Published in final edited form as:

Epidemiology. 2010 July ; 21(4): 540–551. doi:10.1097/EDE.0b013e3181df191c.

Bias formulas for sensitivity analysis for direct and indirect effects

Tyler J. VanderWeele

Departments of Epidemiology and Biostatistics, Harvard University

Abstract

A key question in many studies is how to divide the total effect of an exposure into a component that acts directly on the outcome and a component that acts indirectly, i.e. through some intermediate. For example, one might be interested in the extent to which the effect of diet on blood pressure is mediated through sodium intake and the extent to which it operates through other pathways. In the context of such mediation analysis, even if the effect of the exposure on the outcome is unconfounded, estimates of direct and indirect effects will be biased if control is not made for confounders of the mediator-outcome relationship. Often data are not collected on such mediator-outcome confounding variables; the results in this paper allow researchers to assess the sensitivity of their estimates of direct and indirect effects to the biases from such confounding. Specifically, the paper provides formulas for the bias in estimates of direct and indirect effects due to confounding of the exposure-mediator relationship and of the mediator-outcome relationship. Under some simplifying assumptions, the formulas are particularly easy to use in sensitivity analysis. The bias formulas are illustrated by examples in the literature concerning direct and indirect effects in which mediator-outcome confounding may be present.

The goal of many analyses is to determine whether the effect of an exposure on a particular outcome is mediated by some intermediate variable. That is to say, it is of interest to assess the extent to which the effect of the exposure on the outcome is direct and the extent to which it is mediated through some particular pathway. One way in which these questions are sometimes addressed is by including the intermediate in a regression of the outcome on the exposure and possibly some confounding variables. The estimate of the regression coefficient for the exposure when the intermediate variable is included in the regression is sometimes then interpreted as the direct effect. For such an analysis to be valid, control must be made not only for the variables that confound the exposure-outcome relationship but also for the variables that confound the mediator-outcome relationship. It is now well documented that when control is not made for variables that confound the relationship between the intermediate and the outcome, then controlling for the intermediate in the regression will give biased estimates of the direct effect.¹⁻³ Even when treatment is randomized, there may be confounding of the mediator-outcome relationship and thus mediation analyses must give careful attention to the variables that potentially confound the

mediator-outcome relationship. Often it will not be possible to collect data on all variables that might confound the relationship between the mediator and the outcome.

It is thus important to develop techniques that assess the sensitivity of one's results to unmeasured confounding variables.^{4,5} In this paper a simple sensitivity analysis approach is presented that allows researchers to assess the extent to which an unmeasured mediator-outcome confounder would have to affect both the mediator and the outcome in order to invalidate the qualitative conclusions drawn about direct and indirect effects. The analysis proceeds by analyzing the data under the assumption of no unmeasured confounding, and then assesses the sensitivity of the conclusions about direct and indirect effects to realistic violations of the no-unmeasured-confounding assumptions. The results presented in this paper are very general in the sense of allowing sensitivity analysis under a broad range of estimation approaches and of specifications of the sensitivity parameters. However, under a simplifying set of assumptions, the results yield easy-to-use formulas for sensitivity analysis for direct and indirect effects. Results are given for direct and indirect effects on the additive scale in the main text. Appendix 1 provides results for the direct and indirect effects on the risk ratio scale; these results could also be applied to the odds ratio scale when the outcome is rare.

Direct and Indirect effects: Notation, Definitions and Framework

We will use the following notation. We will let A denote the exposure of interest, Y the outcome of interest, and M the potential mediator of interest. For example, A might denote a particular diet, Y might denote blood pressure, and M might denote sodium intake. We will let C denote a set of baseline covariates not affected by the exposure. We will let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set, possibly contrary to fact, to level a . We will let Y_{am} denote the value of the outcome that would have been observed had the exposure, A , been set (possibly contrary to fact) to level a , and had the mediator, M , been set (possibly contrary to fact) to level m . These counterfactual or potential outcome variables, Y_a , M_a and Y_{am} presuppose that at least hypothetical interventions on A and M are conceivable. We make an assumption, sometimes referred to as the consistency assumption, that when $A = a$, the counterfactual outcomes Y_a and M_a are, respectively, equal to the observed outcomes Y and M . We likewise assume that when $A = a$ and $M = m$, the counterfactual outcome Y_{am} is equal to Y . Further discussion of the interpretation of these consistency assumptions is given elsewhere.^{6,7} When hypothetical interventions are not conceivable, an alternative approach to mediation based on principal stratification is also possible.⁸⁻¹¹ Although principal strata direct effects will not be the focus of this paper, a similar approach to sensitivity analysis to that described here is possible also for principal strata direct effects and is presented in the eAppendix (<http://links.lww.com>).

In this paper, sensitivity analysis for several types of direct and indirect effects are considered. The average controlled direct effect comparing exposure level a to a^* and fixing the mediator to level m is defined by $CDE_{a,a^*}(m) = E[Y_{am} - Y_{a^*m}]$ and captures the effect of exposure A on outcome Y , intervening to fix M to m . If A is a binary exposure, then we would take $a = 1$ and $a^* = 0$ and the controlled direct effect is then simply $E[Y_{1m} - Y_{0m}]$; it

may be different for different levels of m . The so-called natural direct effect³ differs from controlled direct effects in that the intermediate M is set to the level M_{a^*} , the level that it would have naturally been under some reference condition for the exposure, $A = a^*$; individual natural direct effects thus take the form $NDE_{a,a^*}(a^*) = Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$. Similarly, individual natural indirect effects can be defined as $NIE_{a,a^*}(a) = Y_{aM_a} - Y_{aM_{a^*}}$, which compares the effect of the mediator at levels M_a and M_{a^*} on the outcome when exposure A is set to a .

Natural direct and indirect effects have the property that a total effect $Y_a - Y_{a^*}$ decomposes into a natural direct and indirect effect:

$$Y_a - Y_{a^*} = Y_{aM_a} - Y_{a^*M_{a^*}} = (Y_{aM_a} - Y_{aM_{a^*}}) + (Y_{aM_{a^*}} - Y_{a^*M_{a^*}}) = NIE_{a,a^*}(a) + NDE_{a,a^*}(a^*);$$

the decomposition holds even when there are interactions and non-linearities. The approach of Baron and Kenny¹² to direct and indirect effects, common in the social sciences, does not in general allow for effect decomposition in the presence of interactions and non-linearities. If there is no interaction between the effect of the exposure A and the mediator M on the outcome Y (in the sense that $Y_{am} - Y_{a^*m}$ does not vary with m), then the controlled direct effect and the natural direct effect coincide.¹⁵

In general, controlled direct effects cannot be used for effect decomposition unless there is no interaction between the effects of the exposure and the mediator on the outcome. The difference between a total effect and a controlled direct effect cannot in general be interpreted as an indirect effect and thus cannot be used to assess mediation.^{2,3,13,14} This is because, if there is interaction between the effects of exposure A and mediator M on outcome Y , then controlled direct effects may differ from the total effect even if A is not a cause of M . If there is interaction between A and M , then $Y_{am} - Y_{a^*m}$ will differ for different values of m and thus for some m , one of the controlled direct effects $Y_{am} - Y_{a^*m}$ will differ from $Y_a - Y_{a^*}$. Controlled direct effects are often of greater interest in policy evaluation^{3,15}; natural direct and indirect effects are often of greater interest in evaluating etiology.^{3,15-17}

A number of authors have considered various identification strategies for direct and indirect effects.^{2,3,18-20} Here we will follow the exposition given by Pearl³ and some recent overviews in the epidemiologic literature.^{18,21} We note, however, that there are also other subtly different assumptions that suffice for the identification of controlled direct effects and natural direct and indirect effects.¹⁸⁻²⁰

We let $X \perp\!\!\!\perp Y|Z$ denote that X is independent of Y conditional on Z . To identify total effects it is generally assumed that, conditional on some set of measured covariates C , the effect of exposure A on outcome Y is unconfounded given C ; in counterfactual notation this is $Y_a \perp\!\!\!\perp A|C$. In practice, a researcher will attempt to collect data on a sufficiently rich set of covariates C to make the assumption plausible. Controlled direct effects can be identified by including in C all confounders of not only the exposure-outcome relationship but also the mediator-outcome relationship. In counterfactual notation, controlled direct effects will be identified^{3,22,23} if for all a and m ,

$$Y_{am} \perp\!\!\!\perp A|C \quad (1)$$

$$Y_{am} \perp\!\!\!\perp M|\{A, C\}. \quad (2)$$

Assumption (1) can be interpreted as: conditional on C there is no unmeasured confounding for the exposure-outcome relationship (except possibly through M). Assumption (2) can be interpreted as: conditional on $\{A, C\}$ there is no unmeasured confounding for the mediator-outcome relationship. If assumptions (1) and (2) hold, then average controlled direct effects are identified and given by^{3,21}

$$E[Y_{am} - Y_{a^*m}] = \sum_c \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(c)$$

In general if there are variables that confound the mediator-outcome relationship that are not in C such that condition (2) is not satisfied, then the expression given above will be biased.¹⁻³ This bias is sometimes referred to as a form of collider stratification bias.²⁴⁻²⁶ Below we give results to assess the magnitude of the bias. In some cases, there may be an effect of the exposure A that confounds the mediator-outcome relationship; as discussed elsewhere,^{3,18,21,27} controlled direct effects can still be identified in such cases by modifying assumption (2) but non-standard methods are then required.^{21,27} Here, however, we will focus on the simpler setting in which there is no effect of exposure A that confounds the mediator-outcome relationship, i.e. the setting in which none of the mediator-outcome confounders are affected by A .

For the identification of natural direct and indirect effects, additional assumptions are generally needed. Natural direct and indirect effects will be identified if, in addition to assumptions (1) and (2), the following two assumptions hold,³ that for all a , a^* and m ,

$$M_a \perp\!\!\!\perp A|C \quad (3)$$

$$Y_{am} \perp\!\!\!\perp M_{a^*}|C. \quad (4)$$

Assumption (3) can be interpreted as that conditional on C there is no unmeasured confounding of the exposure-mediator relationship. If assumption (2) holds, then assumption (4) will hold if there is no effect L of exposure A that itself affects both M and Y , i.e. no effects of exposure A that confound the mediator-outcome relationship.³ In some cases, assumption (4) may be more plausible if the mediator M occurs shortly after the exposure A .⁷ If, however, assumption 4 is violated so that there is an effect of the exposure that confounds the mediator-outcome relationship, then it has been shown that natural direct and indirect effects will not in general be identified²⁸; in such cases it may still be possible to estimate a controlled direct effect that has been standardized with respect to the distribution of the mediator under a particular treatment level.^{29,30} If assumptions (1)-(4) hold, then the average natural direct effect is identified and is given by^{3,21}:

$$E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}] = \sum_c \sum_m \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(m|a^*, c) P(c)$$

and the average natural indirect effect is identified and is given by

$$E[Y_{aM_a} - Y_{aM_{a^*}}] = \sum_c \sum_m E[Y|a, m, c] \{P(m|a, c) - P(m|a^*, c)\} P(c).$$

If there is no interaction between the effect A and M on Y , then the controlled direct effect and the natural direct effect coincide.¹⁵ Natural direct and indirect effects are then identified under assumptions (1) and (2) alone.

If assumptions (1)-(4) hold, then there are a variety of ways to estimate controlled direct effects and natural direct and indirect effects including regression models,^{7,18,20,31} structural mean models,^{27,32,33} marginal structural models^{21,30,34} and doubly robust estimators.³⁵ For example, regression models for Y conditional on A, M, C and for M conditional on A, C might be used as commonly employed in the social sciences. If there is no interaction between the effects of A and M on Y , then the regression approach of Baron and Kenny¹² utilized extensively in the psychology literature³¹ may be employed. VanderWeele and Vansteelandt⁷ recently showed how the notions of direct and indirect effects from the causal inference literature presented above could be used to extend this regression approach to settings in which there were interactions between A and M . In particular, if assumptions (1)-(4) hold and if Y and M are continuous and the following regression models for Y and M are correctly specified:

$$\begin{aligned} E[Y|a, m, c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c \\ E[M|a, c] &= \beta_0 + \beta_1 a + \beta'_2 c \end{aligned} \quad (5)$$

then the average controlled direct effect and the average natural direct and indirect effects are given by

$$\begin{aligned} E[Y_{am} - Y_{a^*m}] &= \theta_1 (a - a^*) + \theta_3 m (a - a^*) \\ E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}] &= (\theta_1 + \theta_3 \beta_0 + \theta_3 \beta_1 a^* + \theta_3 \beta'_2 E[C]) (a - a^*) \\ E[Y_{aM_a} - Y_{aM_{a^*}}] &= \theta_2 \beta_1 (a - a^*) + \theta_3 \beta_1 a (a - a^*). \end{aligned} \quad (6)$$

Note that the expression above for the controlled direct effect will only in general be valid under assumptions (1) and (2), and the expressions for natural direct and indirect effects will only in general be valid under assumptions (1)-(4). As noted by VanderWeele and Vansteelandt,⁷ if there is no interaction between A and M so that $\theta_3 = 0$, then these expressions reduce to the expressions of Baron and Kenny¹² employed in the psychology literature. The controlled direct effect and the natural direct effect are then both equal to $\theta_1(a - a^*)$ and the natural indirect effect is $\theta_2\beta_1(a - a^*)$.

Note that if the interaction term $\theta_3 am$ is omitted from the regression model for Y (as is not uncommon in epidemiologic analyses) this can lead to very misleading inferences

concerning mediation. In particular, if there is qualitative interaction between A and M so that θ_1 and θ_3 are of different signs, then if the term $\theta_3 am$ is omitted from the regression model for Y , one might then obtain an estimate of θ_1 fairly closely to 0. In this situation one might wrongly conclude that most of the effect of A on Y is mediated by M when in fact there is simply interaction between the effects of A and M on Y . See VanderWeele and Vansteelandt⁷ and Imai et al.²⁰ for further and related discussion on using regression models for mediation analysis, and also for the computation of standard errors. If conditional controlled direct effects, $E[Y_{am} - Y_{a^*m}|c]$, and conditional natural direct and indirect effects, $E[Y_{aMa^*} - Y_{a^*Ma^*}|c]$ and $E[Y_{aMa} - Y_{aMa^*}|c]$, are of interest, the same expressions as in (6) can be used, with the exception that for the conditional natural direct effect, $E[Y_{aMa^*} - Y_{a^*Ma^*}|c]$, the $E[C]$ in (6) is replaced with c . The expressions above are for what are sometimes referred to as the “pure” natural direct effect and the “total” natural indirect effects.¹⁵ If instead the “total” natural direct effect and the “pure” natural indirect effect are of interest, then these are given in terms of the regression coefficients by

$$E[Y_{aMa} - Y_{a^*Ma}] = (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1a + \theta_3\beta_2' E[C]) (a - a^*) \text{ and}$$

$$E[Y_{a^*Ma} - Y_{a^*Ma^*}] = \theta_2\beta_1 (a - a^*) + \theta_3\beta_1a^* (a - a^*).$$

In the next section we will discuss sensitivity analysis to violations of assumption (2), the assumption that the measured covariates C suffice to control for confounding of the mediator-outcome relationship. As noted above, there are a number of different methods and models that can be used to estimate direct and indirect effects. The results described below are applicable irrespective of the method used to estimate direct and indirect effects.

Sensitivity Analysis for Controlled Direct effects

In many settings it may be difficult to include in the covariate set C a sufficient number of covariates to control for confounding of the mediator-outcome relationship. For example, in a randomized trial in which exposure A is randomized, the effect of A on Y and the effect of A on M will be unconfounded by randomization i.e. assumptions (1) and (3) will hold. However, randomization of the exposure A does not imply that the effect of the mediator M on the outcome Y will be unconfounded i.e. assumption (2) may fail. To accommodate this issue of potential mediator-outcome confounding we will give two sensitivity analysis results (one for controlled direct effect and one for natural indirect effects) to assess the sensitivity of inferences about direct and indirect effects to the assumption that, conditional on C , there is no further mediator-outcome confounding. The results will build on related work for total effects.³⁶ The results will assume that the effect of A on Y and the effect of A on M are unconfounded either by randomization or by control for the covariates in C ; that is to say, both results will assume that assumptions (1) and (3) hold. The results will also assume that there are no effects of exposure A that themselves affect both M and Y . Our focus will be violations of assumption (2), i.e. of no unmeasured mediator-outcome confounding, as this is generally the assumption of greatest concern in analyses of direct and

indirect effects. In the eAppendix (<http://links.lww.com>) we also consider possible unmeasured exposure-mediator confounding. We will thus proceed as follows: we will assume that there is some unmeasured variable U such that conditional on C and U there would be no mediator-outcome confounding. In other words, U will denote our unmeasured mediator-outcome confounder. In counterfactual notation we will assume that condition (2) is violated but that

$$Y_{am} \perp\!\!\!\perp M \mid \{A, C, U\} \quad (7)$$

holds, as would be the case in the causal diagram³⁷ in Figure 1. Our first result gives a bias formula for controlled direct effects that can be used in sensitivity analysis when there are one or more unmeasured confounding variables U for the mediator-outcome relationship. Note that if we had only confounding variables C , we might still try to estimate controlled direct effects by the formula given in the previous section,

$$\Sigma_c \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(c) \quad (8)$$

or by using the expression in (6) from the regressions. If there is an unmeasured mediator-outcome confounder U , then this expression will generally be biased for the average controlled direct effect. The bias between the estimator using the observed data and the true average controlled direct effect is then

$$\text{Bias}(CDE_{a,a^*}(m)) = \Sigma_c \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(c) - E[Y_{am} - Y_{a^*m}]$$

Theorem 1 below gives a general bias formula for the difference between the biased estimator using the observed data and the true average controlled direct effect. As discussed below, the general bias formula is not entirely straightforward to use. However we will also give a Corollary to Theorem 1 that shows that under some simplifying assumptions the bias formula takes a very simple form. The proofs of the results are given in Appendix 2.

Theorem 1. Suppose that for all a and m , $Y_{am} \perp\!\!\!\perp A \mid C$ and $Y_{am} \perp\!\!\!\perp M \mid \{A, C, U\}$ then for any reference level u' of U have that the difference between the biased estimator, $\Sigma_c \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(c)$, and the actual average controlled direct effect, $E[Y_{am} - Y_{a^*m}]$, is given by

$$\begin{aligned} \text{Bias}(CDE_{a,a^*}(m)) &= \Sigma_c \Sigma_u \{E[Y|a, m, c, u] \\ &\quad - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a, c)\} P(c) - \Sigma_c \Sigma_u \{E[Y|a^*, m, c, u] \\ &\quad - E[Y|a^*, m, c, u']\} \{P(u|a^*, m, c) \\ &\quad - P(u|a^*, c)\} P(c) \end{aligned}$$

We will return to the interpretation and applicability of the bias formula in Theorem 1 shortly but first we give a corollary which, under simplifying assumptions, simplifies the bias formula considerably.

Corollary 1. Suppose that for all a and m , $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|\{A, C, U\}$. Suppose further that U is binary, that $U \perp\!\!\!\perp A|C$ and that for a particular value m , $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0]$ is constant across strata of a, c so that $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0] = \gamma$ and that $P(U = 1|a^*, m, c) - P(U = 1|a, m, c)$ is constant across strata of c so that $P(U = 1|a, m, c) - P(U = 1|a^*, m, c) = \delta$ then

$$\text{Bias}(CDE_{a,a^*}(m)) = \delta\gamma.$$

Corollary 1 states that if for a particular value m , $P(U = 1|a, m, c) - P(U = 1|a^*, m, c) = \delta$ and $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0] = \gamma$ and if U and A are uncorrelated conditional on C , then the bias constituted by the difference between the estimate using the observed data given in (8) and the true controlled direct effect is simply the product $\delta\gamma$. Under the simplifying assumptions of Corollary 1, this gives rise to a particularly simple sensitivity analysis technique for assessing the sensitivity of estimates of a controlled direct effect to an unmeasured mediator-outcome confounder (see Table 1): we can hypothesize a binary unmeasured mediator-outcome confounding variable U such that the difference in expected outcome Y comparing $U = 1$ and $U = 0$ is γ across strata of A and C conditional on $M = m$, and such that the difference in the prevalence of U , comparing exposure levels a and a^* (e.g. comparing the exposed and unexposed), is δ , across strata of C , conditional on $M = m$. For such an unmeasured mediator-outcome confounding variable, the bias of our estimate in (8) using the observed data is given simply by $\delta\gamma$. We can assess sensitivity to the presence of such an unmeasured confounding variable by varying γ (which under certain assumptions³⁶ can be interpreted as the direct effect of U on Y) and by varying δ , interpreted as the prevalence difference of U , comparing exposure levels a and a^* conditional on $M = m$ and $C = c$. The simple bias formula in Corollary 1 could also be applied to both limits of a confidence interval. Note that the controlled direct effect, $E[Y_{am} - Y_{a^*m}]$, may vary with m , and that for different values of m we will likely want to consider different specifications of the values δ and γ in the sensitivity analysis. If there is no interaction between the effects of A and M on Y , then this simple sensitivity analysis technique based on Corollary 1 will also be applicable to natural direct effects. Below we will also consider general bias formulas for sensitivity analysis for natural direct and indirect effects that will be applicable when there are interactions. Note that Corollary 1 makes three simplifying assumptions to obtain the easy-to-use bias formula. First, Corollary 1 assumes that U and A are uncorrelated conditional on C ; this would hold in Figure 1 but would be violated if U affected A as well as M as is the case in Figure 2. In the eAppendix (<http://links.lww.com>) we consider bias formulas and sensitivity analysis for violations of this assumption as might arise if U affects also A and not just M so that the exposure-outcome and the mediator-outcome relationships are confounded by U . Second, Corollary 1 assumes that for fixed m , $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0]$ is constant across strata of a, c . This would hold if there were no interaction on the additive scale between the effects of U and the effects of A or C ; otherwise it would be violated. Third, Corollary 1 assumes that for fixed m , $P(U = 1|a, m, c) - P(U = 1|a^*, m, c)$ is constant across strata of C . This would hold if U were independent of C conditional on A and M (as might occur if neither U caused C nor C caused U); otherwise the simplifying assumption might be violated. If these assumptions are thought to be unreasonable, then, as discussed below, it will be necessary to use Theorem 1 instead in sensitivity analysis.

It is important to note that δ is the prevalence difference conditional on $M = m$ and $C = c$ and not the unconditional prevalence difference. To see the importance of this distinction, let us assume that A and M are binary and note that U is assumed to be a cause of M . In specifying the conditional prevalence difference $P(U = 1|a, m, c) - P(U = 1|a^*, m, c)$, the variable U might be (unconditionally) of equal or greater prevalence comparing exposure levels $A = 1$ and $A = 0$, but it might still be conditionally less prevalent say, given $M = 1$; this is because $M = 1$ might occur if either $U = 1$ or $A = 1$ and thus conditional on $M = 1$, if $A = 0$ then we would know that $U = 1$ because $M = 1$ only when either $U = 1$ or $A = 1$. The prevalence of U , conditional on $M = m$, in exposure levels $A = 1$ and $A = 0$ will depend on both the unconditional prevalence of U in exposure levels and also on the information that conditioning on $M = m$ gives about the prevalence of U in exposure levels $A = 1$ and $A = 0$. In the language of causal diagrams, the conditional prevalence of U , given $M = m$, depends on the unconditional prevalence and also upon “collider stratification.”²⁴⁻²⁶ When U , A , M are all binary, some intuition can be given concerning such collider stratification (i.e. conditioning on a common effect, M , of U and A). In general, if the mechanism for M is an “and” mechanism (so that M occurs if A and U occur), then this will generally induce positive correlation between A and U , conditional on M ; if the mechanism for M is an “or” mechanism (so that M occurs if A or U occurs) then this will generally induce negative correlation between A and U , conditional on M ; however, exceptions can arise.^{26,38} See Hafeman⁵ for a sensitivity analysis technique involving an alternative specification.

Corollary 1 gives a reasonably straightforward sensitivity analysis technique for controlled direct effects under simplifying assumptions. However, when such simplifying assumptions do not hold, Theorem 1 can still be used. In order to use these bias formulas in sensitivity analysis, one would choose a reference level u' of U and specify (i) $\{E[Y|a, m, c, u] - E[Y|a, m, c, u']\}$ and $\{E[Y|a^*, m, c, u] - E[Y|a^*, m, c, u']\}$, i.e. the relation between U and Y , amongst those with exposure level $A = a$ and $A = a^*$, within strata of c and m . One would also specify (ii) $\{P(u|a, m, c) - P(u|a, c)\}$ and $\{P(u|a^*, m, c) - P(u|a^*, c)\}$, i.e. how within each stratum of C , the distribution of the unmeasured confounder U among those with exposure level $A = a$ and $A = a^*$, conditional on $M = m$, compares with the distribution of U , not conditional on M among those with exposure level $A = a$ and $A = a^*$. Although it will in general be difficult to specify these differences directly, it may be possible to use subject-matter knowledge to postulate plausible values of $P(u|a, m, c)$, $P(u|a^*, m, c)$, $P(u|a, c)$, and $P(u|a^*, c)$, from which differences could be computed. From the quantities (i) and (ii) specified in the sensitivity analysis, one can then calculate from Theorem 1 the bias between the true controlled direct effect and the estimator based on the observed data. In the following section we will discuss a data analysis example and apply this sensitivity analysis techniques for controlled direct effects using Corollary 1. Note that if sensitivity analysis for the conditional controlled direct effect, $E[Y_{am} - Y_{a^*m}|c]$, is of interest, then the same expression as in Theorem 1 can be used, removing the sum over c and the term $P(c)$ from the expression.

Sensitivity Analysis for Natural Direct and Indirect effects

Our second set of results gives a bias formula for natural direct and indirect effects that can be used in sensitivity analysis when there are one or more unmeasured confounding

variables U for the mediator-outcome relationship as in Figure 1. Using observed data, one might attempt to estimate natural direct and indirect effects using the standard formulas, namely,

$$\Sigma_c \Sigma_m \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(m|a^*, c) P(c) \quad (9)$$

and

$$\Sigma_c \Sigma_m E[Y|a, m, c] \{P(m|a, c) - P(m|a^*, c)\} P(c) \quad (10)$$

respectively. The bias for the natural direct effect can then be defined as the difference between the estimators in (9) and (10) and the true natural direct and indirect effects respectively:

$$Bias(NDE_{a,a^*}(a^*)) = \Sigma_c \Sigma_m \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(m|a^*, c) P(c) - E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$$

$$Bias(NIE_{a,a^*}(a)) = \Sigma_c \Sigma_m E[Y|a, m, c] \{P(m|a, c) - P(m|a^*, c)\} P(c) - E[Y_{aM_a} - Y_{a^*M_{a^*}}]$$

Theorem 2 gives formulas for these biases that can be used in sensitivity analysis.

Theorem 2. If Figure 1 represents a causal directed acyclic graph then for all a , a^* , and m , $Y_{am} \perp\!\!\!\perp A|C$, $Y_{am} \perp\!\!\!\perp M|\{A, C, U\}$, $M_a \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M_{a^*}|\{C, U\}$ and for any reference level u' of U the bias formula for the natural direct effect is given by

$$Bias(NDE_{a,a^*}(a^*)) = \Sigma_c \Sigma_m \Sigma_u \{E[Y|a, m, c, u] - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a^*, m, c)\} P(m|a^*, c) P(c).$$

and the bias formula for the natural indirect effect is given by

$$Bias(NIE_{a,a^*}(a)) = -\Sigma_c \Sigma_m \Sigma_u \{E[Y|a, m, c, u] - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a^*, m, c)\} P(m|a^*, c) P(c).$$

Note that $Bias(NIE_{a,a^*}(a)) = -Bias(NDE_{a,a^*}(a^*))$. If sensitivity analysis for natural direct and indirect effects, conditional on $C = c$, are of interest, then the same expressions as in Theorem 2 can be used, removing the sum over c and the term $P(c)$ from the expression. As before, under certain simplifying assumptions, the bias formulas in Theorem 2 reduce to simpler expressions that are relatively straightforward to use in sensitivity analysis. Corollary 2 below gives these simple expressions for natural direct and indirect effects; it uses the same simplifying assumptions as in Corollary 1.

Corollary 2. Suppose Figure 1 represents a causal directed acyclic graph. Suppose further that U is binary, that $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0]$ is constant across strata of a, m, c so that $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0] = \gamma$ and that $P(U = 1|a, m, c) - P(U =$

$1|a^*, m, c)$ is constant across strata of c so that $P(U = 1|a, m, c) - P(U = 1|a^*, m, c) = \delta_m$; then

$$\text{Bias}(NDE_{a,a^*}(a^*)) = \gamma \sum_c \sum_m \delta_m P(m|a^*, c) P(c)$$

$$\text{Bias}(NIE_{a,a^*}(a)) = -\gamma \sum_c \sum_m \delta_m P(m|a^*, c) P(c)$$

If $\delta_m = P(u|a, m, c) - P(u|a^*, m, c)$ is constant across strata of M taking value δ , then $\text{Bias}(NDE_{a,a^*}(a^*)) = \delta\gamma$ and $\text{Bias}(NIE_{a,a^*}(a)) = -\delta\gamma$.

If sensitivity analysis for natural direct and indirect effects, conditional on $C = c$, are of interest, then the same expressions as in Corollary 2 can be used, removing the sum over c and the term $P(c)$ from the expression. The use of Corollary 2 in sensitivity analysis for natural direct and indirect effects is fairly similar to that of Corollary 1 described above for controlled direct effects. However, if $\delta_m = P(u|a, m, c) - P(u|a^*, m, c)$ is not constant in m then one will have to use the empirical distribution $P(m|a^*, c)$, estimated from the data, in applying Corollary 2 rather than relying on the simple expression δ_γ . Discussion of the class of distributions of variables for which an expression such as $\delta_m = P(u|a, m, c) - P(u|a^*, m, c)$ would be constant in m can be found elsewhere.³⁹ As an alternative approach to conducting sensitivity analysis for natural indirect effects, one could use equation (9) to estimate natural direct effects, and then instead of using (10) to estimate natural indirect effects one could use the difference between the total effect, estimated by $\sum_c E[Y|a, c] - E[Y|a^*, c]P(c)$, and the natural direct effect, estimated by equation (9), to give an estimate of the natural indirect effect. The bias for the natural indirect effect obtained in this way is then simply $\text{Bias}(NIE_{a,a^*}(a)) = -\text{Bias}(NDE_{a,a^*}(a^*))$ since the total effect of A on Y is unconfounded and the true natural direct effect, $E[Y_{aM_a} - Y_{a^*M_a^*}]$, is equal to the difference between the total effect and the natural direct effect.

For settings in which the simplifying assumptions of Corollary 2 are considered unreasonable, one can use Theorem 2. The use of Theorem 2 is similar to Theorem 1 except that one must also use the empirical distributions of $P(m|a, c)$ and $P(m|a^*, c)$, estimated from the data, in applying the bias formulas in Theorem 2. As with Theorem 1, one can use the bias formulas in Theorem 2 for natural direct and indirect effects once one specifies (i) $\{E[Y|a, m, c, u] - E[Y|a, m, c, u']\}$ and $\{E[Y|a^*, m, c, u] - E[Y|a^*, m, c, u']\}$ and (ii) $P(u|a, m, c) - P(u|a^*, m, c)$ for each m . Once one has calculated $\text{Bias}(NDE_{a,a^*}(a^*))$, the approach described above for sensitivity analysis for the natural indirect effect is applicable. An alternative sensitivity analysis technique has also been developed by Hafeman.⁵ Hafeman's technique involves specifying the prevalence of the unmeasured confounder and also specifying risk ratios for the effect of the unmeasured confounder U on the mediator and on the outcome. The advantage of Hafeman's technique is that the prevalence of the unmeasured confounder and the risk ratio for the effect of the unmeasured confounder on the mediator are more intuitive quantities to specify than the prevalence difference of the unmeasured confounder within strata of the exposure, mediator and measured confounders.

Hafeman's technique is, however, limited to binary exposures, mediators and outcomes whereas the technique presented here can be applied more generally.

Imai et al.⁴⁰ has also recently provided an alternative sensitivity analysis technique for direct and indirect effects in settings in which linear or probit regression is utilized. This sensitivity analysis technique has been implemented using statistical software⁴¹ and will likely be useful in a number of settings employing regression analysis. Nevertheless the sensitivity analysis technique presented in this section has a number of advantages. First, under the simplifying assumptions presented in this section, the sensitivity analysis technique for direct and indirect effects can be implemented in a straightforward manner without relying on software at all; the simple sensitivity analysis presented above can be conducted by hand in a very straightforward manner. A second advantage of the sensitivity analysis approach presented in this section is the following. The sensitivity analysis of Imai et al.⁴⁰ and the straightforward sensitivity analysis in Corollaries 1 and 2 all make strong simplifying assumptions concerning functional form and homogeneity of effects. When these assumptions fail, the aforementioned sensitivity techniques will not work. However, even when the functional form assumptions in Imai et al.⁴⁰ or the simplifying assumptions presented above fail, the bias formulas in Theorems 1 and 2 can still be used to conduct sensitivity analysis in such settings; the bias formulas in Theorems 1 and 2 are very general. A final and related point is that the sensitivity analysis techniques of Imai et al.⁴⁰ can be applied only when regression is used in the estimation of direct and indirect effects. As noted above, recent work on mediation has considered not only the use of regression models for mediation^{7,12,18,40} but also the use of marginal structural models^{21,30,34} and structural mean models.^{27,32,33,35} The sensitivity analysis technique of Imai et al.⁴⁰ is inapplicable when other sorts of models are used, whereas sensitivity analysis using Theorem 1 is applicable irrespective of what method or model is used in the final estimation process.

Application

In this section we consider an application of the sensitivity analysis techniques presented in this paper. Caffo et al.⁴² studied the extent to which the effect of cumulative lead dose, A , on cognitive function, Y , is mediated by brain volumes, M . Prior research had indicated that occupational exposure to organic and inorganic lead was longitudinally associated both with cognitive decline⁴³ and with a decrease in the volume of brain structures as measured by MRI⁴⁴. Caffo et al.⁴² note that lead dose is associated with persistent changes in the brain structure long after lead levels decline; past cumulative absorption of lead is thus longitudinally associated with a decrease in the volume of total brain, parietal white and gray matter, temporal white matter, and two paralimbic system structures. Caffo et al.⁴² hypothesized that the effect of lead exposure on cognitive function would be at least partially mediated by brain volumes; to examine this, they used data for 2001-2003 from a study of 513 former organolead manufacturing workers in their analyses. Brain volume was measured using magnetic resonance imaging which captures only brain volume differences and not more subtle neurobiologic changes to brain structure. Caffo et al. control for a number of covariates, C , including age, education, smoking, and alcohol consumption; they consider a number of cognitive domains and both white and grey matter volumes and use a linear regression model of Y on A , M , C without a $A \times M$ product term i.e. $\theta_3 = 0$ in

regression model (5) above. Under the assumption that the regression model is correctly specified and that there is no unmeasured exposure-outcome or mediator-outcome confounding, they obtain estimates of a direct effect of 3.79 point decline (95% confidence interval [CI] = -7.40 to -0.18) in executive functioning cognitive test scores per 1- μ g/g increase in peak tibia lead exposure, controlling for white matter in brain regions associated with lead; the indirect effect of lead exposure as mediated through white matter brain volume was a 1.21 ($P = 0.01$) point decline in executive functioning cognitive test scores per 1- μ g/g increase in peak tibia lead exposure; the total effect of lead exposure is thus a 5.00 point decline (95% CI = -8.57 to -1.42) in executive functioning cognitive test scores per 1- μ g/g increase in peak tibia lead exposure.

Note under the assumption of no interaction between the effects of A and M on Y , controlled and natural direct effects coincide. Theorem 2 implies that the bias for the natural indirect effect is simply the negation of the bias of the natural direct effect and thus (under no interaction) the negation of the bias of the controlled direct effect. Suppose now that there is an unmeasured confounding variable U affecting both white matter brain volume and executive function; U might denote an unknown genetic factor. We could employ Corollary 1 to assess the extent to which U might change our conclusions about the direct and indirect effect. The simplifying assumptions of Corollary 1 would be reasonable if U did not affect cumulative lead dose, if U did not interact with cumulative lead dose or with the measured confounding variables in its effects on cognitive function, and if neither U nor the measured confounding factors affected one another. Suppose then that individuals with the genetic factor present had on average 7-point lower executive functioning test scores in that $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0] = -7$; Corollary 1 implies if, conditional on white matter brain volume, individuals with a 1- μ g/g increase in peak tibia lead had a 0.54 higher probability of having the genotype present (i.e. $P(U = 1|a + 1, m, c) - P(U = 1|a, m, c) = 0.54$), then the true direct effect might in fact be $-3.79 - (-7)(0.54) = 0$ (95% CI = -3.61 to 3.61). This degree of confounding seems implausible. Figure 3 gives the values of $\gamma = E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0]$ and $\delta = P(U = 1|a + 1, m, c) - P(U = 1|a, m, c)$ that would be required to completely eliminate the direct effect; values of γ and δ that lie below the curve would reverse the sign of the direct effect point estimate. We also have by Corollary 1 and Theorem 2 that if individuals with the genetic factor had on average 7-point lower executive functioning test scores and if, conditional on white matter brain volume, individuals with a 1- μ g/g increase in peak tibia lead had a 0.17 lower probability of having the genotype present, then the true indirect effect might in fact be $-1.21 + (-7)(-0.17) = 0$. This is arguably somewhat less implausible but perhaps still unlikely. The presence of a direct and indirect effect in the analyses of Caffo et al. thus seem unlikely due simply to such mediator-outcome confounding. A second example,^{25,45,46} employing direct effect risk ratios, is given in the eAppendix (<http://links.lww.com>), and illustrates a case in which the unmeasured mediator-outcome confounder may completely explain away the apparent direct effect.

Discussion

This paper considers sensitivity analysis for direct and indirect effects in the presence of one or more unmeasured mediator-outcome confounders. The results give bias formulas when

the data are analyzed as though there were no missing mediator-outcome confounding variable. Corollaries 1 and 2 give rise to particularly straightforward sensitivity analysis techniques that involve only the product of two sensitivity parameters. However, the use of these straightforward techniques requires some simplifying assumptions: roughly (i) that the unmeasured confounding variable affects only the mediator and the outcome, not the exposure; (ii) that the effect of the unmeasured confounder does not interact on the additive scale with the effects of the exposure or with the measured confounding variables and (iii) that the unmeasured confounding variable is neither a cause of nor caused by the measured confounding variables. Under these simplifying assumptions, sensitivity analysis for mediation is very straightforward (see Table 1). When these assumptions are violated, Theorems 1 and 2 and the results given in the eAppendix (<http://links.lww.com>) can still be used, and cover a wide range of settings for mediation. However, as described above, the use of these more general results in sensitivity analysis is somewhat more complex than the use of Corollaries 1 and 2. Although Theorems 1 and 2 and the results in the eAppendix are quite general, they are limited to settings in which there is no effect of the exposure that confounds the mediator-outcome relationship. Future work will consider sensitivity analysis for direct and indirect effects in cases in which there is an effect of exposure that is a mediator-outcome confounder.

Two further extensions are worth noting. First, although we have considered direct and indirect effects, defined as expected counterfactual outcome differences, the basic approach applies more generally to other measures of effect. In particular, in Appendix 1, it is shown that simple sensitivity analysis techniques for direct and indirect effects are also possible on the risk-ratio scale. Second, the focus here has been on the setting of non-clustered data; the basic sensitivity analysis presented in this paper is applicable also to clustered data and multilevel settings. Discussion of how concepts of mediation extends to a multilevel setting is provided elsewhere.⁴⁷

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks Yasutaka Chiba, Jamie Robins, the editor, and two anonymous referees for helpful comments.

Appendix 1. Bias formulas for risk ratios and odds ratios for direct and indirect effects

For a dichotomous outcome Y , the conditional controlled direct effect risk ratio and odds ratio can be defined as:

$$RR_{a,a^*|c}^{CDE}(m) = \frac{P(Y_{am}|c)}{P(Y_{a^*m}|c)}$$

$$OR_{a,a^*|c}^{CDE}(m) = \frac{\frac{P(Y_{am}|c)}{\{1-P(Y_{am}|c)\}}}{\frac{P(Y_{a^*m}|c)}{\{1-P(Y_{a^*m}|c)\}}}$$

where for a dichotomous variable Y we will use $P(Y = 1)$, $P(Y)$ and $E[Y]$ interchangeably. The conditional natural direct effect risk ratio and odds ratio can be defined as:

$$RR_{a,a^*|c}^{NDE}(a^*) = \frac{P(Y_{aM_{a^*}}|c)}{P(Y_{a^*M_{a^*}}|c)}$$

$$OR_{a,a^*|c}^{NDE}(a^*) = \frac{P(Y_{aM_{a^*}}|c) / \{1 - P(Y_{aM_{a^*}}|c)\}}{P(Y_{a^*M_{a^*}}|c) / \{1 - P(Y_{a^*M_{a^*}}|c)\}}.$$

The conditional natural indirect effect risk ratio and odds ratio can be defined as:

$$RR_{a,a^*|c}^{NIE}(a) = \frac{P(Y_{aM_a}|c)}{P(Y_{a^*M_{a^*}}|c)}$$

$$OR_{a,a^*|c}^{NIE}(a) = \frac{P(Y_{aM_a}|c) / \{1 - P(Y_{aM_a}|c)\}}{P(Y_{a^*M_{a^*}}|c) / \{1 - P(Y_{a^*M_{a^*}}|c)\}}.$$

The risk ratio or odds ratio for the conditional total effect decomposes into the product of risk ratios or odds ratios for the natural direct and indirect effect

$$\frac{P(Y_a|c)}{P(Y_{a^*}|c)} = RR_{a,a^*|c}^{NIE}(a) \times RR_{a,a^*|c}^{NDE}(a^*)$$

$$\frac{P(Y_a|c) / \{1 - P(Y_a|c)\}}{P(Y_{a^*}|c) / \{1 - P(Y_{a^*}|c)\}} = OR_{a,a^*|c}^{NIE}(a) \times OR_{a,a^*|c}^{NDE}(a^*).$$

In this appendix, we give bias formulas for the controlled direct effect and natural direct and indirect effect risk ratios for settings in which there is an unmeasured confounder U of the mediator-outcome relationship under certain simplifying assumptions. More general results are given in the eAppendix (<http://links.lww.com>). If the outcome is rare in all strata of exposure A , mediator M , covariates C and unmeasured confounder U , then these bias formulas for the risk ratio can also be used for the odds-ratio scale. We define the bias for the conditional controlled direct effect and natural direct and indirect effect odds ratios as

$$Bias(CDE_{a,a^*|c}^{RR}(m)) = \frac{P(Y|a,m,c) / RR_{a,a^*|c}^{CDE}(m)}{P(Y|a^*,m,c) / RR_{a,a^*|c}^{CDE}(m)}$$

$$Bias(NDE_{a,a^*|c}^{RR}(a^*)) = \frac{\sum_m P(Y|a,m,c) P(m|a^*,c) / RR_{a,a^*|c}^{NDE}(a^*)}{\sum_m P(Y|a^*,m,c) P(m|a^*,c) / RR_{a,a^*|c}^{NDE}(a^*)}$$

$$Bias(NIE_{a,a^*|c}^{RR}(a)) = \frac{\sum_m P(Y|a,m,c) P(m|a,c) / RR_{a,a^*|c}^{NIE}(a)}{\sum_m P(Y|a^*,m,c) P(m|a^*,c) / RR_{a,a^*|c}^{NIE}(a)}$$

Suppose that for all a and m , $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|\{A, C, U\}$ and suppose further that U

is binary and that $\frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U=0)} = \gamma$ is constant across strata of a then

$$Bias(CDE_{a,a^*|c}^{RR}(m)) = \frac{1 + (\gamma - 1) P(U=1|a, m, c)}{1 + (\gamma - 1) P(U=1|a^*, m, c)}.$$

To make use of this simple bias formula for controlled direct effect risk ratios, one must

specify $\gamma = \frac{P(Y|a, m, c, U=1)}{P(Y|a, m, c, U=0)}$ (informally interpreted as the direct effect of U on Y) and also the prevalence of the unmeasured confounder $P(U=1|a, m, c)$ for exposure levels a and a^* . Once again, for different levels of m , different sensitivity analysis parameters will likely be used. The proof of this result is given in the eAppendix (<http://links.lww.com>); also further results concerning bias formulas for controlled direct effect risk ratios are given in

the eAppendix in settings in which the simplifying assumption that $\frac{P(Y|a, m, c, U=1)}{P(Y|a, m, c, U=0)} = \gamma$ is constant across strata of a is not reasonable.

We now also give simple bias formulas for natural direct and indirect effect risk ratios under simplifying assumptions; again more general results are given in the eAppendix. Suppose that Figure 1 represents a casual directed acyclic graph and suppose further that U is binary

and that $\frac{P(Y|a, m, c, U=1)}{P(Y|a, m, c, U=0)} = \gamma$ is constant across strata of m then

$$\begin{aligned} \text{Bias} \left(NDE_{a,a^*|c}^{RR} (a^*) \right) &= \frac{\sum_m \{1 + (\gamma - 1)\pi_{a,m}\} P(m|a^*, c)}{\sum_m \{1 + (\gamma - 1)\pi_{a^*,m}\} P(m|a^*, c)} \\ \text{Bias} \left(NIE_{a,a^*|c}^{RR} (a) \right) &= 1 / \text{Bias} \left(NDE_{a,a^*|c}^{RR} (a^*) \right) \end{aligned}$$

where $\pi_{a,m} = P(U=1|a, m, c)$ and $\pi_{a^*,m} = P(U=1|a^*, m, c)$. If $\pi_{a,m}$ and $\pi_{a^*,m}$ are constant across m so that $\pi_{a,m} = \pi_a$ and $\pi_{a^*,m} = \pi_{a^*}$ then

$$\begin{aligned} \text{Bias} \left(NDE_{a,a^*|c}^{RR} (a^*) \right) &= \frac{1 + (\gamma - 1)\pi_a}{1 + (\gamma - 1)\pi_{a^*}} \\ \text{Bias} \left(NIE_{a,a^*|c}^{RR} (a) \right) &= \frac{1 + (\gamma - 1)\pi_{a^*}}{1 + (\gamma - 1)\pi_a} \end{aligned}$$

To make use of these simple bias formulas for natural direct and indirect effect risk ratios,

one must specify $\gamma = \frac{P(Y|a, m, c, U=1)}{P(Y|a, m, c, U=0)}$ and also the prevalence of the unmeasured confounder $\pi_{a,m} = P(U=1|a, m, c)$ and $\pi_{a^*,m} = P(U=1|a^*, m, c)$ in each stratum of m .

Appendix 2. Proofs

Proof of Theorem 1.

We have that,

$$\begin{aligned} \text{Bias} (CDE_{a,a^*} (m)) &= \sum_c \{E[Y|a, m, c] - E[Y|a^*, m, c]\} P(c) - E[Y_{am} - Y_{a^*m}] \\ &= \sum_c \sum_u E[Y|a, m, c, u] P(u|a, m, c) P(c) - \sum_c \sum_u E[Y|a^*, m, c, u] P(u|a^*, m, c) P(c) - \sum_c \sum_u E[Y|a, m, c, u] P(u|a, c) P(c) + \\ &= \sum_c \sum_u E[Y|a, m, c, u] \{P(u|a, m, c) - P(u|a, c)\} P(c) - \sum_c \sum_u E[Y|a^*, m, c, u] \{P(u|a^*, m, c) - P(u|a^*, c)\} P(c) \\ &= \sum_c \sum_u \{E[Y|a, m, c, u] - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a, c)\} P(c) - \sum_c \sum_u \{E[Y|a^*, m, c, u] - E[Y|a^*, m, c, u']\} \end{aligned}$$

where the second equality follows because $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|A, C, U$ (see for example Pearl³ or the eAppendix in VanderWeele²¹) and the final equality follows because

for a fixed reference value u' of U , $E[Y|a, m, c, u']$ and $E[Y|a^*, m, c, u']$ are constants and thus $\sum_u E[Y|a, m, c, u']P(u|a, m, c) = E[Y|a, m, c, u'] = \sum_u E[Y|a, m, c, u']P(u|a, c)$ and similarly

$$\sum_u E[Y|a^*, m, c, u'] P(u|a^*, m, c) = E[Y|a^*, m, c, u'] = \sum_u E[Y|a^*, m, c, u'] P(u|a^*, c).$$

Proof of Corollary 1.

For $u' = 0$, we have by Theorem 1 that,

$$\begin{aligned} \text{Bias}(CDE_{a,a^*}(m)) &= \sum_c \sum_u \{E[Y|a, m, c, u] - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a, c)\} P(c) - \sum_c \sum_u \{E[Y|a^*, m, c, u] \\ &= \sum_c \{E[Y|a, m, c, U=1] - E[Y|a, m, c, U=0]\} \{P(U=1|a, m, c) - P(U=1|a, c)\} P(c) - \sum_c \{E[Y|a^*, m, c, U=1] \\ &= \sum_c \gamma \{P(U=1|a, m, c) - P(U=1|c)\} P(c) - \sum_c \gamma \{P(U=1|a^*, m, c) - P(U=1|c)\} P(c) \\ &= \sum_c \gamma \{P(U=1|a, m, c) - P(U=1|a^*, m, c)\} P(c) \\ &= \sum_c \gamma \delta P(c) = \gamma \delta. \end{aligned}$$

Proof of Theorem 2.

Because of its length, the proof of Theorem 2 is given in the eAppendix (<http://links.lww.com>).

Proof of Corollary 2.

For $u' = 0$, we have by Theorem 2 that

$$\begin{aligned} \text{Bias}(NDE_{a,a^*}(a^*)) &= \sum_c \sum_m \sum_u \{E[Y|a, m, c, u] - E[Y|a, m, c, u']\} \{P(u|a, m, c) - P(u|a^*, m, c)\} P(m|a^*, c) P(c). \\ &= \sum_c \sum_m \{E[Y|a, m, c, U=1] - E[Y|a, m, c, U=0]\} \times \{P(U=1|a, m, c) - P(U=1|a^*, m, c)\} P(m|a^*, c) P(c) \\ &= \sum_c \sum_m \delta_m P(m|a^*, c) P(c). \end{aligned}$$

If for all m , δ_m takes value, then $\text{Bias}(NDE_{a,a^*}(a^*)) = \gamma \sum_c \sum_m \delta_m P(m|a^*, c) P(c) = \delta \gamma$.

Because, by Theorem 2, $\text{Bias}(NIE_{a,a^*}(a)) = -\text{Bias}(NDE_{a,a^*}(a^*))$, we immediately have $\text{Bias}(NIE_{a,a^*}(a)) = -\sum_c \sum_m \delta_m P(m|a^*, c) P(c)$ and $\text{Bias}(NIE_{a,a^*}(a)) = -\delta \gamma$ if for all m , δ_m takes value δ . This completes the proof.

References

1. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Evaluation Review*. 1981; 5:602–619.
2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3:143–155. [PubMed: 1576220]
3. Pearl, J. Direct and indirect effects; Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence; San Francisco. 2001; p. 411–420. Morgan Kaufmann
4. Blakely T. Estimating direct and indirect effects - fallible in theory, but in the real world? *International Journal of Epidemiology*. 2002; 31:166–167. [PubMed: 11914315]
5. Hafeman, D. Opening the Black Box: A Reassessment of Mediation From a Counterfactual Perspective [dissertation]. Columbia University; New York: 2008.

6. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009; 20:880–883. [PubMed: 19829187]
7. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface - Special Issue on Mental Health and Social Behavioral Science*. in press.
8. Rubin DB. Direct and indirect effects via potential outcomes. *Scandinavian Journal of Statistics*. 2004; 31:161–170.
9. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*. 2008; 78:2957–2962.
10. Gallop R, Small DS, Lin JY, Elliott MR, Joffe M, Ten Have TR. Mediation analysis with principal stratification. *Statistics in Medicine*. 2009; 28:1108–1130. [PubMed: 19184975]
11. Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*. 2009; 65:514–520. [PubMed: 18759834]
12. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51:1173–1182. [PubMed: 3806354]
13. Kaufman JS, MacLehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives and Innovations*. 2004; 1:4. [PubMed: 15507130]
14. VanderWeele TJ. Mediation and mechanism. *European Journal of Epidemiology*. 2009; 24:217–224. [PubMed: 19330454]
15. Robins, JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P.; Hjort, NL.; Richardson, S., editors. *Highly structured stochastic systems*. Oxford University Press; New York: 2003. p. 70–81.
16. Joffe M, Small D, Hsu C-Y. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*. 2007; 22:74–97.
17. Hafeman DM, Schwartz S. Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology*. 2009; 38:838–845. [PubMed: 19261660]
18. Peterson ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology*. 2006; 17:276–84. [PubMed: 16617276]
19. Hafeman DM, VanderWeele TJ. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*. in press.
20. Imai, K.; Keele, L.; Yamamoto, T. Identification, inference, and sensitivity analysis for causal mediation effects. Princeton Technical Report. <http://imai.princeton.edu/research/files/mediation.pdf>
21. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*. 2009; 20:18–26. [PubMed: 19234398]
22. Robins JM. A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986; 7:1393–1512.
23. Robins JM. Addendum to a new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications*. 1987; 14:923–945.
24. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004; 15:615–625. [PubMed: 15308962]
25. Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight “paradox” uncovered? *American Journal of Epidemiology*. 2006; 164:1115–1120. [PubMed: 16931543]
26. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology*. 2007; 166:1096–1104. [PubMed: 17702973]
27. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology*. 2009; 20:851–860. [PubMed: 19806060]

28. Avin, C.; Shpitser, I.; Pearl, J. Identifiability of path-specific effects; Proceedings of the International Joint Conferences on Artificial Intelligence; 2005; p. 357-363.
29. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B.* 2008; 69:199–216.
30. van der Laan MJ, Petersen ML. Direct effect models. *International Journal of Biostatistics.* 2008
31. MacKinnon, DP. *An Introduction to Statistical Mediation Analysis.* Lawrence Erlbaum Associates; New York: 2008.
32. Robins, JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour, C.; Cooper, GF., editors. *Computation, Causation, and Discovery.* AAAI Press/The MIT Press; Menlo Park, CA, Cambridge, MA: 1999. p. 349-405.
33. Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. Causal mediation analyses with rank preserving models. *Biometrics.* 2007; 63:926–934. [PubMed: 17825022]
34. Robins JM. Association, causation, and marginal structural models. *Synthese.* 1999; 121:151–179.
35. Goetghebeur S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B.* 2008; 70:1049–1066.
36. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis and for general outcomes, treatments and measured and unmeasured confounding variables. Under review.
37. Pearl J. Casual diagrams for empirical research (with discussion). *Biometrika.* 1995; 82:669–710.
38. VanderWeele TJ, Robins JM. Minimal sufficient causation and directed acyclic graphs. *Annals of Statistics.* 2009; 37:1437–1465.
39. VanderWeele TJ. Sensitivity analysis: distributional assumptions and confounding assumptions. *Biometrics.* 2008; 64:645–649. [PubMed: 18482060]
40. Imai, K.; Keele, L.; Tingley, D. A general approach to causal mediation analysis. Princeton Technical Report. <http://imai.princeton.edu/research/files/BaronKenny.pdf>
41. Imai, K.; Keele, L.; Tingley, D.; Yamamoto, T. Causal Mediation Analysis in R. Princeton Technical Report. <http://imai.princeton.edu/research/files/mediationR.pdf>
42. Caffo B, Chen S, Stewart W, Bolla K, Yousem D, Davatzikos C, Schwartz BS. Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function? *American Journal of Epidemiology.* 2008; 167:429–437.
43. Schwartz BS, Stewart WF, Bolla KI, et al. Past adult lead exposure is associated with longitudinal decline in cognitive function. *Neurology.* 2000; 55:1144–1150. [PubMed: 11071492]
44. Stewart WF, Schwartz BS, Davatzikos C, et al. Past adult lead exposure is linked to neurodegeneration measured by brain MRI. *Neurology.* 2006; 66:1476–1484. [PubMed: 16717205]
45. Wilcox A. On the importance -and the unimportance- of birthweight. *Int J Epidemiol.* 2001; 30:1233–1241. [PubMed: 11821313]
46. Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. *Paediatric and Perinatal Epidemiology.* 2009; 23:394–402. [PubMed: 19689488]
47. VanderWeele TJ. Direct and indirect effects of neighborhood-based clustered and longitudinal treatments. *Sociological Research and Methods.* in press.

Figure 1.
Causal diagram in which the unmeasured confounder U confounds only the mediator-outcome relationship.

Figure 2. Causal diagram in which the unmeasured confounder U confounds the exposure-outcome, mediator-outcome and exposure-mediator relationship.

Figure 3.
Values of δ and γ that lie below the curve would reverse the sign of the direct effect point estimate.

Table 1

Summary of Sensitivity Analysis for Controlled Direct Effects

Effect of Interest: In many studies, researchers are interested in the extent to which the effect of an exposure on some outcome is direct and the extent to which it is mediated by some intermediate variable; to obtain direct effect estimates, often the mediator variable is included as a covariate in a regression of the outcome on the exposure
Problem: To ensure the estimate of the direct effect is valid, control needs to be made for confounding of the mediator-outcome relationship and not just the exposure-outcome relationship; in many epidemiologic studies, mediator-outcome confounders are not controlled for or even measured
Sensitivity Analysis: To address this issue, one can assess the sensitivity of an estimate of a direct effect to an unmeasured confounder of the mediator-outcome relationship
Method: Consider some mediator level m . Under simplifying assumptions described in the text and summarized in the Discussion, if γ denotes the effect of the binary unmeasured confounder U on the outcome for individuals with mediator level m and if δ denotes the difference in the prevalence of U between the exposed subjects with mediator level m and the unexposed subjects with mediator level m , then one can subtract the quantity $\delta\gamma$ from the potentially confounded estimate to obtain a valid estimate of the direct effect of the exposure on the outcome with the mediator set to level m
Extensions: Sensitivity analysis techniques for other mediation effects and for settings in which the simplifying assumptions do not hold are described in the text and in the eAppendix