Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-Residuals Approach*

Geoffrey T. Wodtke
University of Toronto

Xiang Zhou
Harvard University

May 15, 2019

Abstract

Analyses of causal mediation are often complicated by treatment-induced confounders of the mediator-outcome relationship. In the presence of such confounders, the natural direct and indirect effects of treatment on the outcome, into which the total effect can be additively decomposed, are not identified. An alternative but similar set of effects, known as randomized intervention analogues to the natural direct effect (R-NDE) and the natural indirect effect (R-NIE), can still be identified in this situation, but existing estimators for these effects require a complicated weighting procedure that is difficult to use in practice. In this paper, we introduce a new method for estimating the R-NDE and R-NIE that involves only a minor adaption of the comparatively simple regression methods used to perform effect decomposition in the absence of treatment-induced confounding. It involves fitting linear models for (a) the conditional mean of the mediator given treatment and a set of baseline confounders and (b) the conditional mean of the outcome given the treatment, mediator, baseline confounders, and the treatment-induced confounders after first residualizing them with respect to the observed past. The R-NDE and R-NIE are simple functions of the parameters in these models when they are correctly specified and when there are no unobserved variables that confound the treatment-outcome, treatment-mediator, or mediator-outcome relationships. We illustrate the method by decomposing the effect of education on depression symptoms at midlife into components operating through income versus alternative factors. R and Stata packages are available for implementing the proposed method.

*Direct all correspondence to Geoffrey T. Wodtke, Department of Sociology, University of Toronto, 725 Spadina Ave, Toronto, ON M6E 3K8, Canada; email: geoffrey.wodtke@utoronto.ca.

1 Introduction

Researchers have become increasingly interested in uncovering the mediating pathways through which one variable affects another ([Hafeman and Schwartz 2009](#)). A common approach to assessing causal mediation involves decomposing a total effect of a treatment on an outcome into an indirect component operating through a mediator of interest and a direct component operating through alternative pathways. This is typically accomplished via an additive decomposition in which the total effect is separated into so-called natural direct and indirect effects ([Pearl 2001](#), [VanderWeele 2015](#), [VanderWeele and Vansteelandt 2009](#)).

The natural direct effect (NDE) is the expected difference in an outcome of interest if each individual were exposed, rather than unexposed, to treatment and then were subsequently exposed to the level of the mediator they would have experienced had they not received treatment. It measures the effect of treatment on the outcome operating through all pathways other than the mediator by comparing outcomes under different levels of treatment after fixing the mediator to the level it would have “naturally” been for each individual under the reference level of treatment. This deactivates causal pathways operating through the mediator while leaving all other pathways intact.

The natural indirect effect (NIE), by contrast, is the expected difference in the outcome if each individual were exposed to treatment and then were subsequently exposed to the level of the mediator they experience as a result of being treated rather than the level of mediator they would have experienced had they not been treated. It measures the effect of treatment operating specifically through the mediator by fixing the level of treatment for each individual and then comparing outcomes under the different levels of the mediator that individuals would have “naturally” experienced if they had previously been exposed, rather than unexposed, to treatment. This deactivates all causal pathways except for that operating through the mediator.

Although the NDE and NIE neatly separate the effects of treatment operating through the mediator versus alternative pathways, they can only be non-parametrically identified under a set of highly restrictive assumptions. In particular, the NDE and NIE can only

be identified if there is (i) no unobserved treatment-outcome confounding, (ii) no unobserved treatment-mediator confounding, (iii) no unobserved mediator-outcome confounding, and (iv) no treatment-induced mediator-outcome confounding (VanderWeele 2015). This last assumption is especially restrictive because it requires that there must not be *any* variables that affect both the mediator and outcome and that are affected by treatment, whether they are observed or not. It is therefore unreasonable in many analyses of causal mediation, where treatment-induced confounding is ubiquitous.

To circumvent this challenge, VanderWeele and colleagues (VanderWeele et al. 2014, VanderWeele 2015) proposed an alternative set of estimands known as randomized intervention analogues to the natural direct effect (R-NDE) and the natural indirect effect (R-NIE), which can be identified in the presence of treatment-induced confounding (see also Geneletti 2007). The R-NDE and R-NIE are similar to the NDE and NIE except that, instead of setting the mediator to the level it would have naturally been for each individual under a particular treatment status, these estimands involve setting the mediator to a value randomly drawn from its population distribution under a given treatment status. Identifying these versions of direct and indirect effects, which sum to a version of the total effect, requires less restrictive assumptions that may be easier to satisfy in practice. Specifically, identifying these effects requires assumptions (i) to (iii) above but not assumption (iv).

Estimating the R-NDE and R-NIE, however, remains difficult. VanderWeele and colleagues (VanderWeele et al. 2014, VanderWeele 2015) outlined an estimator based on inverse probability weighting (IPW) that requires correct models for the probability of treatment given a set of baseline confounders, the joint probability of the treatment-induced confounders given treatment and the baseline confounders, as well as the probability of the mediator given treatment, the baseline confounders, and the treatment-induced confounders. Because IPW estimators are relatively inefficient, highly sensitive to model misspecification, and difficult to use with continuous variables (Howe et al. 2011, Naimi et al. 2014, Wodtke 2018), this approach may be challenging to implement with confidence outside of stylized applications. It is also cumbersome to implement with standard software, and it lacks the intuitive appeal of regression-based estimators commonly used to

analyze causal mediation in the absence of treatment-induced confounding (e.g., [VanderWeele and Vansteelandt 2009](#)).

In this article, we introduce a new method, termed “regression-with-residuals” (RWR), for estimating the R-NDE and R-NIE. It involves only a minor adaption of the familiar regression-based approaches to effect decomposition that are widely used when treatment-induced confounding can be safely assumed away. Briefly, the method involves fitting models for (a) the conditional mean of the mediator given treatment and a set of baseline confounders and (b) the conditional mean of the outcome given the treatment, mediator, baseline confounders, and the treatment-induced confounders after first residualizing them with respect to the observed past. By appropriately residualizing these confounders, which purges them of their association with treatment, RWR avoids the bias that typically results from conditioning on post-treatment variables ([Elwert and Winship 2014](#), [Greenland 2003](#), [Pearl 2009](#)). The models described above can be easily fit using standard software, and estimates of the R-NDE and R-NIE are given by simple functions of their coefficients.

In the sections that follow, we begin by formally defining the R-NDE and R-NIE and outlining the conditions under which they can be identified. Then, we introduce the method of RWR and show that it can be used to estimate these effects in the presence of treatment-induced confounders. Finally, with data from the 1979 National Longitudinal Survey of Youth (NLSY79), we illustrate the proposed method by decomposing the effect of college completion on depression symptoms at midlife into components operating through family income versus alternative pathways.

2 Notation, Estimands, and Identification

We adopt the notation used by [VanderWeele et al. \(2014\)](#). Let Y denote the outcome of interest, A the treatment, M a putative mediator, C a set of baseline confounders, and L a set of confounders for the mediator-outcome relationship that may be affected by treatment. In addition, let Y_a and M_a denote the values of the outcome and mediator, respectively, that would have been observed had an individual previously been exposed

to treatment a , possibly contrary to fact. Similarly, let Y_{am} denote the value of the outcome had an individual been exposed to the levels of treatment and the mediator given by a and m . Finally, let $G_{a|C}$ denote a value of the mediator randomly selected from the population distribution under exposure to treatment a conditional on the baseline confounders C .

With this notation, the randomized intervention analogue of the natural direct effect can be defined as follows:

$$\text{R-NDE} = \mathbb{E}(Y_{a^*G_{a|C}} - Y_{aG_{a|C}}). \quad (1)$$

This estimand represents the expected difference in the outcome if all individuals in some target population were exposed to treatment status a^* rather than a and if they were subsequently exposed to a level of the mediator randomly selected from the distribution under treatment status a among those with baseline confounders C . It captures the effect of treatment on the outcome that is not due to mediation via M . This is achieved by comparing outcomes under different levels of treatment with the mediator randomly selected from the distribution under the reference level of treatment.

Similarly, the randomized intervention analogue of the natural indirect effect can be defined as follows:

$$\text{R-NIE} = \mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{a^*G_{a|C}}). \quad (2)$$

This estimand represents the expected difference in the outcome if all individuals were exposed to treatment status a^* and then were subsequently exposed to a level of the mediator randomly selected from the distribution under treatment status a^* rather than a . It captures the effect of treatment on the outcome due to mediation via M , that is, the effect operating through a causal chain from treatment to the mediator to the outcome. This is achieved by fixing treatment at a reference level (a^*) and then comparing outcomes with the mediator randomly selected from the population distribution under different levels of treatment.

The sum of the R-NDE and R-NIE is equal to the randomized intervention analogue of the total effect:

$$\text{R-ATE} = \text{R-NDE} + \text{R-NIE} = \mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{aG_{a|C}}). \quad (3)$$

This estimand is similar to an average total effect except that it is defined in terms of both a contrast between different levels of treatment and a randomized intervention on the mediator. It gives the expected difference in the outcome if all individuals were exposed to treatment a^* rather than a with the mediator randomly selected from the distribution under each of these alternative treatments.

The R-NDE and R-NIE can be non-parametrically identified from observed data under the following conditional independence assumptions: (i) $Y_{am} \perp\!\!\!\perp A|C$, (ii) $M_a \perp\!\!\!\perp A|C$, and (iii) $Y_{am} \perp\!\!\!\perp M|C, A, L$ (VanderWeele et al. 2014). Several additional conditions referred to as the consistency and stable unit treatment value assumptions are also needed to identify these effects. The consistency assumption here requires that $Y = Y_{am}$ and $M = M_a$ when $A = a$ and $M = m$. The assumption of stable unit treatment values requires that there must not be any interference between individuals in the target population or multiple versions of treatment. Detailed discussions of these assumptions are available elsewhere (VanderWeele 2009).

In words, assumption (i) requires that there must not be any unobserved treatment-outcome confounders conditional on C . Assumption (ii) requires that there must not be any unobserved treatment-mediator confounders conditional on C . And assumption (iii) requires that there must not be any unobserved mediator-outcome confounders conditional on C , A , and L . Figure 1 presents a directed acyclic graph in which all of these assumptions are satisfied, as there are not any unobserved variables that jointly affect treatment, the mediator, or the outcome (Pearl 2009). In this situation, the R-NDE and R-NIE can be expressed in terms of the observed data as follows:

$$\text{R-NDE} = \sum_c \sum_m \sum_l [\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*) - \mathbb{E}(Y|c, a, l, m)P(l|c, a)]P(m|c, a)P(c) \quad (4)$$

$$\text{R-NIE} = \sum_c \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)]\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*)P(c). \quad (5)$$

Although the assumptions outlined previously are strong, they are still considerably

Figure 1: Causal Graph with Treatment A , Mediator M , outcome Y , baseline confounders C , and post-treatment confounders L .

weaker than those needed to identify the components of more conventional effect decompositions (VanderWeele and Vansteelandt 2009), which additionally require that (iv) $Y_{am} \perp\!\!\!\perp M_{a^*} | C$. Known as a “cross-world independence assumption” because it involves a restriction on the joint distribution of two variables, Y_{am} and M_{a^*} , that can never be observed together, this condition is violated anytime there are mediator-outcome confounders affected by treatment (Robins and Greenland 1992, VanderWeele et al. 2014). For example, it is violated in Figure 1 because L affects both M and Y and is also affected by A .

3 Regression-with-residuals Estimation

Regression-with-residuals (RWR) has been previously used to examine whether the effects of a time-varying treatment are modified by time-varying covariates (Almirall et al. 2010, Wodtke and Almirall 2017), to estimate the marginal effects of a time-varying treatment (Wodtke 2018, Wodtke et al. 2018), and to estimate controlled direct effects (Zhou and Wodtke 2018). In this section, we show that RWR can also be used to decompose causal effects in the presence of treatment-induced confounding.

Specifically, randomized intervention analogues of natural direct and indirect effects

can be estimated from the following set of linear models. The first model is for the conditional mean of the mediator given treatment and the baseline confounders. It can be expressed as

$$\mathbb{E}(M|C, A) = \theta_0 + \theta_1^T C^\perp + \theta_2 A, \quad (6)$$

where $C^\perp = C - \mathbb{E}(C)$. This model is nearly identical to a conventional linear regression except that the baseline confounders C have been centered around their marginal means.

The second model is for the conditional mean of the outcome given treatment, the mediator, the baseline confounders, and the post-treatment confounders. It can be expressed as

$$\mathbb{E}(Y|C, A, L, M) = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM, \quad (7)$$

where $L^\perp = L - \mathbb{E}(L|C, A)$. This model is also nearly identical to a conventional linear regression except that, as before, the baseline confounders C have been centered around their marginal means and, in addition, the post-treatment confounders L have been centered around their conditional means given C and A . Thus, $L^\perp$ is a vector of residual terms that can be obtained from a third set of models for the conditional mean of each post-treatment confounder given treatment and the baseline confounders. In this way, equation (7) subsumes another set of models for $\mathbb{E}(L|C, A)$, which are used to residualize the post-treatment confounders with respect to the observed past. These models may be linear, logistic, Poisson, or any other parametric or semi-parametric model, as appropriate depending on the level of measurement for each element of L .

Under assumptions (i) to (iii) and provided that the models for $\mathbb{E}(M|C, A)$, $\mathbb{E}(L|C, A)$, and $\mathbb{E}(Y|C, A, L, M)$ are correctly specified, the R-NDE and R-NIE are equal to

$$\text{R-NDE} = [\beta_2 + \beta_5(\theta_0 + \theta_2 a)](a^* - a) \quad (8)$$

$$\text{R-NIE} = \theta_2(\beta_4 + \beta_5 a^*)(a^* - a), \quad (9)$$

and the R-ATE is equal to their sum. A derivation of these parametric expressions is

provided in Part A of the Appendix.

RWR estimation of these effects proceeds according to the following steps:

1. For each of the baseline confounders, compute $\hat{C}^\perp = C - \bar{C}$, where the overbar denotes a sample mean.
2. For each of the post-treatment confounders, compute $\hat{L}^\perp = L - \hat{\mathbb{E}}(L|C, A)$ by fitting a model for L given C and A and then extracting the residuals.
3. Compute least squares estimates of equation (6) with $\hat{C}^\perp$ substituted for $C^\perp$, which can be expressed as $\hat{\mathbb{E}}(M|C, A) = \hat{\theta}_0 + \hat{\theta}_1^T \hat{C}^\perp + \hat{\theta}_2 A$.
4. Compute least squares estimates of equation (7) with $\hat{C}^\perp$ and $\hat{L}^\perp$ substituted for $C^\perp$ and $L^\perp$, respectively, which can be expressed as $\hat{\mathbb{E}}(Y|C, A, L, M) = \hat{\beta}_0 + \hat{\beta}_1^T \hat{C}^\perp + \hat{\beta}_2 A + \hat{\beta}_3^T \hat{L}^\perp + \hat{\beta}_4 M + \hat{\beta}_5 A M$.
5. Compute $\widehat{\text{R-NDE}} = [\hat{\beta}_2 + \hat{\beta}_5(\hat{\theta}_0 + \hat{\theta}_2 a)](a^* - a)$ and $\widehat{\text{R-NIE}} = \hat{\theta}_2(\hat{\beta}_4 + \hat{\beta}_5 a^*)(a^* - a)$.

These estimates are consistent under the assumptions outlined previously (Almirall et al. 2010, Wodtke and Almirall 2017). Standard errors and confidence intervals can be computed using the non-parametric bootstrap (Efron and Tibshirani 1994). Alternatively, Part B of the Appendix provides analytic standard errors obtained using the delta method, although this approach requires linear models for all of the post-treatment confounders in L .

Adjustment for post-treatment confounders in a conventional regression model would typically engender bias due to over-control of intermediate pathways and collider stratification (Elwert and Winship 2014, Greenland 2003, Pearl 2009). These problems occur because conditioning on a variable that is affected by treatment may inappropriately block causal pathways and unblock non-causal pathways from treatment to the outcome. RWR avoids these problems by adjusting only for residual transformations of the post-treatment confounders. Because the residualized confounders are purged of their association with treatment, adjusting for them in a regression model for Y is unproblematic, and it appropriately controls for mediator-outcome confounding via observed post-treatment variables.

RWR requires correctly specified models for the outcome, mediator, and post-treatment confounders. These models must be linear for the mediator and outcome, and thus RWR is best suited for applications in which M and Y are both continuous. Nevertheless, like conventional least squares regression, RWR may also be used when these variables are binary, ordinal, or counts if a linear model provides a defensible approximation for the true conditional expectation function. Moreover, when the mediator is binary or counts but a linear model does not provide a defensible approximation, a generalized linear model may still be used to estimate $\mathbb{E}(M|C, A)$. In this case, parametric expressions for the R-NDE and R-NIE will depend on levels of the baseline confounders C and the model used for the mediator M . In general, they are given by

$$\text{R-NDE}(c) = [\beta_2 + \beta_5 \mathbb{E}(M|c, a)](a^* - a) \quad (10)$$

$$\text{R-NIE}(c) = (\beta_4 + \beta_5 a^*)[E(M|c, a^*) - E(M|c, a)]. \quad (11)$$

A derivation of these expressions is provided in Part C of the Appendix.

The flexibility of RWR is also evident in the ease with which it can accommodate effect modification (Wodtke et al. 2018). This is achieved by incorporating, for example, two-way interactions between $C^\perp$ and A , $C^\perp$ and M , or $L^\perp$ and M , which allow the effects of treatment and the mediator to vary across levels of the confounders. As long as these interaction terms are constructed with the residualized confounders, computation of the R-NIE and R-NDE proceeds exactly as outlined previously, which we illustrate in the following example.

4 Empirical Illustration

In this section, we decompose the total effect of post-secondary education on depression symptoms into direct and indirect components using RWR. Education may improve mental health by providing access to higher economic status and greater financial resources, or it may affect mental health through other channels—for example, by providing greater access to health information and improving health behaviors (Heckman et al. 2018, Lee

2011). To investigate the extent to which income mediates the effect of post-secondary education on depression, we use data from $n = 2,924$ individuals in the NLSY79. In these data, the outcome, Y , represents scores on the Center for Epidemiologic Studies Depression Scale (CES-D) when respondents were age 40. We standardize CES-D scores to have mean zero and unit variance, where higher scores imply more depressive symptoms. The treatment, A , is defined as completion of a four-year college degree by age 25. The mediator of interest, M , is the percentile rank of a respondent's equivalized family income averaged over age 36-40. The vector of baseline confounders, C , which includes gender, race, Hispanic ethnicity, mother's years of schooling, father's presence in the home, number of siblings, urban residence, educational expectations, and percentile scores on the Armed Forces Qualification Test, were measured when respondents were age 13-17. Finally, the vector of post-treatment confounders, L , includes CES-D scores measured when respondents were age 27-30, the proportion of time a respondent was married between 1990 and 1998, and the number of relationship transitions experienced by a respondent between 1990 and 1998. These variables capture mental health and family stability during young adulthood, which may be affected by treatment (college completion) and also affect both the mediator (family income between age 36 and 40) and the outcome (depression at age 40).

We adopt the following models for the mediator and outcome:

$$\mathbb{E}(M|C, A) = \theta_0 + \theta_1^T C^\perp + \theta_2 A + \theta_3 C^\perp A \quad (12)$$

$$\mathbb{E}(Y|C, A, L, M) = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 A M + \beta_6 C^\perp A, \quad (13)$$

which allow the effects of college completion on family income and depression to vary across levels of the baseline confounders. As mentioned previously, two-way interactions between $C^\perp$ and M or $L^\perp$ and M may also be included in the outcome model, but for parsimony, we do not include these additional terms in the present analysis.

To estimate the R-NDE and R-NIE, we first compute residuals for each of the baseline confounders C and post-treatment confounders L , which involves centering the elements of C around their sample means and centering the elements of L around their estimated

conditional means given the past. For consistency with the mediator and outcome models, the models for the post-treatment confounders L include C , A , and all two-way interactions between them. We then compute least squares estimates of equations (12) and (13), and finally, we use their coefficients to construct RWR estimates of the R-NDE, R-NIE, and R-ATE.

Our results suggest that completing a post-secondary education has a sizable total effect on depression. Specifically, completing college is estimated to lower depression scores by 0.16 standard deviations on average (95% CI: [-0.29, -0.03]). The R-NDE and R-NIE are estimated to be -0.11 (95% CI: [-0.25, 0.03]) and -0.05 (95% CI: [-0.1, 0.03]), respectively. This suggests that about 30% ($0.05/0.16 \approx 0.3$) of the total effect of college completion on depression is mediated via family income. Note, however, that all of the effect estimates reported here are somewhat imprecise, as indicated by their wide confidence intervals.

5 Discussion

In the presence of treatment-induced confounding, natural direct and indirect effects are not identified, which often complicates analyses of causal mediation. Analogues to natural direct and indirect effects that are defined in terms of randomized interventions on the mediator can still be identified in this situation, but estimating them with IPW methods is challenging in practice. We described a new method – RWR – for estimating the R-NDE and R-NIE that involves only a minor adaption of the regression-based methods commonly used to decompose effects into direct and indirect components in the absence of treatment-induced confounding. It involves, first, fitting a linear regression of the mediator on treatment and a set of baseline confounders that have been centered around their sample means, and second, fitting another linear regression of the outcome on treatment, the mediator, the mean-centered confounders at baseline, and a set of post-treatment confounders that have been residualized with respect to the observed past. Estimates of the R-NDE and R-NIE are constructed with simple functions of the coefficients in these models.

The method's simplicity is in part premised on a set of strong modeling assumptions. In particular, RWR requires correct models for the conditional mean of the mediator, the outcome, and each of the post-treatment confounders. If any of these models is misspecified, then estimates of the R-NDE and R-NIE may be biased. Although simulation studies suggest that RWR is not especially sensitive to at least some types of misspecification ([Almirall et al. 2010](#), [Wodtke 2018](#), [Wodtke et al. 2018](#)), an important direction for future research will be to explore the possibility of combining RWR with modern methods of model selection and regularization in an effort to improve its robustness.

RWR is also premised on a set of strong identification assumptions, which require that all relevant confounders of the treatment-outcome, treatment-mediator, and mediator-outcome relationships have been observed and appropriately controlled. In observational studies where treatment has not been randomly assigned, all of these assumptions must be carefully scrutinized. If any are violated, then RWR estimates of direct and indirect effects will be biased. In experimental studies where treatment has been randomly assigned, the assumptions of no unobserved treatment-outcome or treatment-mediator confounding are met by design, but it remains possible that the mediator-outcome relationship is still confounded by unobserved factors. Thus, no matter the research design, it is important to critically evaluate the assumption that there are not any unobserved mediator-outcome confounders. To this end, we have developed a method for assessing the sensitivity of RWR estimates to hypothetical patterns of unobserved mediator-outcome confounding, as detailed in Part D of the Appendix.

We focused on a two-way decomposition of a total effect into randomized intervention analogues of natural direct and indirect effects, which is designed to evaluate mediation. The methods discussed previously can, however, also be used to estimate the components of more complex decompositions that evaluate the degree to which a total effect is due to mediation versus interaction ([VanderWeele 2014, 2013](#), [VanderWeele and Tchetgen Tchetgen 2014](#)). [VanderWeele \(2014\)](#), for example, decomposes a total effect into components due to mediation, interaction, both, or neither. In Part E of the Appendix, we show that the components of this four-way decomposition, when defined in terms of randomized interventions on the mediator, can also be estimated with RWR.

Because RWR involves only a minor adaption of conventional least squares regression, it is based on computations that should be familiar and intuitive to most applied researchers. Moreover, the method can be easily implemented with off-the-shelf statistical software. We therefore expect that it will find wide application in analyses of causal mediation where treatment-induced confounding is present. To this end, we have developed an open-source R package, `rwrmed`, as well as a Stata package by the same name with similar functionality, for decomposing total effects with RWR. The R package is available at <https://github.com/xiangzhou09/rwrmed>, and the Stata package at <https://github.com/gtwodtke/rwrmed>. In addition, Part F of the Appendix provides the R code for implementing RWR in our empirical example.

References

- Almirall, D., Ten Have, T. and Murphy, S. A.: 2010, Structural nested mean models for assessing time-varying effect moderation, *Biometrics* **66**(1), 131–139.
- Efron, B. and Tibshirani, R. J.: 1994, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, New York.
- Elwert, F. and Winship, C.: 2014, Endogenous selection bias: The problem of conditioning on a collider variable, *Annual Review of Sociology* **40**, 31–53.
- Geneletti, S.: 2007, Identifying direct and indirect effects in a non-counterfactual framework, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2), 199–215.
- Greenland, S.: 2003, Quantifying biases in causal models: Classical confounding vs collider-stratification bias, *Epidemiology* **14**(3), 300–306.
- Hafeman, D. M. and Schwartz, S.: 2009, Opening the black box: A motivation for the assessment of mediation, *International Journal of Epidemiology* **38**(3), 838–845.
- Heckman, J. J., Humphries, J. E. and Veramendi, G.: 2018, The nonmarket benefits of education and ability, *Journal of Human Capital* **12**(2), 282–304.
- Howe, C. J., Cole, S. R., Chmiel, J. S. and Munoz, A.: 2011, Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias, *American Journal of Epidemiology* **173**(5), 569–577.
- Lee, J.: 2011, Pathways from education to depression, *Journal of Cross-cultural Gerontology* **26**(2), 121–135.
- Naimi, A. I., Moodie, E. M., Auger, N. and Kaufman, J. S.: 2014, Constructing inverse probability weights for continuous exposures: A comparison of methods, *Epidemiology* **25**(2), 292–299.

- Pearl, J.: 2001, Direct and indirect effects, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 411–420.
- Pearl, J.: 2009, *Causality: Models, Reasoning, and Inference*, Cambridge University Press.
- Pearl, J.: 2010, The foundations of causal inference, *Sociological Methodology* **40**(1), 75–149.
- Robins, J. M. and Greenland, S.: 1992, Identifiability and exchangeability for direct and indirect effects, *Epidemiology* **3**(2), 143–155.
- Sobel, M. E.: 1982, Asymptotic confidence intervals for indirect effects in structural equation models, *Sociological Methodology* **13**(1), 290–312.
- VanderWeele, T. J.: 2009, Concerning the consistency assumption in causal inference, *Epidemiology* **20**(6), 880–883.
- VanderWeele, T. J.: 2013, A three-way decomposition of a total effect into direct, indirect, and interactive effects, *Epidemiology* **24**(2), 224–232.
- VanderWeele, T. J.: 2014, A unification of mediation and interaction: a four-way decomposition, *Epidemiology* **25**(5), 749.
- VanderWeele, T. J.: 2015, *Explanation in Causal Inference: Methods for Mediation and Interaction*, Oxford University Press.
- VanderWeele, T. J. and Tchetgen Tchetgen, E.: 2014, Attributing effects to interactions, *Epidemiology* **25**(5), 711–722.
- VanderWeele, T. J. and Vansteelandt, S.: 2009, Conceptual issues concerning mediation, interventions and composition, *Statistics and its Interface* **2**(4), 457–468.
- VanderWeele, T. J., Vansteelandt, S. and Robins, J. M.: 2014, Effect decomposition in the presence of an exposure-induced mediator-outcome confounder, *Epidemiology* **25**(2), 300–306.
- Wodtke, G. T.: 2018, Regression-based adjustment for time-varying confounders, *Sociological Methods & Research* (online access ahead of print) .

- Wodtke, G. T., Alaca, Z. and Zhou, X.: 2018, Regression-with-residuals estimation of marginal effects: A method of adjusting for treatment-induced confounders that are also moderators, *arXiv preprint* (<https://arxiv.org/abs/1808.07795>) .
- Wodtke, G. T. and Almirall, D.: 2017, Estimating moderated causal effects with time-varying treatments and time-varying moderators: Structural nested mean models and regression with residuals, *Sociological Methodology* **47**(1), 212–245.
- Zhou, X. and Wodtke, G. T.: 2018, A regression-with-residuals method for estimating controlled direct effects, *Political Analysis* (online access ahead of print) .

Appendix

A: Derivation of Parametric Expressions for the R-NDE and R-NIE under Linear Models for M and Y

Under assumptions (i) to (iii) and the assumption that equations (6) and (7) from the main text are both correctly specified, then the R-NDE is equal to

$$\begin{aligned}
\text{R-NDE} &= \sum_c \sum_m \sum_l [\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*) - \mathbb{E}(Y|c, a, l, m)P(l|c, a)]P(m|c, a)P(c) \\
&= \sum_c \sum_m \sum_l [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T l^\perp + \beta_4 m + \beta_5 a^* m)P(l|c, a^*) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a + \beta_3^T l^\perp + \beta_4 m + \beta_5 a m)P(l|c, a)]P(m|c, a)P(c) \\
&= \sum_c \sum_m [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a^*)|c, a^*) + \beta_4 m + \beta_5 a^* m) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a)|c, a) + \beta_4 m + \beta_5 a m)]P(m|c, a)P(c) \\
&= \sum_c \sum_m [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a + \beta_4 m + \beta_5 a m)]P(m|c, a)P(c) \\
&= \sum_c [(\beta_2 a^* + \beta_5 a^* m) - (\beta_2 a + \beta_5 a m)]P(m|c, a)P(c) \\
&= \sum_c [(\beta_2 a^* + \beta_5 a^* \mathbb{E}(M|c, a)) - (\beta_2 a + \beta_5 a \mathbb{E}(M|c, a))]P(c) \\
&= \sum_c [(\beta_2 a^* + \beta_5 a^* (\theta_0 + \theta_1^T c^\perp + \theta_2 a)) - (\beta_2 a + \beta_5 a (\theta_0 + \theta_1^T c^\perp + \theta_2 a))]P(c) \\
&= (\beta_2 a^* + \beta_5 a^* (\theta_0 + \theta_1^T \mathbb{E}(C - \mathbb{E}(C)) + \theta_2 a)) - (\beta_2 a + \beta_5 a (\theta_0 + \theta_1^T \mathbb{E}(C - \mathbb{E}(C)) \\
&\quad + \theta_2 a)) \\
&= (\beta_2 a^* + \beta_5 a^* (\theta_0 + \theta_2 a)) - (\beta_2 a + \beta_5 a (\theta_0 + \theta_2 a)) \\
&= [\beta_2 + \beta_5 (\theta_0 + \theta_2 a)](a^* - a),
\end{aligned}$$

and the R-NIE is equal to

$$\begin{aligned}
\text{R-NIE} &= \sum_c \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)] \mathbb{E}(Y|c, a^*, l, m) P(l|c, a^*) P(c) \\
&= \sum_c \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T l^\perp + \beta_4 m + \beta_5 a m) \times \\
&\quad P(l|c, a^*) P(c) \\
&= \sum_c \sum_m [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a^*)|c, a^*) + \\
&\quad \beta_4 m + \beta_5 a^* m) P(c) \\
&= \sum_c \sum_m [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) P(c) \\
&= \sum_c \sum_m [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) P(m|c, a^*) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) P(m|c, a)] P(c) \\
&= \sum_c [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + E(M|c, a^*)(\beta_4 + \beta_5 a^*)) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + E(M|c, a)(\beta_4 + \beta_5 a^*))] P(c) \\
&= \sum_c [(E(M|c, a^*) - E(M|c, a))(\beta_4 + \beta_5 a^*) P(c) \\
&= \sum_c [(\theta_0 + \theta_1^T c^\perp + \theta_2 a^*) - (\theta_0 + \theta_1^T c^\perp + \theta_2 a)] (\beta_4 + \beta_5 a^*) P(c) \\
&= \sum_c (\theta_2 a^* - \theta_2 a) (\beta_4 + \beta_5 a^*) P(c) \\
&= \theta_2 (\beta_4 + \beta_5 a^*) (a^* - a).
\end{aligned}$$

B: Analytic Standard Errors for RWR

In this appendix, we outline an approach to obtaining analytic standard errors for the RWR estimates of the R-NDE and R-NIE. With this approach, we assume that the variables $\{C, A, L, M, Y\}$ satisfy the Causal Markov assumption, that is, they are represented by a recursive system of equations with independent errors (Pearl 2010). Let X denote a $p \times 1$ vector of ones, the treatment A , baseline confounders centered at their sample means $\hat{C}^\perp$, and any interactions between A and $\hat{C}^\perp$; let L denote a $q \times 1$ vector of post-

treatment confounders; and finally, let Z denote a $r \times 1$ vector containing the mediator M and any of its interactions with X . A “naive” least squares regression of the outcome Y on $\{X, L, Z\}$ can be expressed as follows:

$$\begin{aligned} Y &= \hat{\alpha}^T X + \hat{\eta}^T L + \hat{\gamma}^T Z + \hat{Y}^\perp \\ &= \hat{\alpha}^T X + \sum_j \hat{\eta}_j L_j + \hat{\gamma}^T Z + \hat{Y}^\perp, \end{aligned} \quad (14)$$

where L_j is the j th element of L and $\hat{Y}^\perp$ denotes the residual. Similarly, a least squares regression of each L_j on X can be expressed as follows

$$L_j = \hat{\lambda}_j^T X + \hat{L}_j^\perp, \quad (15)$$

where $\hat{L}_j^\perp$ denotes the residual.

Substituting (15) into (14) yields the following expression for the outcome:

$$Y = (\hat{\alpha}^T + \sum_j \hat{\eta}_j \hat{\lambda}_j^T) X + \sum_j \hat{\eta}_j \hat{L}_j^\perp + \hat{\gamma}^T Z + \hat{Y}^\perp. \quad (16)$$

Since $\hat{Y}^\perp$ is the least squares residual for regression (14), it is orthogonal to the span of $\{X, L, Z\}$. Because each $\hat{L}_j^\perp$ is a linear combination of X and L_j , $\{X, \hat{L}_j^\perp, Z\}$ and $\{X, L, Z\}$ span the same space. Thus, equation (16) represents the least squares fit of Y on $\{X, \hat{L}_j^\perp, Z\}$, meaning that $\hat{\alpha}_{RWR}^T = (\hat{\alpha}^T + \sum_j \hat{\eta}_j \hat{\lambda}_j^T)$ are the RWR estimates of the coefficients on treatment, the baseline confounders, and any interactions between them, $\hat{\eta}$ are the RWR estimates of the coefficients on the post-treatment confounders, and $\hat{\gamma}$ are the RWR estimates of the coefficients on the mediator and any of its interactions with treatment and/or the baseline confounders. Therefore, the asymptotic variance-covariance matrix for the RWR estimates $(\hat{\eta}, \hat{\gamma})$ can be obtained directly via conventional methods after fitting the naive regression (14).

The asymptotic variance-covariance matrix for $\hat{\alpha}_{RWR}$ can be obtained with the delta method. Given the assumption that Y and L have mutually independent errors, each $\hat{\lambda}_j$ is independent of $\hat{\alpha}$ and $\hat{\eta}$. The variance-covariance matrix for $\hat{\alpha}_{RWR}$ can then be estimated

as

$$\hat{\mathbf{V}}(\hat{\alpha}_{RWR}) = \hat{\mathbf{V}}(\hat{\alpha}) + \widehat{\text{Cov}}(\hat{\alpha}, \hat{\eta})\hat{\Lambda}^T + \hat{\Lambda}\hat{\mathbf{V}}(\hat{\eta})\hat{\Lambda}^T + \sum_{j,k} \hat{\eta}_j \hat{\eta}_k \widehat{\text{Cov}}(\hat{\lambda}_j, \hat{\lambda}_k), \quad (17)$$

where $\hat{\Lambda} = [\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_q]$ is a $p \times q$ matrix of estimated coefficients from model (15), $\mathbf{V}(\cdot)$ is the variance-covariance matrix of a random vector, and $\text{Cov}(\cdot, \cdot)$ is the covariance matrix between two random vectors. In equation (17), the first three terms can be obtained directly from the naive regression (14), and the covariance matrix between $\hat{\lambda}_j$ and $\hat{\lambda}_k$ in the last term can be estimated as

$$\widehat{\text{Cov}}(\hat{\lambda}_j, \hat{\lambda}_k) = \frac{(\hat{l}_j^\perp)^T \hat{l}_k^\perp}{n - p} (X^T X)^{-1},$$

where n is the sample size, and $\hat{l}_j^\perp$ and $\hat{l}_k^\perp$ are $n \times 1$ vectors of the residualized confounders $\hat{L}_j^\perp$ and $\hat{L}_k^\perp$. Similarly, the covariance matrix between $\hat{\alpha}_{RWR}$ and $\hat{\gamma}$ can be estimated as

$$\widehat{\text{Cov}}(\hat{\alpha}_{RWR}, \hat{\gamma}) = \widehat{\text{Cov}}(\hat{\alpha}, \hat{\gamma}) + \hat{\Lambda} \widehat{\text{Cov}}(\hat{\eta}, \hat{\gamma}). \quad (18)$$

Now consider the plug-in estimators of the R-NDE and R-NIE. Without loss of generality, assume that $a^* - a = 1$. Given that the error terms for equations (6) and (7) are independent, asymptotic variances for $\widehat{\text{R-NDE}}$ and $\widehat{\text{R-NIE}}$ can be estimated using the delta method (Sobel 1982):

$$\hat{\mathbf{V}}[\widehat{\text{R-NDE}}] = \hat{\mathbf{V}}(\hat{\beta}_2) + (\hat{\theta}_0 + \hat{\theta}_2 a)^2 \hat{\mathbf{V}}(\hat{\beta}_5) + \hat{\beta}_5^2 \hat{\mathbf{V}}(\hat{\theta}_0 + \hat{\theta}_2 a) + (\hat{\theta}_0 + \hat{\theta}_2 a) \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_5) \quad (19)$$

$$\hat{\mathbf{V}}[\widehat{\text{R-NIE}}] = (\hat{\beta}_4 + \hat{\beta}_5 a^*)^2 \hat{\mathbf{V}}(\hat{\theta}_2) + \hat{\theta}_2^2 \hat{\mathbf{V}}(\hat{\beta}_4 + \hat{\beta}_5 a^*). \quad (20)$$

In these equations, the terms involving $\hat{\theta}_0$ and $\hat{\theta}_2$ can be estimated by applying equations (17-18) to the RWR regression of model (6) from the main text, and the terms involving $\hat{\beta}_2$, $\hat{\beta}_4$, and $\hat{\beta}_5$ can be estimated by applying equations (17-18) to the RWR regression of model (7) from the main text.

C: Derivation of Parametric Expressions for the R-NDE and R-NIE under a Nonlinear Model for M

When the mediator is binary or counts, a generalized linear model may be preferred for estimating $\mathbb{E}[M|A, C]$. Under assumptions (i) to (iii) and the assumption that both this mediator model and the outcome model (7) from the main text is correctly specified, then the R-NDE conditional on $C = c$ is equal to

$$\begin{aligned}
\text{R-NDE}(c) &= \sum_m \sum_l [\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*) - \mathbb{E}(Y|c, a, l, m)P(l|c, a)]P(m|c, a) \\
&= \sum_m \sum_l [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T l^\perp + \beta_4 m + \beta_5 a^* m)P(l|c, a^*) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a + \beta_3^T l^\perp + \beta_4 m + \beta_5 a m)P(l|c, a)]P(m|c, a) \\
&= \sum_m [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a^*)|c, a^*) + \beta_4 m + \beta_5 a^* m) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a)|c, a) + \beta_4 m + \beta_5 a m)]P(m|c, a) \\
&= \sum_m [(\beta_2 a^* + \beta_5 a^* m) - (\beta_2 a + \beta_5 a m)]P(m|c, a) \\
&= [(\beta_2 a^* + \beta_5 a^* \mathbb{E}(M|c, a)) - (\beta_2 a + \beta_5 a \mathbb{E}(M|c, a))] \\
&= [\beta_2 + \beta_5 \mathbb{E}(M|c, a)](a^* - a)
\end{aligned}$$

and the R-NIE conditional on $C = c$ is equal to

$$\begin{aligned}
\text{R-NIE}(c) &= \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)] \mathbb{E}(Y|c, a^*, l, m) P(l|c, a^*) \\
&= \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T l^\perp + \beta_4 m + \beta_5 a m) P(l|c, a^*) \\
&= \sum_m [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_3^T \mathbb{E}(L - \mathbb{E}(L|c, a^*)|c, a^*) + \\
&\quad \beta_4 m + \beta_5 a^* m) \\
&= \sum_m [P(m|c, a^*) - P(m|c, a)] (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) \\
&= \sum_m [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) P(m|c, a^*) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + \beta_4 m + \beta_5 a^* m) P(m|c, a)] \\
&= [(\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + E(M|c, a^*)(\beta_4 + \beta_5 a^*)) - \\
&\quad (\beta_0 + \beta_1^T c^\perp + \beta_2 a^* + E(M|c, a)(\beta_4 + \beta_5 a^*))] \\
&= (\beta_4 + \beta_5 a^*) [E(M|c, a^*) - E(M|c, a)].
\end{aligned}$$

D: Sensitivity Analysis for Unobserved Confounding

Assumptions (i) to (iii) require that there must not be any unobserved confounding of the treatment-outcome, treatment-mediator, or mediator-outcome relationships. The first two of these assumptions are similar to the conventional “exogeneity of treatment” assumption required in observational studies, where it is justified by adjusting for a sufficient set of baseline confounders, or in experimental studies, where it is met by design via random assignment. The third assumption, however, may fail to hold even in randomized experiments, and if unobserved confounding exists for the mediator-outcome relationship, RWR estimates of the R-NDE and R-NIE will be biased and inconsistent. In this section, we outline a simple sensitivity analysis that permits an assessment of whether RWR estimates are robust to this type of unobserved confounding.

Consider the following set of linear structural equations:

$$M = \theta_0 + \theta_1^T C^\perp + \theta_2 A + \epsilon_M, \quad (21)$$

$$Y = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM + \epsilon_Y. \quad (22)$$

The assumption of no unobserved mediator-outcome confounding implies that the error terms, ϵ_M and ϵ_Y , are statistically independent. When this assumption is violated, ϵ_M and ϵ_Y are correlated. A linear projection of ϵ_Y on ϵ_M can be expressed as

$$\epsilon_Y = \phi \epsilon_M + \psi. \quad (23)$$

Under the assumption that $\mathbb{E}[\psi|C, A, L, M] = 0$, substituting (23) into (22) and taking the conditional expectation of Y yields

$$\begin{aligned} \mathbb{E}[Y|C, A, L, M] &= (\beta_0 - \phi\theta_0) + (\beta_1 - \phi\theta_1)^T C^\perp + (\beta_2 - \phi\theta_2)A + \beta_3^T L^\perp + \\ &\quad (\beta_4 + \phi)M + \beta_5 AM. \end{aligned} \quad (24)$$

Thus, in this case, the RWR estimates of $(\beta_0, \beta_1, \beta_2, \beta_4)$ suffer from an asymptotic bias of $\phi(-\theta_0, -\theta_1, -\theta_2, 1)$. Accordingly, the biases for the RWR estimators of the R-NDE and R-NIE can be expressed as

$$\text{Bias}[\text{R-NDE}] = -\phi\theta_2(a^* - a) \quad (25)$$

$$\text{Bias}[\text{R-NIE}] = \phi\theta_2(a^* - a). \quad (26)$$

The biases for the R-NDE and R-NIE are equal in magnitude but opposite in direction. This implies that the total effect, defined as the sum of the R-NDE and R-NIE, is not affected by unobserved mediator-outcome confounding, as expected. These expressions also imply that the R-NIE, and thus the mediating role of M , will be overstated if ϕ and θ_2 are in the same direction and understated if they are in the opposite direction.

In practice, neither the sign nor the magnitude of ϕ is known. Moreover, ϕ is not on an interpretable scale. To circumvent this problem, ϕ can be re-expressed in terms of the

correlation between ϵ_Y and ϵ_M as follows:

$$\phi = \frac{\text{sd}(\psi)}{\text{sd}(\epsilon_M)} \frac{\rho}{\sqrt{1 - \rho^2}}, \quad (27)$$

where $\rho = \text{Corr}[\epsilon_Y, \epsilon_M]$. Substituting (27) into (26) yields

$$\text{Bias}[\text{R-NIE}] = \frac{\theta_2 \cdot \text{sd}(\psi)}{\text{sd}(\epsilon_M)} \frac{\rho}{\sqrt{1 - \rho^2}} (a^* - a). \quad (28)$$

The bias for the R-NDE can be expressed analogously. Under the assumptions of no treatment-mediator or treatment-outcome confounding, θ_2 , $\text{sd}(\epsilon_M)$, and $\text{sd}(\psi)$ can be consistently estimated from the mediator and outcome regressions described in the main text. Thus, we can evaluate the bias terms as functions of ρ and construct a range of bias-adjusted RWR estimates for the R-NDE and R-NIE across different values of ρ . In addition, we can identify the value of ρ that would suffice to reduce the estimated R-NIE to zero. Alternatively, we can also identify the value of ρ that would suffice to render the estimated R-NIE statistically insignificant.

E: A Four-way Decomposition

As shown by [VanderWeele \(2014\)](#) the R-NDE can be further decomposed into the following two components:

$$\begin{aligned} \text{R-NDE} &= \mathbb{E}(Y_{a^*m} - Y_{am}) + [\mathbb{E}(Y_{a^*G_{a|C}} - Y_{aG_{a|C}}) - \mathbb{E}(Y_{a^*m} - Y_{am})] \\ &= \text{CDE}(m) + \text{R-INT}_{ref}(m). \end{aligned} \quad (29)$$

The first term in (29), $\text{CDE}(m) = \mathbb{E}(Y_{a^*m} - Y_{am})$, is a controlled direct effect that gives the expected difference in the outcome under treatment a^* rather than a if the mediator were set to m for all individuals. It represents the component of the total effect due to neither mediation nor interaction. The second term, $\text{R-INT}_{ref}(m) = [\mathbb{E}(Y_{a^*G_{a|C}} - Y_{aG_{a|C}}) - \mathbb{E}(Y_{a^*m} - Y_{am})]$, is a so-called reference interaction effect. It represents the component of the total effect due to an interaction between treatment and the mediator occurring in

the absence of any mediation. Under assumptions (i) to (iii) and the assumption that equations (6) and (7) from the main text are both correctly specified, the controlled direct effect is equal to

$$\begin{aligned} \text{CDE}(m) &= \sum_c \sum_l [\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*) - \mathbb{E}(Y|c, a, l, m)P(l|c, a)]P(c) \\ &= (\beta_2 + \beta_5 m)(a^* - a). \end{aligned} \quad (30)$$

By extension, the reference interaction effect is equal to

$$\text{R-INT}_{ref}(m) = \text{R-NDE} - \text{CDE}(m) = \beta_5(\theta_0 + \theta_2 a - m)(a^* - a). \quad (31)$$

[VanderWeele \(2014\)](#) also shows that the R-NIE can be further decomposed as follows:

$$\begin{aligned} \text{R-NIE} &= \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}}) + [\mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{a^*G_{a|C}}) - \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}})] \\ &= \text{R-PIE} + \text{R-INT}_{med}. \end{aligned} \quad (32)$$

The first term in (32), $\text{R-PIE} = \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}})$, is a randomized intervention analogue of a so-called pure indirect effect, which captures the component of the total effect due to mediation in the absence of any interaction between the effects of treatment and the mediator on the outcome. The second term, $\text{R-INT}_{med} = [\mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{aG_{a^*|C}}) - \mathbb{E}(Y_{a^*G_{a|C}} - Y_{aG_{a|C}})]$, is a randomized intervention analogue of a so-called mediated interaction effect. It captures the component of the total effect due to mediation and interaction operating jointly. Under the same assumptions outlined previously, the R-PIE is equal to

$$\begin{aligned} \text{R-PIE} &= \sum_c \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)]\mathbb{E}(Y|c, a, l, m)P(l|c, a)P(c). \\ &= \theta_2(\beta_4 + \beta_5 a)(a^* - a) \end{aligned} \quad (33)$$

By extension, the mediated interaction effect is equal to

$$\text{R-INT}_{med} = \text{R-NIE} - \text{R-PIE} = \theta_2 \beta_5 (a^* - a)^2. \quad (34)$$

And thus the randomized intervention analogue to the average total effect can be expressed as

$$\text{R-ATE} = \text{R-NDE} + \text{R-NIE} = \text{CDE}(m) + \text{R-INT}_{ref}(m) + \text{R-PIE} + \text{R-INT}_{med}. \quad (35)$$

RWR estimates of equations (6) and (7) from the main text can be used to construct estimates for each component of this four-way decomposition. This is accomplished merely by substituting the appropriate parameter estimates from these models into formulas (30-31) and (33-34). Standard errors and confidence intervals can be computed using either the non-parametric bootstrap or the analytic approach outlined in Part B of the Appendix.

F: Implementation of RWR in R

Below, we illustrate the implementation of RWR in R for estimating the R-NDE and R-NIE of college education on depression.

```
# R code #
rm(list=ls(all=TRUE))
devtools::install_github("xiangzhou09/rwrmed")
library(rwrmed)

# load data #
load("depression.RData")

# baseline confounders #
pre_cov <- c("male", "black", "test_score", "educ_exp", "father", "hispanic",
            "urban", "educ_mom", "num_sibs")

# mediator and outcome equations #
m_form <- ses ~ (male + black + test_score + educ_exp + father + hispanic +
                urban + educ_mom + num_sibs) * college
y_form <- cesd40 ~ (male + black + test_score + educ_exp + father + hispanic + urban +
                  educ_mom + num_sibs) * college + ses + college * ses +
cesd92 + prmarr98 + transitions98
```

```

# models for the post-treatment confounders #
m1 <- lm(cesd92 ~ (male + black + test_score + educ_exp + father + hispanic +
                  urban + educ_mom + num_sibs) * college, weights = weights,
        data = depression)
m2 <- lm(prmarr98 ~ (male + black + test_score + educ_exp + father + hispanic +
                   urban + educ_mom + num_sibs) * college, weights = weights,
        data = depression)
m3 <- lm(transitions98 ~ (male + black + test_score + educ_exp + father + hispanic +
                         urban + educ_mom + num_sibs) * college, weights = weights,
        data = depression)

# RWR estimation #
fit <- rwrmed(treatment = "college", pre_cov = pre_cov, zmodels = list(m1, m2, m3),
             y_form = y_form, m_form = m_form, weights = weights, data = depression)

# effect decomposition #
out <- decomp(fit, rep = 500)
print(out, digits = 2)

```