Research Article

Soojin Park*, Suyeon Kang, Chioun Lee, and Shujie Ma

Sensitivity analysis for causal decomposition analysis: Assessing robustness toward omitted variable bias

https://doi.org/10.1515/jci-2022-0031
received April 27, 2022; accepted February 03, 2023

Abstract: A key objective of decomposition analysis is to identify a factor (the "mediator") contributing to disparities in an outcome between social groups. In decomposition analysis, a scholarly interest often centers on estimating how much the disparity (e.g., health disparities between Black women and White men) would be reduced/remain if we set the mediator (e.g., education) distribution of one social group equal to another. However, causally identifying disparity reduction and remaining depends on the no omitted mediator–outcome confounding assumption, which is not empirically testable. Therefore, we propose a set of sensitivity analyses to assess the robustness of disparity reduction to possible unobserved confounding. We derived general bias formulas for disparity reduction, which can be used beyond a particular statistical model and do not require any functional assumptions. Moreover, the same bias formulas apply with unobserved confounding measured before and after the group status. On the basis of the formulas, we provide sensitivity analysis techniques based on regression coefficients and R^2 values by extending the existing approaches. The R^2-based sensitivity analysis offers a straightforward interpretation of sensitivity parameters and a standard way to report the robustness of research findings. Although we introduce sensitivity analysis techniques in the context of decomposition analysis, they can be utilized in any mediation setting based on interventional indirect effects when the exposure is randomized (or conditionally ignorable given covariates).

Keywords: interventional indirect effect, unobserved confounding, disparity reduction, disparity remaining, robustness value

MSC 2020: 62D20

1 Introduction

Decomposition analysis aims to identify factors that may close the observed gap in social, psychological, behavioral, or health outcomes between groups defined by social-demographic factors, such as gender/sex, race/ethnicity, socioeconomic status (SES), etc. Such factors are called "mediators" because they are believed to lie between the exposure (one's social position) and the outcomes. Traditional decomposition analysis based on the difference-in-coefficients approach [1,2] provides a straightforward way to estimate

* **Corresponding author: Soojin Park,** School of Education, University of California, Riverside, California, United States of America, e-mail: soojin.park@ucr.edu
Suyeon Kang: Department of Statistics, University of California, Riverside, California, United States of America, e-mail: skang062@ucr.edu
Chioun Lee: Department of Sociology, University of California, Riverside, California, United States of America, e-mail: chiounl@ucr.edu
Shujie Ma: Department of Statistics, University of California, Riverside, California, United States of America, e-mail: shujie.ma@ucr.edu

the degree to which the observed disparity is reduced or remains after controlling for the mediator. However, the traditional method is limited to a specific statistical model that cannot be readily generalizable to discrete mediators or outcomes and nonlinear relationships [3]. Recently, several researchers developed and applied decomposition analysis within the counterfactual framework of causal inference, namely, "causal decomposition analysis," that overcomes the limitations of traditional decomposition analysis. VanderWeele and Ronbinson [4] and Jackson and VanderWeele [5] conceptualized decomposition analysis within the counterfactual framework, and many articles have appeared on this topic in the past 5 years [6–11].

A central goal of causal decomposition analysis is to estimate the degree to which an observed disparity would be reduced or remain if we equalize the mediator distribution across social groups. For example, how much would health disparities decrease if we equalize the education level between Black women and White men? The causal identification of the disparity reduction and remaining hinges on the strong assumption of no unobserved confounding in the mediator–outcome relationship. One way to assess the robustness of findings against possible violations of this assumption is to conduct a sensitivity analysis, yet few sensitivity analysis techniques are available in the causal decomposition framework. Previously, Park et al. [10] developed a preliminary sensitivity analysis that assesses the robustness of disparity reduction and remaining estimates. However, Park et al. method is restricted to a certain setting that requires conditional independence between unobserved and observed intermediate confounders (i.e., the effects of social groups confounding the mediator–outcome relationship). Another limitation is that the interpretation of the sensitivity parameter is not straightforward since the prevalence difference in an unobserved confounder (e.g., being discriminated), comparing individuals in different groups, is conditioned on the mediator (e.g., education), which is a descendant of (a variable affected by) the group status.

In this study, we propose a set of sensitivity analyses for causal decomposition analysis, consisting of sensitivity parameters that are easy to interpret without making the restrictive assumption. After introducing our motivating example using data from Midlife Development in the U.S. (MIDUS) (Section 2), we review literature on causal decomposition analysis as a statistical framework that identifies contributing factors to disparities (Section 3). This review highlights how causal decomposition analysis differs from the related definitions in the causal mediation literature, including natural direct and indirect effects [12,13]. As a result, we show that sensitivity analysis developed for natural indirect effects can assess the robustness of disparity reduction when no intermediate confounders exist, which is unlikely in many studies that investigate contributing factors to disparities.

Therefore, we develop a set of sensitivity analyses for disparity reduction and remaining that incorporate "observed" intermediate confounders. We begin by deriving general bias formulas for disparity reduction and remaining which are the basis of our proposed sensitivity analyses (Section 4). Since general bias formulas do not rely on any statistical models, they apply to various situations, including linear and nonlinear relationships as well as different types of mediator, outcome, or omitted confounder variables. We show that same bias formulas apply for "unobserved" confounding measured before and after the group status. Second, we provide simplified bias formulas given linear models specified for the outcome and the unobserved confounder (Section 5). The simplified bias formulas offer a sensitivity analysis that is straightforward to use if the linearity assumption is met. Finally, we reparameterize the regression-based sensitivity analysis to R^2 values (the proportion of variance explained) by extending the method of Cinelli and Hazlett [14] (Section 6). A critical advantage of this reparameterization to R^2 values is that we can estimate the correct standard errors with a varying amount of unobserved confounding. Another advantage is to provide a standard way of reporting the degree to which research findings are robust against the no unobserved confounding assumption. The standard way is referred to as the "robustness value" (RV) [14], which is the minimum strength of the confounder on the mediator and outcome, assuming an equal strength, to change research findings. The RV conveniently summarizes how sensitive the conclusions are to unobserved confounding.

In Section 7, we conclude with a discussion. Our sensitivity analysis is implemented in the "causal.decomp" R package. Code to replicate all analyses can be found at https://github.com/soojinpark33/Sensitivity-Analysis-for-CDA/blob/main/README.md.

2 Running example

To motivate the concepts and methods that we present, we rely on an epidemiological example; studies have consistently observed that racial and gender minorities, particularly Black women, show poorer cardiovascular health (CVH) than other race–gender groups. SES which is a fundamental cause and a key determinant of access to resources, may operate via many mechanisms to affect multiple disease outcomes [15], including cardiovascular disease [16]. Educational attainment plays a key role in explaining racial and gender disparities in health, and education also affects other subsequent SES measures, such as income and wealth. Therefore, we hypothesize that the observed disparity in CVH between race–gender groups would decrease if we equalize the education levels between the groups. Causal decomposition analysis can be used to test this hypothesis.

From the hypothesis, we define four social groups: White men ($R = 0$), Black women ($R = 1$), Black men ($R = 2$), and White women ($R = 3$); the mediator is education (M); the outcome is CVH (Y). One concern is that education status is not randomized, and the relationship between education and later CVH could be confounded by various life course factors. Therefore, on the basis of literature [17,18], we identified possible confounders, such as age (C_1), genetic vulnerability (C_2), and SES (X_1) and adverse experience in childhood (X_2).

We encode our understanding of data-generation process involving these variables in Figure 1(a). We assume racial and gendered disparities in CVH arise through childhood abuse (X_2) and education (M). We also assume that these disparities could arise through historical processes that include racism and sexism [20]. For example, due to historical processes, Blacks are more likely than Whites to be born into a family with low childhood SES (X_1) and suffer from a particular genetic vulnerabilities (C_2, parental history of cardiovascular and metabolic diseases). Age (C_1) also interacts with the historical process and in turn affects all other variables.

However, one may argue that the relationship between education and later CVH could still be confounded by unobserved confounders (e.g., discrimination based on race, gender, and social class) even after controlling for the observed variables. Figure 1(b) shows a scenario in which an unobserved variable U confounds the relationship between the mediator and the outcome. Depending on a kind of unobserved variable, it could be measured before, concurrently with, or after the group status. For example, if perceived discrimination were the unobserved confounder, U would be measured after the group status, as shown in Figure 1(b). As another example, if an unknown genetic factor that affects later education and CVH were the unobserved confounder, U would be measured before the group status.

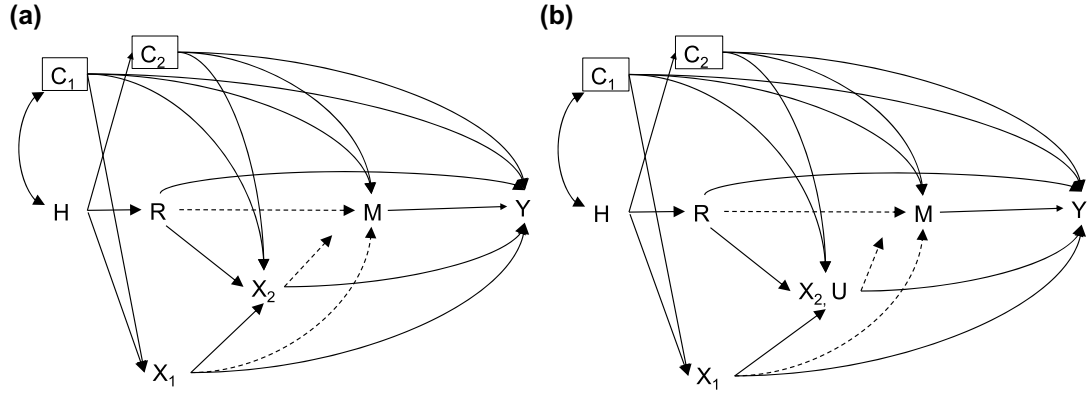

Figure 1: Directed acyclic graph [19]: (a) when no unobserved confounder exists and (b) when unobserved confounder U exists. Note. (1) Diagram represents the relationship between race and gender intersectional status R, cardiovascular health Y, and education M, as well as history H, age C_1, genetic vulnerability C_2, childhood SES X_1, and childhood abuse X_2. (2) Placing a box around the conditioning variables implies that a disparity is considered within levels of these variables. (3) Dotted lines indicate intervening on M given baseline covariates. (4) In figure b, X_2 and U are separate nodes but presented together because the result remains valid irrespective of the direction of causality between X_2 and U, or the absence thereof.

How can we validate our conclusions on disparity reduction and remaining to the possible omission of unobserved variables? In Section 4, we present a set of sensitivity analyses that applied researchers can use to validate their findings to possible omitted variables. To proceed, investigators should further clarify baseline covariates and intermediate confounders, in addition to identifying the group status, the mediator (education), and the outcome (CVH). Baseline covariates characterize demographics through which CVH or education differences are considered equitable [8], which are age or genetic vulnerability in our example. Intermediate confounders represent the effects of race–gender status that confound education and CVH, which are childhood SES and abuse in the example. We use the following notation to represent baseline covariates and intermediate confounders. Intermediate confounders are denoted as $\mathbf{X} = (X_1, X_2)$, where $\mathbf{x} \in \mathcal{X}$; baseline covariates are denoted as $\mathbf{C} = (C_1, C_2)$, where $\mathbf{c} \in C$.

3 Review and implications on existing sensitivity analyses

This section provides an overview of causal decomposition analysis based on interventional analogues of natural direct and indirect effects (interventional effects [21,22]). We then compare this approach with the related definitions in the causal mediation literature, including natural direct and indirect effects. Finally, we show the conditions under which sensitivity analyses developed for natural indirect effects can be used to assess the sensitivity of disparity reduction.

3.1 Causal decomposition analysis

Causal decomposition analysis does not contribute to the argument of whether socially defined characteristics such as race and gender can be given a causal interpretation [5]. Rather, it focuses on estimating a causal effect of potentially manipulable factors (mediators) in reducing the observed disparity. Therefore, causal decomposition analysis aims to estimate how much the disparity would be reduced by intervening to equalize the mediator distribution between social groups. In the motivating example, this intervention implies increasing the Black women's education to the level of White men among those with the same baseline covariates. To illustrate, we use the example of comparing Black women (comparison group) and White men (reference group). Statistically, it requires imputing each Black women's mediator with a randomly drawn value from the mediator distribution of White men among those with the same level of baseline covariates.

3.1.1 Definitions

Here, we fix the reference group to $R = 0$ and compare it with different comparison groups $R = r$, where $r \in \{1, 2, 3\}$. Let $G_{m|R=0,\mathbf{c}}$ denote a random draw from the mediator distribution ($M = m$) of the reference group ($R = 0$) given $\mathbf{C} = \mathbf{c}$. Let $Y_i(m)$ denote the potential value of the outcome for individual i under $M = m$. Then, $E[Y_i(G_{m|R=0,\mathbf{c}})|R_i = r, \mathbf{c}]$ is the average counterfactual outcome for a comparison group r given baseline covariates of $\mathbf{c}$, if their education level was set to a random draw from that of the reference group with baseline covariates of $\mathbf{c}$.

Using this notation, disparity reduction is defined as, given baseline covariates, the difference between the average CVH of a comparison group and the average counterfactual CVH if their education level was the same as the reference group among those with the same baseline covariates levels. Formally,

$$\delta(r, 0|r) \equiv E[Y_i|R_i = r, \mathbf{c}] - E[Y_i(G_{m|R=0,\mathbf{c}})|R_i = r, \mathbf{c}] \quad \text{for} \quad r \in \{1, 2, 3\} \quad \text{and} \quad \mathbf{c} \in C. \tag{1}$$

Likewise, the disparity remaining is defined as, given baseline covariates, the difference between the average counterfactual CVH of a comparison group after the hypothetical intervention and the average CVH outcome of the reference group. Formally,

$$\zeta(0|r, 0) \equiv E[Y_i(G_{m|R=0,\mathbf{c}})|R_i = r, \mathbf{c}] - E[Y_i|R_i = 0, \mathbf{c}] \quad \text{for} \quad r \in \{1, 2, 3\} \quad \text{and} \quad \mathbf{c} \in C. \tag{2}$$

The observed disparity is equal to the sum of disparity reduction and disparity remaining due to hypothetically intervening on the mediator as $\tau(r, 0) = \delta(r, 0|r) + \zeta(0|r, 0)$ for $r \in \{1, 2, 3\}$.

3.1.2 Identification assumptions and results

Defining disparity reduction and remaining requires an unobservable quantity, i.e., $E[Y_i(G_{m|R=0,\mathbf{c}})|R_i = r, \mathbf{c}]$, which is an inherent problem of causal inference [23]. Therefore, we invoke the following assumptions to identify disparity reduction and disparity remaining.

- **A1: Conditional independence**: $Y_i(m) \perp M_i|R_i = r, \mathbf{X}_i = \mathbf{x}, \mathbf{C}_i = \mathbf{c}$, for all $r \in \{0, 1, 2, 3\}$, $\mathbf{x} \in \mathcal{X}$, $m \in \mathcal{M}$ and $\mathbf{c} \in C$. There is no omitted confounding in the mediator–outcome relationship given the race–gender status and measured confounders ($\mathbf{X}$ and $\mathbf{C}$).
- **A2: Positivity**: $0 < P(M_i = m|R_i = r, \mathbf{x}, \mathbf{c}) < 1$ or all $r \in \{0, 1, 2, 3\}$, $\mathbf{x} \in \mathcal{X}$, $m \in \mathcal{M}$ and $\mathbf{c} \in C$. The conditional probability of m for a comparison group is positive given measured confounders.
- **A3: Consistency**: if $M_i = m$, then $Y_i = Y_i(m)$ for all $m \in \mathcal{M}$, where $Y_i(m)$ is the potential value of the outcome for individual i under $M = m$. Consistency states that the observed outcome under the exposure history is the same as the potential outcome after setting the exposure to that level.

These three assumptions are all strong, and the plausibility of the assumptions depends on the context of the study. This study aims to address the possible violation of conditional independence.

Given the assumptions, the disparity reduction and remaining are nonparametrically identified as follows:

$$\begin{aligned}
\delta(r, 0|r) &= E[Y_i|R_i = r, \mathbf{c}] - \sum_{\mathbf{x},m} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|R_i = r, \mathbf{c})P(m|R_i = 0, \mathbf{c}) \quad \text{and} \\
\zeta(0|r, 0) &= \sum_{\mathbf{x},m} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|R_i = r, \mathbf{c})P(m|R_i = 0, \mathbf{c}) - E[Y_i|R_i = 0, \mathbf{c}],
\end{aligned} \tag{3}$$

where $r \in \{1, 2, 3\}$, $x \in \mathcal{X}$, $m \in \mathcal{M}_1$, and $\mathbf{c} \in C$.

3.1.3 Estimation

If the assumptions hold, there are many ways to estimate disparity reduction and disparity remaining including regression-based estimators [5,24][1], weighting-based estimators [8], imputation-based estimators [10,25][2], and doubly robust estimators [11]. Although limited due to its modeling assumptions, regression-based methods are perhaps the most straightforward way to estimate disparity reduction and remaining. Consider the following models fitted to the intermediate confounders, mediator, and outcome as follows:

$$X_i = \gamma + \sum_r \gamma_r I_i(r) + \gamma_c \mathbf{C}_i + \varepsilon_{x,i}, \tag{4}$$

$$M_i = \alpha + \sum_r \alpha_r I_i(r) + \alpha_c \mathbf{C}_i + \varepsilon_{m,i}, \quad \text{and} \tag{5}$$

$$Y_i = \beta + \sum_r \beta_r I_i(r) + \beta_x \mathbf{X}_i + \beta_m M_i + \beta_c \mathbf{C}_i + \varepsilon_{y,i}, \tag{6}$$

1 Regression-based estimators refer to a method utilizing regression coefficients directly, such as the difference in coefficients and product of coefficients.

2 Imputation-based estimators refer to a method that predicts or imputes counterfactual outcome values.

where $r \in \{1, 2, 3\}$. Under these models, disparity reduction is estimated as $\hat{\delta}(r, 0|r) = \hat{\alpha}_r\hat{\beta}_m$; disparity remaining is estimated as $\hat{\zeta}(0|r, 0) = \hat{\beta}_r + \hat{\beta}_x\hat{\gamma}_r$, where $\hat{A}$ indicates a consistent estimator of parameter A. Note that this regression-based estimator hinges on the functional form assumptions made in the models shown in equations (4)–(6). If the interaction effect exists between the social group and the mediator, one should include the interaction term in the outcome model as follows:

$$Y_i = \beta + \sum_r \beta_r I_i(r) + \beta_x \mathbf{X}_i + \beta_m M_i + \sum_r \beta_{rm} I_i(r) M_i + \beta_c \mathbf{C}_i + \varepsilon_{y,i}. \tag{7}$$

Then, the disparity reduction is estimated as $\hat{\delta}(r, 0|r) = \hat{\alpha}_r(\hat{\beta}_m + \hat{\beta}_{rm})$; disparity remaining is estimated as $\hat{\zeta}(0|r, 0) = \hat{\beta}_r + \hat{\beta}_x\hat{\gamma}_r + \hat{\beta}_{rm}(\hat{\alpha} + \hat{\alpha}_c\hat{E}[\mathbf{C}_i])$. A proof is given in Park et al. [24] and is, thus, omitted here. As far as investigators are willing to make the modeling assumptions along with A1–A3, these estimators provide a valid estimate of disparity reduction and disparity remaining.

3.2 Comparison with the related definitions in the literature

In this section, we compare disparity reduction and remaining defined in equations (1) and (2) with the related definitions in the literature: conditional average treatment effects (CATE), controlled direct effects (CDEs), natural in/direct effects, and interventional in/direct effects.

CATE and CDE. Defining disparity reduction and remaining requires stochastic interventions on the mediator to follow an alternative distribution. Instead of stochastic interventions, we could intervene to fix the mediator to a prespecified value. Then, disparity reduction and remaining will similarly correspond to CATE and CDE, respectively. The CATE is the effect of an intervention (M) on the outcome (Y) conditional on R, as $E[Y_i|R_i = 1, M_i = m, c] - E[Y_i|R_i = 1, M_i = m', c]$. In our example, the CATE is the expected change in CVH in response to the change in education from $M = m$ to $M = m'$ among Black women. The CDE (e.g., [26,27]) is the effect of exposure (R) on the outcome (Y) after fixing the mediator to a prespecified value $M_i = m$ over the entire sample as $E[Y_i|R_i = 1, M_i = m, c] - E[Y_i|R_i = 0, M_i = m, c]$. The CDE is of interest when the exposure effect at a prespecified value is meaningful. For example, it would be interesting to ask how much of the Black women–White men disparity would remain if every individual attended college. While interesting, these effects based on a prespecified mediator value may be less realistic for interventions. Intervening to have every individual in the population to take a certain value of a mediator may not be feasible in practice. One alternative to avoid this global intervention is to use a stochastic intervention that follows the observed distribution of a different group, as shown in the definitions in equations (1) and (2).

Natural in/direct effects. First, we argue that defining natural indirect effects is not straightforward in disparities research. In the example, the natural indirect effect defined between Black women and White men in the motivating example is the expected difference, comparing each Black woman's actual and potential CVH after setting her education level to a value that would have naturally been observed had she been born a White man. Given that gender and race are essentially nonmodifiable, it is somewhat strange to consider fixing each individual's mediator to a value that would have naturally been observed had the individual been born a White man [28]. In contrast, interpreting disparity reduction defined in equation (1) is straightforward because we (1) do not define counterfactual outcomes with respect to social groups such as race and gender and (2) assign Black women's education to a randomly drawn value from the observed distribution of White men's education, rather than assigning education to the value that would have been realized in a counterfactual world in which she was a White man. Disparity reduction compares the average Black woman's CVH to the average counterfactual CVH outcome of Black women after equalizing their education level to that of White men as a group.

Next, we compare the conditional independence assumption (A1) with the related assumptions required to identify natural indirect effects. We don't compare assumptions regarding the exposure (treatment ignorability) because causal decomposition analysis does not attempt to estimate the causal effect of social groups (race and gender) and thus does not make any assumption regarding the exposure

(group status). Pearl [19] made the following assumption to identify natural indirect and direct effects: $Y_i(r, m) \perp M_i(r')|\mathbf{C}_i = c$, which implies (1) no unobserved pre-exposure confounders (i.e., variables measured before the exposure that confound the mediator–outcome relationship) given the race–gender status and baseline covariates and (2) no intermediate confounders. The first assumption might be met in some disparities research since there are hardly any variables that can be measured before race and gender status. However, the second assumption is unlikely to be met in many disparities research settings, since a myriad of life-course factors contribute to disparities in education and CVH.

Alternatively, Robins [13] made the following assumption: $Y_i(r, m) \perp M_i(r)|R_i = r, \mathbf{X}_i = x, \mathbf{C}_i = c$, which allows intermediate confounders $\mathbf{X}$. This is an important advantage considering that the assumption of no intermediate confounders is unlikely to be met in many disparities research settings. However, this relaxed assumption comes with the cost of requiring no interactions between the exposure and the mediator at the individual level. Unfortunately, assuming no interaction effects between the group status and the mediator is also a strong assumption that is unlikely to be met in many studies. For example, prior studies documented diminished returns of human capital as a result of discrimination for minorities [29].

Instead of the no interaction assumption, Petersen et al. [30] require the following assumption: $E[Y_i(1, m) - Y_i(0, m)|M_i(0) = m, C_i = c] = E[Y_i(1, m) - Y_i(0, m)|C_i = c]$, which implies that the exposure effect at the controlled level of $M = m$ is independent of the potential mediator under $R = 0$. Although this assumption is weaker than the previously discussed no intermediate confounding assumption, it is still difficult to check whether this assumption is met in practice.

Interventional in/direct effects. Even when these additional assumptions are not met, previous studies [30,31] have noted that the natural direct effect still estimates an interesting causal parameter: the average of CDEs with respect to the conditional distribution of M for $R = 0$ given C. The related idea was discussed in the noncounterfactual literature [32,33], which described a target trial using a randomized intervention that follows an observed alternative distribution. The interventional analogs were formalized within the counterfactual outcomes framework and used to accommodate intermediate confounders [21] and time-varying exposures and mediators [34,35]. In the literature, the interventional indirect effects were based on defining counterfactual outcomes with regard to both exposure and mediator.

By applying the interventional analogs, VanderWeele and Robinson [4] investigated contributing factors to racial disparities. They suggested focusing on the causal effect of manipulable factors (mediators) rather than the causal effect of race (exposure), thereby defining counterfactual outcomes with respect to the mediators, not the exposure. Following their approach, Jackson and VanderWeele [5] proposed formal definitions of disparity reduction and disparity remaining, one of which is shown in equations (1) and (2). The causal estimands were further extended by other scholars, including Lundberg [11] who allowed intervening to fix the mediator to a prespecified value or the mediator to follow various distributions (rather than the observed distribution of the reference group), and Park et al. [10] who accommodated the case of intervening on multiple mediators.

3.3 Implications on existing sensitivity analyses

3.3.1 Under no intermediate confounders

If no intermediate confounders exist, interventional direct and indirect effects coincide with natural direct and indirect effects, respectively. Although it may be a rare situation, suppose that the intermediate confounders (e.g., childhood SES and abuse) do not exist. Then, one can use a nonparametric sensitivity analysis [36,37] or parametric sensitivity analysis [3] developed for natural direct and indirect effects to assess the robustness of estimated disparity reduction and remaining to possible unobserved pre-exposure confounding.

3.3.2 Under no interaction between *R* and *M*

Now, we consider a situation where intermediate confounders exist, but no interaction exists between social groups and the mediator. This situation is perhaps plausible in some studies that examine disparities depending on a type of mediator. Then, the estimator of disparity reduction $\alpha_r\beta_m$ is consistent for the natural indirect effect given equations (5) and (6). However, the sensitivity analysis developed for natural direct and indirect effects to possible violations of unobserved intermediate confounding [38,39] is not appropriate for assessing the robustness of estimated disparity reduction and remaining. For instance, a sensitivity analysis developed by Imai and Yamamoto [38] addresses the bias due to incorporating interaction effects given that all intermediate confounders are observed. VanderWeele and Chiba [39] do not assume that all intermediate confounders are observed; however, the bias formulas do not incorporate existing intermediate confounders.

In conclusion, a new sensitivity analysis is necessary for disparity reduction and remaining that incorporates existing intermediate confounders and possible interaction effects.

3.4 An application to MIDUS

We estimate disparity reduction and remaining with varying assumptions. Table 1 shows the estimated quantities of interest between Black women and White men. Other comparisons are available, but we only present the results between Black women and White men for simplicity. The initial disparity between Black women and White men is −0.965 (equivalent to 0.420 SD of CVH), and the 95% confidence interval (CI) is bounded away from zero, meaning that Black women have significantly worse CVH than White men after controlling for age and genetic vulnerability.

First, we estimate disparity reduction and remaining, assuming no interaction between education and race–gender status (first column in Table 1). These estimators are consistent for natural direct and indirect effects defined by Robins [13]. Disparity reduction is negative (−0.401) and is significant at the 95% confidence level. The initial disparity would decrease by 41.6% if Black women's education level were the same as White men's among those with the same age and genetic vulnerability level.

We compare this result after relaxing the no interaction assumption (third column), which is consistent with disparity reduction and remaining defined in equation (3). The difference in disparity reduction and remaining estimates is small. Disparity reduction changes from −0.401 (first column) to −0.360 (third column); disparity remaining changes from −0.564 (first column) to −0.604 (third column). This slight change after relaxing the no interaction assumption implies that there is little evidence for the presence of the interaction effect.

Table 1: Estimates of the disparity reduction and disparity remaining

	Estimate (95% CI)		
***R–M* interaction**	**No**	**Yes**	**Yes**
Intermediate confounders	**Yes**	**No**	**Yes**
Initial disparity ($\tau(1, 0)$)	−0.965	−0.965	−0.965
(95% CI)	(−1.259, −0.662)	(−1.257, −0.675)	(−1.238, −0.658)
Disparity remaining ($\zeta(0\|1, 0)$)	−0.564	−0.480	−0.604
(95% CI)	(−0.863, −0.278)	(−0.791, −0.143)	(−0.868, −0.187)
Disparity reduction ($\delta(1, 0\|1)$)	−0.401	−0.485	−0.360
(95% CI)	(−0.561, −0.239)	(−0.727, −0.261)	(−0.712, −0.219)
% Reduction	41.6%	50.2%	37.3%

Note: Black women: $R = 1$, White men: $R = 0$.

Finally, we examine the result after relaxing the no interaction assumption but without intermediate confounders (second column), which is consistent with the natural indirect effect defined by Pearl [19]. Not controlling for intermediate confounders results in an overestimation or underestimation of disparity reduction and remaining when compared to controlling for intermediate confounders (third column). After including observed intermediate confounders, disparity reduction changes from −0.485 (second column) to −0.360 (third column); disparity remaining changes from −0.480 (second column) to −0.604 (third column). This result suggests overestimation of disparity reduction due to education when intermediate confounders were not accounted for. In the next section, we present a sensitivity analysis to possible unobserved confounding that incorporate observed intermediate confounders and the interaction effect.

4 General bias formulas

We begin by calculating the bias for disparity reduction and remaining when the unobserved confounder U exists, on which the two sensitivity analyses that we propose in Sections 5 and 6 are based. We first derive bias formulas when intermediate unobserved confounding exist and later show that the same bias formulas apply, with an additional assumption, to pre-exposure unobserved confounding.

The conditions required to calculate the bias are as follows: (1) no omitted confounding exists in the mediator and outcome relationship given group status (r), observed confounders ($\mathbf{x}$, $\mathbf{c}$), and the unobserved confounder (u), as $M_i \perp Y_i(m)|R_i = r, \mathbf{X}_i = \mathbf{x}, \mathbf{C}_i = \mathbf{c}, U_i = u$ and (2) the unobserved confounder (U) is an effect of group status (R) and thereby is considered as an intermediate unobserved confounder as shown in Figure 1(b). An example of intermediate unobserved confounder includes discrimination, which is an effect of group status and adversely affects education and CVH.

Suppose that we only had observed data. Then, researchers would often estimate the disparity reduction using observed data, $\delta_{\text{res}}(r, 0|r) = E[Y_i|R_i = r, \mathbf{c}] - \sum_{\mathbf{x},m} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{X} = \mathbf{x}|R_i = r, \mathbf{c})P(M_i = m|R_i = 0, \mathbf{c})$, as equation (3). However, if the unobserved confounder U exists, this expression will lead to a biased estimate of disparity reduction. The bias is, therefore, defined as the difference between the expected value of this estimator using observed data and the true effect of disparity reduction as follows:

$$\begin{aligned}\text{bias}(\delta(r, 0|r)) &= \delta_{\text{res}}(r, 0|r) - \delta(r, 0|r) \\ &= E[Y_i|R_i = r, \mathbf{c}] - \sum_{\mathbf{x},m} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|R_i = r, \mathbf{c})P(m|R_i = 0, \mathbf{c}) \\ &\quad - E[Y_i|R_i = r, \mathbf{c}] + \sum_{\mathbf{x},m,u} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}, u]P(\mathbf{x}, u|R_i = r, \mathbf{c})P(m|R_i = 0, \mathbf{c}).\end{aligned} \tag{8}$$

Note that we used $P(\mathbf{x}, u|R_i = r, \mathbf{c})$ to accommodate both cases: (1) when U is measured before $\mathbf{X}$ and (2) when U is measured after $\mathbf{X}$. The bias for disparity remaining is defined the same way ($\text{bias}(\zeta(0)) = \zeta_{\text{res}}(0) - \zeta(0)$).

Then, for a particular value of $U = u'$, the biases for disparity reduction and remaining are given by

$$\begin{aligned}\text{bias}(\delta(r, 0|r)) &= \sum_{\mathbf{x},m,u} \{E[Y|r, \mathbf{x}, m, \mathbf{c}, u] - E[Y|r, \mathbf{x}, m, \mathbf{c}, u']\}\{P(u|r, \mathbf{x}, c) - P(u|r, \mathbf{x}, m, \mathbf{c})\} \\ &\quad \times P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}), \\ \text{bias}(\zeta(0|r, 0)) &= -\sum_{\mathbf{x},m,u} \{E[Y|r, \mathbf{x}, m, \mathbf{c}, u] - [Y|r, \mathbf{x}, m, \mathbf{c}, u']\}\{P(u|r, \mathbf{x}, c) - P(u|r, \mathbf{x}, m, \mathbf{c})\} \\ &\quad \times P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}).\end{aligned} \tag{9}$$

A proof is given in Appendix A. These general bias formulas do not require any assumptions regarding functional forms, variable types of the mediator, outcome, or unobserved confounders. While these bias formulas can be used in general settings, their applicability may be limited due to too many moving parts (i.e., sensitivity parameters). Therefore, we provide simplified bias formulas under linear models specified for the outcome and the unobserved confounder that are straightforward to use in Section 6. We will discuss the applicability of the general bias formulas later.

Note that the same bias formulas apply with pre-exposure unobserved confounding under the additional assumption of $R \perp U|\mathbf{C}$, where U is pre-exposure unobserved confounder. The assumption states that the pre-exposure confounder and group status do not affect each other given baseline covariates. An example of such confounder would be an unknown genetic factor that is not related to race or gender, but affects later education and CVH. Given the pre-exposure confounder U, the bias for disparity reduction is defined as follows:

$$\begin{aligned}\text{bias}(\delta(r, 0|r)) = E[Y_i|R_i = r, \mathbf{c}] - \sum_{\mathbf{x},m} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|R_i = r, \mathbf{c})P(m|R_i = 0, \mathbf{c}) \\ - E[Y_i|R_i = r, \mathbf{c}] + \sum_{\mathbf{x},m,u} E[Y_i|R_i = r, \mathbf{x}, m, \mathbf{c}, u]P(\mathbf{x}|R_i = r, \mathbf{c}, u)P(m|R_i = 0, \mathbf{c})P(u|\mathbf{c}).\end{aligned}$$

This equation leads to the same expression shown in equation (8) since $P(\mathbf{x}|R_i = r, \mathbf{c}, u)P(u|\mathbf{c}) = P(\mathbf{x}|R_i = r, \mathbf{c}, u)P(u|R_i = r, \mathbf{c})$ due to the assumption of $R \perp U|\mathbf{C}$.

5 Sensitivity analysis using regression coefficients

Unlike the general bias formulas, the proposed sensitivity analyses in this study are developed under a particular statistical model that assumes linearity. However, we show that an extension is possible when linearity is violated.

5.1 Under the linearity assumption

The general bias formulas in equation (9) can be simplified under linear models specified for the outcome and the unobserved confounder as follows:

$$\begin{aligned}E[Y_i|r, x, m, c, u] = \beta + \beta_r + \beta_x x + \beta_m m + \beta_c c + \beta_u u, \quad \text{and} \\ E[U_i|r, x, m, c] = \delta + \delta_r + \delta_x x + \delta_m m + \delta_c c,\end{aligned} \tag{10}$$

for $r \in \{1, 2, 3\}$. We first assume the simplest models for the outcome and the unobserved confounder and discuss later how to relax the linearity assumption. Given equation (10), the nonparametric bias formulas for disparity reduction and remaining are considerably simplified as follows:

$$\text{bias}(\delta(r, 0|r)) = \alpha_r\delta_m\beta_u, \quad \text{and} \quad \text{bias}(\zeta(0|r, 0)) = -\alpha_r\delta_m\beta_u \quad \text{for} \quad r \in \{1, 2, 3\}, \tag{11}$$

where an unbiased estimate of α_r can be obtained by fitting the mediator model in equation (5). A proof is given in Appendix B.

We offer several remarks about these bias formulas. First, the bias for disparity reduction and remaining is the same except for the sign. This is because the initial disparity is an observed quantity conditional on specified covariates, and hence, no bias exists due to the unobserved mediator–outcome confounder.[3]

Second, the bias is zero when either β_u or δ_m is zero, meaning that the unobserved confounder is not associated with the outcome or the mediator given observed confounders. The bias is also zero when the mediator does not differ by group given baseline covariates.

---

3 No bias exists for the initial disparity as shown below.

$$\text{bias}(\tau(r, 0)) = \text{bias}(\delta(r, 0|r)) + \text{bias}(\zeta(0|r, 0)) = \alpha_r\delta_m\beta_u - \alpha_r\delta_m\beta_u = 0. \tag{12}$$

Third, the unobserved variable confounds the intermediate confounders–outcome and intermediate confounders–mediator relationships ($\mathbf{X} - Y$ and $\mathbf{X} - M$) in addition to the mediator–outcome relationship ($M - Y$). However, confounding relationships in the $\mathbf{X} - Y$ and $\mathbf{X} - M$ relationships do not contribute to the bias of disparity reduction and remaining. The only confounding path that matters for disparity reduction and remaining is $M \leftarrow U \rightarrow Y$. Intuitively, it makes sense since conditional independence (A1) only concerns unobserved confounding in the mediator–outcome relationship.

Finally, the intermediate confounder and outcome models specified in equation (10) imply that the effect of unobserved confounder (U) on the outcome (Y) is constant across social groups. However, it may be too restrictive to assume the constant effect of U across social groups. For example, discrimination has a differential effect on CVH by race [40]. To address this differential impact of U on Y, we can add the interaction effect between u and $I(r)$ in the outcome model as $\sum_r \beta_{ru} I(r)u$. Then the bias for disparity reduction is given by $\text{bias}(\delta(r, 0|r)) = \alpha_r \delta_m (\beta_u + \beta_{ru})$.

The bias formulas expressed in regression coefficients give us an intuitive idea of sensitivity analysis. The basic idea is to find a combination of β_u and δ_m that will explain the disparity reduction and remaining, and change the significance of the effects. Given the combinations, researchers are required to determine whether the amount of confounding expressed in these sensitivity parameters is plausible or not. To do so, understanding the precise meaning of the sensitivity parameters is essential. One sensitivity parameter (β_u) is the difference in the outcome, comparing individuals who differ by one unit on the confounder U, after controlling for the group status, the mediator, and observed confounders. Another sensitivity parameter (δ_m) is the difference in the unobserved confounder U, comparing individuals who differ by one unit on the mediator M, after controlling for race–gender status and observed confounders. We illustrate the use of this sensitivity analysis with the MIDUS example in Section 5.3.

5.2 When the linearity assumption is violated

When the linearity assumption made in equation (10) is violated, one can either (1) derive their bias formulas based on modified models for Y and U or (2) use the original bias formula shown in equation (9). We showed earlier how to address the differential impact of the unobserved confounder U on the outcome Y by group status. However, addressing another differential impact (e.g., the interaction between the unobserved confounder and the mediator) leads to a more complex form of bias formulas. In this case, researchers can directly use the nonparametric bias formulas. To use the nonparametric bias formulas, one should choose a reference value for $U = u'$ and specify $E[Y_i|r, \mathbf{x}, m, \mathbf{c}, u] - E[Y_i|r, \mathbf{x}, m, \mathbf{c}, u']$, which is the difference in the outcome among the comparison group $R = r$, comparing $U = u$ and $U = u'$, across strata of $\mathbf{x}$, m, and $\mathbf{c}$. Also, one should specify $P(u|r, \mathbf{x}, c) - P(u|r, \mathbf{x}, m, \mathbf{c})$, which is the distribution of the unobserved confounder U for the comparison group $R = r$, conditional on $\mathbf{X} = \mathbf{x}$ and $\mathbf{C} = \mathbf{c}$, compared with the distribution of the unobserved confounder U conditional on $\mathbf{X} = \mathbf{x}$, $\mathbf{C} = \mathbf{c}$, and $M = m$. As mentioned earlier, using the original bias formulas is not entirely straightforward since specifying these values may be difficult. However, in some cases, it may be possible to draw these values based on substantive knowledge and use the bias formulas directly.

5.3 Illustration using regression-based sensitivity analysis

This section uses the example provided in Section 3.4 to describe and interpret the sensitivity analysis using regression coefficients. In Section 3.4, we estimated the disparity reduction and disparity remaining between Black women and White men. We here focus on the case after assuming differential effects of education by group (R–M interaction) and accounting for existing intermediate confounders. As shown in Table 1, the initial disparity would be reduced by about 37.3% if Black women's education level was the same as that of White men among those with the same age and genetic vulnerability. This result can be given a causal interpretation under the assumptions described in A1–A3. Suppose that conditional

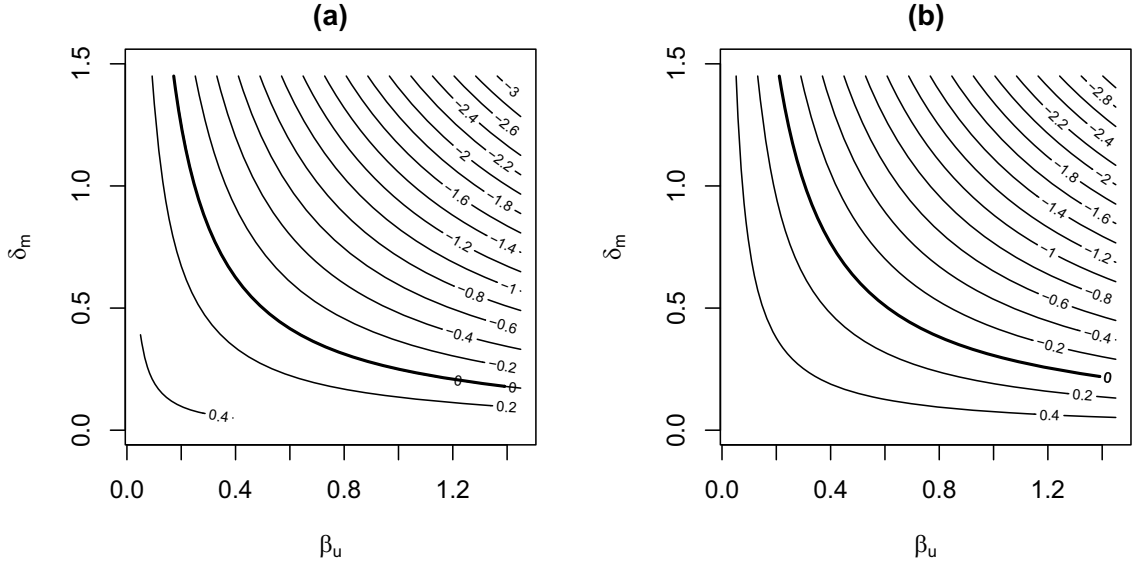

Figure 2: Sensitivity contour plots using regression coefficients. (a) Disparity reduction. (b) Disparity remaining. Note. (1) Bold lines represent the points at which the estimates become zero. (2) Standard lines represent the points at which the estimates become the respective value (e.g., 0.4, 0.2, −0.2, −0.4).

independence (A1) is violated because an unobserved variable, that is, whether each individual perceives to be discriminated, confounds education and CVH. How strongly does this unobserved confounder have to be to explain away the estimates?

In Figure 2, hypothetical values for β_u and δ_m lie on the horizontal and vertical axis, respectively. The contour lines show the true disparity reduction and remaining values at hypothesized values of β_u and δ_m. In addition to the figure, we can also calculate the true disparity reduction and remaining given $\hat{\alpha} = -1.736$. Suppose that the difference in CVH between those who are discriminated and those not discriminated is, say, $\beta_u = 0.993$ (as strong as the effect of education for White men, which is the strongest among four race–gender groups) after controlling for the group status, the mediator, and observed confounders. Then to completely explain the estimated disparity reduction ($\delta(1, 0|1) = -0.360$), δ_m would need to be close to 0.208 (i.e., $-0.360 - (-1.736 \times 0.993 \times 0.208) \approx 0$). This δ_m value indicates that the probability difference of being discriminated is 20.8%, when comparing individuals with the same values for the group status and observed confounders (childhood SES and abuse, education, age, and genetic vulnerability), but differing by one unit on education. Given the same β_u, to completely explain the estimated disparity remaining ($\zeta(0|1, 0) = -0.604$), δ_m would need to be close to 0.350 (i.e., $-0.604 - (-1.736 \times 0.993 \times 0.350) \approx 0$). These δ_m values seem unlikely large given that we compare those who have the same group status, age, and genetic vulnerability, but differing by one unit in education (e.g., no schooling vs graduating junior high school).

6 Sensitivity analysis using R^2 values

The regression-based sensitivity analysis has often been used with a binary unobserved confounder, such as whether discriminated. It is perhaps because the interpretation of sensitivity parameters with a binary unobserved confounder is straightforward. Specifically, δ_m is the *prevalence difference* in the unobserved confounder U, comparing individuals with one unit increase in the mediator M, given controls; β_u is the outcome difference between *two levels of the unobserved confounder*, given controls. However, the interpretation of sensitivity parameters is no longer straightforward if the unobserved confounder is continuous, such as discrimination intensity. With a continuous unobserved confounder, sensitivity parameters depend

on the scale of the unobserved confounder. To see this, δ_m is the *difference in the unobserved confounder*, comparing individuals with a one-unit increase in the mediator M, given controls; β_u is the outcome difference, comparing individuals with a *one-unit increase in the unobserved confounder*, given controls. If sensitivity parameters depend on the scale of the unobserved confounder, it is challenging to determine whether the confounding amount expressed by the sensitivity parameters is large or small even with the substantial knowledge of possible unobserved confounders (e.g., degree of discrimination). One solution is to standardize the unobserved confounder as shown in Section 5.3. Another solution is to reparameterize sensitivity parameters to R^2 values.

In this section, we re-express the simplified bias formulas shown in equation (11) using scale-free R^2 values. By using the R^2 values, we can obtain correct standard errors for disparity reduction and remaining with a varying amount of unobserved confounding. We also present a standard way to report the amount of confounding using the RVs. Cinelli and Hazlett [14] developed a way to express biases using R^2 values in the context of estimating a treatment effect, and we extend this reparameterization of R^2 values to our bias formulas for disparity reduction and remaining.

This reparameterization to R^2 values is based on a particular statistical model specified in equations (5) and (6) that assume linearity. We later provide the extension that can be used when linearity is violated. Note that we rewrite equation (6) as follows to differentiate outcome coefficients before and after including the unobserved confounder U.

$$E[Y_i|r, x, m, c] = \beta_{\text{res}} + \beta_{\text{res},r} + \beta_{\text{res},x}x + \beta_{\text{res},m}m + \beta_{\text{res},c}c. \tag{13}$$

6.1 Under the linearity assumption

6.1.1 Point estimates

Let the partial R^2 value of the unobserved confounder (U) with the outcome (Y) for the comparison group ($R = r$) given mediator (M) and observed confounders ($\mathbf{X}$ and $\mathbf{C}$) be denoted as $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$; and the partial R^2 value of the unobserved confounder (U) with the mediator (M) for the comparison group ($R = r$) given observed confounders ($\mathbf{X}$ and $\mathbf{C}$) be denoted as $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$. Then, the absolute value of bias for disparity reduction and remaining is expressed as follows:

$$|\text{bias}(\delta(r, 0|r))| = |\text{bias}(\zeta(0|r, 0))| = |\alpha_r|\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}\sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1 - R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\text{df}}, \tag{14}$$

for $r \in \{1, 2, 3\}$. We can obtain an unbiased estimate of α_r by fitting the mediator model shown in equation (5); $\text{Var}(\hat{\beta}_{\text{res},m})$ is obtained from the sample variance of $\hat{\beta}_{\text{res},m}$ and df is obtained from the degrees of freedom of the outcome model shown in equation (13). A proof is given in Appendix C.

6.1.2 Standard errors

Some investigators might also be interested in quantifying the amount of confounding that would change the significance of the effects. The standard error for disparity reduction for $R = r$ can be calculated approximately using the Delta method [41] as follows:

$$\text{Var}(\hat{\delta}(r, 0|r)) \approx \alpha_r^2\,\text{Var}(\hat{\beta}_m) + \beta_m^2\,\text{Var}(\hat{\alpha}_r), \tag{15}$$

where $\text{Var}(\hat{\beta}_m) = \text{Var}(\hat{\beta}_{\text{res},m})\left(\frac{1 - R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}}{1 - R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\frac{\text{df}}{\text{df} - 1}\right)$ and

$$\beta_m = \begin{cases} \beta_{\text{res},m} - \sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}\sqrt{\dfrac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1 - R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}}\text{df}, & \text{if} \quad \beta_{\text{res},m} \text{is positive,} \\ \beta_{\text{res},m} + \sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}\sqrt{\dfrac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1 - R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}}\text{df}, & \text{if} \quad \beta_{\text{res},m} \text{is negative.} \end{cases}$$

We can obtain an unbiased estimate of α_r by fitting a mediator model shown in equation (5). Given the estimate, $\sqrt{\text{Var}(\hat{\alpha}_r)}$ can be consistently estimated by computing the sample variance of $\hat{\alpha}_r$. Likewise, we can obtain the estimate of $\beta_{\text{res},m}$ by fitting an outcome model as shown in equation (13). Given the estimate, $\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}$ can be consistently estimated by computing the sample variance of $\hat{\beta}_{\text{res},m}$. A proof is shown in Appendix D.

Note that the standard error of disparity reduction shown in equation (15) is the same as the sample standard deviation of the disparity reduction estimate ($\hat{\delta}_{\text{res}}(r, 0|r)$) if and only if the two sensitivity parameters are zero, meaning that there is no confounding in the mediator–outcome relationship given the intersectional group and measured confounders.

Calculating the standard error for disparity remaining is more complex. Therefore, we use $\tau(r, 0) = \delta(r, 0|r) + \zeta(0|r, 0)$ to approximately calculate the standard error for disparity remaining as follows:

$$\begin{aligned} \text{Var}(\hat{\zeta}(0|r, 0)) \approx\ & \text{Var}(\hat{\tau}(r, 0)) + \text{Var}(\hat{\delta}(r, 0|r)) - 2\text{Cov}(\hat{\tau}(r, 0), \hat{\delta}_{\text{res}}(r, 0|r)) \\ & + 2kE[\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}]\text{Cov}(\hat{\tau}(r, 0), \hat{\alpha}_r) + 2kE[\hat{\alpha}_r]\text{Cov}(\hat{\tau}(r, 0), \sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}), \end{aligned} \tag{16}$$

where $k = \sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1 - R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}}\text{df}$. Here, the estimates of $\text{Var}(\hat{\tau}(r, 0))$ and $\text{Cov}(\hat{\tau}(r, 0), \hat{\delta}_{\text{res}}(r, 0|r))$ can be obtained from sample variance and covariance matrix of the initial disparity and disparity reduction estimates; $\text{Cov}(\hat{\tau}(r, 0), \hat{\alpha}_r)$ can be obtained from sample covariance of the initial disparity and the regression coefficient $\hat{\alpha}_r$; $\text{Cov}(\hat{\tau}(r, 0), \sqrt{\text{Var}(\hat{\beta}_{\text{res},m})})$ can be obtained from sample covariance of the initial disparity and the standard error of $\hat{\beta}_{\text{res},m}$.

Again, the standard error of disparity remaining shown in equation (16) is the same as the sample standard deviation of the remaining disparity estimate ($\hat{\zeta}_{\text{res}}(0|r, 0)$) if and only if the two sensitivity parameters are zero, meaning that there is no omitted confounding existing in the mediator and outcome relationship.

This bias formulas in equation (14) and standard errors for disparity reduction and remaining in equations (15) and (16) can be used as sensitivity analysis that depends on two sensitivity parameters $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$ and $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$. The two sensitivity parameters imply the degree to which an unobserved confounder is associated with the mediator and the outcome for the comparison group $R = r$ after conditioning on appropriate controls, expressed in R^2 values. Larger R^2 values of the sensitivity parameters indicate a larger bias for disparity reduction and remaining due to unobserved confounder U.

We conduct a Monte Carlo simulation study to evaluate how well the proposed standard error estimators perform in varying sample sizes and effect sizes of sensitivity parameters. In this simulation, we find that the 95% CI coverage rate exceeds 0.91 regardless of sample sizes and effect sizes of sensitivity parameters. The coverage rate tends to be closer to the nominal level (0.95) with a sample size of 500 or larger. The coverage rate is lower than expected (0.916) with a small sample size (e.g., 100) and a small effect size of sensitivity parameters (e.g., 0.02). The results and details of the simulation study are reported in Appendix E. In addition, we provide a standard error estimator for a percent reduction in Appendix F.

6.2 When the linearity assumption is violated

The expression of point estimates and standard errors becomes more cumbersome when the interaction effect exists in the group–mediator relationships in the outcome model. In such a case, the bias for disparity reduction and remaining estimates is the same as equation (14), except that $\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}$ should be

replaced with $\sqrt{\{\text{Var}(\hat{\beta}_{\text{res},m}) + \text{Var}(\hat{\beta}_{\text{res},rm}) + 2\text{Cov}(\hat{\beta}_{\text{res},m}, \hat{\beta}_{\text{res},rm})\}}$. For calculating standard errors for the disparity reduction and remaining, $\beta^2_{\text{res},m}$ should be replaced with $(\beta_{\text{res},m} + \beta_{\text{res},rm})^2$.

Perhaps, a more straightforward way to address differential effects is to use an algorithm to change the reference group to $R = r$ when computing the effect of the mediator on the outcome ($\beta_{\text{res},m}$) and its corresponding standard error for each comparison group ($\sqrt{\text{Var}}(\hat{\beta}_{\text{res},m})$). For example, we fit the outcome model with the group–mediator interaction effect, setting the reference group to Black women. Then, the mediator effect on the outcome and its corresponding standard error for Black women can be obtained, respectively, as $\hat{\beta}_{\text{res},m}$ and $\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}$ in the outcome model even after adding the interaction effect. This algorithm provides a convenient and flexible way to address differential effects of the mediator or intermediate confounders by group status.

However, this algorithm does not address other nonlinear effects, such as the interaction effect in the mediator and intermediate confounder relationship. In this case, one should consider using the general bias formulas directly.

6.3 RVs

Despite the development of numerous sensitivity analyses, not many social or medical scientists have used sensitivity analysis to assess the robustness of their findings. Recent literature on sensitivity analysis [14,42] emphasized the use of standard way of reporting the robustness of research findings, which is expected to facilitate the discussions regarding how credible the estimated effect is to possible violations of no omitted confounding. For example, Ding and VanderWeele [42] advanced the *E-value* that reports the robustness of research findings measured in the risk ratio; Cinelli and Hazlett [14] advanced the RV that reports the robustness of research findings derived from linear regressions. In this section, we extend the RV computed for the treatment effect to disparity reduction and remaining.

We define the RV as the strength of association that will explain the estimated disparity reduction or remaining, assuming an equal association to the mediator and the outcome, as $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} = R^2_{M\sim U|r,\mathbf{X},\mathbf{C}} = \text{RV}$. Then, the RV for disparity reduction and remaining are, respectively, given by

$$\text{RV}_{\delta(r,0|r)} = \frac{1}{2}\left(\sqrt{g^4_{\delta(r,0|r)} + 4g^2_{\delta(r,0|r)}} - g^2_{\delta(r,0|r)}\right) \quad \text{and} \quad \text{RV}_{\zeta(0|r,0)} = \frac{1}{2}\left(\sqrt{g^4_{\zeta(0|r,0)} + 4g^2_{\zeta(0|r,0)}} - g^2_{\zeta(0|r,0)}\right), \tag{17}$$

where $g_{\delta(r,0|r)} = \frac{|\hat{\delta}_{\text{res}}(r,0 \mid r)|}{|\alpha_r|\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})\text{df}}}$ and $g_{\zeta(0|r,0)} = \frac{|\hat{\zeta}_{\text{res}}(0 \mid r,0)|}{|\alpha_r|\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})\text{df}}}$. A proof is given in Appendix G.

Next, we define $\text{RV}_{\alpha=0.05}$ as the strength of association that will change the significance of the estimated disparity reduction or remaining at the $\alpha = 0.05$ level, assuming an equal association to the mediator and the outcome. While the RV for disparity reduction and remaining can be computed easily from regression results, the RV_α cannot be computed easily. Therefore, we use a computational approach to obtain an approximate value of RV_α. Specifically, we find combinations of two sensitivity parameters ($R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$ and $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$) that make the 95% CI of disparity reduction (or remaining) to cover approximately zero (i.e., $\delta(r, 0|r) \pm t_{0.05,\text{df}}se(\delta(r, 0|r)) < 0.001$). Once the combinations of two sensitivity parameters are identified that will make the CI approximately cover zero, we compute the average value of the two sensitivity parameters.

6.4 Illustration using R^2-based sensitivity analysis

Section 5.3 presents sensitivity analysis using regression coefficients. This section presents the same sensitivity analysis, but it is parameterized in R^2 values as defined in equation (14). Figure 3 presents the results

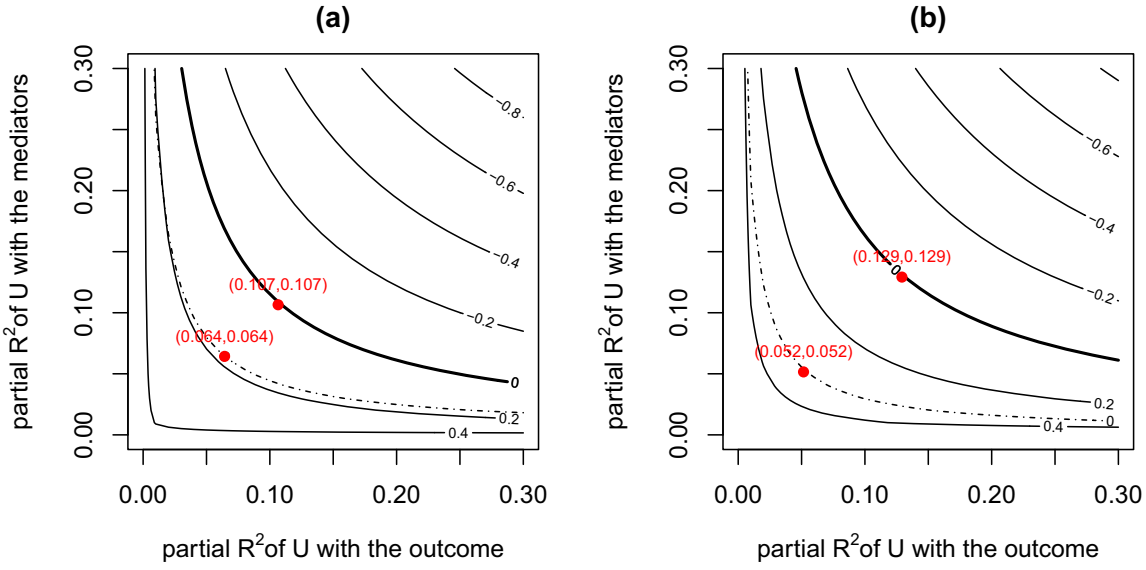

Figure 3: Sensitivity contour plots using R^2 values. (a) Disparity reduction. (b) Disparity remaining. Note (1) Bold lines represent the points at which the estimates become zero. (2) Standard lines represent the points at which the estimates become the respective value (e.g., −0.2, −0.1, 0.1, 0.2). (3) Dashed lines represent the points at which the upper and lower CIs include zero. (4) Red points represents the RVs, i.e., the partial R^2 values that make the estimates or upper/lower limits of CIs zero, assuming equal R^2 values between the two sensitivity parameters.

for the sensitivity analysis of disparity reduction (A) and disparity remaining (B) based on two sensitivity parameters. The two sensitivity parameters are (1) the partial R^2 value of discrimination with CVH given the group status, mediator, and observed confounders, namely, $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$ (x-axis), and (2) the partial R^2 value of discrimination with the education level given the group status and observed confounders, namely, $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$ (y-axis). We plot the points at which the estimated disparity reduction (remaining) becomes zero (bold line), and the 95% CIs cover zero (dashed line) given the combination of two sensitivity parameters.

The analysis indicates that the disparity reduction would still be negative (i.e., education significantly reduces the CVH gap between Black women and White men) if the unobserved confounder explains less than 10.7% of the variance of the mediator and the outcome after accounting for the existing confounders. The disparity reduction would still be significant at the 95% CI if the unobserved confounder explains less than 6.4% of the variance of the mediator and the outcome after accounting for the existing confounders.

Similarly, the disparity remaining would still be negative if the unobserved confounder explains less than 12.9% of the variance of the mediator and the outcome after accounting for the existing confounders, respectively. The disparity remaining would still be significant at the 95% CI if the unobserved confounder explains less than 5.2% of the variance of the mediator and the outcome after accounting for the existing confounders.

Although the qualitative classifications of effect size depend on the context of studies, we use Cohen's [43] guideline to judge how large the amount of confounding is. The amount of confounding required to change the disparity reduction (10.7%) and remaining (12.9%) estimates from negative to positive is considered medium. The amount of confounding required to change the significance of the disparity reduction (6.4%) and remaining (5.2%) estimates is considered small. These results indicate that the conclusion regarding disparity reduction and remaining due to the mediator (education) could be changed with an unobserved confounder that has a small effect on the mediator and the outcome after controlling for the existing confounders.

7 Discussion

In this study, we developed a set of sensitivity analyses that assesses the sensitivity of disparity reduction and remaining to possible violations of no-unobserved mediator–outcome confounding. Although we used the example of estimating disparity reduction and remaining after intervening on a mediator, the proposed sensitivity analyses can be used in any settings based on interventional indirect and direct effects when the exposure is randomized (or conditionally ignorable given covariates).

Our study contributes to the fast-growing causal decomposition literature in several ways. First, we compared causal decomposition analysis based on interventional effects with causal mediation analysis based on natural effects. More importantly, we clarified that sensitivity analyses developed for natural indirect effects to possible pre-exposure confounding can only assess the sensitivity of disparity reduction when no intermediate confounding exists. We argued that the assumption (no intermediate confounding) is restrictive and may only be met when the mediator is measured shortly after the group status, for example, childhood poverty. Yet, such an assumption would be unrealistic in many studies that examine life-long health disparities between social groups.

Second, we derived general bias formulas for disparity reduction and remaining, which serve as the basis of our proposed sensitivity analyses. The general bias formulas can be used beyond a particular statistical model and are applicable to any variable type of mediator, outcome, or intermediate confounder and do not require any assumptions. Moreover, the same bias formulas apply with pre-exposure and intermediate unobserved confounding. Within the causal mediation framework based on natural indirect effects, it has long been desired by many researchers to address both pre-exposure and intermediate unobserved confounding [21,38,44]. Researchers can handle both types of confounding by utilizing the proposed sensitivity analyses, as the causal estimand is based on interventional direct and indirect effects. We acknowledge that not every situation is suitable for interventional direct and indirect effects. However, in some cases, such as the motivating example, defining interventional indirect effects makes more sense than natural indirect effects.

Third, we proposed a sensitivity analysis based on regression coefficients and R^2 values by extending existing approaches. Our contribution is to derive the conditions under which the existing approaches to sensitivity analysis with unobserved confounding can be applied to disparity reduction and remaining. The regression-based sensitivity analysis provides a straightforward way to assess the sensitivity of the effect estimates even without a specific software program. Compared to Park et al. [10], there are two advantages: (1) the sensitivity analysis does not require a restrictive conditional independence assumption, and (2) sensitivity parameters are not conditional on the descendent of the exposure (group status). Although it hinges on the linearity assumption, we provided extensions that relax this assumption. We reparameterized regression-based sensitivity analysis to the scale-free R^2 values. The R^2-based sensitivity analysis is particularly useful for evaluating the sensitivity of the estimates' statistical inferences. In addition, RVs provide a standard way to compare our findings' sensitivity with that of other studies.

We also acknowledge the limitations of the proposed sensitivity analyses. First, the R^2-based sensitivity analysis is only available for continuous outcomes. An extension to discrete outcomes is left for future research. Second, the proposed sensitivity analysis addresses unobserved confounding fixed in time. Addressing unobserved time-varying confounding would be an important generalization of this research. The time-varying confounding issue cannot be easily resolved within the causal mediation framework based on natural indirect effects, as VanderWeele et al. [22] have pointed out. An alternative would be to use interventional indirect effects. Third, in addition to addressing unobserved confounders, it is crucial to address measurement errors in a mediator. A sensitivity analysis that simultaneously addresses unobserved confounders and measurement errors would be particularly beneficial to researchers investigating the role of psycho-social factors, which are susceptible to measurement errors, in reducing disparities.

Acknowledgments: We would like to thank the editor and reviewers for their comments and suggestions, which helped us to improve the quality of the manuscript.

Funding information: This research was supported by a grant from the American Educational Research Association which receives funds for its 'AERA Grants Program' from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those AERA or NSF.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The data and code utilized for the current study are accessible in the personal GitHub repository at https://github.com/soojinpark33/Sensitivity-Analysis-for-CDA.

References

[1] Olkin I, Finn JD. Correlations redux. Psychol Bulletin. 1995 Jul;118(1):155.
[2] Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials or observational studies. Amer J Epidemiol. 1992 Nov 1;136(9):1148–59.
[3] Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Meth. 2010 Dec;15(4):309–34.
[4] VanderWeele TJ, Robinson WR. On causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology (Cambridge, Mass.). 2014 Jul;25(4):473.
[5] Jackson JW, VanderWeele TJ. Decomposition analysis to identify intervention targets for reducing disparities. Epidemiology (Cambridge, Mass.). 2018 Nov;29(6):825–35.
[6] Jackson JW. Explaining intersectionality through description, counterfactual thinking, and mediation analysis. Social Psychiat Psychiat Epidemiol. 2017 Jul;52(7):785–93.
[7] Jackson JW. On the interpretation of path-specific effects in health disparities research. Epidemiology. 2018 Jul 1;29(4):517–20.
[8] Jackson JW. Meaningful causal decompositions in health equity research: definition, identification, and estimation through a weighting framework. Epidemiology. 2020 Nov 5;32(2):282–90.
[9] Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: defining effects based on what we want to learn. Psychol Met. 2021 Apr;26(2):255.
[10] Park S, Qin X, Lee C. Estimation and sensitivity analysis for causal decomposition in health disparity research. Sociol Meth Res. 2020 Aug;28:00491241211067516.
[11] Lundberg I. The gap-closing estimand: a causal approach to study interventions that close disparities across social categories. Sociol Meth Res. 2022:00491241211055769.
[12] Pearl J. The causal mediation formula–a guide to the assessment of pathways and mechanisms. Prevent Sci. 2012 Aug;13(4):426–36.
[13] Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. Oxf Stat Sci Ser. 2003 Jan 1;70–82.
[14] Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. J R Stat Soc Ser B (Statist Methodol). 2020 Feb;82(1):39–67.
[15] Link BG, Phelan J. Social conditions as fundamental causes of disease. J Health Soc Behav. 1995 Jan 1;80–94.
[16] Glymour MM, Clark CR, Patton KK. Socioeconomic determinants of cardiovascular disease: recent findings and future directions. Current Epidemiol Reports. 2014 Jun;1(2):89–97.
[17] Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. Amer J Public Health. 1992 Jun;82(6):816–20.
[18] Suglia SF, Koenen KC, Boynton-Jarrett R, Chan PS, Clark CJ, Danese A, et al. Childhood and adolescent adversity and cardiometabolic outcomes: a scientific statement from the American Heart Association. Circulation. 2018 Jan 30;137(5):e15–28.
[19] Pearl J. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. New York: Morgan Kaufmann Publishers Inc. 2022; p. 411–20.
[20] Kaufman JS. Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. Soc Sci Med. 2008 Apr 1;66(8):1659–69.
[21] VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. Epidemiol Meth. 2014 Jan 3;2(1):95–115.
[22] VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator–outcome confounder. Epidemiology (Cambridge, Mass.). 2014 Mar;25(2):300.
[23] Holland PW. Statistics and causal inference. J Amer Stat Assoc. 1986 Dec 1;81(396):945–60.
[24] Park S, Kang S, Lee C. Choosing an optimal method for causal decomposition analysis: a better practice for identifying contributing factors to health disparities. 2021 Sep 14. arXiv: 2109.06940.

[25] Sudharsanan N, Bijlsma MJ. Educational note: causal decomposition of population health differences using Monte Carlo integration and the g-formula. Int J Epidemiol. 2021 Dec;50(6):2098–107.
[26] Robins JM. Model with that of DAGs representing the non-parametric structural equations. Highly Struct Stochastic Sys. 2003;27:70.
[27] Pearl J. Causality. Cambridge University Press; 2009 Sep 14.
[28] Jackson JW, VanderWeele TJ. Intersectional decomposition analysis with differential exposure, effects, and construct. Soc Sci Med. 2019 Apr 1;226 254–9.
[29] Assari S. Blacks' diminished health returns of educational attainment: health and retirement study. J Med Res Innov. 2020 May 31;4(2):e000212.
[30] Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. Epidemiology. 2006 May;1:276–84.
[31] van der Laan MJ, Petersen ML. Direct effect models. Int J Biostat. 2008 Oct 28;4(1).
[32] Geneletti S. Identifying direct and indirect effects in a non-ï£¡counterfactual framework. J R Stat Soc Ser B (Stat Meth). 2007 Apr;69(2):199–215.
[33] Didelez V, Dawid P, Geneletti S. Direct and indirect effects of sequential treatments. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence. 2006. p. 138–146.
[34] Zheng W, van der Laan M. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. J Causal Inference. 2017 Sep 1;5(2). p. 20160006.
[35] VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. J R Stat Soc Ser B (Statist Meth). 2017 Jun;79(3):917–38.
[36] VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology (Cambridge, Mass.). 2010 Jul;21(4):540.
[37] Hong G, Qin X, Yang F. Weighting-based sensitivity analysis in causal mediation studies. J Educat Behav Stat. 2018 Feb;43(1):32–56.
[38] Imai K, Yamamoto T. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Politic Anal. 2013;21(2):141–71.
[39] VanderWeele TJ, Chiba Y. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator–outcome confounders. Epidemiol Biostat Public Health. 2014;11(2).
[40] Bey GS, Jesdale B, Forrester S, Person SD, Kiefe C. Intersectional effects of racial and gender discrimination on cardiovascular health vary among black and white women and men in the CARDIA study. SSM-population Health. 2019 Aug 1;8:100446.
[41] Sobel ME. Asymptotic intervals for indirect effects in structural equations models. S. Leinhart Sociol Methodol. 1982;13:290–312.
[42] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology (Cambridge, Mass.). 2016 May;27(3):368.
[43] Cohen J. Statistical power analysis for the behavior science. New York: Routledge; 2013.
[44] Park S, Esterling KM. Sensitivity analysis for pretreatment confounding with multiple mediators. J Educ Behav Stat. 2021 Feb;46(1):85–108.
[45] Goodman LA. On the exact variance of products. J Amer Stat Assoc. 1960 Dec 1;55(292):708–13.
[46] Kendall MG, Alan S. The advanced theory of statistics. Vol. II and III. New York: Hafner; 1961.

Appendix

A General bias formulas of $\delta(r, 0|r)$ and $\zeta(0|r, 0)$

The bias for disparity reduction for r (bias$(\delta(r, 0|r))$) is defined as the difference between the expected estimate and the true value. The bias equals

$$\begin{aligned}
&E[Y|r, \mathbf{c}] - \sum_{\mathbf{x},m} E[Y|r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) - E[Y|r, \mathbf{c}] \\
&\quad + \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(\mathbf{x}, u|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&= -\sum_{\mathbf{x},m} E[Y|r, \mathbf{x}, m, \mathbf{c}]P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) + \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(\mathbf{x}|r, \mathbf{c}, u)P(m|R \\
&= 0, \mathbf{c})P(u|r, \mathbf{c}) \\
&= -\sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(u|r, \mathbf{x}, m, \mathbf{c})P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&\quad + \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(\mathbf{x}|r, \mathbf{c}, u)P(m|R = 0, \mathbf{c})P(u|r, \mathbf{c}) \\
&= -\sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(u|r, \mathbf{x}, m, \mathbf{c})P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&\quad + \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]\frac{P(u|r, \mathbf{x}, c)}{P(u|r, c)}P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c})P(u|r, \mathbf{c}) \\
&= \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]\{P(u|r, \mathbf{x}, c) - P(u|r, \mathbf{x}, m, \mathbf{c})\}P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}).
\end{aligned} \tag{A1}$$

The second equality holds because of the law of total probability. The third equality is due to Bayes' theorem.

The last expression of equation (A1) is the general bias formula for $\delta(r, 0|r)$. Since the bias for $\tau(r, 0)$ due to omitted variable U is zero, the bias for $\zeta(0|r, 0) = -\text{bias}(\delta(r, 0|r))$. This completes the proof.

B Sensitivity analysis using regression coefficients

Suppose that the following regression models for Y and U are correctly specified as equation (10). Then the bias for disparity reduction is given by

$$\begin{aligned}
&= \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]\{P(u|r, \mathbf{x}, c) - P(u|r, \mathbf{x}, m, \mathbf{c})\}P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&= \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]\left\{\sum_{m} P(u|r, \mathbf{x}, m, c)P(m|r, \mathbf{x}, c)\right\}P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&\quad - \sum_{\mathbf{x},m,u} E[Y|r, \mathbf{x}, m, \mathbf{c}, u]P(u|r, \mathbf{x}, m, \mathbf{c})P(\mathbf{x}|r, \mathbf{c})P(m|R = 0, \mathbf{c}) \\
&= \beta + \beta_r + \beta_x E[X|r, c] + \beta_m E[M|R = 0, c] + \beta_c c + \beta_u\{\delta + \delta_r + \delta_x E[X|r, c] + \delta_m(\alpha + \alpha_r + \alpha_c c) + \delta_c c\} \\
&\quad - [\beta + \beta_r + \beta_x E[X|r, c] + \beta_m E[M|R = 0, c] + \beta_c c + \beta_u\{\delta + \delta_r + \delta_x E[X|r, c] + \delta_m(\alpha + \alpha_c c) + \delta_c c\}] \\
&= \beta_u \delta_m \alpha_r.
\end{aligned} \tag{A2}$$

This completes the proof.

C Bias formulas based on the coefficients of determination

This proof is a straightforward extension of Cinelli and Hazlett [14]. By using equation (11), the bias for disparity reduction is given by

$$\begin{aligned}\text{bias}(\delta(r,0|r)) &= \alpha_r\beta_u\delta_m \\ &= \alpha_r \times \frac{\text{Cov}(Y,U|r,\mathbf{X},M,\mathbf{C})}{\text{Var}(U|r,\mathbf{X},D,M,\mathbf{C})} \times \frac{\text{Cov}(U,M|r,\mathbf{X},\mathbf{C})}{\text{Var}(M|r,\mathbf{X},\mathbf{C})} \\ &= \alpha_r \times \frac{\text{Cor}(Y,U|r,\mathbf{X},M,\mathbf{C})\sqrt{\text{Var}(Y|r,\mathbf{X},M,\mathbf{C})}}{\sqrt{\text{Var}(U|r,\mathbf{X},M,\mathbf{C})}} \frac{\text{Cor}(U,M|r,\mathbf{X},\mathbf{C})\sqrt{\text{Var}(U|r,\mathbf{X},\mathbf{C})}}{\sqrt{\text{Var}(M|r,\mathbf{X},\mathbf{C})}},\end{aligned} \tag{A3}$$

where Cov and Cor represent the covariance and correlation, respectively. The third equality is because $\text{Cov}(A,B|C) = \text{Cor}(A,B|C)\sqrt{\text{Var}(A|C)}\sqrt{\text{Var}(B|C)}$.

Suppose that the partial R^2 value of unmeasured confounder U with the outcome for $R = r$ given $\mathbf{X}$, M, and $\mathbf{C}$ be denoted as $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$, and suppose also that the partial R^2 value of unmeasured confounder U with the mediators for $R = r$ given $\mathbf{X}$ and $\mathbf{C}$ be denoted as $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$. Then, given equations (5) and (13), the absolute value of the bias can be expressed as follows:

$$\begin{aligned}|\text{bias}(\delta(r,0|r))| &= |\alpha_r|\sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}} \times \frac{\sqrt{\text{Var}(Y|r,\mathbf{X},M,\mathbf{C})}}{\sqrt{\text{Var}(M|r,\mathbf{X},\mathbf{C})}} \\ &= |\alpha_r|\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}\sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\text{df}}.\end{aligned} \tag{A4}$$

The first equality is derived from $\frac{\text{Var}(U\mid r,\mathbf{X},\mathbf{C})}{\text{Var}(U\mid r,\mathbf{X},M,\mathbf{C})} = \frac{1}{1-R^2_{U\sim M|r,\mathbf{X},\mathbf{C}}} = \frac{1}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}$ and $\text{Cor}^2(A,B|C) = R^2_{A\sim B|C}$. The second equality holds because $\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})} = \frac{\sqrt{\text{Var}(Y\mid r,\mathbf{X},M,\mathbf{C})}}{\sqrt{\text{Var}(M\mid r,\mathbf{X},\mathbf{C})}}\sqrt{\frac{1}{\text{df}}}$, where β_m and df are obtained from the outcome model shown in equation (13). This completes the proof.

D Standard errors of disparity reduction and remaining

We first calculate the standard error for disparity reduction. The standard errors for $\hat{\beta}_{\text{res},m}$ and $\hat{\beta}_m$ can be obtained, respectively, as follows:

$$\begin{aligned}\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})} &= \frac{\sqrt{\text{Var}(Y|r,\mathbf{X},M,C)}}{\sqrt{\text{Var}(M|r,\mathbf{X},C)}}\sqrt{\frac{1}{\text{df}}} \\ \sqrt{\text{Var}(\hat{\beta}_m)} &= \frac{\sqrt{\text{Var}(Y|r,\mathbf{X},M,C,U)}}{\sqrt{\text{Var}(M|r,\mathbf{X},C,U)}}\sqrt{\frac{1}{\text{df}-1}}.\end{aligned} \tag{A5}$$

The ratio of these standard errors is

$$\frac{\sqrt{\text{Var}(\hat{\beta}_m)}}{\sqrt{\text{Var}(\hat{\beta}_{\text{res},m})}} = \frac{\sqrt{(1-R^2_{Y\sim U|r,\mathbf{X},M,C})}}{\sqrt{(1-R^2_{M\sim U||r,\mathbf{X},C})}}\sqrt{\frac{\text{df}}{\text{df}-1}}, \tag{A6}$$

because $\sqrt{(1-R^2_{Y\sim U|r,\mathbf{X},M,C})} = \frac{\sqrt{\text{Var}(Y\mid r,\mathbf{X},M,C,U)}}{\sqrt{\text{Var}(Y\mid r,\mathbf{X},M,C)}}$ and $\sqrt{(1-R^2_{M\sim U||r,\mathbf{X},C})} = \frac{\sqrt{\text{Var}(M\mid r,\mathbf{X},C,U)}}{\sqrt{\text{Var}(M\mid r,\mathbf{X},C)}}$. Therefore, we have $\text{Var}(\hat{\beta}_m) = \text{Var}(\hat{\beta}_{\text{res},m})(\frac{1-R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\frac{\text{df}}{\text{df}-1})$.

Using the parametric identification result shown in Section 3.2, the standard error for disparity reduction for $R = r$ can be calculated approximately using the Delta method [41] as follows:

$$\mathrm{Var}(\hat{\delta}(r,0|r)) \approx \alpha_r^2 \mathrm{Var}(\hat{\beta}_m) + \beta_m^2 \mathrm{Var}(\hat{\alpha}_r), \tag{A7}$$

where $\beta_m = \beta_{\mathrm{res},m} - \sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})}\sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} \times R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\mathrm{df}}$ and $\mathrm{Var}(\hat{\beta}_m) = \mathrm{Var}(\hat{\beta}_{\mathrm{res},m})(\frac{1-R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\frac{\mathrm{df}}{\mathrm{df}-1})$.

Next, we calculate the standard error for disparity remaining. By using $\tau(r,0) = \delta(r,0|r) + \zeta(0|r,0)$, we have

$$\begin{aligned}
\mathrm{Var}(\hat{\zeta}(0|r,0)) &= \mathrm{Var}(\hat{\tau}(r,0) - \hat{\delta}(r,0|r)) \\
&= \mathrm{Var}(\hat{\tau}(r,0)) + \mathrm{Var}(\hat{\delta}(r,0|r)) - 2\mathrm{Cov}(\hat{\tau}(r,0), \hat{\delta}(r,0|r)) \\
&= \mathrm{Var}(\hat{\tau}(r,0)) + \mathrm{Var}(\hat{\delta}(r,0|r)) - 2\mathrm{Cov}(\hat{\tau}(r,0), \hat{\delta}_{\mathrm{res}}(r,0|r)) \\
&\quad + 2\mathrm{Cov}(\hat{\tau}(r,0), \mathrm{bias}(\delta(r,0|r))) \\
&\approx \mathrm{Var}(\hat{\tau}(r,0)) + \mathrm{Var}(\hat{\delta}(r,0|r)) - 2\mathrm{Cov}(\hat{\tau}(r,0), \hat{\delta}_{\mathrm{res}}(r,0|r)) \\
&\quad + 2kE[\sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})}]\mathrm{Cov}(\hat{\tau}(r,0), \hat{\alpha}_r) + 2kE[\hat{\alpha}_r]\mathrm{Cov}(\hat{\tau}(r,0), \sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})}),
\end{aligned} \tag{A8}$$

where $k = \sqrt{\frac{R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}{1-R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}}\mathrm{df}}$. The third equality holds because $\hat{\delta}(r,0|r) = \hat{\delta}_{\mathrm{res}}(r,0|r) - \mathrm{bias}(\delta(r,0|r))$. The fourth equality follows Goodman [45] and uses the conventional asymptotic approximation procedure [46]. This completes the proof.

E A simulation study

A Monte Carlo simulation study is conducted to investigate the performance of the proposed standard error estimators of disparity reduction and remaining given in equations (15) and (16). We consider varying sample sizes (100, 500, and 1,000) and effect sizes of sensitivity parameters, i.e., partial R^2 values of U with M and U with Y (0.02, 0.13, and 0.26). The sample sizes of 100, 500, and 1,000 cover small, medium, and large data and the sensitivity parameters of 0.02, 0.13, and 0.26 represent small, medium, and large effect sizes, respectively, according to Cohen [43].

For the simulation study, we first generate synthetic population data that have similar characteristics with the MIDUS data in terms of the distributions of variables R, X, M, Y, and C, and the relationships between them. To be specific, we first generated C and R using the distributions of a covariate (age) and race–gender status of the MIDUS data. Then we used equations (4), (5), and (7) to extract the estimates of regression coefficients and generate X, M, and Y, respectively. We then generated U by adjusting the coefficient values to a certain degree so that $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}}$ and $R^2_{M\sim U|r,\mathbf{X},\mathbf{C}}$ becomes 0.02, 0.13, or 0.26. Once the population data are generated, we took a random sample of size $n = \{100, 500, 1{,}000\}$ from the population data and estimated the standard error estimators of disparity reduction and remaining based on which

Table A1: Simulation results: 95% CI coverage rates of the standard error estimators of disparity reduction and remaining

Effect size	Sample size	Disparity reduction	Disparity remaining
Small	100	0.916	0.931
	500	0.934	0.942
	1,000	0.927	0.935
Medium	100	0.921	0.930
	500	0.946	0.940
	1,000	0.941	0.941
Large	100	0.932	0.935
	500	0.959	0.942
	1,000	0.951	0.951

the 95% CIs are constructed. The results are reported in Table A1 and are based on 1,000 simulations. We compute 95% CI coverage for nine different conditions.

The 95% CI coverage rate reaches the nominal level (0.95) with a sample size of 500 or larger, yet it reaches more quickly with medium and large effect sizes of sensitivity parameters. With a small effect size and a sample size of 100, the coverage is as low as 0.916 for disparity reduction. This result suggests that the coverage rate could be lower than expected with a small sample size (e.g., 100) and a small effect size of sensitivity parameters (e.g., 0.02).

F Standard errors for a percent reduction

As the variance of a ratio of two random variables with nonzero means can be approximated using the delta method and a first-order Taylor expansion, an uncertainty measure of a percent reduction can be derived in a similar way as

$$\mathrm{Var}\left(\frac{\hat{\delta}(r,0|r)}{\hat{\tau}(r,0)}\right) \approx \frac{\delta(r,0|r)^2}{\tau(r,0)^2}\left\{\frac{\mathrm{Var}(\hat{\delta}(r,0|r))}{\delta(r,0|r)^2} - 2\frac{\mathrm{Cov}(\hat{\tau}(r,0),\hat{\delta}(r,0|r))}{\delta(r,0|r)\tau(r,0)} + \frac{\mathrm{Var}(\hat{\tau}(r,0))}{\tau(r,0)^2}\right\}, \tag{A9}$$

where an unbiased estimate of $\tau(r,0)$ and $\mathrm{Var}[\hat{\tau}(r,0)]$ can be obtained from the data; $\delta(r,0|r)$ is given by $\delta_{\mathrm{res}}(r,0|r) - \mathrm{bias}(\delta(r,0|r))$ where $\mathrm{bias}(\delta(r,0|r))$ is given by equation (14); $\mathrm{Var}(\hat{\delta}(r,0|r))$ is given by equation (A7); $\mathrm{Cov}(\hat{\tau}(r,0),\hat{\delta}(r,0|r))$ is given by equation (16).

G RVs for disparity reduction and remaining

We define the RV as the strength of association that will explain away the estimated disparity reduction (or remaining), assuming an equal association to the mediator and the outcome, as $R^2_{Y\sim U|r,\mathbf{X},M,\mathbf{C}} = R^2_{M\sim U|r,\mathbf{X},\mathbf{C}} = \mathrm{RV}$. We find the RV such that $|\mathrm{bias}(\delta(r,0|r))| = |\hat{\delta}(r,0|r)|$. Using equation 14, we have

$$|\alpha_r|\sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})}\sqrt{\frac{\mathrm{RV}^2_{\delta(r,0|r)}}{1-\mathrm{RV}_{\delta(r,0|r)}}\mathrm{df}} = |\hat{\delta}(r,0|r)| \tag{A10}$$

Dividing both sides by $|\alpha_r|\sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})}\sqrt{\mathrm{df}}$, we have

$$\sqrt{\frac{\mathrm{RV}^2_{\delta(r,0|r)}}{1-\mathrm{RV}_{\delta(r,0|r)}}} = \frac{|\hat{\delta}(r,0|r)|}{|\alpha_r|\sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})\mathrm{df}}} \tag{A11}$$

We define $g_{\delta(r,0|r)} \equiv \frac{|\hat{\delta}(r,0\,|\,r)|}{|\alpha_r|\sqrt{\mathrm{Var}(\hat{\beta}_{\mathrm{res},m})\mathrm{df}}}$ and solve for $\mathrm{RV}_{\delta(r)}$. Then, $\mathrm{RV}_{\delta(r,0|r)} = \frac{1}{2}(\sqrt{g^4_{\delta(r,0|r)} + 4g^2_{\delta(r,0|r)}} - g^2_{\delta(r,0|r)})$. The RV for $\zeta(0|r,0)$ can be obtained the same way and thus is omitted. This completes the proof.