Biometrika (2016), **103**, 2, *pp.* 483–490
Printed in Great Britain
doi: 10.1093/biomet/asw012
Advance Access publication 29 April 2016

Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding

BY PENG DING

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

AND TYLER J. VANDERWEELE

Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.
tvanderw@hsph.harvard.edu

SUMMARY

It is often of interest to decompose the total effect of an exposure into a component that acts on the outcome through some mediator and a component that acts independently through other pathways. Said another way, we are interested in the direct and indirect effects of the exposure on the outcome. Even if the exposure is randomly assigned, it is often infeasible to randomize the mediator, leaving the mediator-outcome confounding not fully controlled. We develop a sensitivity analysis technique that can bound the direct and indirect effects without parametric assumptions about the unmeasured mediator-outcome confounding.

Some key words: Bounding factor; Causal inference; Collider; Natural direct effect; Natural indirect effect.

1. INTRODUCTION

Researchers often conduct mediation analysis to assess the extent to which an effect of an exposure on the outcome is mediated through a particular pathway and the extent to which the effect operates directly. Mediation analysis initially developed within genetics and psychology based on linear structural equation models (Wright, 1934; Baron & Kenny, 1986), and has been formalized by the notions of natural direct and indirect effects under the potential outcomes framework (Robins & Greenland, 1992; Pearl, 2001) and the decision-theoretic framework (Didelez et al., 2006; Geneletti, 2007). However, identification of natural direct and indirect effects used in that literature relies on strong assumptions, including the assumption of no unmeasured mediator-outcome confounding (Pearl, 2001; Imai et al., 2010; VanderWeele, 2010). Even if we can rule out unmeasured exposure-mediator and exposure-outcome confounding by randomly assigning the exposure, full control of mediator-outcome confounding is often impossible because it is infeasible to randomize the mediator. Therefore, it is crucial in applied mediation analyses to investigate the sensitivity of the conclusions to unmeasured mediator-outcome confounding. Previous sensitivity analysis techniques rely on restrictive modelling assumptions (Imai et al., 2010), use sensitivity parameters involving counterfactual terms (Tchetgen Tchetgen & Shpitser, 2012), or require the specification of a large number of sensitivity parameters (VanderWeele, 2010). Other work (Sjölander, 2009; Robins & Richardson, 2010) has provided bounds for natural direct and indirect effects without imposing assumptions, but these consider the most extreme scenarios and the bounds are often too broad to be useful in practice. We develop a sensitivity analysis technique which has only two sensitivity parameters and does not make any modelling assumptions or any assumptions about the type of the unmeasured mediator-outcome confounder or confounders. Our results imply Cornfield-type inequalities (Cornfield et al., 1959; Ding & VanderWeele,

2014) that the unmeasured confounder must satisfy to reduce the observed natural direct effect to a certain level or explain it away.

2. NOTATION AND FRAMEWORK FOR MEDIATION ANALYSIS

Let A denote the exposure, Y the outcome, M the mediator, C a set of observed baseline covariates not affected by the exposure, and U a set of unmeasured baseline covariates not affected by the exposure. In order to define causal effects, we invoke the potential outcomes framework (Neyman, 1923; Rubin, 1974) and apply it in the context of mediation (Robins & Greenland, 1992; Pearl, 2001). If a hypothetical intervention on A is well-defined, we let Y_a and M_a denote the potential values of the outcome and the mediator that would have been observed had the exposure A been set to level a. If hypothetical interventions on A and M are both well-defined, we further let Y_{am} denote the potential value of the outcome that would have been observed had the exposure A been set to level a and the mediator M been set to level m (Robins & Greenland, 1992; Pearl, 2001). Following Pearl (2009) and VanderWeele (2015), we need the following consistency assumption for all a and m: $Y_a = Y$ and $M_a = M$ if $A = a$; and $Y_{am} = Y$ if $A = a$ and $M = m$. We further need the composition assumption that $Y_{aM_a} = Y_a$ for $a = 0, 1$.

We will assume that the exposure A is binary, but all the results in this paper are also applicable to a categorical or continuous exposure and could be used to compare any two levels of A. In the main text we consider a binary outcome Y, but in § 6 we note that all the results hold for count and continuous positive outcomes and time-to-event outcomes with rare events. The mediator M, the observed covariates C, and the unmeasured confounder or confounders U can be of general types, i.e., categorical, continuous or mixed, and scalar or vector. For notational simplicity, in the main text we assume that (M, C, U) are categorical, and in the Supplementary Material we present results for general types.

On the risk ratio scale, the conditional natural direct and indirect effects, comparing the exposure levels $A = 1$ and $A = 0$ within the observed covariate level $C = c$, are defined as

$$\text{NDE}^{\text{true}}_{\text{RR}|c} = \frac{\text{pr}\,(Y_{1M_0} = 1 \mid c)}{\text{pr}\,(Y_{0M_0} = 1 \mid c)}, \quad \text{NIE}^{\text{true}}_{\text{RR}|c} = \frac{\text{pr}\,(Y_{1M_1} = 1 \mid c)}{\text{pr}\,(Y_{1M_0} = 1 \mid c)}. \tag{1}$$

The conditional natural direct effect compares the distributions of the potential outcomes when the exposure level changes from $A = 0$ to $A = 1$ but the mediator is fixed at M_0. The conditional natural indirect effect compares the distributions of the potential outcomes when the exposure level is fixed at $A = 1$ but the mediator changes from M_0 to M_1. The conditional total effect can be decomposed as a product of the conditional direct and indirect effects as follows:

$$\text{TE}^{\text{true}}_{\text{RR}|c} = \frac{\text{pr}\,(Y_1 = 1 \mid c)}{\text{pr}\,(Y_0 = 1 \mid c)} = \text{NDE}^{\text{true}}_{\text{RR}|c} \times \text{NIE}^{\text{true}}_{\text{RR}|c}.$$

On the risk difference scale, the conditional natural direct and indirect effects are defined as

$$\text{NDE}^{\text{true}}_{\text{RD}|c} = \text{pr}\,(Y_{1M_0} = 1 \mid c) - \text{pr}\,(Y_{0M_0} = 1 \mid c), \tag{2}$$

$$\text{NIE}^{\text{true}}_{\text{RD}|c} = \text{pr}\,(Y_{1M_1} = 1 \mid c) - \text{pr}\,(Y_{1M_0} = 1 \mid c), \tag{3}$$

and the conditional total effect has the decomposition

$$\text{TE}^{\text{true}}_{\text{RD}|c} = \text{pr}\,(Y_1 = 1 \mid c) - \text{pr}\,(Y_0 = 1 \mid c) = \text{NDE}^{\text{true}}_{\text{RD}|c} + \text{NIE}^{\text{true}}_{\text{RD}|c}.$$

3. IDENTIFICATION OF CONDITIONAL NATURAL DIRECT AND INDIRECT EFFECTS

Here we follow the identification strategy of Pearl (2001) for natural direct and indirect effects. A number of authors have provided other subtly different sufficient conditions (see, e.g., Imai et al., 2010; Vansteelandt & VanderWeele, 2012; Lendle et al., 2013). Let $\perp\!\!\!\perp$ denote independence of random variables. To

identify the conditional natural direct and indirect effects by the joint distribution of the observed variables (A, M, Y, C), Pearl (2001) assumed that for all a, a^* and m,

$$Y_{am} \perp\!\!\!\perp A \mid C, \quad Y_{am} \perp\!\!\!\perp M \mid (A, C), \quad M_a \perp\!\!\!\perp A \mid C, \quad Y_{am} \perp\!\!\!\perp M_{a^*} \mid C. \tag{4}$$

The four assumptions in (4) require that the observed covariates C control exposure-outcome confounding, control mediator-outcome confounding, control exposure-mediator confounding, and ensure cross-world counterfactual independence, respectively; on a nonparametric structural equation model (Pearl, 2009), this fourth assumption is essentially that none of the mediator-outcome confounders are themselves affected by the exposure (Pearl, 2001; VanderWeele, 2015). In particular, on the risk ratio scale, we can identify the conditional natural direct and indirect effects by

$$\text{NDE}^{\text{obs}}_{\text{RR}|c} = \frac{\sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\,\text{pr}\,(m \mid A=0, c)}{\sum_m \text{pr}\,(Y=1 \mid A=0, m, c)\,\text{pr}\,(m \mid A=0, c)}, \tag{5}$$

$$\text{NIE}^{\text{obs}}_{\text{RR}|c} = \frac{\sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\,\text{pr}\,(m \mid A=1, c)}{\sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\,\text{pr}\,(m \mid A=0, c)}. \tag{6}$$

On the risk difference scale, we can identify the conditional natural direct and indirect effects by

$$\text{NDE}^{\text{obs}}_{\text{RD}|c} = \sum_m \{\text{pr}\,(Y=1 \mid A=1, m, c) - \text{pr}\,(Y=1 \mid A=0, m, c)\}\,\text{pr}\,(m \mid A=0, c), \tag{7}$$

$$\text{NIE}^{\text{obs}}_{\text{RD}|c} = \sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\{\text{pr}\,(m \mid A=1, c) - \text{pr}\,(m \mid A=0, c)\}. \tag{8}$$

Proofs of (5)–(8) can be found in Pearl (2001) and VanderWeele (2015).

If we replace $Y_{aM_{a^*}}$ in definitions (1)–(3) by $Y_{a,G_{a^*|c}}$, with $G_{a^*|c}$ being a random draw from the conditional distribution $\text{pr}\,(M_{a^*} \mid c)$, then we can drop the cross-world counterfactual independence assumption $Y_{am} \perp\!\!\!\perp M_{a^*} \mid C$ (VanderWeele, 2015). This view is related to the decision-theoretic framework without using potential outcomes (Didelez et al., 2006; Geneletti, 2007). We show in the Supplementary Material that because the alternative frameworks lead to the same empirical identification formulae as in (5)–(8), all our results below can be applied.

4. Sensitivity analysis with unmeasured mediator-outcome confounding

4·1. *Unmeasured mediator-outcome confounding*

The assumptions in (4) are strong and untestable. If the exposure is randomly assigned given the values of the observed covariates C, as in completely randomized experiments or randomized block experiments, then the first and third assumptions of (4) hold automatically owing to the randomization. In observational studies, we may have background knowledge to collect adequate covariates C to control the exposure-outcome and exposure-mediator confounding such that the first and third assumptions in (4) are plausible. However, direct intervention on the mediator is often infeasible, and it may not be possible to randomize. Therefore, the second assumption in (4), the absence of mediator-outcome confounding, may be violated in practice. Furthermore, the fourth assumption in (4) cannot be guaranteed even under randomization of both A and M, and thus it is fundamentally untestable (Robins & Richardson, 2010).

For sensitivity analysis, we assume that (C, U) jointly ensure (4), that is,

$$Y_{am} \perp\!\!\!\perp A \mid (C, U), \quad Y_{am} \perp\!\!\!\perp M \mid (A, C, U), \quad M_a \perp\!\!\!\perp A \mid (C, U), \quad Y_{am} \perp\!\!\!\perp M_{a^*} \mid (C, U). \tag{9}$$

When C controls the exposure-mediator and exposure-outcome confounding, we further assume that

$$A \perp\!\!\!\perp U \mid C. \tag{10}$$

The independence relationships in (9) impose no restrictions on the unmeasured confounders U, and they become assumptions if we require at least one of the sensitivity parameters introduced in § 4·2 to be finite. Figure 1 illustrates such a scenario with the assumptions in (9) and (10) holding, where U contains the

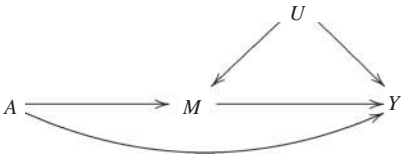

Fig. 1. Directed acyclic graph with mediator-outcome confounding within strata of observed covariates C.

common causes of the mediator and the outcome, and A and U are conditionally independent given C. In § 6 and the Supplementary Material, we comment on the applicability of our results under violations of the assumption in (10).

Under the assumptions in (9) and (10), we can express conditional natural direct and indirect effects using the joint distribution of (A, M, Y, C, U). In particular, on the risk ratio scale,

$$\text{NDE}^{\text{true}}_{\text{RR}|c} = \frac{\sum_u \sum_m \text{pr}\,(Y=1 \mid A=1, m, c, u)\,\text{pr}\,(m \mid A=0, c, u)\,\text{pr}\,(u \mid c)}{\sum_u \sum_m \text{pr}\,(Y=1 \mid A=0, m, c, u)\,\text{pr}\,(m \mid A=0, c, u)\,\text{pr}\,(u \mid c)}, \tag{11}$$

$$\text{NIE}^{\text{true}}_{\text{RR}|c} = \frac{\sum_u \sum_m \text{pr}\,(Y=1 \mid A=1, m, c, u)\,\text{pr}\,(m \mid A=1, c, u)\,\text{pr}\,(u \mid c)}{\sum_u \sum_m \text{pr}\,(Y=1 \mid A=1, m, c, u)\,\text{pr}\,(m \mid A=0, c, u)\,\text{pr}\,(u \mid c)}. \tag{12}$$

On the risk difference scale,

$$\text{NDE}^{\text{true}}_{\text{RD}|c} = \sum_u \sum_m \{\text{pr}\,(Y=1 \mid A=1, m, c, u) - \text{pr}\,(Y=1 \mid A=0, m, c, u)\} \times \text{pr}\,(m \mid A=0, c, u)\,\text{pr}\,(u \mid c), \tag{13}$$

$$\text{NIE}^{\text{true}}_{\text{RD}|c} = \sum_u \sum_m \text{pr}\,(Y=1 \mid A=1, m, c, u)\,\text{pr}\,(u \mid c) \times \{\text{pr}\,(m \mid A=1, c, u) - \text{pr}\,(m \mid A=0, c, u)\}. \tag{14}$$

The proofs of (11)–(14) follow from Pearl (2001) and VanderWeele (2015). Unfortunately, however, (11)–(14) depend not only on the joint distribution of the observed variables (A, M, Y, C) but also on the distribution of the unobserved variable U. In the following, we will give sharp bounds on the true conditional direct and indirect effects in terms of the observed conditional natural direct and indirect effects and two measures of the mediator-outcome confounding that can be taken as sensitivity parameters.

4·2. *Sensitivity parameters and the bounding factor*

First, we introduce a conditional association measure between U and Y given $(A = 1, M, C = c)$, and define our first sensitivity parameter as

$$\text{RR}_{UY|(A=1,M,c)} = \max_m \text{RR}_{UY|(A=1,m,c)} = \max_m \frac{\max_u \text{pr}\,(Y=1 \mid A=1, m, c, u)}{\min_u \text{pr}\,(Y=1 \mid A=1, m, c, u)},$$

where $\text{RR}_{UY|(A=1,m,c)}$ is the maximum divided by the minimum of the probabilities $\text{pr}\,(Y=1 \mid A=1, m, c, u)$ over u. When U is binary, $\text{RR}_{UY|(A=1,m,c)}$ reduces to the usual conditional risk ratio of U on Y, and $\text{RR}_{UY|(A=1,M,c)}$ is the maximum of these conditional risk ratios over m. If U and Y are conditionally independent given (A, M, C), then $\text{RR}_{UY|(A=1,M,c)} = 1$.

Second, we introduce a conditional association measure between A and U given M. As illustrated in Fig. 1, although $A \perp\!\!\!\perp U \mid C$, an association between A and U conditional on M arises from conditioning on the common descendant M of A and U, also called the collider bias. Our second sensitivity parameter will assess the magnitude of this association generated by collider bias. We define our second sensitivity

parameter as

$$\text{RR}_{AU|(M,C)} = \max_m \text{RR}_{AU|(m,c)} = \max_m \max_u \frac{\text{pr}\,(u \mid A=1, m, c)}{\text{pr}\,(u \mid A=0, m, c)}, \tag{15}$$

where $\text{RR}_{AU|(m,c)}$ is the maximum of the risk ratio of A on U taking value u given $M=m$ and $C=c$. When U is binary, $\text{RR}_{AU|(m,c)}$ reduces to the usual conditional risk ratio of A on U given $M=m$ and $C=c$. The second sensitivity parameter can be viewed as the maximum of the collider bias ratios conditioning over the stratum $M=m$. We give an alternative form

$$\text{RR}_{AU|(m,c)} = \max_u \left\{ \frac{\text{pr}\,(m \mid A=1, c, u)}{\text{pr}\,(m \mid A=0, c, u)} \right\} \Big/ \left\{ \frac{\text{pr}\,(m \mid A=1, c)}{\text{pr}\,(m \mid A=0, c)} \right\}, \tag{16}$$

which is the maximum conditional relative risk of A on $M=m$ within stratum $U=u$ divided by the unconditional relative risk of A on $M=m$. The relative risk unconditional on U is identifiable from the observed data, and therefore the second sensitivity parameter depends crucially on the relative risk conditional on U.

Nonparametrically, we can specify the second sensitivity parameter using expression (15) or (16). If we would like to impose parametric assumptions, for example that $\text{pr}\,(m \mid a, c, u)$ follows a log-linear model, then it reduces to a function of the regression coefficients, which will depend explicitly on the A-M and U-M associations, as shown in the Supplementary Material.

To aid interpretation, Lemma S4 in the Supplementary Material shows that

$$\text{RR}_{AU|(m,c)} \leqslant \max_{u \neq u'} \frac{\text{pr}\,(m \mid A=1, c, u)\,\text{pr}\,(m \mid A=0, c, u')}{\text{pr}\,(m \mid A=0, c, u)\,\text{pr}\,(m \mid A=1, c, u')},$$

which measures the interaction of A and U on M taking value m given $C=c$ on the risk ratio scale (Piegorsch et al., 1994; Yang et al., 1999).

To further aid specification of this second parameter, we note that Greenland (2003) showed that, depending on the magnitude of the association, in most but not all settings the magnitude of the ratio association measure relating A and U introduced by conditioning on M is smaller than the ratios relating A and M and relating U and M. Thus, the lower of these two ratios can help to specify the second parameter. In particular, when the exposure is weakly associated with the mediator, the collider bias is small. If $A \perp\!\!\!\perp M \mid C$, then the collider bias is zero, i.e., $\text{RR}_{AU|(M,c)} = 1$.

Finally, we introduce the bounding factor

$$\text{BF}_{U|(M,c)} = \frac{\text{RR}_{AU|(M,c)} \times \text{RR}_{UY|(A=1,M,c)}}{\text{RR}_{AU|(M,c)} + \text{RR}_{UY|(A=1,M,c)} - 1},$$

which is symmetric and monotone in both $\text{RR}_{AU|(M,c)}$ and $\text{RR}_{UY|(A=1,M,c)}$, and is no larger than either sensitivity parameter. If one of the sensitivity parameters equals unity, then the bounding factor also equals unity. The bounding factor, a measure of the strength of unmeasured mediator-outcome confounding, plays a central role in bounding the natural direct and indirect effects in the following theorems.

4·3. *Bounding natural direct and indirect effects on the risk ratio scale*

THEOREM 1. *Under the assumptions in* (9) *and* (10)*, the true conditional natural direct effect on the risk ratio scale has the sharp bound* $\text{NDE}^{\text{true}}_{\text{RR}|c} \geqslant \text{NDE}^{\text{obs}}_{\text{RR}|c}/\text{BF}_{U|(M,c)}$.

The sharp bound is attainable when U is binary, $\text{pr}\,(m \mid A=0, c)$ is degenerate, and some other conditions hold as discussed in the Supplementary Material. Theorem 1 provides an easy-to-use sensitivity analysis technique. After specifying the strength of the unmeasured mediator-outcome confounder, we can calculate the bounding factor and then divide the point and interval estimates of the conditional natural direct effect by this bounding factor. This yields lower bounds on the conditional natural direct effect estimates. We can analogously apply the theorems below.

As shown in § 2, the conditional total effect can be decomposed as the product of the conditional natural direct and indirect effects on the risk ratio scale, which, coupled with Theorem 1, implies the following bound on the conditional natural indirect effects.

THEOREM 2. *Under the assumptions in* (9) *and* (10)*, the true conditional natural indirect effect on the risk ratio scale has the sharp bound* $\text{NIE}^{\text{true}}_{\text{RR}|c} \leqslant \text{NIE}^{\text{obs}}_{\text{RR}|c} \times \text{BF}_{U|(M,c)}$.

Even if a researcher does not feel comfortable specifying the sensitivity parameters, Theorems 1 and 2 can still be used to report how large the sensitivity parameters would have to be for an estimate or lower confidence limit to lie below its null hypothesis value. We illustrate this in § § 4·5 and 5.

If the natural direct effect averaged over C is of interest, the true unconditional natural direct effect must be at least as large as the minimum of $\text{NDE}^{\text{obs}}_{\text{RR}|c}/\text{BF}_{U|(M,c)}$ over c. If we further assume a common conditional natural direct effect among levels of C, as in the log-linear or logistic model for rare outcomes (cf. VanderWeele, 2015), then the true unconditional natural direct effect must be at least as large as the maximum of $\text{NDE}^{\text{obs}}_{\text{RR}|c}/\text{BF}_{U|(M,c)}$ over c. Similar arguments hold for the unconditional natural indirect effect.

4·4. *Bounding natural direct and indirect effects on the risk difference scale*

THEOREM 3. *Under the assumptions in* (9) *and* (10)*, the true conditional natural direct effect on the risk difference scale has the sharp bound*

$$\text{NDE}^{\text{true}}_{\text{RD}|c} \geqslant \sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\,\text{pr}\,(m \mid A=0, c)\big/\text{BF}_{U|(M,c)} - \text{pr}\,(Y=1 \mid A=0, c).$$

Because the conditional total effect can be decomposed as the sum of the conditional natural direct and indirect effects on the risk difference scale as shown in § 2, the identifiability of the conditional total effect and Theorem 3 imply the following bound on the conditional natural indirect effect.

THEOREM 4. *Under the assumptions in* (9) *and* (10)*, the true conditional natural indirect effect on the risk difference scale has the sharp bound*

$$\text{NIE}^{\text{true}}_{\text{RD}|c} \leqslant \text{pr}\,(Y=1 \mid A=1, c) - \sum_m \text{pr}\,(Y=1 \mid A=1, m, c)\,\text{pr}\,(m \mid A=0, c)\big/\text{BF}_{U|(M,c)}.$$

Because of the linearity of the risk difference, the true unconditional direct and indirect effects can be obtained by averaging the bounds in Theorems 3 and 4 over the distribution of the observed covariates C.

4·5. *Cornfield-type inequalities for unmeasured mediator-outcome confounding*

We can equivalently state Theorem 1 in terms of the smallest value of the bounding factor to reduce an observed conditional natural direct effect to a true conditional causal natural direct effect, i.e., $\text{BF}_{U|(M,c)} \geqslant \text{NDE}^{\text{obs}}_{\text{RR}|c}/\text{NDE}^{\text{true}}_{\text{RR}|c}$, which further implies the following Cornfield-type inequalities (Cornfield et al., 1959; Ding & VanderWeele, 2014).

THEOREM 5. *Under the assumptions in* (9) *and* (10)*, to reduce an observed conditional natural direct effect* $\text{NDE}^{\text{obs}}_{\text{RR}|c}$ *to a true conditional natural direct effect* $\text{NDE}^{\text{true}}_{\text{RR}|c}$*, both* $\text{RR}_{AU|(M,c)}$ *and* $\text{RR}_{UY|(A=1,M,c)}$ *must exceed* $\text{NDE}^{\text{obs}}_{\text{RR}|c}/\text{NDE}^{\text{true}}_{\text{RR}|c}$*, and the larger of them must exceed*

$$\left[\text{NDE}^{\text{obs}}_{\text{RR}|c} + \left\{\text{NDE}^{\text{obs}}_{\text{RR}|c}(\text{NDE}^{\text{obs}}_{\text{RR}|c} - \text{NDE}^{\text{true}}_{\text{RR}|c})\right\}^{1/2}\right] \Big/ \text{NDE}^{\text{true}}_{\text{RR}|c}.$$

To explain away an observed conditional natural direct effect $\text{NDE}^{\text{obs}}_{\text{RR}|c}$, i.e., $\text{NDE}^{\text{true}}_{\text{RR}|c} = 1$, both sensitivity parameters must exceed $\text{NDE}^{\text{obs}}_{\text{RR}|c}$, and the maximum of them must exceed $\text{NDE}^{\text{obs}}_{\text{RR}|c} + \{\text{NDE}^{\text{obs}}_{\text{RR}|c}(\text{NDE}^{\text{obs}}_{\text{RR}|c} - 1)\}^{1/2}$. In Theorem S1 in the Supplementary Material, we present the inequalities derived from Theorem 3 on the risk difference scale.

5. ILLUSTRATION

VanderWeele et al. (2012) conducted mediation analysis to assess the extent to which the effect that variants on chromosome 15q25.1 have on lung cancer is mediated through smoking and the extent to

which it operates through other causal pathways. The exposure levels correspond to changes from zero to two C alleles; smoking intensity is measured by the square root of number of cigarettes smoked per day; and the outcome is the lung cancer indicator. The analysis of VanderWeele et al. (2012) was on the odds ratio scale using a lung cancer case-control study, but for a rare disease the odds ratios approximate risk ratios. After controlling for observed sociodemographic covariates, they found that the natural direct effect estimate is 1·72 with 95% confidence interval [1·34, 2·21], and the natural indirect effect estimate is 1·03 with 95% confidence interval [0·99, 1·07]. Their analysis used logistic regression models, requiring all the odds ratios to be the same across different levels of the measured covariates.

The evidence for the indirect effect is weak, because the confidence interval covers the null hypothesis of no effect. However, the direct effect deviates significantly from the null. According to § 4·5, to reduce the point estimate of the conditional natural direct effect to below unity, both $\text{RR}_{AU|(M,c)}$ and $\text{RR}_{UY|(A=1,M,c)}$ must exceed 1·72, and the maximum of them must exceed $1{\cdot}72 + (1{\cdot}72 \times 0{\cdot}72)^{1/2} = 2{\cdot}83$. For a binary confounder U under parametric models with main effects, to explain away the direct effect estimate it would generally have to (Greenland, 2003, cf. Supplementary Material) increase the likelihood of Y and increase M by at least 1·72-fold, and it would have to increase at least one of Y and M by 2·83-fold. To reduce the lower confidence limit to below unity, both sensitivity parameters must exceed 1·34, and the maximum of them must exceed $1{\cdot}34 + (1{\cdot}34 \times 0{\cdot}34)^{1/2} = 2{\cdot}02$. For a binary confounder U under parametric models with main effects, to explain away the lower confidence limit for the direct effect it would generally have to increase the likelihood of Y and increase M by at least 1·34-fold, and it would have to increase at least one of Y and M by 2·02-fold. This would constitute fairly substantial confounding.

Previous studies have found that the exposure-mediator association in this context is weak (Saccone et al., 2010). Suppose that the risk ratio relating A and M is less than 1·40. If we assume that the collider bias is smaller than this in magnitude, e.g., $\text{RR}_{AU|(M,c)} \leqslant 1{\cdot}40$, as indicated by Greenland (2003), then $\text{RR}_{UY|(A=1,M,c)}$ must be at least as large as 11·47 to reduce the point estimate to below unity, and be at least as large as 8·93 to reduce the lower confidence limit to below unity. In general, when $\text{RR}_{AU|(M,c)}$ is relatively small, we require an extremely large $\text{RR}_{UY|(A=1,M,c)}$ to reduce the conditional natural direct effect estimate to below unity. In fact, if $\text{RR}_{AU|(M,c)}$ is smaller than the lower confidence limit of the conditional natural direct effect, it is impossible to reduce it to below unity because the bounding factor is always smaller than $\text{RR}_{AU|(M,c)}$.

6. Discussion

Theorems 1–5 are most useful when the conditional natural direct effect is greater than unity. We can also simply relabel the exposure levels and all the results will still hold.

In § 4 we derived sensitivity analysis formulae for causal parameters on the risk ratio and risk difference scales. If we have rare outcomes, as in most case-control studies, we can approximate causal parameters on the odds ratio scale by those on the risk ratio scale, and all the results about risk ratio also apply to the odds ratio. We have illustrated this in § 5. Furthermore, we comment in the Supplementary Material that similar results hold for count and continuous positive outcomes and rare time-to-event outcomes, if we replace the relative risks on the outcome by the hazard ratios and mean ratios.

The assumption $A \perp\!\!\!\perp U \mid C$ may be violated if U affects (A, M, Y) simultaneously, i.e., if unmeasured exposure-mediator, exposure-outcome and mediator-outcome confounding all exist. Even if $A \perp\!\!\!\perp U \mid C$ is violated, we show in Theorem S2 in the Supplementary Material that Theorems 1 and 3 can be interpreted as the bounds of the conditional natural direct effects for the unexposed population, which is also of interest in other contexts (Vansteelandt & VanderWeele, 2012; Lendle et al., 2013).

Acknowledgement

The authors thank the editor, associate editor and two referees for helpful comments. This research was funded by the U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theorems and more details about the discussions in § § 3, 4·2, 4·5 and 6.

REFERENCES

BARON, R. M. & KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–82.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. & WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

DIDELEZ, V., DAWID, A. P. & GENELETTI, S. (2006). Direct and indirect effects of sequential treatments. In *Proc. 22nd Conf. Uncert. Artif. Intel.*, R. Dechter & T. S. Richardson, eds. Corvallis: Association for Uncertainty in Artificial Intelligence Press, pp. 138–46.

DING, P. & VANDERWEELE, T. J. (2014). Generalized Cornfield conditions for the risk difference. *Biometrika* **101**, 971–7.

GENELETTI, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc.* B **69**, 199–215.

GREENLAND, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14**, 300–6.

IMAI, K., KEELE, L. & YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25**, 51–71.

LENDLE, S. D., SUBBARAMAN, M. S. & VAN DER LAAN, M. J. (2013). Identification and efficient estimation of the natural direct effect among the untreated. *Biometrics* **69**, 310–7.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5**, 465–72.

PEARL, J. (2001). Direct and indirect effects. In *Proc. 17th Conf. Uncert. Artif. Intel.*, J. S. Breese & D. Koller, eds. San Francisco: Morgan Kaufmann, pp. 411–20.

PEARL, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

PIEGORSCH, W. W., WEINBERG, C. R. & TAYLOR, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statist. Med.* **13**, 153–62.

ROBINS, J. M. & GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–55.

ROBINS, J. M. & RICHARDSON, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, P. Shrout, ed. Oxford: Oxford University Press, pp. 103–58.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

SACCONE, N. L., CULVERHOUSE, R. C., SCHWANTES-AN, T.-H., CANNON, D. S., CHEN, X., CICHON, S., GIEGLING, I., HAN, S., HAN, Y., KESKITALO-VUOKKO, K. ET AL. (2010). Multiple independent loci at chromosome 15q25.1 affect smoking quantity: A meta-analysis and comparison with lung cancer and COPD. *PLoS Genetics* **6**, e1001053.

SJÖLANDER, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statist. Med.* **28**, 558–71.

TCHETGEN TCHETGEN, E. J. & SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40**, 1816–45.

VANDERWEELE, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21**, 540–51.

VANDERWEELE, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.

VANDERWEELE, T. J., ASOMANING, K., TCHETGEN TCHETGEN, E. J., HAN, Y., SPITZ, M. R., SHETE, S., WU, X., GABORIEAU, V., WANG, Y., MCLAUGHLIN, J. ET AL. (2012). Genetic variants on 15q25.1, smoking, and lung cancer: An assessment of mediation and interaction. *Am. J. Epidemiol.* **175**, 1013–20.

VANSTEELANDT, S. & VANDERWEELE, T. J. (2012). Natural direct and indirect effects on the exposed: Effect decomposition under weaker assumptions. *Biometrics* **68**, 1019–27.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5**, 161–215.

YANG, Q., KHOURY, M. J., SUN, F. & FLANDERS, W. D. (1999). Case-only design to measure gene-gene interaction. *Epidemiology* **10**, 167–70.

[*Received April* 2015. *Revised March* 2016]