Practice of Epidemiology

Mediation Analysis for Health Disparities Research

Ashley I. Naimi*, Mireille E. Schnitzer, Erica E. M. Moodie, and Lisa M. Bodnar

* Correspondence to Dr. Ashley I. Naimi, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, 503 Parran Hall, Pittsburgh, PA 15261 (e-mail: ashley.naimi@pitt.edu).

Social epidemiologists often seek to determine the mechanisms that underlie health disparities. This work is typically based on mediation procedures that may not be justified with exposures of common interest in social epidemiology. In this analysis, we explored the consequences of using standard approaches, referred to as the difference and generalized product methods, when mediator-outcome confounders are associated with the exposure. We compared these with inverse probability-weighted marginal structural models, the structural transformation method, doubly robust g-estimation of a structural nested model, and doubly robust targeted minimum loss-based estimation. We used data on 900,726 births from 2003 to 2007 in the Penn Moms study, conducted in Pennsylvania, to assess the extent to which breastfeeding prior to hospital discharge explained the racial disparity in infant mortality. Overall, for every 1,000 births, 3.36 more infant deaths occurred among non-Hispanic black women relative to all other women (95% confidence interval: 2.78, 3.93). Using the difference and generalized product methods to assess the disparity that would remain if everyone breastfed prior to discharge suggested a complete elimination of the disparity (risk difference = −0.87 per 1,000 births; 95% confidence interval: −1.39, −0.35). In contrast, doubly robust methods suggested a reduction in the disparity to 2.45 (95% confidence interval: 2.20, 2.71) more infant deaths per 1,000 births among non-Hispanic black women. Standard approaches for mediation analysis in health disparities research can yield misleading results.

health disparities; marginal structural models; mediation analysis; proportion explained; social epidemiology; structural nested models; targeted minimum loss-based estimation

Abbreviations: CDM, counterfactual disparity measure; MSM, marginal structural model; TMLE, targeted minimum loss-based estimation.

Mediation analysis has long been of interest in medical research. Often, the approach is used to assess the extent to which an exposure-outcome relationship is attributable to a third variable. Such questions are routine in social epidemiology. Commonly used racial/ethnic classifications, measures of socioeconomic position, or characterizations of the neighborhood environment are all associated with several health outcomes throughout the life course (1, 2). Variables representing these constructs are often taken to designate "fundamental" (3) or "upstream" (4) causes that shape the distribution of more proximal risk factors leading to health disparities.

Many studies have sought to determine how these upstream determinants of health are "explained" by more proximal risk factors. Examples are numerous and include serum potassium concentrations in the relationship between race and diabetes (5), cancer stage at diagnosis in the relationship between socioeconomic position and mortality (6), and tobacco consumption in the relationship between neighborhood socioeconomic status and lung cancer (7). Most of these studies rely on a procedure for mediation (the difference method) that yields valid inferences under rather strict conditions (8, 9). Moreover, the additional assumptions required for causal inference are difficult to justify with exposures of typical interest in social epidemiology (10). This has led to some debate on the causal status of "nonmanipulable" exposures, such as race, sex, or socioeconomic position (10–15).

From this debate, some progress has emerged. Recent work has shown how mediation methods can be used to

Am J Epidemiol. 2016;184(4):315–324

assess the magnitude of the disparity that would remain if a downstream risk factor were changed (16). This work focused on key conceptual and analytical issues when race is the exposure of interest. Here, we review several methods for mediation analysis and discuss other complications that can arise in social epidemiology. In particular, we focus on the interpretation of parameter estimates from different approaches when mediator-outcome confounders are associated with the exposure. We provide an overview of 6 approaches, and we present technical details and annotated software code in Web Appendices 1 and 2 (available at http://aje.oxfordjournals.org/).

We explore 2 broad classes of methods. The first is more commonly used but applicable in a narrower range of settings. This class includes the difference method (17) and a generalization of the product method (18, 19). The second is more general in that methods in this class can accommodate exposure-mediator interactions and mediator-outcome confounders associated with the exposure. These include inverse probability-weighted marginal structural models (MSMs) (20), the structural transformation method (also known as sequential g-estimation) (21), g-estimation of a structural nested model (22), and targeted minimum loss-based estimation (TMLE) (23).

We present an example in which we determine the proportion of the racial disparity in infant mortality explained by breastfeeding prior to hospital discharge. We explain how different methods can be used to better exploit existing knowledge about the particular mediator and outcome under study, and we show how these methods yield different results in an empirical analysis of data on 900,726 births in Pennsylvania between 2003 and 2011.

COUNTERFACTUAL DISPARITY MEASURES: DEFINITION

Figures 1A and 1B are causal diagrams (24), where X represents an exposure, M a mediator, and Y an outcome of interest. We represent exposure-outcome confounders as C_{XY} and mediator-outcome confounders as C_{MY}. Four assumptions may be required to estimate direct causal exposure effects (18):

1. No uncontrolled exposure-outcome confounding
2. No uncontrolled mediator-outcome confounding
3. No mediator-outcome confounders affected by the exposure
4. No exposure-mediator interaction on the scale of interest

Assumptions 1–3 are encoded in Figure 1A, which shows that adjusting for C_{XY} and C_{MY} leaves no open back-door path from X to Y (assumptions 1 and 2), and where there is no arrow from X to C_{MY} (assumption 3). If the stable unit treatment value assumption (25) is met for both the exposure and the mediator, if there is no selection or information bias, and if assumptions 1–4 hold, one can use any of the mediation methods we subsequently discuss to estimate the direct causal effect of the exposure. If assumption 4 is violated, the difference method can no longer be used (8, 9). If assumption 3 is violated (as in Figure 1B), neither the difference method nor the generalized product method will yield valid causal inferences.
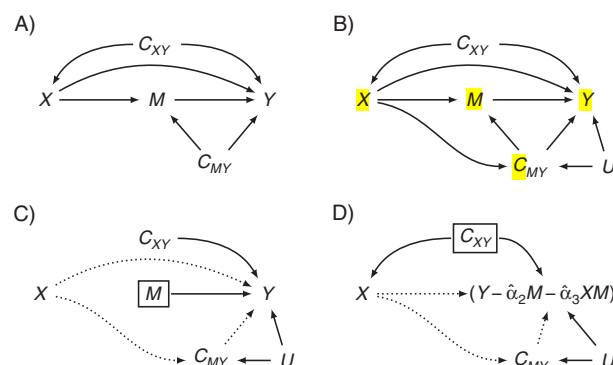

Figure 1. Mediation diagrams for health disparities settings with a social exposure of interest (X), a potential mediator of the health disparity (M), and a health outcome (Y). A) Standard setting with an exposure X, a mediator M, an outcome Y, and measured confounders of the exposure-outcome C_{XY} and mediator-outcome C_{MY}. This diagram assumes that the exposure X is not associated with confounders of the mediator-outcome relationship C_{MY}. B) Setting in which the exposure is associated with confounders of the mediator-outcome relationship. In this diagram, conditioning on C_{MY} using standard regression methods will induce collider bias along the $X \rightarrow C_{MY} \leftarrow U \rightarrow Y$ path and block the portion of the association of interest along the $X \rightarrow C_{MY} \rightarrow Y$ path. C) Scenario encountered after applying inverse probability weights for the mediator and exposure. D) Scenario encountered after a structural transformation by subtracting the effect of the mediator from the outcome.

Fewer assumptions are required to estimate counterfactual disparity measures (CDMs) (16). These might be of interest in cases where X is a social exposure of interest, such as an indicator of maternal race (1 if non-Hispanic black, 0 otherwise), M denotes whether a woman breastfed her infant prior to discharge from the hospital (1 if no, 0 if yes), and Y denotes whether an infant born alive died within the first year of life (1 if yes, 0 if no). We define a CDM of this association on the difference scale as

$$\text{CDM}(m = 0) \equiv E[Y(m = 0)|X = 1] \\ - E[Y(m = 0)|X = 0], \quad (1)$$

where $Y(m = 0)$ is the potential outcome that would be observed if, possibly contrary to fact, a woman began breastfeeding her infant prior to discharge from the place of delivery (26). In equation 1, $\text{CDM}(m = 0)$ represents the magnitude of the racial disparity in infant mortality that would be observed if all women breastfed prior to discharge. To identify and estimate the CDM defined in equation 1, fewer assumptions are required for the exposure, but one still requires assumptions for the mediator.

Because interest lies in the association between race and the outcome, assumption 1 is not required to estimate the CDM defined in equation 1, because part of the association may be due to exposure-outcome confounding (16). In fact, if nonmanipulable exposures cannot be construed as counterfactually defined causes, variables denoted C_{XY} cannot represent confounders in the strict causal sense (27), and our use of arrows emanating from and entering into X in Figure 1

represents a misuse of causal diagrams ([24]). However, one may want to standardize the disparity measure in equation 1 by the distribution of certain covariates associated with the exposure and the outcome. Furthermore, accounting for the uncertain causal status of race (designated X) in Figure [1] is beyond the scope of this article, and thus we maintain this incorrect usage throughout.

We additionally denote confounders of the relationship between breastfeeding prior to discharge and infant mortality as C_{MY}, which includes markers of 1) year of birth, 2) urbanicity, 3) maternal education, 4) paternal education, 5) maternal marital status at pregnancy, 6) maternal participation in the Special Supplemental Nutrition Program for Women, Infants, and Children, 7) birth weight (kg), 8) gestational age at birth (weeks), 9) the interaction between birth weight and gestational age at birth, 10) 5-minute Apgar score, 11) parity, 12) prepregnancy smoking status, 13) gestational smoking status, 14) week at first prenatal visit, 15) total number of prenatal visits, 16) maternal age, and 17) paternal age. These will be used as confounders of the relationship between breastfeeding status prior to discharge and infant mortality in our empirical analysis. Importantly, the manner in which these variables relate to race determines whether standard approaches or more general approaches should be used to estimate the CDM.

COUNTERFACTUAL DISPARITY MEASURES: ESTIMATION

Standard approaches

The difference method. If there is no association between any variables in C_{MY} and race and there is no interaction between race and breastfeeding status, then one may use the difference method to estimate the CDM. To do this, we fit a model for the relationship between race and infant mortality with adjustment for the variables in C_{XY} and C_{MY} as

$$E(Y|X, C_{XY}, C_{MY}) = \alpha_0 + \alpha_1 X + \alpha'_2 C_{XY} + \alpha'_3 C_{MY}. \quad (2)$$

In the absence of exposure-mediator interactions, adding the mediator to this model yields

$$E(Y|X, M, C_{XY}, C_{MY}) = \beta_0 + \beta_1 X + \beta_2 M + \beta'_3 C_{XY} + \beta'_4 C_{MY}. \quad (3)$$

It is easy to show (Appendix) that under certain additional assumptions required for causal inference (counterfactual consistency ([28]), no interference ([29]), correct model specification ([30]), positivity ([31])), $\beta_1 \equiv \text{CDM}(m = 0)$, and thus one can calculate the proportion of the disparity that would be eliminated as $(\alpha_1 - \beta_1)/\alpha_1$ ([32]). In the medical literature, this measure is often interpreted as the extent to which the disparity can be prevented or reduced if the mediator is changed ([5]). However, if race is associated with confounders of the mediator-outcome relationship (as in Figure [1]B), then adjusting for C_{MY} in model 2 will yield a parameter α_1 that does not meaningfully correspond to a total association, because part of the association between race and the outcome occurs through C_{MY}.

The generalized product method. If none of the variables in C_{MY} are associated with race but there is an interaction between race and breastfeeding status, then one may use a generalization of the product method ([18], [19]) to estimate the CDM. To implement the approach, we fit a simple linear regression model as

$$E(Y|X, M, C_{XY}, C_{MY}) = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_3 XM + \gamma'_4 C_{XY} + \gamma'_5 C_{MY}. \quad (4)$$

The parameter estimate for race (γ_1) from this model can then be interpreted as the CDM($m = 0$). However, as with the difference method, if race is associated with confounders of the mediator-outcome relationship (as in Figure [1]B), then adjusting for C_{MY} in model 4 will yield a parameter γ_1 that must be interpreted as the association that would remain if all women breastfed *and* if all the variables in C_{MY} were set to their referent levels.

More general approaches

As with many variables of common interest in social epidemiology, race will often be associated with variables that act as mediator-outcome confounders. In particular, it is in fact the case that birth weight and gestational age, which are included in C_{MY}, are strongly associated with race and are also strong determinants of the outcome in our empirical example. Thus, the CDM($m = 0$) can be thought of as the combined magnitude of the arrows in Figure [1]B from X to Y if M is set to its referent level. When mediator-outcome confounders are associated with the exposure, more general approaches that include the methods we subsequently discuss must be used. To estimate the CDM, these methods require that there be no uncontrolled confounding of the mediator-outcome relationship. When controlled direct effects are of interest, no exposure-outcome confounding is also a requirement.

We next discuss several methods for dealing with such issues in the context of health disparities research. Following Robins et al. ([33]), we comment on how the variation in the assumptions underlying these methods might be used to take advantage of prior knowledge on the particular mediator and outcome under study. In effect, we consider when to adjust for confounding by either 1) modeling the mediator, 2) modeling the outcome, or 3) modeling both the mediator and the outcome to obtain a doubly robust estimate of the CDM.

Inverse probability-weighted MSMs. Inverse probability-weighted MSMs ([20]) can be used to estimate CDMs. The approach proceeds by modeling the mediator and exposure, generating inverse probability weights from these models, and fitting a weighted regression model of the outcome against the exposure, the mediator, and their interaction. The weights can be obtained as

$$sw = \frac{f_X(X)}{f_X(X|C_{XY})} \times \frac{f_M(M)}{f_M(M|X, C_{XY}, C_{MY})}, \quad (5)$$

where $f_X(\cdot)$ and $f_M(\cdot)$ are the probability density functions for X and M, respectively. As explained elsewhere ([20], [34]), correct specification of the models for the denominator of sw yields an unbiased estimate of the parameter of interest. To correctly estimate the CDM, this approach relies on the

assumption that the mediator model is correctly specified as a function of all mediator-outcome confounders.

Conceptually, inverse probability weighting removes the association between C_{XY} and X, and between C_{MY} and M in Figure 1B, enabling one to "erase" arrows, yielding Figure 1C. Using this approach, one estimates the combined magnitude of dashed arrows in Figure 1C. This approach enables estimation of the CDM on the risk difference, risk ratio, or odds ratio scale using weighted binomial regression models with the appropriate link function (20).

Problems may be encountered if the predicted probabilities of the observed exposure or mediator are small. This may occur if, for example, the mediator is continuous, or if very few (or very many) individuals in a given stratum of C_{MY} have $M = 1$ (or $M = 0$), leading to inverse probability weights for these individuals that would be very large. Other authors have discussed ways to manage such positivity violations when possible (35–37).

Structural transformation. Goetgeluk et al. (38) introduced an approach we call the structural transformation method. To estimate the CDM on the risk difference scale using this approach, one proceeds by modeling the outcome using an identity link function, as

$$E(Y|X, M, C_{XY}, C_{MY}) = \alpha_0 + \alpha_1 X + \alpha_2 M + \alpha_3 XM$$
$$+ \alpha'_4 C_{XY} + \alpha'_5 C_{MY}. \quad (6)$$

Using parameter estimates from this model, one creates a transformed outcome as

$$\tilde{Y} = Y - \hat{\alpha}_2 M - \hat{\alpha}_3 XM.$$

Finally, one regresses this transformed outcome against the exposure and exposure-outcome confounders:

$$E(\tilde{Y}|X, C_{XY}) = \beta_0 + \psi_1 X + \beta'_1 C_{XY}. \quad (7)$$

The parameter estimate, $\hat{\psi}_1$, can be interpreted as the CDM defined in equation 1. To estimate standard errors for $\hat{\psi}_1$, one can use a modified sandwich variance estimator (21, 39) or the bootstrap. The approach can be used with linear or log-linear risk models, but it cannot be used with logistic models due to noncollapsibility (21).

Conceptually, the structural transformation method subtracts the effect of the mediator from the outcome and regresses the remainder against the exposure (39). The approach yields an estimate of the combined magnitude of the dashed arrows in Figure 1D (40). Vansteelandt (21) discusses use of this approach (termed "sequential g-estimation") in different settings, including with case-control designs. Unlike inverse probability-weighted MSMs, this approach relies on correct specification of the outcome model (models 6 and 7). This distinction serves as a basis for choosing whether to use inverse probability weighting (i.e., modeling the mediator) or the structural transformation method (modeling the outcome) to estimate the CDM (33). For example, because infant mortality is rare, nonconvergence may be an issue with a flexible linear model for the outcome. In this case, it may be preferable to adjust for confounding using inverse probability-weighted MSMs. In contrast, if specifying a flexible model for the mediator led to highly variable weights, use of the structural transformation method might be preferred.

G-estimation of a structural nested mean model. Alternatively, one may choose to posit a structural nested mean model (22) for the CDM as

$$E[Y(m) - Y(m = 0)|M = m, X = x] = \gamma_2 m + \gamma_3 xm,$$
$$E[Y(m = 0)|X = x] - E[Y(m = 0)|X = 0] = \psi_1 x$$

where, for a binary x, ψ_1 represents the CDM($m = 0$). These parameters can be estimated using a modified doubly robust g-estimator (33, 41). The approach can be divided into 2 stages, each with 2 steps. The first stage is to estimate γ_2 and γ_3 in the structural nested model. In step 1, we obtain predicted mediator probabilities (denoted $\hat{p}(M)$) from, for example, a logistic model such as

$$\hat{p}(M) = \{1 + \exp[-\hat{\delta}_0 - \hat{\delta}_1 X - \hat{\delta}'_2 C_{XY} - \hat{\delta}'_3 C_{MY}]\}^{-1}, \quad (8)$$

which can be used to create mediator residuals $\hat{r}(M) = [M - \hat{p}(M)]$. Step 2 is to regress the outcome against the mediator residuals, the interaction between the mediator residuals and the exposure, and confounders C_{MY} and C_{XY} using linear regression:

$$E[Y|X, M, C_{XY}, C_{MY}] = \beta_0 + \gamma_2 \hat{r}(M) + \gamma_3 X \hat{r}(M)$$
$$+ \beta_1 C_{XY} + \beta_2 C_{MY}, \quad (9)$$

where γ_2 and γ_3 correspond to the desired parameters for the mediator's effect in the structural nested mean model above.

Once γ_2 and γ_3 are estimated, one can proceed with the second stage to estimate ψ_1 using the same procedure. In step 1, estimate

$$\hat{p}(X) = \{1 + \exp[-\hat{\xi}_0 - \hat{\xi}'_1 C_{XY}]\}^{-1}. \quad (10)$$

Create a transformed outcome as

$$\tilde{Y} = Y - \hat{\gamma}_2 M - \hat{\gamma}_2 XM, \quad (11)$$

which is the outcome with the mediator effect removed. Finally, regress the transformed outcome against the exposure residuals and confounders of the exposure-outcome relationship in a linear regression model. Parameters for this model can be obtained using, for example, a standard ordinary least squares estimator:

$$E(\tilde{Y}|X, C_{XY}) = \theta_0 + \psi_1 \hat{r}(X) + \theta C_{XY}. \quad (12)$$

By regressing the transformed outcome against the residual exposure and confounders of the exposure-outcome relationship, one can implement a modified doubly robust g-estimator using standard software routines (21, 33).

In large data sets, the standard (i.e., uncorrected) robust variance estimator (42, 43) will yield conservative confidence intervals for ψ_1 (44), but the bootstrap may also be used. In contrast to inverse probability-weighted MSMs or the structural transformation method, the OLS-based g-estimator we present is doubly robust. This means that the CDM($m = 0$) will be consistently estimated if either 1) the model used to generate $\hat{r}(M)$ is correctly specified, 2) the models that regress the outcome against the residuals and other covariates are correctly specified, or 3) both conditions (1 and 2) hold (45).

Targeted minimum loss-based estimation. A final approach for estimating the CDM is TMLE (23). One can

implement this procedure by first fitting the following regression models:

$$\text{logit}\{E[X|C_{XY}]\} = \theta_0 + \theta'_1 C_{XY}, \qquad (13)$$

$$\text{logit}\{E[M|X, C_{MY}, C_{XY}]\} = \beta_0 + \beta_1 X + \beta'_2 C_{MY} \\ + \beta'_3 C_{XY}, \qquad (14)$$

and

$$\text{logit}\{E[Y|X, M, C_{MY}, C_{XY}]\} \\ = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_3 XM + \gamma'_4 C_{MY} + \gamma'_5 C_{XY}, \quad (15)$$

where $\text{logit}\{\cdot\} = \dfrac{\log(\cdot)}{\log(1-\cdot)}$. Once these models have been fitted, we obtain predicted values from each model under $X = 1$ and $M = 0$ for all individuals. We refer to the outcome model predictions (equation 15) under $X = x$ and $M = m$ as $Q_2^{x,m}$. Using these outcome model predictions, we estimate a "fluctuation parameter," denoted ε_2, from a no-intercept logistic regression model:

$$\text{logit}\{E[Y|X = 1, M = 0, C_{MY}, C_{XY}]\} = \\ \varepsilon_2 \frac{I(X = 1, M = 0)}{\{1 - \hat{E}[M|X = 1, C_{MY}, C_{XY}]\}\hat{E}[X|C_{XY}]} + \text{logit}[Q_2^{1,0}], \qquad (16)$$

where $\text{logit}[Q_2^{1,0}]$ is entered as an offset in the model. The fluctuation parameter, ε_2, captures the degree of residual confounding in the predictions $Q_2^{1,0}$. If these predictions are unbiased, $\hat{\varepsilon}_2$ will be zero (46). The predictions $\hat{E}[\bullet]$ in the denominator of

$$\frac{I(X = 1, M = 0)}{\{1 - \hat{E}[M|X = 1, C_{MY}, C_{XY}]\}\hat{E}[X|C_{XY}]}$$

are obtained from the models for M and X, where $I(\bullet)$ denotes an indicator function equal to 1 if $\bullet$ is true (0 otherwise). Once fitted, we generate updated predictions, denoted $Q_2^{*1,0}$, from model 16, setting $X = 1$ and $M = 0$ for all subjects. We regress these updated predictions against X and C_{XY} using logistic regression:

$$\text{logit}\{E[Q_2^{*1,0}|X, C_{XY}]\} = \beta_0 + \beta_1 X + \beta_2 C_{XY}, \qquad (17)$$

and we obtain predicted values for each individual from this model under $X = 1$, which we denote $Q_1^{1,0}$. We then estimate a second fluctuation parameter, denoted ε_1, from the following model:

$$\text{logit}\{E[Q_2^{*1,0}|X = 1, C_{XY}]\} \\ = \varepsilon_1 \frac{I(X = 1)}{\hat{E}[X|C_{XY}]} + \text{logit}[Q_1^{1,0}], \qquad (18)$$

and produce predicted values $Q_1^{*1,0}$ from this model under $X = 1$. Taking the average of these predicted values over all individuals yields an estimate of $E[Y(0)|X = 1]$. To obtain

an estimate of $E[Y(0)|X = 0]$, we repeat the entire process, replacing all instances of $X = 1$ in the above equations with $X = 0$. We can then take the difference of the averages of $Q_1^{*1,0}$ and $Q_1^{*0,0}$ as estimates of the CDM($m = 0$). Analytical standard errors can be obtained by means of the delta method using the equations provided in section 1 of Web Appendix 1. Alternatively, the bootstrap may be used.

Conceptually, TMLE starts by obtaining initial outcome predictions under the desired exposure-mediator combination, such as $X = 1$ and $M = 0$. However, these predictions may be biased. The next TMLE step is to allow the outcome predictions to "fluctuate" against a variable that enables us to target the same outcome prediction in another way. This variable, referred to as a "clever covariate" (47, p. 73), is an inverse probability weight that serves as an independent variable in regression models 16 and 18. This clever covariate confers TMLE with a doubly robust property, potentially reducing the bias of the initial predictions. Like g-estimation, TMLE is a consistent (doubly robust) estimator of the CDM if either model 17 (for the mediator) or model 15 (for the outcome) is correct.

Several techniques are available for estimating the conditional probabilities from models 13, 14, and 15. We used standard regression. However, the cross-validated Super Learner (48) can combine parametric and nonparametric (e.g., machine learning) methods into a single algorithm to generate the predicted values required for all TMLE steps.

It is important to note that the logistic regression models 16, 17, and 18 should be estimated by solving the score of the binomial likelihood function (49). Epidemiologists regularly accomplish this when fitting logistic regression models via standard software routines, such as PROC LOGISTIC in SAS (SAS Institute, Inc., Cary, North Carolina). However, these routines cannot be used for fitting models 17 or 18, because the dependent variables ($Q_2^{*1,0}$) can take on values other than 0 or 1. Thus, to solve the binomial likelihood function for models 17 or 18, less standard routines, such as PROC MODEL or PROC NLMIXED, must be used.

Table 1 presents a brief summary and comparison of these general methods. Additional details are available in section 2 of Web Appendix 1. Tables 2 and 3 present an outline of the steps required for g-estimation and TMLE.

Example data and empirical analysis

To facilitate implementation, we provide example software code and simulated data to fit each of the 4 general methods to estimate the CDM on the difference scale in Web Appendix 2. In these simulated data, the true CDM($m = 0$) = 0.05 on the difference scale. When the exposure is manipulable and counterfactual outcomes are well-defined, this same code can be used to estimate the controlled direct effect on the difference scale (50).

To examine how different methods affect interpretation in an empirical setting, we estimated the magnitude of the racial disparity in infant mortality that would remain if every woman breastfed her infant prior to discharge from the place of birth. Our sample consisted of 900,726 live-born singletons from Pennsylvania. Data were obtained from linked birth/infant-death records from 2003 to 2011. We used complete case analysis

Table 1.   Comparison of Methods for Estimation of Counterfactual Disparity Measures or Controlled Direct Effects

| Method | Modeling Requirement for Consistency | | Doubly Robust? |
	CDM	CDE	
IPW MSMs	Correct specification of mediator model	Correct specification of both mediator and exposure models	No
Structural transformation	Correct specification of full outcome model	Correct specification of both full outcome model and exposure-outcome model	No
G-estimation	Correct specification of the outcome model or correct specification of mediator model	Correct specification of the outcome model or correct specification of the exposure/mediator models	Yes
TMLE	Correct specification of both the models for the outcome and $E[Q_2|X,C_{XY}]$ or correct specification of the mediator model	Correct specification of both the models for the outcome and $E[Q_2|X,C_{XY}]$ or correct specification of the exposure/mediator models	Yes

Abbreviations: CDE, controlled direct effect; CDM, counterfactual disparity measure; IPW MSMs, inverse probability-weighted marginal structural models; TMLE, targeted minimum loss-based estimation.

to deal with all missing data. Previous reports have shown high sensitivity and moderate false discovery rates for the birth certificate–based measure of breastfeeding status (51). Additional cohort details are available elsewhere (52). In these analyses, C_{XY} was the empty set. We adjusted for confounders of the mediator-outcome relationship outlined in the section on the definition of the CDM. We adjusted for continuous confounders with restricted quadratic splines and for categorical confounders with using dummy variables.

Of the 900,726 children born in our sample, 3,555 infants died between 2003 and 2011. There were 84,405 births to non-Hispanic black women (590 infant deaths) and 816,321 births among women of all other races/ethnicities (2,965 infant deaths). The overall relationship between race and infant mortality was 3.36 (95% confidence interval: 2.78, 3.93) more infant deaths for every 1,000 births among non-Hispanic black women compared with all other women. This estimate was used as our measurement of the total association for all proportion-eliminated calculations.

Table 4 presents the results of our mediation analyses. These values are estimates of the CDM($m = 0$) obtained

using each of the 6 different methods. When using the difference and generalized product methods, we estimated that a hypothetical intervention in which all women breastfed prior to discharge would completely eliminate the disparity in infant mortality between non-Hispanic black and all other women (Table 2). In contrast, inverse probability-weighted MSM estimates suggested that the disparity would be reduced to 2.24 (95% confidence interval: 1.41, 3.07) more infant deaths per 1,000 livebirths among non-Hispanic black women. Use of the structural transformation method suggested a reduction in the disparity to 3.08 (95% confidence interval: 2.51, 3.66) more infant deaths per 1,000 livebirths. Both doubly robust approaches yielded estimates that were no different from the inverse probability-weighted MSM estimate.

DISCUSSION

Social epidemiologists have long been interested in elucidating mechanisms underlying health disparities. Methods for mediation analysis are often used for this purpose. Using

Table 2.   Steps for Implementing a Modified Doubly Robust G-Estimator of a Structural Nested Mean Model

Step	Model	Objective
		Stage 1
Step 1	Model 8[a]	Obtain predicted mediator values and create mediator residuals.
Step 2	Model 9[b]	Regress outcome against predicted mediator residuals, interaction between exposure and mediator residuals, and mediator-outcome confounders using ordinary least squares to obtain mediator effect estimates.
		Intermediate Transformation Between Stages
	Equation 11[c]	Create transformed outcome by removing estimated mediator effect from observed outcome.
		Stage 2
Step 1	Model 10[d]	Obtain predicted exposure values and create exposure residuals.
Step 2	Model 12[e]	Regress transformed outcome against predicted exposure residuals and exposure-outcome confounders using ordinary least squares to estimate counterfactual disparity measure.

[a] $\hat{p}(M) = \{1 + \exp[-\hat{\xi}_0 - \hat{\xi}_1 X - \hat{\xi}'_2 C_{XY} - \hat{\xi}'_3 C_{MY}]\}^{-1}$.
[b] $E[Y|X,M,C_{XY},C_{MY}] = \beta_0 + \gamma_2\hat{r}(M) + \gamma_3 X\hat{r}(M) + \beta_1 C_{XY} + \beta_2 C_{MY}$.
[c] $\tilde{Y} = Y - \hat{\gamma}_2 M - \hat{\gamma}_2 XM$.
[d] $\hat{p}(X) = \{1 + \exp[-\hat{\xi}_0 - \hat{\xi}'_1 C_{XY}]\}^{-1}$.
[e] $E(\tilde{Y}|X,C_{XY}) = \theta_0 + \psi_1\hat{r}(X) + \theta C_{XY}$.

Table 3. Steps for Implementing Doubly Robust Targeted Minimum Loss-Based Estimation[a]

Step	Model	Objective
1	Models 13,[b] 14,[c] and 15[d]	Fit exposure, mediator, and outcome models. Create the inverse probability weights, referred to as a clever covariate, required for models 16 and 18 using predicted values from exposure and mediator models.
2	Model 16[e]	Using logistic regression, regress the observed outcome against the clever covariate, with predicted values from model 15[d] entered as an offset. Generate predictions from this model.
3	Model 17[f]	Regress predicted values from model 16[e] in step 2 against exposure and exposure-outcome confounders. Generate predictions from this model.
4	Model 18[g]	Regress predicted values from model 16[e] in step 2 against the clever covariate, with predicted values from model 17[f] in step 3 entered as an offset. Generate predictions from this model.
5	No model	Take the average of the predictions from model 18[g] in step 4 as the desired estimate.

[a] To estimate $E[Y(m)|X=x]$, predicted values for each step must be obtained by setting $X=x$ and $M=m$ for all individuals in the sample.

[b] $\text{logit}\{E[X|C_{XY}]\} = \theta_0 + \theta'_1 C_{XY}$.

[c] $\text{logit}\{E[M|X, C_{MY}, C_{XY}]\} = \beta_0 + \beta_1 X + \beta'_2 C_{MY} + \beta'_3 C_{XY}$.

[d] $\text{logit}\{E[Y|X, M, C_{MY}, C_{XY}]\} = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_3 XM + \gamma'_4 C_{MY} + \gamma'_5 C_{XY}$.

[e] $\text{logit}\{E[Y|X=x, M=m, C_{MY}, C_{XY}]\} = \varepsilon_2 \dfrac{I(X=x, M=m)}{\{1 - \hat{E}[M|X=x, C_{MY}, C_{XY}]\}\hat{E}[X|C_{XY}]} + \text{logit}[Q_2^{x,m}]$.

[f] $\text{logit}\{E[Q_2^{*x,m}|X, C_{XY}]\} = \beta_0 + \beta_1 X + \beta_2 C_{XY}$.

[g] $\text{logit}\{E[Q_2^{*x,m}|X=x, C_{XY}]\} = \varepsilon_1 \dfrac{I(X=x)}{\hat{E}[X|C_{XY}]} + \text{logit}[Q_1^{x,m}]$.

the racial disparity in infant mortality as an example, we examined complications that can arise when quantifying the extent to which a health disparity is explained by a risk factor of interest using standard regression approaches for mediation analysis. In an analysis of 900,726 births in Pennsylvania,

Table 4. Estimated Counterfactual Disparity and Corresponding Proportion Explained by Breastfeeding for the Association Between Race and Infant Mortality Among 900,726 Births in Pennsylvania, United States, 2003–2011[a]

Method	CDM[b]	95% Confidence Interval	Proportion Explained,[c] %
Difference	−0.87	−1.39, −0.35	126
Generalized product	−0.85	−1.39, −0.30	125
IPW MSMs	2.24	1.41, 3.07	33
Structural transformation	3.08	2.51, 3.66	8
G-estimation	2.33	1.68, 2.97	31
TMLE	2.45	2.20, 2.71	27

Abbreviations: CDM, counterfactual disparity measure; IPW MSMs, inverse probability-weighted marginal structural models; TMLE, targeted minimum loss-based estimation.

[a] All results are expressed as the risk difference per 1,000 livebirths. The referent for the mediator ($m=0$) corresponds to breastfeeding, and thus the CDM corresponds to the racial disparity that would be observed if all women breastfed.

[b] Adjusted for year of birth, urbanicity, maternal and paternal education, birth weight, gestational age, 5-minute Apgar score, parity, prepregnancy smoking, smoking during pregnancy, maternal and paternal ages, prenatal visits (total number and gestational week at initiation), and maternal participation in the Special Supplemental Nutrition Program for Women, Infants, and Children.

[c] Proportion explained = (3.36 − CDM($m=0$))/3.36 × 100.

use of standard methods suggested that a complete reduction of the disparity would be observed if every woman in the population were to breastfeed prior to discharge from the place of birth. This is because the parameter estimate for race from both standard approaches is obtained from a model that simultaneously conditions on mediator-outcome confounders. These confounders include gestational age at birth and birth weight, which are strong determinants of the outcome and are also strongly correlated with race. This accounts for most of the elimination of the racial disparity in our example.

Mishandling of mediator-outcome confounders associated with the exposure has important implications for social epidemiology. Typical exposures in this field are associated with myriad downstream risk factors that may confound a mediator-outcome relationship, making assumption 3, as outlined in the Introduction, unrealistic in most settings. More general approaches separate the adjustment for mediator-outcome confounders from the estimation of the parameter for race. Among them, inverse probability-weighted MSMs and the structural transformation method are easier to use. However, they are singly robust, in that consistent estimation depends on a single regression model for either the outcome or the mediator. When choosing between them, one should first determine whether background knowledge and the available data would allow one to more confidently specify a model for the mediator or for the outcome.

Often, however, the trade-offs between modeling either the exposure or the outcome will be difficult to gauge. G-estimation of a structural nested mean model and TMLE are doubly robust because these estimators are consistent if either the outcome model or the mediator model is correctly specified. In our example analysis, because the 95% confidence intervals for both doubly robust approaches fell within the range of the corresponding confidence intervals for the inverse probability-weighted MSM parameter, we surmise that the

exposure model is driving the consistency of the doubly robust approaches (45). Thus, our analysis also demonstrates the benefits of using doubly robust techniques in addition to more commonly employed singly robust methods.

In this work, we defined a CDM as a causal effect metric with the exposure of interest. This measure is counterfactual insofar as the assumptions required for causal inference for the mediator are met. Furthermore, it is based on the fact that the statistical association between race and the outcome carries relevant disparity-related (but not necessarily causal) information. CDMs correspond directly to research questions about how an intervention might affect a health disparity (16, 53). When the exposure of interest is not subject to counterfactual consistency violations and other assumptions required for causal inference are also met, CDMs are controlled direct effects in the literature on causal mediation analysis (50).

Our example analysis is subject to limitations in that we used a complete case analysis to deal with missing data, which may have biased our results (54). Additionally, we did not have information on postdischarge breastfeeding practices. However, while important, these limitations are tangential to our primary purpose. We sought to illustrate different methods for mediation analysis and to demonstrate potential problems with using standard approaches for effect decomposition in social epidemiology.

A more controversial issue involves our interpretation of the parameter for maternal race/ethnicity, measured via self-report. Use of such exposure variables to infer causality has been the source of some debate (13, 16), and some have questioned the validity of claiming that such racial/ethnic classification schemes are nonmanipulable or noncausal (14, 15). We take this self-reported racial classification scheme to reflect historical and contemporary features of the social, political, and economic systems that characterize the lives of the women in our cohort rather than innate biological characteristics (55–57). In this framework, differences in the probability of breastfeeding prior to discharge (our mediator) and the probability of infant mortality (our outcome) between racial/ethnic groups can be construed as biological expressions of long-standing race relations (58). Such race relations have deep and complicated historical roots (59, 60), making it difficult to identify the infant mortality risk in, for example, non-Hispanic black women that would have been observed under a different set of race relations throughout history. These complications underlie much of the debate on the causal status of race (13). However, by expressing interest in the associations for race, CDMs avoid these issues, while enabling causal inference in a counterfactual framework.

While methods for mediation have proliferated in the last decade, use of standard approaches continues in situations where they may not be applicable. This has important consequences for interpretation. Importantly, the presence of mediator-outcome confounders associated with the exposure precludes the use of standard approaches for mediation analysis. This issue is particularly relevant for research on health disparities.

ACKNOWLEDGMENTS

REFERENCES

1. Gee GC, Walsemann KM, Brondolo E. A life course perspective on how racism may be related to health inequities. *Am J Public Health*. 2012;102(5):967–974.
2. Williams DR, Costa MV, Odunlami AO, et al. Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities. *J Public Health Manag Pract*. 2008;14(suppl): S8–S17.
3. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav*. 1995;Spec No:80–94.
4. Gehlert S, Sohmer D, Sacks T, et al. Targeting health disparities: a model linking upstream determinants to downstream interventions. *Health Aff (Millwood)*. 2008;27(2):339–349.
5. Chatterjee R, Yeh HC, Shafi T, et al. Serum potassium and the racial disparity in diabetes risk: the Atherosclerosis Risk in Communities (ARIC) study. *Am J Clin Nutr*. 2011;93(5):1087–1091.
6. Ibfelt EH, Kjær SK, Høgdall C, et al. Socioeconomic position and survival after cervical cancer: influence of cancer stage, comorbidity and smoking among Danish women diagnosed between 2005 and 2010. *Br J Cancer*. 2013;109(9):2489–2495.
7. Hystad P, Carpiano RM, Demers PA, et al. Neighbourhood socioeconomic status and individual lung cancer risk: evaluating long-term exposure measures and mediating mechanisms. *Soc Sci Med*. 2013;97:95–103.
8. Jiang Z, VanderWeele TJ. When is the difference method conservative for assessing mediation? *Am J Epidemiol*. 2015; 182(2):105–108.
9. Naimi AI. Invited commentary: boundless science—putting natural direct and indirect effects in a clearer empirical context. *Am J Epidemiol*. 2015;182(2):109–114.
10. Naimi AI, Kaufman JS. Counterfactual theory in social epidemiology: reconciling analysis and action for the social determinants of health. *Curr Epidemiol Rep*. 2015;2(1):52–60.
11. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
12. Bollen KA, Pearl J. Eight myths about causality and structural equation models. In: Morgan SL, ed. *Handbook of Causal Analysis for Social Research*. New York, NY: Springer Publishing Company; 2013:301–328.
13. Kaufman JS, Cooper RS. Seeking causal explanations in social epidemiology. *Am J Epidemiol*. 1999;150(2):113–120.
14. Glymour C, Glymour MR. Commentary: race and sex are causes. *Epidemiology*. 2014;25(4):488–490.
15. Krieger N. On the causal interpretation of race. *Epidemiology*. 2014;25(6):937.
16. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*. 2014;25(4):473–484.
17. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Eval Rev*. 1981;5(5):602–619.

18. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2(4):457–468.

19. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*. 2013;18(2):137–150.

20. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects [erratum in *Epidemiology*. 2009;20(4):629]. *Epidemiology*. 2009;20(1):18–26.

21. Vansteelandt S. Estimating direct effects in cohort and case-control studies [erratum in *Epidemiology* 2010:21(2)]. *Epidemiology*. 2009;20(6):851–860.

22. Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour C, Cooper G, eds. *Computation, Causation, and Discovery*. Menlo Park, CA /Cambridge, MA: AAAI Press/The MIT Press; 1999:349–405.

23. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*. 2012;8(1):Article 9.

24. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–688.

25. Rubin DB. Comment: which ifs have causal answers. *J Am Stat Assoc*. 1986;81(396):961–962.

26. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc*. 2005;100(469):322–331.

27. VanderWeele TJ, Shpitser I. On the definition of a confounder. *Ann Stat*. 2013;41(1):196–220.

28. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009; 20(1):3–5.

29. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc*. 2008;103(482):832–842.

30. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol*. 1986; 123(3):392–402.

31. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171(6):674–677.

32. VanderWeele TJ. Policy-relevant proportions for direct effects. *Epidemiology*. 2013;24(1):175–176.

33. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48(2):479–495.

34. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561–570.

35. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.

36. Xiao Y, Abrahamowicz M, Moodie EE. Accuracy of conventional and marginal structural Cox model estimators: a simulation study. *Int J Biostat*. 2010;6(2):Article 13.

37. Naimi AI, Moodie EE, Auger N, et al. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*. 2014;25(2):292–299.

38. Goetgeluk S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. *J R Stat Soc Series B Stat Methodol*. 2008;70(5):1049–1066.

39. Vansteelandt S. Estimation of direct and indirect effects. In: Berzuini C, Dawid P, Bernardinelli L, eds. *Causality: Statistical Perspectives and Applications*. West Sussex, United Kingdom: John Wiley & Sons, Ltd.; 2012:126–150.

40. Loeys T, Moerkerke B, Raes A, et al. Estimation of controlled direct effects in the presence of exposure-induced confounding and latent variables. *Struct Equ Modeling*. 2014;21(3):396–407.

41. Robins J. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Service Research Methodology: A Focus on AIDS*. Washington, DC: National Center for Health Services Research, US Public Health Service; 1989:113–159.

42. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; 1967:221–233.

43. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817–838.

44. Henmi M, Eguchi S. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*. 2004;91(4):929–941.

45. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962–973.

46. Gruber S, van der Laan MJ. tmle: An R package for targeted maximum likelihood estimation. *J Stat Softw*. 2012;51(13):1–35.

47. Rose S, van der Laan MJ. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer-Verlag New York; 2011.

48. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):Article 25.

49. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am J Epidemiol*. 2014;179(2):252–260.

50. Pearl J. Direct and indirect effects. In: Breese J, Koller D, eds. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers; 2001:411–420.

51. Martin JA, Wilson EC, Osterman MJ, et al. Assessing the quality of medical and health data from the 2003 birth certificate revision: results from two states. *Natl Vital Stat Rep*. 2013; 62(2):1–19.

52. Bodnar LM, Abrams B, Bertolet M, et al. Validity of birth certificate-derived maternal weight data. *Paediatr Perinat Epidemiol*. 2014;28(3):203–212.

53. Naimi AI, Moodie EE, Auger N, et al. Stochastic mediation contrasts in epidemiologic research: interpregnancy interval and the educational disparity in preterm delivery. *Am J Epidemiol*. 2014;180(4):436–445.

54. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2002.

55. Krieger N. Refiguring "race": epidemiology, racialized biology, and biological expressions of race relations. *Int J Health Serv*. 2000;30(1):211–216.

56. Graves JL. *The Emperor's New Clothes: Biological Theories of Race at the Millennium*. New Brunswick, NJ: Rutgers University Press; 2001.

57. Krieger N. Stormy weather: race, gene expression, and the science of health disparities. *Am J Public Health*. 2005;95(12):2155–2160.

58. Krieger N. Does racism harm health? Did child abuse exist before 1962? On explicit questions, critical science, and current controversies: an ecosocial perspective. *Am J Public Health*. 2003;93(2):194–199.

59. Jacobson MF. *Whiteness of a Different Color: European Immigrants and the Alchemy of Race*. Cambridge, MA: Harvard University Press; 1999.

60. Zinn H. *A People's History of the United States: 1492–2001*. New York, NY: HarperCollins Publishers; 2003.

APPENDIX

Assume that Figure 1A in the main text holds and that the functional form of model 3 in the main text is correct. Then, under counterfactual consistency (cc) and conditional exchangeability (e):

$$E(Y|x, m, c_{XY}, c_{MY}) \overset{\text{by } cc}{=} E(Y(m)|x, m, c_{XY}, c_{MY})$$

$$\overset{\text{by } e}{=} E(Y(m)|x, c_{XY}, c_{MY}).$$

Thus, it follows that:

$$CDM(m) \equiv E(Y(m)|x = 1, c_{XY}, c_{MY}) - E(Y(m)|x = 0, c_{XY}, c_{MY})$$
$$= (\beta_0 + \beta_1 \times 1 + \beta_2 m + \beta_3 c_{XY} + \beta_4 c_{MY}) - (\beta_0 + \beta_1 \times 0 + \beta_2 m + \beta_3 c_{XY} + \beta_4 c_{MY}) = \beta_1.$$