

Effect Decomposition in the Presence of Treatment-induced Confounding

A Regression-with-residuals Approach

Geoffrey T. Wodtke^a and Xiang Zhou^b

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 (rNDE) and the natural indirect effect (rNIE), 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 rNDE and rNIE 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 rNDE and rNIE 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

(*Epidemiology* 2020;31: 369–375)

Submitted May 14, 2019; accepted January 23, 2020.

From the ^aDepartment of Sociology, University of Chicago, Chicago, IL; and ^bDepartment of Sociology, Harvard University, Cambridge, MA.

Supported by a grant from the Social Sciences and Humanities Research Council of Canada (Grant No. 435-2018-0736).

Disclosure: The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

The study is exempt from IRB review because it involves only minimal risk and anonymous secondary data. Replication files are available from the corresponding author upon request.

Correspondence: Geoffrey T. Wodtke, Department of Sociology, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637. E-mail: wodtke@uchicago.edu.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 1044-3983/20/3103-0369

DOI: 10.1097/EDE.0000000000001168

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 (1) no unobserved treatment–outcome confounding, (2) no unobserved treatment–mediator confounding, (3) no unobserved mediator–outcome confounding, and (4) 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 (rNDE) and the natural indirect effect (rNIE), which can be identified in the presence of treatment-induced confounding (see also Didelez et al⁷ and Geneletti⁸). The rNDE and rNIE 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 (1) to (3) above but not assumption (4).

Estimating the rNDE and rNIE, however, remains difficult. VanderWeele and colleagues⁵ 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., VanderWeele⁴).

In this article, we introduce a new method termed “regression-with-residuals” for estimating the rNDE and rNIE. 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 (1) a generalized linear model for the conditional mean of the mediator given treatment and a set of baseline confounders, (2) 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, (3) 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 rNDE and rNIE are given by simple functions of their coefficients. Regression-with-residuals estimates are consistent and asymptotically unbiased when assumptions (1) to (3) 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 rNDE and rNIE and outlining the conditions under which they can be identified. Then, we introduce regression-with-residuals 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.

NOTATION, ESTIMANDS, AND IDENTIFICATION

We adopt the notation used by VanderWeele et al.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

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

This estimand represents the expected difference in the outcome if all individuals in some target population were exposed to treatment 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 an 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

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

This estimand represents the expected difference in the outcome if all individuals were exposed to treatment 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 rNDE and rNIE is equal to the randomized intervention analogue of the total effect:

$$\text{rATE} = \text{rNDE} + \text{rNIE} = \mathbb{E}(Y_{a^*G_{a^*|C}} - Y_{aG_{a|C}}). \quad (3)$$

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.^{5,7,8}

The rNDE and rNIE can be identified from observed data under the following conditional independence assumptions: (1) $Y_{am} \perp\!\!\!\perp A \mid C$, (2) $M_a \perp\!\!\!\perp A \mid C$, and (3) $Y_{am} \perp\!\!\!\perp M \mid C, A, L$.⁵ In words, assumption (1) requires that there must not be any unobserved treatment–outcome confounders conditional on C . Assumption (2) requires that there must not be any unobserved treatment–mediator confounders conditional on C . And assumption (3) requires that there must not be any unobserved mediator–outcome confounders conditional on C , A , and L . 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 rNDE and rNIE 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 presents a directed acyclic graph in which all of the conditional independence assumptions are satisfied, as there are not any unobserved variables that jointly affect treatment,

the mediator, or the outcome.¹² In this situation, the rNDE and rNIE can be expressed in terms of the observed data as follows:

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

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

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 (4) $Y_{am} \perp\!\!\!\perp M_{a^*} \mid C$. 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 rNDE and rNIE 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 hypothetical modifications to treatment, for example “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.

REGRESSION-WITH-RESIDUALS ESTIMATION

Regression-with-residuals 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 regression-with-residuals can also be used to decompose causal effects in the presence of treatment-induced confounding into direct and indirect components. For simplicity, we introduce regression-with-residuals by focusing on its implementation with linear models for the outcome, mediator, and treatment-induced confounders. Later, we explain how regression-with-residuals can also be implemented with a more general class of models for the mediator and treatment-induced confounders.

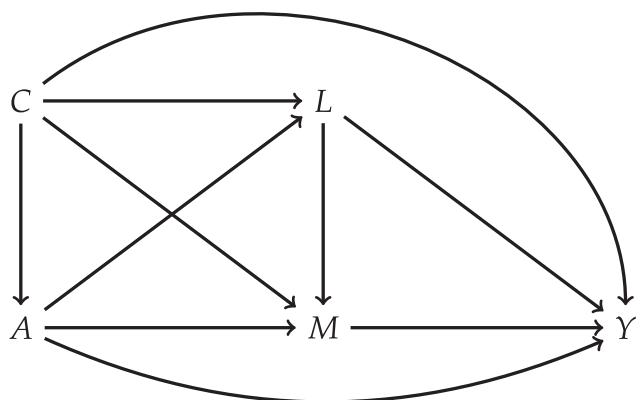


FIGURE 1. Causal graph with treatment A , mediator M , outcome Y , baseline confounders C , and posttreatment confounders L .

Linear Regression-with-residuals

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

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

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 posttreatment 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 posttreatment confounder given treatment and the baseline confounders. These models can be expressed as

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

Under assumptions (1) to (3) 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 rNDE and rNIE are equal to

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

$$\text{rNIE} = \theta_2(\beta_4 + \beta_5 a^*)(a^* - a), \quad (10)$$

and the rATE is equal to their sum. A derivation of these parametric expressions is provided in eAppendix 1; <http://links.lww.com/EDE/B651>.

Regression-with-residuals 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 posttreatment 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 $\hat{\mathbb{E}}(Y | C, A, L, M) = \hat{\beta}_0 + \hat{\beta}_1^T \hat{C}^\perp + \hat{\beta}_2 A + \hat{\beta}_3^T \hat{L}^\perp + \hat{\beta}_4 M + \hat{\beta}_5 A M$.
5. Compute $\widehat{\text{rNDE}} = [\hat{\beta}_2 + \hat{\beta}_5(\hat{\theta}_0 + \hat{\theta}_2 a)](a^* - a)$ and $\widehat{\text{rNIE}} = \hat{\theta}_2(\hat{\beta}_4 + \hat{\beta}_5 a^*)(a^* - a)$.

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, eAppendix 2; <http://links.lww.com/EDE/B651> provides analytic standard errors obtained using the delta method.

Adjustment for posttreatment 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. Regression-with-residuals avoids these problems by adjusting only for residual transformations of the posttreatment confounders. Because the residualized confounders are purged of their association with treatment, adjusting for them in a regression model for Y is unproblematic.

Extensions

Regression-with-residuals requires correctly specified models for the outcome, mediator, and posttreatment confounders. The model for the outcome must be linear, and thus regression-with-residuals 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, regression-with-residuals 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 rNIE and rNDE proceeds as outlined previously.

Regression-with-residuals 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 rNDE and rNIE will depend on levels of the baseline confounders C and the model used for the mediator M . In general, they are given by

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

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

A derivation of these expressions is provided in eAppendix 3; <http://links.lww.com/EDE/B651>.

Similarly, regression-with-residuals 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 regression-with-residuals is that the parametric expressions for the rNDE and rNIE are insensitive to the choice of models for the posttreatment confounders. Thus, regardless of the models used to residualize L , computation of the rNIE and rNDE proceeds exactly as outlined previously.

Because regression-with-residuals 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 regression-with-residuals.

EMPIRICAL ILLUSTRATION

In this section, we decompose the effect of postsecondary education on depression into direct and indirect components using regression-with-residuals. 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 40 years of age. We standardize CES-D scores to have mean zero and unit variance, where higher scores imply more depressive symptoms (in eAppendix 4; <http://links.lww.com/EDE/B651>, we present a parallel analysis in which the outcome is coded instead as a binary variable). The treatment, A , is defined as completion of a 4-year college degree by 25 years of age. The mediator, M , is the inverse hyperbolic sine of a respondent's equivalized family income averaged over 35–40 years of age (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). 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 13–17 years of age. Finally, the vector of posttreatment confounders, L , includes CES-D scores measured when respondents were 27–30 years of age, 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:

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

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

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 posttreatment 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 rNDE, rNIE, and rATE from their coefficients.

We estimate 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 rNDE and rNIE 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 the eAppendix; <http://links.lww.com/EDE/B651>. We find that our estimate of the rNIE is 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 rNIE 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.

DISCUSSION

Treatment-induced confounding complicates analyses of causal mediation. We proposed the method of regression-with-residuals 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 posttreatment confounders that have been residualized with respect to the observed past.

Estimates of the rNDE and rNIE 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, regression-with-residuals requires correct models for the conditional mean of the mediator, the outcome, and each of the posttreatment confounders. If any of these models are misspecified, then estimates of direct and indirect effects may be biased. In eAppendix 5; <http://links.lww.com/EDE/B651>, we present simulations that evaluate the sensitivity of regression-with-residuals to incorrect model specification. An important direction for future research will be to explore the possibility of combining regression-with-residuals with methods of model selection and regularization in an effort to improve its robustness. Another option would be to explore combining regression-with-residuals with propensity score adjustment in a procedure similar to sequential g-estimation.²⁴

Regression-with-residuals 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 regression-with-residuals 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 regression-with-residuals is based. To this end, we have developed methods for assessing the sensitivity of regression-with-residuals to hypothetical patterns of unobserved confounding, as detailed in eAppendix 6; <http://links.lww.com/EDE/B651>.

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 eAppendix 7; <http://links.lww.com/EDE/B651>, we show that the components of this four-way decomposition, when defined in terms of randomized interventions, can also be estimated with regression-with-residuals.

Because regression-with-residuals 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 regression-with-residuals. The R package is available at <https://github.com/xiangzhou09/rwrmed> and the Stata package at <https://github.com/gtwodtke/rwrmed>.

In addition, eAppendix 8; <http://links.lww.com/EDE/B651> provides the R code for implementing regression-with-residuals in our empirical example.

REFERENCES

1. Hafeman DM, Schwartz S. Opening the black box: a motivation for the assessment of mediation. *Int J Epidemiol*. 2009;38:838–845.
2. Pearl J. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. 411–420. Morgan Kaufmann Publishers Inc; 2001.
3. VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press; 2015.
4. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*. 2009;2:457–468.
5. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014;25:300–306.
6. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)*. 2017;28:258.
7. Didelez V, Dawid AP, Geneletti S. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 138–146. AUAI Press; 2006.
8. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J R Stat Soc Series B Stat Methodol*. 2007;69:199–215.
9. Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am J Epidemiol*. 2011;173:569–577.
10. Naimi AI, Moodie EM, Auger N, Kaufman JS. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*. 2014;25:292–299.
11. Wodtke GT. Regression-based adjustment for time-varying confounders. *Sociological Methods & Research (online access ahead of print)*. 2018.
12. Pearl J. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press; 2009.
13. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3:143–155.
14. Nguyen TQ, Schmid I, Stuart EA. 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. Zhou X, Wodtke GT. A regression-with-residuals method for estimating controlled direct effects. *Political Analysis*. 2019;27:360–369.
16. Almirall D, Have TT, Murphy SA. Structural nested mean models for assessing time-varying effect moderation. *Biometrics*. 2010;66:131–139.
17. Wodtke GT, Almirall D. Estimating moderated causal effects with time-varying treatments and time-varying moderators: Structural nested mean models and regression with residuals. *Sociological Methodology*. 2017;47:212–245.
18. Wodtke GT, Alaca Z, Zhou X. Regression-with-residuals estimation of marginal effects: a method of adjusting for treatment-induced confounders that may also be effect modifiers. *J R Stat Soc Series A Stat Soc*. 2019;18:311–332.
19. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall/CRC; New York: 1994.
20. Elwert F, Winship C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*. 2014;40:31–53.
21. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300–306.

22. Heckman JJ, Humphries JE, Veramendi G. The nonmarket benefits of education and ability. *Journal of Human Capital*. 2018;12:282–304.
23. Lee J. Pathways from education to depression. *J Cross cult Gerontol*. 2011;26:121–135.
24. Vansteelandt S, Sjolander A. Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods*. 2016;5:37–56.
25. VanderWeele TJ. A unification of mediation and interaction: a four-way decomposition. *Epidemiology*. 2014;25:749.
26. VanderWeele T. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*. 2013;24:224–232.
27. VanderWeele T, Tchetgen ET. Attributing effects to interactions. *Epidemiology*. 2014;25:711–722.