Combining weighting with covariance or prognostic score adjustment may improve precision in estimation. In addition, estimation instability might arise due to extreme values of weights. Besides a common remedy to trim the weights, machine learning techniques may potentially improve the estimation of the weight (Lee et al., 2010). Fourth, because the minimum/maximum values of the bounds for NIE and NDE cannot be explicitly derived, they are empirically established when the grids for ρ_{10} are fine-grained and when the patterns are smooth. The peak values might be overlooked in the absence of these conditions. Finally, some posttreatment confounders precede the focal mediator M, whereas some others are concurrent to M. This paper is restricted to the former case. Qin et al. (2021) proposed a sensitivity analysis strategy for causal mediation studies that involve two concurrent mediators that are not conditionally independent. Readers may also consider alternative strategies proposed by Imai and Yamamoto (2013), Daniel et al. (2015), and Albert et al. (2019).

ACKNOWLEDGMENTS

The first two authors have made equal contributions to this paper. The research reported here was supported by the National Science Foundation (SES 1659935) and by the Institute of Education Sciences (IES), U.S. Department of Education, through a Statistical and Research Methodology Grant (R305D200031). The opinions expressed are those of the authors and do not represent views of the National Science Foundation or the Institute of Education Sciences, U.S. Department of Education. The authors would like to thank Li Cai, Donna Coffman, Trang Nguyen, Stephen Raudenbush, Geoff Wodtke, and participants at the University of California-Los Angeles Social Research Methodology - Human Development & Psychology Joint Brown Bag Speaker Series and at the University of Pittsburgh Department of Biostatistics Colloquium for helpful comments. The authors also thank the Associate Editor and three anonymous reviewers for their valuable suggestions.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available at the National Center for Health Statistics (NCHS) Research Data Center (RDC) and can be downloaded from https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//169986/full2r.zip. The R code and the data file are included in the online Supporting Information.

ORCID

Guanglei Hong https://orcid.org/0000-0002-8254-4655
Fan Yang https://orcid.org/0000-0003-3671-4745
Xu Qin https://orcid.org/0000-0002-5907-1511

REFERENCES

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

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

Avin, C., Shpitser, I. & Pearl, J. (2005) Identifiability of path-specific effects. *Proceedings of the International Joint Conference on Artificial Intelligence*, 19, 357–363 .

Bein, E., Deutsch, J., Hong, G., Porter, K., Qin, X. & Yang, C. (2018) Two-step estimation in rmpw analysis. *Statistics in Medicine*, 37, 1304–1324.

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

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

Hong, G. (2010) Ratio of mediator probability weighting for estimating natural direct and indirect effects. In: *Proceedings of the American Statistical Association, Biometrics Section* (pp. 2401–2415). Alexandria, VA: American Statistical Association.

Hong, G. (2015). *Causality in a social world: moderation, mediation and spill-over*. Chichester: John Wiley.

Hong, G., Deutsch, J. & Hill, H.D. (2015) Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics*, 40, 307–340.

Hong, G. & Nomi, T. (2012) Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness*, 5, 261–289.

Hong, G., Qin, X. & Yang, F. (2018) Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43, 32–56.

Huber, M. (2014) Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29, 920–943.

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

Imai, K. & Yamamoto, T. (2013) Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis*, 21, 141–171.

Lange, T., Vansteelandt, S. & Bekaert, M. (2012) A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, 176, 190–195.

Lee, B.K., Lessler, J. & Stuart, E.A. (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29, 337–346.

Little, R.J. & Rubin, D.B. (2019) *Statistical analysis with missing data* (Vol. 793). Hoboken, NJ: John Wiley & Sons.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. & 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, 1443–1452.

Olkin, I. (1981) Range restrictions for product-moment correlation matrices. *Psychometrika*, 46, 469–472.

Pearl, J. (2001) Direct and indirect effects. In: Breese, J. & Koller, D. (Eds.) *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

Qin, X., Deutsch, J. & Hong, G. (2021) Unpacking complex mediation mechanisms and their heterogeneity between sites in a job corps evaluation. *The Journal of Policy Analysis and Management*, 40, 158–190.

Radloff, L. S. (1977) The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385–401.

Robins, J. M. (2003) Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P.J., Hjort, N.L. & Richardson, S. (Eds.) *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.

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

Rosenbaum, P.R. (1984) The consequence of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A (General)*, 147, 656–666.

Rudolph, K.E., Sofrygin, O., Schmidt, N.M., Crowder, R., Glymour, M.M., Ahern, J. & Osypuk, T.L. (2018) Mediation of neighborhood effects on adolescent substance use by the school and peer environments. *Epidemiology*, 29, 590–598.

Stanley, J.C. & Wang, M.D. (1969) Restrictions on the possible values of r12, given r13 and r23. *Educational and Psychological Measurement*, 29, 579–581.

Tchetgen Tchetgen, E.J. & Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40, 1816–1845.

U.S. *Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation*. (1997) National Evaluation of Welfare-to-Work Strategies. https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//169986/full2r.zip

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

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

Wodtke, G.T. & Zhou, X. (2020) Effect decomposition in the presence of treatment-induced confounding: a regression-with-residuals approach. *Epidemiology*, 31, 369–375.

Yang, H., Wong, W.H., Bradley, K.D. & Toland, M.D. (2017) Partial and semi-partial correlations for categorical variables in educational research: addressing two common misconceptions. *General Linear Model Journal*, 43, 1–15.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 4 as well as data/code are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Hong, G., Yang, F., & Qin, X. (2023) Posttreatment confounding in causal mediation studies: A cutting-edge problem and a novel solution via sensitivity analysis. *Biometrics*, 79, 1042–1056. https://doi.org/10.1111/biom.13705