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

Geoffrey T. Wodtke
University of Chicago

Xiang Zhou
Harvard University

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. We introduce a new method for estimating the R-NDE and R-NIE that involves only a minor adaptation of the comparatively simple regression methods used to perform effect decomposition in the absence of treatment-induced confounding. It involves fitting (a) a generalized linear model for the conditional mean of the mediator given treatment and a set of baseline confounders and (b) a linear model for the conditional mean of the outcome given the treatment, mediator, baseline confounders, and a set of treatment-induced confounders that have been residualized 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 at midlife into components operating through income versus alternative factors. R and Stata packages are available for implementing the proposed method.

Keywords: mediation, effect decomposition, causal inference, confounding.
Running head: effect decomposition using regression-with-residuals.

*Direct all correspondence to Geoffrey T. Wodtke, Department of Sociology, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637; email: wodtke@uchicago.edu.

†The authors certify that they have no conflicts of interest. This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada (Grant No. 435-2018-0736). The study is exempt from IRB review because it involves only minimal risk and anonymous secondary data. Replication files are available as online supplementary materials.

1 Introduction

Researchers have become increasingly interested in uncovering the mediating pathways through which one variable affects another.[1] A common approach to assessing causal mediation involves decomposing a total effect of 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 natural direct and indirect effects.[2–4]

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.

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.

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.[3] 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[3,5,6] 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 Didelez et al.[7] and Geneletti[8]). 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 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[5] 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,[9–11] 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[4]).

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 adaptation of the familiar regression-based approaches to effect decomposition that are widely used when treatment-induced confounding is assumed away. Briefly, the method involves fitting (a)

a generalized linear model for the conditional mean of the mediator given treatment and a set of baseline confounders, (b) a generalized linear model for the conditional mean of each treatment-induced confounder given treatment and the baseline confounders, which are used to compute residual terms, and finally, (c) a linear model for the conditional mean of the outcome given the treatment, mediator, baseline confounders, and treatment-induced confounders that have been residualized with respect to the observed past. These models can be fit using standard software, and estimates of the R-NDE and R-NIE are given by simple functions of their coefficients. RWR estimates are consistent and asymptotically unbiased when assumptions (i) to (iii) are satisfied and when all of the models mentioned previously are correctly specified; otherwise, they may be biased.

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 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 at midlife into components operating through family income versus alternative pathways.

2 Notation, Estimands, and Identification

We adopt the notation used by VanderWeele, Vansteelandt, and Robins.[5] 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

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

This estimand represents the expected difference in the outcome if all individuals in some target population were exposed to treatment a^* rather than a and if they were subsequently exposed to a level of the mediator randomly selected from the distribution under treatment a among those with baseline confounders C.[5,7,8] 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

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

This estimand represents the expected difference in the outcome if all individuals were exposed to treatment a^* and then were subsequently exposed to a level of the mediator randomly selected from the distribution under treatment a^* rather than a.[5,7,8] It captures an effect of treatment on the outcome due to mediation via M. This is achieved by fixing treatment at 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:

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

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.[5,7,8]

The R-NDE and R-NIE can be 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) $Y_{am} \perp\!\!\!\perp M|C, A, L$.[5] 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.[a] 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.[12] 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 weaker than those needed to identify the components of more conventional effect decompositions,[4] 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.[5,13] For example, it is violated in Figure 1 because L affects both M and Y and is also affected by A.

The R-NDE and R-NIE evaluate idealized interventions on treatment and the distribution of a putative mediator. Such interventions may not be practical or even feasible in many applications. Nevertheless, these estimands can still inform the development of more effective interventions in practice by answering "what if" questions about hy-

[a] Several other 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. In addition, non-parametric identification of the R-NDE and R-NIE requires $G_{a|C}$ and $G_{a^*|C}$ to have the same support; otherwise, model-based extrapolation is needed to identify $\mathbb{E}[Y_{a^*G_{a|C}}]$ and/or $\mathbb{E}[Y_{aG_{a^*|C}}]$, as the potential outcomes $Y_{a^*G_{a|C}}$ and $Y_{aG_{a^*|C}}$ may not exist for certain values of the mediator.

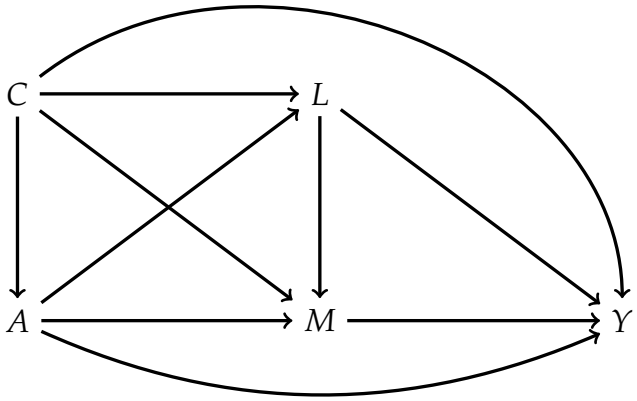

Figure 1: Causal Graph with Treatment A, Mediator M, outcome Y, baseline confounders C, and post-treatment confounders L.

pothetical modifications to treatment, like "what would be the effect of treatment if its components that only serve to improve a mediator were eliminated?" Answers to such questions can guide researchers in imagining and then constructing alternative worlds where the effects of treatment might be attenuated, neutralized, or amplified.[14] If researchers are interested in answering other types of questions about causal mediation, then they should consider focusing instead on different estimands, such as controlled direct effects or path-specific effects,[5,15] that may better correspond with the particular query of interest.

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,[16,17] to estimate the marginal effects of a time-varying treatment,[11,18] and to estimate controlled direct effects.[15] In this section, we show that RWR can also be used to decompose causal effects in the presence of treatment-induced confounding into direct and indirect components. For simplicity, we introduce RWR by focusing on its implementation with linear models for the outcome, mediator, and treatment-induced confounders. Later, we explain how

RWR can also be implemented with a more general class of models for the mediator and treatment-induced confounders.

3.1 Linear RWR

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

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

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 the treatment, mediator, baseline confounders, and post-treatment confounders. It can be expressed as

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

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 linear models for the conditional mean of each post-treatment confounder given treatment and the baseline confounders. These models can be expressed as

$$\mathbb{E}(L|C,A) = \tau_0 + \tau_1^T C^{\perp} + \tau_2 A. \tag{8}$$

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

$$\text{R-NDE} = [\beta_2 + \beta_5(\theta_0 + \theta_2 a)](a^* - a) \tag{9}$$

$$\text{R-NIE} = \theta_2(\beta_4 + \beta_5 a^*)(a^* - a), \tag{10}$$

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

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

1. For each of the baseline confounders, compute $\hat{C}^{\perp} = C - \overline{C}$, 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 linear regression of L on 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 $\hat{\mathbb{E}}(Y|C, A, L, M) = \hat{\beta}_0 + \hat{\beta}_1^T \hat{C}^{\perp} + \hat{\beta}_2 A + \hat{\beta}_3^T \hat{L}^{\perp} + \hat{\beta}_4 M + \hat{\beta}_5 AM$.

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.[16,17] Standard errors and confidence intervals can be computed using the non-parametric bootstrap.[19] Alternatively, Part B of the eAppendix provides analytic standard errors obtained using the delta method.

Adjustment for post-treatment confounders in a conventional regression model would typically engender bias due to over-control of intermediate pathways and collider stratification.[12,20,21] 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.

3.2 Extensions

RWR requires correctly specified models for the outcome, mediator, and post-treatment confounders. The model for the outcome must be linear, and thus RWR is best suited for applications in which Y is continuous. It may also be used when the outcome is binary or counts, provided that a linear model represents a defensible approximation for the true conditional expectation function in any particular application.

Although linearity in the outcome model is restrictive, RWR is flexible in other ways. For example, it can easily accommodate effect modification.[15,18] This is achieved by incorporating 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 as outlined previously.

RWR is also flexible in that it can be easily used with nonlinear models for the mediator. Specifically, when the mediator is binary or counts, a generalized linear model, such as logistic or Poisson regression, may 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

$$\text{R-NDE}(c) = [\beta_2 + \beta_5 \mathbb{E}(M|c, a)](a^* - a) \tag{11}$$

$$\text{R-NIE}(c) = (\beta_4 + \beta_5 a^*)[E(M|c, a^*) - E(M|c, a)]. \tag{12}$$

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

Similarly, RWR can be implemented with a large class of models for the treatment-induced confounders. 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 in L. A convenient feature of RWR is that the parametric expressions for the R-NDE and R-NIE are insensitive to the choice of models for the post-treatment

confounders. Thus, regardless of the models used to residualize L, computation of the R-NIE and R-NDE proceeds exactly as outlined previously.

Because RWR may be biased under incorrect models for $\mathbb{E}(M|C,A)$, $\mathbb{E}(L|C,A)$, or $\mathbb{E}(Y|C,A,L,M)$, analysts should attempt to avoid misspecification. This might be achieved by using diagnostic procedures for detecting non-linearity (e.g., partial residual plots), by incorporating a large number of interaction terms, and/or by using conventional model selection techniques (e.g., information criteria) for adjudicating between competing models. When it is available, subject matter knowledge could also guide the choice of models used with RWR.

4 Empirical Illustration

In this section, we decompose the effect of post-secondary education on depression into direct and indirect components using RWR. Education may improve mental health by providing access to 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.[22,23] To investigate whether income mediates the effect of education on depression, we use data from $n = 2,988$ individuals in the NLSY79. 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 (in Part D of the eAppendix, we present a parallel analysis in which the outcome is coded instead as binary variable). The treatment, A, is defined as completion of a four-year college degree by age 25. The mediator, M, is the inverse hyperbolic sine of a respondent's equivalized family income averaged over age 35-40.[b] The vector of baseline confounders, C, 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, which were measured when respondents were age

[b]The inverse hyperbolic sine is a normalizing transformation for right-skewed variables, like income, that is similar to the natural log except that it is defined at 0 and therefore accommodates respondents who report earning no income.

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 college completion and may also affect family income and depression at midlife.

We adopt the following models for the mediator and outcome:

$$\mathbb{E}(M|C,A) = \theta_0 + \theta_1^T C^\perp + \theta_2 A + \theta_3 C^\perp A \tag{13}$$

$$\mathbb{E}(Y|C,A,L,M) = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM + \beta_6 C^\perp A, \tag{14}$$

which allow the effects of college completion on family income and depression to vary across levels of the baseline confounders. We estimate these models by first computing residuals for each of the baseline confounders C and post-treatment confounders L. This involves centering the elements of C around their sample means and centering the elements of L around their estimated conditional means given the past, which we compute from linear models that include C, A, and two-way interactions between C and A as predictors. We then compute least squares estimates of equations (13) and (14) using these residual terms, and finally, we construct estimates of the R-NDE, R-NIE, and R-ATE from their coefficients.

We find that completing college has a sizable overall effect on depression. Specifically, completing college is estimated to lower depression scores by 0.15 standard deviations on average (95% CI: [-0.28, -0.01]). The R-NDE and R-NIE are estimated to be -0.11 (95% CI: [-0.25, 0.03]) and -0.04 (95% CI: [-0.10, 0.005]), respectively. This suggests that only about 27% ($-0.04/-0.15 = 0.27$) of the overall effect is mediated by family income, although all of the estimates reported here are imprecise, as indicated by their wide confidence intervals.

To assess the robustness of our estimates to unobserved confounding, we also conducted a sensitivity analysis using methods outlined in Part F of the eAppendix. We find that our estimate of the R-NIE is highly sensitive to unobserved confounding of the

mediator-outcome relationship. Specifically, if the error terms from our models of family income and depression were negatively correlated, our estimate of the R-NIE would be biased downward, and a bias-adjusted estimate would reach zero under an error correlation as small as -0.12. This suggests that the effect of college completion on depression likely operates through pathways other than family income.

5 Discussion

Treatment-induced confounding complicates analyses of causal mediation. We proposed the method of RWR for decomposing an overall effect of treatment into direct and indirect components when treatment-induced confounding is present. The method involves, first, fitting a generalized linear model for the mediator with treatment and a set of baseline confounders as predictors, and second, fitting a linear regression of the outcome on treatment, the mediator, the 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 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 are misspecified, then estimates of direct and indirect effects may be biased. Part E of the eAppendix presents simulations that evaluate the sensitivity of RWR to incorrect model specification. An important direction for future research will be to explore the possibility of combining RWR with methods of model selection and regularization in an effort to improve its robustness. Another option would be to explore combining RWR with propensity score adjustment in a procedure similar to sequential g-estimation.[24]

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 and 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 identification assumptions on which RWR is based. To this end, we have developed methods for assessing the sensitivity of RWR to hypothetical patterns of unobserved confounding, as detailed in Part F of the eAppendix.

We focused on a two-way decomposition of an overall effect into randomized intervention analogues of natural direct and indirect effects, which is designed to evaluate mediation. The methods discussed previously can also be used to estimate more nuanced decompositions that evaluate the degree to which an effect is due to mediation versus interaction.[25–27] VanderWeele,[25] for example, decomposes a total effect into components due to mediation, interaction, both, or neither. In Part G of the eAppendix, we show that the components of this four-way decomposition, when defined in terms of randomized interventions, 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 to most applied researchers. Moreover, the method can be easily implemented with off-the-shelf software. We therefore expect that it will find wide application in analyses of causal mediation. 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 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 H of the eAppendix provides the R code for implementing RWR in our empirical example.

References

1. Danella M Hafeman and Sharon Schwartz. "Opening the Black Box: A Motivation for the Assessment of Mediation". *International Journal of Epidemiology* 38.3 (2009), pp. 838–845.

2. Judea Pearl. "Direct and Indirect Effects". *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. 2001, pp. 411–420.

3. Tyler J VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

4. Tyler J VanderWeele and Stijn Vansteelandt. "Conceptual Issues Concerning Mediation, Interventions and Composition". *Statistics and its Interface* 2.4 (2009), pp. 457–468.

5. Tyler J VanderWeele, Stijn Vansteelandt, and James M Robins. "Effect Decomposition in the Presence of an Exposure-induced Mediator-outcome Confounder". *Epidemiology* 25.2 (2014), pp. 300–306.

6. Stijn Vansteelandt and Rhian M Daniel. "Interventional Effects for Mediation Analysis with Multiple Mediators". *Epidemiology (Cambridge, Mass.)* 28.2 (2017), p. 258.

7. Vanessa Didelez, A Philip Dawid, and Sara Geneletti. "Direct and Indirect Effects of Sequential Treatments". *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2006, pp. 138–146.

8. Sara Geneletti. "Identifying Direct and Indirect Effects in a Non-counterfactual Framework". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (2007), pp. 199–215.

9. Chanelle J Howe et al. "Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias". *American Journal of Epidemiology* 173.5 (2011), pp. 569–577.

10. Ashley I Naimi et al. "Constructing Inverse Probability Weights for Continuous Exposures: A Comparison of Methods". *Epidemiology* 25.2 (2014), pp. 292–299.

11. Geoffrey T Wodtke. "Regression-based Adjustment for Time-varying Confounders". *Sociological Methods & Research (online access ahead of print)* (2018).

12. Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

13. James M Robins and Sander Greenland. "Identifiability and Exchangeability for Direct and Indirect Effects". *Epidemiology* 3.2 (1992), pp. 143–155.

14. Trang Q Nguyen, Ian Schmid, and Elizabeth A Stuart. "Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects based on What we Want to Learn". *arXiv preprint (https://arxiv.org/abs/1904.08515)* (2019).

15. Xiang Zhou and Geoffrey T Wodtke. "A Regression-with-residuals Method for Estimating Controlled Direct Effects". *Political Analysis* 27.3 (2019), pp. 360–369.

16. Daniel Almirall, Thomas Ten Have, and Susan A Murphy. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation". *Biometrics* 66.1 (2010), pp. 131–139.

17. Geoffrey T Wodtke and Daniel Almirall. "Estimating Moderated Causal Effects with Time-Varying Treatments and Time-Varying Moderators: Structural Nested Mean Models and Regression with Residuals". *Sociological Methodology* 47.1 (2017), pp. 212–245.

18. Geoffrey T Wodtke, Zahide Alaca, and Xiang Zhou. "Regression-with-residuals Estimation of Marginal effects: a Method of Adjusting for Treatment-induced Confounders that may also be Effect Modifiers". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2019).

19. Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC, 1994.

20. Felix Elwert and Christopher Winship. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable". *Annual Review of Sociology* 40 (2014), pp. 31–53.

21. Sander Greenland. "Quantifying Biases in Causal Models: Classical Confounding vs Collider-stratification Bias". *Epidemiology* 14.3 (2003), pp. 300–306.

22. James J Heckman, John Eric Humphries, and Gregory Veramendi. "The Nonmarket Benefits of Education and Ability". *Journal of Human Capital* 12.2 (2018), pp. 282–304.

23. Jinkook Lee. "Pathways from Education to Depression". *Journal of Cross-cultural Gerontology* 26.2 (2011), pp. 121–135.

24. Stijn Vansteelandt and Arvid Sjolander. "Revisiting G-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding". *Epidemiologic Methods* 5.1 (2016), pp. 37–56.

25. Tyler J VanderWeele. "A Unification of Mediation and Interaction: a Four-Way Decomposition". *Epidemiology* 25.5 (2014), p. 749.

26. Tyler VanderWeele. "A Three-way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects". *Epidemiology* 24.2 (2013), pp. 224–232.

27. Tyler VanderWeele and Eric Tchetgen Tchetgen. "Attributing Effects to Interactions". *Epidemiology* 25.5 (2014), pp. 711–722.

28. Judea Pearl. "The Foundations of Causal Inference". *Sociological Methodology* 40.1 (2010), pp. 75–149.

29. Michael E Sobel. "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models". *Sociological Methodology* 13.1 (1982), pp. 290–312.

eAppendix

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

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

and the R-NIE is equal to

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

B: Analytic Standard Errors for RWR

In this appendix, we outline an approach to obtaining analytic standard errors for 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, we assume that they are represented by a recursive system of equations with independent errors.[28] 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:

$$\begin{aligned} Y &= \hat{\alpha}^T X + \hat{\eta}^T L + \hat{\gamma}^T Z + \hat{Y}^{\perp} \\ &= \hat{\alpha}^T X + \sum_j \hat{\eta}_j L_j + \hat{\gamma}^T Z + \hat{Y}^{\perp}, \end{aligned} \tag{15}$$

where L_j is the jth 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

$$L_j = \hat{\lambda}_j^T X + \hat{L}_j^{\perp}, \tag{16}$$

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

Substituting (16) into (15) yields the following expression for the outcome:

$$Y = (\hat{\alpha}^T + \sum_j \hat{\eta}_j \hat{\lambda}_j^T) X + \sum_j \hat{\eta}_j \hat{L}_j^{\perp} + \hat{\gamma}^T Z + \hat{Y}^{\perp}. \tag{17}$$

Since $\hat{Y}^{\perp}$ is the least squares residual for regression (15), 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, $\{X, \hat{L}^{\perp}, Z\}$ and $\{X, L, Z\}$ span the same space. Thus, equation (17) represents the least squares fit of Y on $\{X, \hat{L}^{\perp}, Z\}$, 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 (15).

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

$$\hat{\mathbb{V}}(\hat{\alpha}_{RWR}) = \hat{\mathbb{V}}(\hat{\alpha}) + \widehat{\text{Cov}}(\hat{\alpha}, \hat{\eta})\hat{\Lambda}^T + \hat{\Lambda}\hat{\mathbb{V}}(\hat{\eta})\hat{\Lambda}^T + \sum_{j,k} \hat{\eta}_j \hat{\eta}_k \widehat{\text{Cov}}(\hat{\lambda}_j, \hat{\lambda}_k), \tag{18}$$

where $\hat{\Lambda} = [\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_q]$ is a $p \times q$ matrix of estimated coefficients from model (16), $\mathbb{V}(\cdot)$ is the variance-covariance matrix of a random vector, and $\text{Cov}(\cdot, \cdot)$ is the covariance matrix between two random vectors. In equation (18), the first three terms can be obtained directly from the naive regression (15), and the covariance matrix between $\hat{\lambda}_j$ and $\hat{\lambda}_k$ in the last term can be estimated as

$$\widehat{\text{Cov}}(\hat{\lambda}_j, \hat{\lambda}_k) = \frac{(\hat{\boldsymbol{l}}_j^{\perp})^T \hat{\boldsymbol{l}}_k^{\perp}}{n - p} (X^T X)^{-1}, \tag{19}$$

where n is the sample size, and $\hat{\boldsymbol{l}}_j^{\perp}$ and $\hat{\boldsymbol{l}}_k^{\perp}$ 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

$$\widehat{\text{Cov}}(\hat{\alpha}_{RWR}, \hat{\gamma}) = \widehat{\text{Cov}}(\hat{\alpha}, \hat{\gamma}) + \hat{\Lambda}\widehat{\text{Cov}}(\hat{\eta}, \hat{\gamma}). \tag{20}$$

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[29]:

$$\hat{\mathbb{V}}[\widehat{\text{R-NDE}}] = \hat{\mathbb{V}}(\hat{\beta}_2) + (\hat{\theta}_0 + \hat{\theta}_2 a)^2 \hat{\mathbb{V}}(\hat{\beta}_5) + \hat{\beta}_5^2 \hat{\mathbb{V}}(\hat{\theta}_0 + \hat{\theta}_2 a) + (\hat{\theta}_0 + \hat{\theta}_2 a)\widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_5) \tag{21}$$

$$\hat{\mathbb{V}}[\widehat{\text{R-NIE}}] = (\hat{\beta}_4 + \hat{\beta}_5 a^*)^2 \hat{\mathbb{V}}(\hat{\theta}_2) + \hat{\theta}_2^2 \hat{\mathbb{V}}(\hat{\beta}_4 + \hat{\beta}_5 a^*). \tag{22}$$

In these equations, the terms involving $\hat{\theta}_0$ and $\hat{\theta}_2$ can be estimated by applying equations (18-20) 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 (18-20) 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 are 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

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

D: Empirical Illustration with a Binary Outcome

In this section, we present a parallel analysis of the NLSY79 in which the outcome, Y, is coded as a binary variable for illustrative purposes. Specifically, Y is coded 1 if a respondent scored in the top quintile of the CES-D distribution, indicating he or she is among the most depressed 20% of the population, and 0 otherwise. All other variables are defined as in Section 4 from the main text. We use the following models for the mediator and outcome:

$$\mathbb{E}(M|C,A) = \theta_0 + \theta_1^T C^\perp + \theta_2 A + \theta_3 C^\perp A \tag{23}$$

$$\mathbb{E}(Y|C,A,L,M) = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM + \beta_6 C^\perp A, \tag{24}$$

where $\mathbb{E}(Y|C,A,L,M) = P(Y = 1|C,A,L,M)$ and thus our model for the outcome is a linear probability model. 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. Specifically, we estimate conditional means for each post-treatment confounder from a linear regression that includes C, A, and all two-way interactions between them. We then compute least squares estimates of equations (13) and (14), and finally, we use their coefficients to construct RWR estimates of the R-NDE, R-NIE, and R-ATE.

Consistent with our analysis of continuous scores from the CES-D, results based on the binary measure described previously also suggest that completing a post-secondary education has a sizable overall effect on the risk of depression. Specifically, completing college is estimated to lower the risk of depression by 7.7 percentage points (95% CI: [-0.129,-0.021]). The R-NDE is estimated to be -0.058 (95% CI: [-0.105, -0.003]), which suggests that attending college would still reduce the risk of depression by 5.8 percentage points even after an intervention to fix the income distribution to that observed when nobody receives a post-secondary education. The R-NIE is estimated to be -0.020 (95% CI: [-0.042,-0.003]. This suggests that, if everyone already attended college, the risk of depression would be further reduced by only about 2 percentage points after an intervention to shift the income distribution to that observed when everyone attends college from that observed when nobody attends. Estimates of the R-NDE and R-NIE provide some minimal evidence of mediation, although they are fairly imprecise, as indicated by their wide confidence intervals.

These results are based on a linear probability model for the outcome. As with any model, researchers should consider the possibility of bias due to misspecification and take steps to avoid it (e.g., by using regression diagnostics, interaction terms, model selection techniques, and subject matter knowledge). With a linear model for a strictly bounded outcome, researchers should take additional precautions to ensure that it is a reasonable approximation for the true conditional expectation function and does not suffer from severe misspecification. For example, with a linear probability model, researchers should confirm that it does not yield many nonsensical predictions well outside the logical $[0, 1]$ range.

E: Simulation Study of Bias due to Model Misspecification

Even when assumptions (i) to (iii) are satisfied, RWR may still yield biased estimates if models for the outcome, mediator, and/or treatment-induced confounders are incorrectly specified. In this section, we use a series of simulation experiments to investigate the sensitivity of RWR to several types of model misspecification, including incorrectly modeled effect modification and non-linearity.

Specifically, we simulate $n = 500$ observations and estimate the R-NDE and R-NIE of a binary treatment A on a continuous outcome Y via a continuous mediator M in the presence of a baseline confounder C, a treatment-induced confounder L, and an "unobserved" variable U that affects both L and Y but not A or M. In each simulation, we generate these variables as follows: $U \sim N(\mu_U = 0, \sigma_U = 1)$; $C \sim N(\mu_C = 0, \sigma_C = 1)$; $A|C \sim Bernoulli(\pi_{A|C} = \Phi(-0.3 + 0.5C))$; $L|U, C, A \sim N(\mu_{L|U,C,A} = 0.5U + C(0.5 - \eta C) + A(0.5 + \alpha C), \sigma_{L|U,C,A} = 1)$; $M|C, A, L \sim N(\mu_{M|C,A,L} = C(0.5 - \eta C) + A(0.5 + \alpha C) + 0.5(L - \mu_{L|C,A}), \sigma_{L|U,C,A} = 1)$; and $Y|U, C, A, L, M \sim N(\mu_{Y|U,C,A,L,M} = 0.5U + C(0.5 - \eta C) + A(0.5 + \alpha C) + (L - \mu_{L|C,A})(0.5 + \lambda C) + M(1 + \gamma C - 0.2A), \sigma_{Y|U,C,A,L,M} = 1)$. Here, Φ is the standard normal cumulative distribution function; α is a parameter that controls the degree to which C modifies the effects of A on L, M, and Y; γ is a parameter that controls the degree to which C modifies the effect of M on Y; λ is a parameter that controls the degree to which C modifies the effect of L on Y; and finally, η is a parameter that controls whether the effects of C on L, M, and Y are linear versus parabolic. In all simulations, the R-NDE and R-NIE are identified, and their true values are 0.5 and 0.4, respectively.

With these data, we implement RWR exactly as outlined in Section 3.1. That is, in all simulations, we implement RWR by fitting a model for L that is linear and additive in A and C, a model for M that is linear and additive in A and C, and a model for Y that is linear and additive in C and L but multiplicative in A and M. These modeling constraints are satisfied in some simulations but not in others, as we vary the values of $\{\alpha, \gamma, \lambda, \eta\}$ to introduce different types of effect modification and non-linearity. We then evaluate the performance of RWR in terms of its absolute bias, the magnitude of its absolute bias relative to the true effect of interest, its root mean squared error (RMSE), and the magnitude

of its RMSE relative to the RMSE of the RWR estimator with correctly specified models, which are computed from 10,000 simulated datasets in each experiment.[c]

Table 1 presents results from a set of simulation experiments that evaluate the performance of RWR when its models for L, M, and Y are incorrectly specified because the effects of treatment on these variables are constrained to be invariant when in fact they differ across levels of C. The effects of A on L, M, and Y are made to differ across C by varying the value of α from 0.0 to 0.5 while setting all other tuning parameters equal to zero. When $\alpha = 0.0$, the effects of treatment are invariant, and there is no model misspecification. When $\alpha = 0.5$, by contrast, the unit-specific effects of treatment have a standard deviation as large as their mean, and models that constrain these effects to be invariant are badly misspecified. Results show that RWR is biased for the R-NDE and R-NIE when its models for L, M, and Y incorrectly constrain the effects of A to be invariant in C, as expected. The magnitude of bias, however, is not especially large in this particular scenario.

Table 2 presents results from a second set of simulation experiments that evaluate the performance of RWR when its model for Y is incorrectly specified because the effects of the mediator on the outcome are constrained to be invariant when in fact they differs across levels of C. The effect of M on Y is made to differ across C by varying the value of γ from 0.0 to 0.5 while setting all other tuning parameters equal to zero. When $\gamma = 0.0$, the effect of the mediator is invariant, and there is no model misspecification. When $\gamma = 0.5$, the unit-specific effects of the mediator vary considerably in C, and thus an outcome model that constrains these effects to be invariant is badly misspecified. Consistent with findings from Table 1, the results in Table 2 also demonstrate that RWR is biased when its outcome model is misspecified, in this case because it incorrectly constrains the effects of M on Y to be invariant across C.

Table 3 presents results from a third set of simulation experiments that evaluate the performance of RWR when its model for Y is incorrectly specified because the effects of L on Y are constrained to be invariant when in fact they differ across levels of C. The effects of L on Y are made to differ across C by varying the value of λ from 0.0, in which case there is no effect modification, to 0.5, in which case the unit-specific effects of L have a

[c]The relative bias is computed as the ratio of the absolute bias to the true effect of interest.

Table 1: Misspecification bias in RWR due to incorrectly modeled $A \rightarrow Y$, $A \rightarrow M$, and $A \rightarrow L$ effect modification by C

	α					
	0.0	0.1	0.2	0.3	0.4	0.5
R-NDE						
Absolute Bias	0.003	0.009	0.020	0.026	0.034	0.046
Relative Bias	0.005	0.019	0.041	0.052	0.069	0.091
RMSE	0.140	0.142	0.142	0.143	0.144	0.148
Relative RMSE	1.000	1.016	1.012	1.018	1.029	1.053
R-NIE						
Absolute Bias	0.002	0.014	0.032	0.049	0.067	0.086
Relative Bias	0.005	0.035	0.079	0.121	0.167	0.216
RMSE	0.096	0.099	0.104	0.112	0.123	0.135
Relative RMSE	1.000	1.033	1.084	1.164	1.273	1.403

Note: Results are based on 10,000 simulations. Across all simulations, $\gamma = 0$, $\lambda = 0$, and $\eta = 0$.

standard deviation as large as their mean. As before, the remaining tuning parameters are all set to zero. Consistent with the results discussed previously, this set of simulations shows that RWR is also biased when its outcome model incorrectly constrains the effects of L on Y to be invariant across C, although the magnitude of this bias is fairly small across all scenarios.

Finally, Table 4 presents results from a fourth set of simulation experiments that evaluate the performance of RWR when its models for L, M, and Y are incorrectly specified because the effects of C on these variables are assumed to be linear when in fact they are parabolic. The effects of C on L, M, and Y are made to be nonlinear by varying the value of η from 0.0 to 0.5 while setting all other tuning parameters to zero. As η increases from zero, the effects of C become increasingly nonlinear, and the results in Table 4 show that RWR becomes increasingly biased, as expected. Nevertheless, the bias due to nonlinearity in these simulations is generally small. For reference, we also computed a set of naive regression estimates that do not adjust for the treatment-induced confounder L but that are otherwise based on correct models for $\mathbb{E}(M|C, A)$ and $\mathbb{E}(Y|C, A, M)$. These estimates suffer from bias due to uncontrolled mediator-outcome confounding by L but

Table 2: Misspecification bias in RWR due to incorrectly modeled $M \rightarrow Y$ effect modification by C

	γ					
	0.0	0.1	0.2	0.3	0.4	0.5
R-NDE						
Absolute Bias	0.003	-0.025	-0.048	-0.075	-0.099	-0.123
Relative Bias	0.005	-0.050	-0.096	-0.150	-0.198	-0.245
RMSE	0.140	0.143	0.148	0.161	0.176	0.195
Relative RMSE	1.000	1.024	1.060	1.150	1.259	1.390
R-NIE						
Absolute Bias	0.002	0.028	0.059	0.088	0.118	0.149
Relative Bias	0.005	0.070	0.147	0.221	0.296	0.373
RMSE	0.096	0.108	0.126	0.149	0.174	0.201
Relative RMSE	1.000	1.126	1.312	1.544	1.810	2.092

Note: Results are based on 10,000 simulations. Across all simulations, $\alpha = 0$, $\lambda = 0$, and $\eta = 0$.

Table 3: Misspecification bias in RWR due to incorrectly modeled $L \rightarrow Y$ effect modification by C

	λ					
	0.0	0.1	0.2	0.3	0.4	0.5
R-NDE						
Absolute Bias	0.003	-0.012	-0.022	-0.035	-0.048	-0.056
Relative Bias	0.005	-0.024	-0.044	-0.071	-0.095	-0.113
RMSE	0.140	0.143	0.144	0.148	0.154	0.159
Relative RMSE	1.000	1.020	1.026	1.053	1.098	1.137
R-NIE						
Absolute Bias	0.002	0.008	0.018	0.026	0.037	0.046
Relative Bias	0.005	0.019	0.044	0.066	0.091	0.116
RMSE	0.096	0.100	0.104	0.108	0.114	0.120
Relative RMSE	1.000	1.042	1.078	1.125	1.186	1.242

Note: Results are based on 10,000 simulations. Across all simulations, $\alpha = 0$, $\gamma = 0$, and $\eta = 0$.

Table 4: Misspecification bias in RWR due to incorrectly modeled non-linearity in the $C \rightarrow Y$, $C \rightarrow M$, and $C \rightarrow L$ effects

	η					
	0.0	0.1	0.2	0.3	0.4	0.5
R-NDE						
Absolute Bias	0.003	0.012	0.023	0.028	0.032	0.038
Relative Bias	0.005	0.024	0.046	0.056	0.063	0.076
RMSE	0.14	0.144	0.145	0.148	0.153	0.158
Relative RMSE	1.000	1.026	1.036	1.057	1.090	1.125
R-NIE						
Absolute Bias	0.002	-0.011	-0.017	-0.021	-0.023	-0.025
Relative Bias	0.005	-0.029	-0.044	-0.052	-0.059	-0.062
RMSE	0.096	0.098	0.100	0.104	0.108	0.112
Relative RMSE	1.000	1.020	1.039	1.077	1.119	1.164

Note: Results are based on 10,000 simulations. Across all simulations, $\alpha = 0$, $\gamma = 0$, and $\lambda = 0$.

not from bias due to effect modification or nonlinearity that has been incorrectly modeled, as above. Results from this ancillary analysis indicate that naive regression estimates understate the true R-NDE by -0.165, or 32.9 percent, and that they overstate the true R-NIE by 0.169, or 42.3 percent. Thus, in the simulations considered here where the magnitude of confounding is fairly large, bias arising from incorrect model specification is generally less severe than bias arising from uncontrolled mediator-outcome confounding.

F: 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. In this section, we outline a parametric approach to sensitivity analysis that permits an assessment of whether RWR estimates are robust to violations of these three assumptions.

Consider the following set of linear structural equations characterizing the true causal relationships between A, M, and Y:

$$A = \gamma_0 + \gamma_1^T C^\perp + \epsilon_A, \tag{25}$$

$$M = \theta_0 + \theta_1^T C^\perp + \theta_2 A + \epsilon_M, \tag{26}$$

$$Y = \beta_0 + \beta_1^T C^\perp + \beta_2 A + \beta_3^T L^\perp + \beta_4 M + \beta_5 AM + \epsilon_Y. \tag{27}$$

The assumptions of no unobserved confounding imply that the error terms $(\epsilon_A, \epsilon_M, \epsilon_Y)$ are pairwise independent.

When the mediator-outcome relationship is confounded by unobserved factors (but not the treatment-outcome or treatment-mediator relationships), ϵ_M and ϵ_Y are correlated. A linear projection of ϵ_Y on ϵ_M can be expressed as

$$\epsilon_Y = \phi_{MY}\epsilon_M + \psi_{MY}. \tag{28}$$

Under the assumption that $\mathbb{E}[\psi_{MY}|C, A, L, M] = 0$, substituting (28) into (27) and taking the conditional expectation of Y yields

$$\begin{aligned}\mathbb{E}[Y|C, A, L, M] = (\beta_0 - \phi_{MY}\theta_0) + (\beta_1 - \phi_{MY}\theta_1)^T C^\perp + (\beta_2 - \phi_{MY}\theta_2)A + \beta_3^T L^\perp + \\ (\beta_4 + \phi_{MY})M + \beta_5 AM.\end{aligned} \tag{29}$$

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

$$\text{Bias[R-NDE]} = -\phi_{MY}\theta_2(a^* - a) \tag{30}$$

$$\text{Bias[R-NIE]} = \phi_{MY}\theta_2(a^* - a). \tag{31}$$

The biases for the R-NDE and R-NIE are equal in magnitude but opposite in direction. This implies that the overall 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_{MY} 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_{MY} is known. Moreover, ϕ_{MY} is not on an interpretable scale. To circumvent this problem, ϕ_{MY} can be re-expressed in terms of the correlation between ϵ_Y and ϵ_M as follows:

$$\phi_{MY} = \frac{\text{sd}(\psi_{MY})}{\text{sd}(\epsilon_M)} \frac{\rho_{MY}}{\sqrt{1-\rho_{MY}^2}}, \tag{32}$$

where $\rho_{MY} = \text{Corr}[\epsilon_Y, \epsilon_M]$. Substituting (32) into (30) yields

$$\text{Bias[R-NDE]} = -\frac{\theta_2 \cdot \text{sd}(\psi_{MY})}{\text{sd}(\epsilon_M)} \frac{\rho_{MY}}{\sqrt{1-\rho_{MY}^2}} (a^* - a). \tag{33}$$

The bias for the R-NIE can be expressed analogously. Under the assumptions of no unobserved treatment-mediator or treatment-outcome confounding, θ_2, $\text{sd}(\epsilon_M)$, and $\text{sd}(\psi_{MY})$ 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_{MY} and construct a range of bias-adjusted RWR estimates for the R-NDE and R-NIE across different values of ρ_{MY}. In addition, we can identify the value of ρ_{MY} that would suffice to reduce the estimated R-NDE or R-NIE to zero, or alternatively, the value that would suffice to render the estimated R-NDE or R-NIE statistically insignificant.

Next, consider the case where the treatment-outcome relationship is confounded by unobserved factors (but not the treatment-mediator or mediator-outcome relationships). In this case, ϵ_A and ϵ_Y are correlated, and a linear projection of ϵ_Y on ϵ_A can be expressed as

$$\epsilon_Y = \phi_{AY}\epsilon_A + \psi_{AY}. \tag{34}$$

Under the assumption that $\mathbb{E}[\psi_{AY}|C, A, L, M] = 0$, substituting (34) into (27) and taking

the conditional expectation of Y yields

$$\mathbb{E}[Y|C,A,L,M] = (\beta_0 - \phi_{AY}\gamma_0) + (\beta_1 - \phi_{AY}\gamma_1)^T C^{\perp} + (\beta_2 + \phi_{AY})A + \beta_3^T L^{\perp} + \beta_4 M + \beta_5 AM.$$

Thus, in this case, RWR estimates of $(\beta_0, \beta_1, \beta_2)$ suffer from an asymptotic bias of $\phi_{AY}(-\gamma_0, -\gamma_1, 1)$. Accordingly, the bias for the RWR estimator of the R-NDE can be expressed as

$$\text{Bias[R-NDE]} = \phi_{AY}(a^* - a),$$

and because the treatment-mediator and mediator-outcome relationships are unconfounded, the RWR estimator of the R-NIE is asymptotically unbiased. As before, the bias for the R-NDE can also be expressed as a function of $\rho_{AY} = \text{Corr}(\epsilon_A, \epsilon_Y)$:

$$\text{Bias[R-NDE]} = \frac{\text{sd}(\psi_{AY})}{\text{sd}(\epsilon_A)} \frac{\rho_{AY}}{\sqrt{1-\rho_{AY}^2}}(a^* - a).$$

Finally, consider the case where the treatment-mediator relationship is confounded by unobserved factors (but not the treatment-outcome or mediator-outcome relationships). In this case, ϵ_A and ϵ_M are correlated, and a linear projection of ϵ_M on ϵ_A can be expressed as

$$\epsilon_M = \phi_{AM}\epsilon_A + \psi_{AM}. \tag{35}$$

Under the assumption that $\mathbb{E}[\psi_{AM}|C,A] = 0$, substituting (35) into (26) and taking the conditional expectation of M yields

$$\mathbb{E}[M|C,A] = (\theta_0 - \phi_{AM}\gamma_0) + (\theta_1 - \phi_{AM}\gamma_1)^T C^{\perp} + (\theta_2 + \phi_{AM})A$$

In this case, RWR estimates of $(\theta_0, \theta_1, \theta_2)$ suffer from an asymptotic bias of $\phi_{AM}(-\gamma_0, -\gamma_1, 1)$. Accordingly, bias terms for the RWR estimators of the R-NDE and

R-NIE can be expressed as

$$\text{Bias[R-NDE]} = \phi_{AM}\beta_5(a-\gamma_0)(a^*-a),$$
$$\text{Bias[R-NIE]} = \phi_{AM}(\beta_4+\beta_5 a^*)(a^*-a).$$

Defining $\rho_{AM} = \text{Corr}(\epsilon_A, \epsilon_M)$, the above formulas can also be expressed as

$$\text{Bias[R-NDE]} = \frac{\text{sd}(\psi_{AM})}{\text{sd}(\epsilon_A)}\frac{\rho_{AM}}{\sqrt{1-\rho^2_{AM}}}\beta_5(a-\gamma_0)(a^*-a),$$
$$\text{Bias[R-NIE]} = \frac{\text{sd}(\psi_{AM})}{\text{sd}(\epsilon_A)}\frac{\rho_{AM}}{\sqrt{1-\rho^2_{AM}}}(\beta_4+\beta_5 a^*)(a^*-a),$$

where $\text{sd}(\epsilon_A)$ and $\text{sd}(\psi_{AM})$ can be estimated by fitting models (25) and (26) to the observed data.

G: A Four-way Decomposition

As shown by VanderWeele,[25] the R-NDE can be further decomposed into the following two components:

$$\begin{aligned}\text{R-NDE} &= \mathbb{E}(Y_{a^*m} - Y_{am}) + [\mathbb{E}(Y_{a^*G_{a|C}} - Y_{aG_{a|C}}) - \mathbb{E}(Y_{a^*m} - Y_{am})] \\ &= \text{CDE}(m) + \text{R-INT}_{ref}(m). \end{aligned} \tag{36}$$

The first term in (36), $\text{CDE}(m) = \mathbb{E}(Y_{a^*m} - Y_{am})$, 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 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

$$\begin{aligned}\text{CDE}(m) &= \sum_c \sum_l [\mathbb{E}(Y|c, a^*, l, m)P(l|c, a^*) - \mathbb{E}(Y|c, a, l, m)P(l|c, a)]P(c) \\ &= (\beta_2 + \beta_5 m)(a^* - a). \end{aligned} \tag{37}$$

By extension, the reference interaction effect is equal to

$$\text{R-INT}_{ref}(m) = \text{R-NDE} - \text{CDE}(m) = \beta_5(\theta_0 + \theta_2 a - m)(a^* - a). \tag{38}$$

VanderWeele[25] also shows that the R-NIE can be further decomposed as follows:

$$\begin{aligned}\text{R-NIE} &= \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}}) + [\mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{a^*G_{a|C}}) - \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}})] \\ &= \text{R-PIE} + \text{R-INT}_{med}. \end{aligned} \tag{39}$$

The first term in (39), $\text{R-PIE} = \mathbb{E}(Y_{aG_{a^*|C}} - Y_{aG_{a|C}})$, 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

$$\begin{aligned}\text{R-PIE} &= \sum_c \sum_m \sum_l [P(m|c, a^*) - P(m|c, a)]\mathbb{E}(Y|c, a, l, m)P(l|c, a)P(c). \\ &= \theta_2(\beta_4 + \beta_5 a)(a^* - a) \end{aligned} \tag{40}$$

By extension, the mediated interaction effect is equal to

$$\text{R-INT}_{med} = \text{R-NIE} - \text{R-PIE} = \theta_2\beta_5(a^* - a)^2. \tag{41}$$

And thus the randomized intervention analogue to the average total effect can be ex-

pressed as

$$\text{R-ATE} = \text{R-NDE} + \text{R-NIE} = \text{CDE}(m) + \text{R-INT}_{ref}(m) + \text{R-PIE} + \text{R-INT}_{med}. \quad (42)$$

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 (37-38) and (40-41). Standard errors and confidence intervals can be computed using either the non-parametric bootstrap or the analytic approach outlined in Part B of the eAppendix.

H: Implementation of RWR in R

Below, we illustrate the implementation of RWR in R for estimating the R-NDE and R-NIE of college completion on depression. The output also includes the four-component decomposition outlined in Part G of the eAppendix.

```
# 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 transformation #
depression$ihsinc<-log(depression$tfinc_dest_b+sqrt(depression$tfinc_dest_b+1))

# mediator and outcome equations #
m_form <- ihsinc ~ (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 + ihsinc + college * ihsinc +
  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)
```