Zhichao Jiang^{1,2} / Tyler VanderWeele³

Causal Mediation Analysis in the Presence of a Misclassified Binary Exposure

¹ Department of Biostatistics and Epidemiology, University of Massachusetts, 715 North Pleasant Street Amherst, USA, E-mail: zhichaojiang@umass.edu

² Government, Harvard University, 1737 Cambridge Street, CGIS Knafel Building 30, Cambridge, MA, USA, E-mail: zhichaojiang@umass.edu

³ Departments of Epidemiology and Biostatistics, Harvard University, 677 Huntington Avenue, Boston, MA, USA, E-mail: tvanderw@hsph.harvard.edu

Abstract:

Mediation analysis is popular in examining the extent to which the effect of an exposure on an outcome is through an intermediate variable. When the exposure is subject to misclassification, the effects estimated can be severely biased. In this paper, when the mediator is binary, we first study the bias on traditional direct and indirect effect estimates in the presence of conditional non-differential misclassification of a binary exposure. We show that in the absence of interaction, the misclassification of the exposure will bias the direct effect towards the null but can bias the indirect effect in either direction. We then develop an EM algorithm approach to correcting for the misclassification, and conduct simulation studies to assess the performance of the correction approach. Finally, we apply the approach to National Center for Health Statistics birth certificate data to study the effect of smoking status on the preterm birth mediated through pre-eclampsia.

Keywords: EM algorithm, mediation analysis, misclassification, sensitivity analysis

DOI: 10.1515/em-2016-0006

Received: March 2, 2016; **Revised:** August 10, 2019; **Accepted:** October 31, 2019

1 Introduction

Causal mediation analysis investigates the role of an intermediate variable in explaining the relationship between an exposure variable and an outcome of interest. Mediation analysis has been widely used in various fields and was made popular by the seminal work of Baron and Kenny (1986). Recently, mediation analysis has been advanced by applying the counterfactual framework (e. g. Robins and Greenland 1992; Pearl 2001; VanderWeele and Vansteelandt 2009, 2010; Imai, Keele, and Tingley 2010). The use of the counterfactual framework allows the total effect to be decomposed into a direct effect and an indirect effect even in the presence of interactions and non-linearities (VanderWeele and Vansteelandt 2009, 2010).

In practice, variables may not be measured accurately, which leads to the problem of measurement error. For binary variables, this is usually referred to as misclassification. As demonstrated by Aigner (1973), the classification error of a binary regressor may have a non-zero mean and is negatively correlated with the true underlying variable. This is different from the measurement error of a continuous variable which is usually assumed to be centered and uncorrelated with the underlying variable. As a result, for misclassification problems, the traditional ordinary least squares estimators will lead to incorrect conclusions. To deal with this problem, several papers propose methods to correct for the misclassification (e. g. Freeman 1983; Savoca 2000).

In the context of mediation analysis, when a binary mediator is misclassified, Valeri and VanderWeele (2014) deduce the asymptotic bias for the estimators using the observed mediator and propose correction approaches, allowing for the presence of exposure-mediator interaction. In the non-parametric setting, VanderWeele, Valeri, and Ogburn (2012) investigate the sign of the bias due to the misclassification of a binary mediator. There are also many results on the measurement error of a continuous mediator in various settings (e. g. Hoyle and Kenny 1999; le Cessie et al. 2012; Valeri, Lin, and VanderWeele 2014). However, no one has considered work on misclassification of a binary exposure in the context of mediation, though this is a common problem. For example, in our data illustration, the exposure is the smoking status. Because the smoking status of study participants is self-reported, the exposure is typically subject to misclassification, which may jeopardize our inference.

Zhichao Jiang is the corresponding author.

© 2019 Walter de Gruyter GmbH, Berlin/Boston.

The present work makes two main contributions. First, when the mediator is binary, we study the implications of using a misclassified exposure in the non-parametric setting. The traditional non-differential misclassification assumption requires that the classification error is independent of other variables. In contrast, we allow the classification error to depend on the covariates. Thus, our assumption of misclassification is weaker than the traditional non-differential misclassification assumption. We derive the relationships between the naive estimators using the observed exposure and the true natural direct and indirect effects.

Second, we propose an approach to correcting the bias for estimating natural direct and indirect effects. The approach requires a logistic model for the exposure, and pre-specified sensitivity and specificity parameters for the misclassification. We can use this approach in sensitivity analysis by varying the sensitivity and specificity parameters in a pre-specified range. To evaluate the performance of the approach, we conduct a simulation study with different values of the sensitivity and specificity parameters.

The paper is organized as follows. We introduce the counterfactual framework, notation and confounding assumptions in Section 2. In Section 3, we present the results for biases when the exposure is misclassified and the mediator is binary in the non-parametric setting. We propose an approach to correcting the bias for estimating direct and indirect effects in Section 4 and conduct a simulation study in Section 5. In Section 6, we apply the proposed approach to a perinatal epidemiological study. Finally, we conclude with a discussion in Section 7. In the online supplementary materials, we present all the proofs and details for the simulation study and the real data example.

2 Mediation analysis based on the counterfactual framework

Let A denote a binary exposure, M denote a mediator, Y denote an outcome of interest, and $\mathbf{C}$ denote a vector of covariates. Let Y_a and M_a denote the values of the outcome and mediator that would have been observed if the exposure A had been set to level a . Let Y_{am} denote the value of the outcome that would have been observed if the exposure and the mediator had been set to levels a and m , respectively. Let $Y_{aM_{a'}}$ denote the value of the outcome that would have been observed if the exposure and the mediator had been set to a and $M_{a'}$, respectively. The average total effect, conditional on $\mathbf{C} = \mathbf{c}$, comparing exposure levels 1 with 0, is defined by $TE_{\mathbf{c}} = E(Y_1 - Y_0 | \mathbf{C} = \mathbf{c})$, which compares the average outcome in subgroup $\mathbf{C} = \mathbf{c}$ if A had been set to 1 with the average outcome in subgroup $\mathbf{C} = \mathbf{c}$ if A had been set to 0. The natural direct effect, conditional on $\mathbf{C} = \mathbf{c}$, comparing the effect of the exposure levels 1 and 0 while fixing the mediator to the level it would have naturally been under some reference condition for the exposure, $A = 0$, is defined by $NDE_{\mathbf{c}} = E(Y_{1M_0} - Y_{0M_0} | \mathbf{C} = \mathbf{c})$. The natural indirect effect, conditional on $\mathbf{C} = \mathbf{c}$, comparing the effect of the mediator at levels M_1 and M_0 while fixing the exposure at level 1, is defined by $NIE_{\mathbf{c}} = E(Y_{1M_1} - Y_{1M_0} | \mathbf{C} = \mathbf{c})$ (Robins and Greenland 1992; Pearl 2001).

Let $X_1 \perp\!\!\!\perp X_2 | X_3$ denote that X_1 is independent of X_2 conditional on X_3 . To identify natural direct and indirect effects, the following four confounding assumptions are sufficient. Conditioning on covariates $\mathbf{C}$, there is no unmeasured confounding of (i) the exposure-outcome relationship ($Y_a \perp\!\!\!\perp A | \mathbf{C}$), (ii) the mediator-outcome relationship ($Y_{am} \perp\!\!\!\perp M | \mathbf{C}$), (iii) the exposure-mediator relationship ($M_a \perp\!\!\!\perp A | \mathbf{C}$), and (iv) there are no mediator-outcome confounders affected by the exposure ($Y_{am} \perp\!\!\!\perp M_{a'} | \mathbf{C}$). Under these assumptions, we have the mediation formulas (Pearl 2001; VanderWeele 2015):

$$NDE_{\mathbf{c}} = \sum_m \{E(Y | A = 1, M = m, \mathbf{C} = \mathbf{c}) - E(Y | A = 0, M = m, \mathbf{C} = \mathbf{c})\}P(M = m | A = 0, \mathbf{C} = \mathbf{c}), \quad (1)$$

$$NIE_{\mathbf{c}} = \sum_m E(Y | M = m, A = 1, \mathbf{C} = \mathbf{c}) \cdot \{P(M = m | A = 1, \mathbf{C} = \mathbf{c}) - P(M = m | A = 0, \mathbf{C} = \mathbf{c})\}. \quad (2)$$

For a binary outcome, we are often interested in the natural direct and indirect effects on the odds ratio scale (VanderWeele 2015), defined as

$$OR_{\mathbf{c}}^{NDE} = \frac{P(Y_{1M_0} = 1 | \mathbf{C} = \mathbf{c})P(Y_{0M_0} = 0 | \mathbf{C} = \mathbf{c})}{P(Y_{1M_0} = 0 | \mathbf{C} = \mathbf{c})P(Y_{0M_0} = 1 | \mathbf{C} = \mathbf{c})}, \quad (3)$$

$$OR_{\mathbf{c}}^{NIE} = \frac{P(Y_{1M_1} = 1 | \mathbf{C} = \mathbf{c})P(Y_{1M_0} = 0 | \mathbf{C} = \mathbf{c})}{P(Y_{1M_1} = 0 | \mathbf{C} = \mathbf{c})P(Y_{1M_0} = 1 | \mathbf{C} = \mathbf{c})}. \quad (4)$$

Under the confounding assumptions (i)–(iv), we can identify $OR_{\mathbf{C}}^{NDE}$ and $OR_{\mathbf{C}}^{NIE}$ based on the following formula for $a, a' = 0, 1$:

$$\begin{aligned} P(Y_{aM_{a'}} = 1 \mid \mathbf{C} = \mathbf{c}) \\ = \sum_{m=0,1} P(Y = 1 \mid A = a, M = m, \mathbf{C} = \mathbf{c}) P(M = m \mid A = a', \mathbf{C} = \mathbf{c}). \end{aligned}$$

3 Results on direct and indirect effects with a binary mediator when the exposure is misclassified

In this section, we derive the relationships between the true natural direct and indirect effects and the naive estimated effects with a correctly measured binary mediator. We do not provide a closed form expression of the bias for cases with a continuous mediator but will propose an approach to correcting for bias with both binary and continuous mediators in the next section.

Let A^* denote the observed exposure, we assume A^* to be non-differentially misclassified conditional on the covariates $\mathbf{C}$, i. e. $P(A^* \mid A, M, Y, \mathbf{C}) = P(A^* \mid A, \mathbf{C})$. This conditional non-differential misclassification assumption allows the misclassification probability to depend on the covariates. Therefore, it is weaker than the traditional non-differential misclassification assumption, which assumes that the misclassification probability does not depend on the covariates, i. e. $P(A^* \mid A, M, Y, \mathbf{C}) = P(A^* \mid A)$.

For a binary mediator, we do not need any parametric assumptions and we can write (1) and (2) as

$$NDE_{\mathbf{c}} = (\eta_{11\mathbf{c}} - \eta_{01\mathbf{c}})\delta_{0\mathbf{c}} + (\eta_{10\mathbf{c}} - \eta_{00\mathbf{c}})(1 - \delta_{0\mathbf{c}}), \quad (5)$$

$$NIE_{\mathbf{c}} = (\eta_{11\mathbf{c}} - \eta_{10\mathbf{c}})(\delta_{1\mathbf{c}} - \delta_{0\mathbf{c}}), \quad (6)$$

where

$$\delta_{a\mathbf{c}} = P(M = 1 \mid A = a, \mathbf{C} = \mathbf{c}), \quad \eta_{am\mathbf{c}} = E(Y \mid A = a, M = m, \mathbf{C} = \mathbf{c}).$$

Because $\delta_{a\mathbf{c}}$ and $\eta_{am\mathbf{c}}$ are defined in terms of the true exposure A , they cannot be calculated from the observed data. We then define the observed analogue of $\delta_{a\mathbf{c}}$ by replacing A with A^* , $\delta_{a\mathbf{c}}^* = P(M = 1 \mid A^* = a, \mathbf{C} = \mathbf{c})$.

The naive estimators use the observed exposure instead of the true exposure to calculate the direct and indirect effects, and we denote them by $NDE_{\mathbf{c}}^{na}$ and $NIE_{\mathbf{c}}^{na}$, respectively. From (1) with A replaced by A^* , we have

$$NDE_{\mathbf{c}}^{na} = (\text{SN}'_{\mathbf{c}} + \text{SP}'_{\mathbf{c}} - 1) \cdot \left\{ (\eta_{11\mathbf{c}} - \eta_{01\mathbf{c}}) \frac{\delta_{1\mathbf{c}}\delta_{0\mathbf{c}}}{\delta_{1\mathbf{c}}^*} + (\eta_{10\mathbf{c}} - \eta_{00\mathbf{c}}) \frac{(1-\delta_{1\mathbf{c}})(1-\delta_{0\mathbf{c}})}{1-\delta_{1\mathbf{c}}^*} \right\}, \quad (7)$$

where $\text{SN}'_{\mathbf{c}} = P(A = 1 \mid A^* = 1, \mathbf{C} = \mathbf{c})$ and $\text{SP}'_{\mathbf{c}} = P(A = 0 \mid A^* = 0, \mathbf{C} = \mathbf{c})$. We show the detailed proof of (7) in Appendix S1.1. Furthermore, if $\eta_{11\mathbf{c}} - \eta_{01\mathbf{c}} - \eta_{10\mathbf{c}} + \eta_{00\mathbf{c}} = 0$, which is an average no-interaction assumption on the additive scale conditional on $\mathbf{C} = \mathbf{c}$, then we can simplify (7) as

$$NDE_{\mathbf{c}}^{na} = NDE_{\mathbf{c}} \cdot \frac{\delta_{1\mathbf{c}}\delta_{0\mathbf{c}} - \delta_{1\mathbf{c}}^*(\delta_{1\mathbf{c}} + \delta_{0\mathbf{c}} - 1)}{\delta_{1\mathbf{c}}^*(1 - \delta_{1\mathbf{c}}^*)} \cdot (\text{SN}'_{\mathbf{c}} + \text{SP}'_{\mathbf{c}} - 1).$$

In Appendix S1.1, we show that

$$\left| \frac{\delta_{1\mathbf{c}}\delta_{0\mathbf{c}} - \delta_{1\mathbf{c}}^*(\delta_{1\mathbf{c}} + \delta_{0\mathbf{c}} - 1)}{\delta_{1\mathbf{c}}^*(1 - \delta_{1\mathbf{c}}^*)} \cdot (\text{SN}'_{\mathbf{c}} + \text{SP}'_{\mathbf{c}} - 1) \right| \leq 1.$$

Therefore, we can obtain that $|NDE_{\mathbf{c}}^{na}| \leq |NDE_{\mathbf{c}}|$, i. e. the naive estimator will always estimate the true natural direct effect towards the null under the average no-interaction assumption conditional on $\mathbf{C} = \mathbf{c}$. For the natural indirect effect, we show in Appendix S1.2 that

$$NIE_{\mathbf{c}}^{na} = \{E(Y \mid A^* = 1, M = 1, \mathbf{C} = \mathbf{c}) - E(Y \mid A^* = 1, M = 0, \mathbf{C} = \mathbf{c})\}(\delta_{1\mathbf{c}}^* - \delta_{0\mathbf{c}}^*) \quad (8)$$

$$\begin{aligned} &= \left\{ (\eta_{11\mathbf{c}} - \eta_{01\mathbf{c}}) \frac{\delta_{1\mathbf{c}}\text{SN}'_{\mathbf{c}}}{\delta_{1\mathbf{c}}^*} + \eta_{01\mathbf{c}} - (\eta_{10\mathbf{c}} - \eta_{00\mathbf{c}}) \frac{(1-\delta_{1\mathbf{c}})\text{SN}'_{\mathbf{c}}}{1-\delta_{1\mathbf{c}}^*} - \eta_{00\mathbf{c}} \right\} \\ &\quad \cdot (\delta_{1\mathbf{c}} - \delta_{0\mathbf{c}})(\text{SN}'_{\mathbf{c}} + \text{SP}'_{\mathbf{c}} - 1). \end{aligned} \quad (9)$$

Equation (9) is more complex than (7) for NDE_c^{na} , and does not have further implications without additional assumption. Under the average no-interaction assumption conditional on $\mathbf{C} = \mathbf{c}$, we show in Appendix S1.2 that

$$NIE_c^{na} = (SN'_c + SP'_c - 1)NIE_c + (\eta_{11c} - \eta_{01c}) \frac{(\delta_{1c} - \delta_{0c})^2 SN'_c (1 - SN'_c) (SN'_c + SP'_c - 1)}{\delta_{1c}^* (1 - \delta_{1c}^*)}. \quad (10)$$

Unlike the natural direct effect, the naive estimator will not always estimate the true natural indirect effect towards the null. For example, suppose that $\eta_{11c} = \eta_{10c} = 0.6$, $\eta_{01c} = \eta_{00c} = 0.4$, $SN'_c = SP'_c = 3/4$, $\delta_{1c} = 3/4$, $\delta_{0c} = 1/4$. Then there is no interaction and $NIE_c = 0$ and $NIE_c^{na} = 0.02 > 0$; if instead $\eta_{11c} = \eta_{01c} = 0.6$, $\eta_{10c} = \eta_{00c} = 0.4$, and the other parameters remain unchanged, then $NIE_c = 0.1$ and $NIE_c^{na} = 0.05 < 0.1$. Therefore, when the exposure is misclassified, the naive estimator is not always conservative for the true natural indirect effect. This is different from the conclusion in VanderWeele et al. (2012) concerning measurement error on the mediator. They find that the bias of the naive estimator can be only in one direction when the mediator is measured with error, while in our case, when the exposure is misclassified, the bias for the indirect effect can be in either direction.

In practice, if we have some prior knowledge, we can still obtain the direction of bias of the natural indirect effect. For example, if we know that SN' is very close to 1, then the second term in (10) is close to 0. Thus, we have $|NIE_c^{na}| \approx |(SN'_c + SP'_c - 1)NIE_c| \leq |NIE_c|$, i. e. the naive estimator estimates the true natural indirect effect towards the null.

To better understand the relationship between the true effects and the naive estimands, we give two artificial examples under two scenarios without covariates:

1. high specificity ($SP = 0.9$): $\eta_{11} = 0.8$, $\eta_{10} = 0.6$, $\eta_{01} = 0.4$ and $\eta_{00} = 0.2$,
2. low specificity ($SP = 0.6$): $\eta_{11} = 0.6$, $\eta_{10} = 0.6$, $\eta_{01} = 0.7$ and $\eta_{00} = 0.3$.

In these two scenarios, we choose $P(A = 1) = 0.6$, $\delta_1 = 0.8$, $\delta_0 = 0.2$ and vary SN from 0.5 to 1. We calculate NDE and NIE according to (5) and (6), and calculate NDE^{na} and NIE^{na} according to (7) and (10). Figure 1 shows the plots of NDE^{na} and NIE^{na} . From the plots, we can see that both NDE^{na} and NIE^{na} increase as SN increases. In addition, we have $|NDE^{na}| \leq |NDE|$, which is consistent with our theoretical results. In contrast, both $|NIE^{na}| \leq |NIE|$ and $|NIE^{na}| \geq |NIE|$ are possible in the plots.

Figure 1: Plots of NDE^{na} and NIE^{na} when SN varies from 0.5 to 1. The solid line is NDE^{na} and the dashed line is NIE^{na} .

The relationship between the true effects and the naive estimators allows for drawing qualitative conclusions on the risk difference scale. Moreover, Jiang et al. (2015) showed that a positive causal effect on the log odds ratio scale implies that the effect is positive on the risk ratio scale, which further implies that the effect is positive on the risk difference scale. Therefore, our result is also useful for drawing qualitative conclusions on the risk ratio and odds ratio scales. For example, if the naive estimator of the natural direct effect on the odds ratio scale is larger than 1, then we know that the naive estimator on the risk difference scale is positive. Therefore, if there is no interaction on the additive scale, we then have $|NDE_c^{na}| \leq |NDE_c|$, and thus we can obtain that the true natural direct effect on the risk difference scale is positive. As a result, we can conclude that the true natural direct effect on the odds ratio scale is larger than 1. In the application section, we will present the results on both the odds ratio and risk difference scales.

4 Correction approach for direct and indirect effect estimators

Misclassification of the true exposure A can be seen as a type of missing data problem. In this section, we propose an approach based on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) to correcting the bias result from the misclassified exposure. We will focus on the cases with binary mediators and comment on the cases with continuous mediators in the end of this section. The approach uses the EM algorithm to obtain the MLEs of the true direct and indirect effects. It requires SN_c and SP_c to be known. It also requires us to specify the distributions of A , M and Y . This is different from mediation analysis without exposure misclassification, where the distribution of A is not needed and the expectation of Y is needed instead of its distribution. In some cases, we can avoid imposing parametric assumptions by stratifying on the covariates. However, when there are too many covariates, stratification is not possible and we need to assume parametric models for A , M and Y . For the binary exposure A , we assume the following logistic regression model:

$$\text{logit}\{P(A = 1 | C = c)\} = \gamma'c,$$

where γ is estimated from the observed data. For binary mediators, we assume a logistic regression model. For binary outcomes, we assume a logistic regression model, and for continuous outcomes, we assume a linear regression model. After implementing the approach, we can obtain the MLEs of the parameters in the regression models. By plugging the estimated parameters into (1) and (2) or (3) and (4), we then obtain the maximum likelihood estimators of the direct and indirect effects on the risk difference or odds ratio scale.

Before introducing the correction approach, we need to show the validity of the EM algorithm. That is, the joint distribution of (A, M, Y, C) is identifiable from the observed data if the sensitivity and specificity parameters are known. From the law of total probability, we have

$$\begin{aligned} P(A^* = 1, M, Y, C) &= P(A = 1, A^* = 1, M, Y, C) + P(A = 0, A^* = 1, M, Y, C) \\ &= P(A = 1, M, Y, C) \cdot SN_c + P(A = 0, M, Y, C) \cdot (1 - SP_c), \end{aligned} \quad (11)$$

where the last equality follows from $P(A^* | A, M, Y, C) = P(A^* | A, C)$. Similarly, we can obtain

$$P(A^* = 0, M, Y, C) = P(A = 1, M, Y, C) \cdot (1 - SN_c) + P(A = 0, M, Y, C) \cdot SP_c. \quad (12)$$

The probabilities on the left hand side of (11) and (12) can be identified from the observed data. When $SN_c + SP_c > 1$, which is reasonable in most real problems, we can uniquely solve $P(A, M, Y, C)$ from (11) and (12). Therefore, the joint distribution of (A, M, Y, C) is identifiable. As a result, we can use the EM algorithm to obtain the MLEs of the parameters. Note that the identification results do not require any parametric models and thus the EM algorithm is valid for both continuous and binary mediators and outcomes.

Let γ , β and θ denote the parameters of the exposure model, the mediator model and the outcome model, respectively. To implement the EM algorithm, we can write the complete-data log-likelihood as

$$\begin{aligned} (\gamma, \beta, \theta) &= \sum_{i=1}^n \{ \ell_{A|C}(\gamma | A_i, C_i) + \ell_{A^*|A,C}(A_i^*, A_i, C) + \ell_{M|A,C}(\beta | M_i, A_i, C_i) \\ &\quad + \ell_{Y|A,M,C}(\theta | A_i, M_i, Y_i, C_i) \}, \end{aligned}$$

where $\ell_{A|C}$, $\ell_{A^*|A,C}$, $\ell_{M|A,C}$ and $\ell_{Y|A,M,C}$ are log-likelihoods of the exposure model, the misclassification model, the mediator model and the outcome model, respectively. Because SN_c and SP_c are pre-specified, we can ignore the term $\ell_{A^*|A,C}$ in the EM algorithm. As a remark, for a binary outcome, θ includes only the regression coefficients in the models, but for a continuous outcome, it consists of both the regression coefficients and the variance of the error term.

The EM algorithm consists of two steps, the E-step and the M-step. At the E-step, we calculate the expectation of the log-likelihood conditional on the current parameters and the observed data (A^*, M, Y, C) , denoted by $Q(\gamma, \beta, \theta | \gamma^{(k)}, \beta^{(k)}, \theta^{(k)})$, where $\gamma^{(k)}, \beta^{(k)}, \theta^{(k)}$ are the estimates of the parameters after the k -th iteration of the M-step. We then have

$$\begin{aligned} Q(\gamma, \beta, \theta | \gamma^{(k)}, \beta^{(k)}, \theta^{(k)}) &= \sum_{i=1}^n \sum_{a=0,1} h_{ia}(\gamma^{(k)}, \beta^{(k)}, \theta^{(k)}) \ell_{A|C}(\gamma | A_i = a, C_i) \\ &\quad + \sum_{i=1}^n \sum_{a=0,1} h_{ia}(\gamma^{(k)}, \beta^{(k)}, \theta^{(k)}) \ell_{M|A,C}(\beta | M_i, A_i = a, C_i), \\ &\quad + \sum_{i=1}^n \sum_{a=0,1} h_{ia}(\gamma^{(k)}, \beta^{(k)}, \theta^{(k)}) \ell_{Y|A,M,C}(\theta | A_i = a, M_i, Y_i, C_i) \end{aligned} \quad (13)$$

for $a = 0, 1$, where

$$h_{ia}(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}) = P(A_i = a | A_i^*, M_i, Y_i, \mathbf{C}_i) \\ = \frac{P(A_i = a | \mathbf{C}_i)P(A_i^* | A_i = a, \mathbf{C}_i)P(M_i | A_i = a, \mathbf{C}_i)P(Y_i | A_i = a, M_i, \mathbf{C}_i)}{\sum_{a=0,1} P(A_i = a | \mathbf{C}_i)P(A_i^* | A_i = a, \mathbf{C}_i)P(M_i | A_i = a, \mathbf{C}_i)P(Y_i | A_i = a, M_i, \mathbf{C}_i)},$$

and $P(A_i = a | \mathbf{C}_i)$, $P(M_i | A_i = a, \mathbf{C}_i)$ and $P(Y_i | A_i = a, M_i, \mathbf{C}_i)$ are the likelihood functions for the linear model or the logistic regression model. Note that $P(A_i^* | A_i, \mathbf{C}_i)$ depends on the pre-specified SN_c and SP_c . In practice, we can set a plausible range of values for them to conduct sensitivity analysis. At the M-step, we update the parameters by maximizing $Q(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta} | \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)})$. From (13), we can maximize the log-likelihoods for the three regression models separately using weighted linear or logistic regression. The corrected coefficients from the mediator and outcome models immediately give the corrected natural direct and indirect effects.

For a continuous mediator, our correction approach is also applicable. We can replace $\ell_{M|A,C}(\boldsymbol{\beta} | A_i, M_i, \mathbf{C}_i)$ with the likelihood function for the linear model and follow the same procedure to obtain the corrected natural direct and indirect effects.

5 A simulation study

We conducted a simulation study to evaluate the finite sample performance of our correction approach for binary mediators. We also conducted a simulation study for continuous mediators in Appendix S2.1. Denote $\text{expit}(x) = 1/(1 + \exp(-x))$. We did not consider covariates in the simulation study. First, we generated A from a Bernoulli distribution with probability $\text{expit}(\gamma)$, and generated A^* from Bernoulli distributions with conditional probabilities SN and $1 - SP$. We then generated the mediator and outcome according to the following four scenarios.

1. Binary M and continuous Y : $M \sim \text{Bernoulli}(\text{logit}(\beta_0 + \beta_1 A))$ and $Y \sim N(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM, 1)$;
2. Binary M and Y : $M \sim \text{Bernoulli}(\text{logit}(\beta_0 + \beta_1 A))$ and $Y \sim \text{Bernoulli}(\text{logit}(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM))$.

We chose two sets of true values for γ , $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to include cases with and without exposure-mediator interaction in the outcome model.

1. Without interaction: $\gamma = 0.1$, $\boldsymbol{\beta} = (0.1, 0.5)$, $\boldsymbol{\theta} = (0.1, 0.5, 0.5, 0)$;
2. With interaction: $\gamma = 0.1$, $\boldsymbol{\beta} = (0.1, 0.5)$, $\boldsymbol{\theta} = (0.1, 1, 1, 0.5)$.

We calculated the true values of the NDE and NIE from these values of the parameters. For the scenarios with binary outcomes, we focused on the NDE and NIE on the odds ratio scale instead of the risk difference scale. We chose four levels of sample sizes, $N = 200, 1000, 5000, 10000$, and three levels of sensitivity parameters: (1) $SN = SP = 0.80$; (2) $SN = SP = 0.85$; (3) $SN = SP = 0.90$. For each case, we calculated the biases and root mean square errors (RMSEs) of the NDE and NIE using our correction approach as well as the naive NDE and NIE.

Table 1 presents the simulation results without exposure-mediator interaction. When the outcome is continuous, the approach can eliminate the bias for a sample size of $N = 200$. However, when the outcome is binary, a sample size of $N = 200$ is not enough for the approach to eliminate the bias, and we might need sample sizes as large as $N = 5000$ to get more accurate estimates. In general, the approach estimates the natural indirect effect with greater accuracy than the natural direct effect. The RMSEs tend to increase as the probability of misclassification increases.

Table 1: Biases and RMSEs (in parentheses) of the EM algorithm correction approach for binary mediators when sample sizes are $N = 200, 1000, 5000, 10000$, with three levels of sensitivity parameters in the absence of exposure-mediator interaction. For scenarios with binary outcomes, the NDE and NIE are on the odds ratio scale, and we also give the estimates for the total effect (TE). In the table, “con” stands for continuous and “bin” stands for binary. The user time for 100 replications of all the scenarios is 192.8s at a laptop with 2.3 GHz Intel Core i7.

Bin M & con Y		N = 200	N = 1000	N = 5000	N = 10000
NDE=0.5, NIE=0.06					
SP=SN=0.90	NDE	−0.004 (0.174)	0.000 (0.081)	0.002 (0.037)	0.000 (0.025)
	NIE	0.001 (0.054)	0.000 (0.024)	0.000 (0.010)	0.001 (0.007)
SP=SN=0.85	NDE	0.000 (0.199)	0.003 (0.093)	0.002 (0.040)	0.002 (0.030)
	NIE	−0.004 (0.062)	−0.001 (0.025)	0.001 (0.012)	0.001 (0.008)

SP=SN=0.80	NDE	-0.026 (0.231)	-0.007 (0.100)	0.002 (0.045)	0.002 (0.033)
	NIE	0.000 (0.077)	0.001 (0.029)	0.001 (0.013)	0.001 (0.009)
Naive estimator when SP=SN=0.80	NDE	-0.198 (0.248)	-0.201 (0.211)	-0.203 (0.205)	-0.203 (0.204)
	NIE	-0.019 (0.044)	-0.021 (0.027)	-0.021 (0.023)	-0.021 (0.022)

Bin M & bin Y**NDE=1.65, NIE=1.06, TE=1.75 (OR)**

SP=SN=0.90	NDE	0.156 (0.867)	0.037 (0.313)	0.015 (0.132)	0.003 (0.080)
	NIE	-0.004 (0.088)	-0.002 (0.031)	0.000 (0.014)	0.001 (0.010)
	TE	0.152 (0.925)	0.037 (0.337)	0.017 (0.143)	0.005 (0.098)
SP=SN=0.85	NDE	0.350 (1.219)	0.028 (0.342)	0.010 (0.143)	0.009 (0.102)
	NIE	-0.007 (0.096)	0.000 (0.035)	0.000 (0.015)	0.000 (0.011)
	TE	0.343 (1.213)	0.030 (0.374)	0.010 (0.156)	0.010 (0.111)
SP=SN=0.80	NDE	0.317 (1.617)	0.061 (0.402)	0.012 (0.174)	0.014 (0.122)
	NIE	-0.012 (0.128)	-0.001 (0.042)	0.016 (0.018)	0.017 (0.013)
	TE	0.293 (0.889)	0.063 (0.430)	0.005 (0.189)	0.008 (0.111)
Naive estimator when SP=SN=0.80	NDE	-0.233 (1.639)	-0.289 (0.347)	-0.300 (0.310)	-0.302 (0.307)
	NIE	-0.022 (0.052)	-0.022 (0.029)	-0.022 (0.023)	-0.021 (0.022)
	TE	-0.278 (0.551)	-0.336 (0.390)	-0.346 (0.356)	-0.348 (0.353)

Table 2 presents the simulation results in the presence of exposure-mediator interaction. The results are similar to those in Table 1 except for two aspects. First, the RMSEs are larger than the case without exposure-mediator interaction under the same setting. Second, when the mediator and outcome are both binary, the approach may not be able to obtain a accurate estimate for the direct effect even for a sample size of $N = 5000$. The RMSEs in this case are still very large even when the sample size goes to $N = 10000$. However, when the sample size is above 100,000, our correction approach performs well even with binary mediator and outcome (results shown in Appendix S2.2).

Table 2: Biases and RMSEs (in parentheses) of the EM algorithm correction approach for binary mediators when sample sizes are $N = 200, 1000, 5000, 10000$, with three levels of sensitivity parameters in the presence of exposure-mediator interaction. For binary outcomes, the NDE and NIE are on the odds ratio scale, and we also give the results for the total effect (TE). In the table, “con” stands for continuous and “bin” stands for binary. The user time for 100 replication of all the scenarios is 103.4s at a laptop with 2.3 GHz Intel Core i7.

Bin M & con Y		N = 200	N = 1000	N = 5000	N = 10000
NDE=1.26, NIE=0.18					
SP=SN=0.90	NDE	-0.001 (0.177)	0.001 (0.075)	0.001 (0.034)	0.001 (0.025)
	NIE	-0.005 (0.131)	-0.005 (0.055)	-0.001 (0.025)	-0.001 (0.018)
SP=SN=0.85	NDE	-0.013 (0.192)	0.003 (0.085)	0.000 (0.036)	0.001 (0.026)
	NIE	-0.001 (0.142)	-0.004 (0.066)	-0.004 (0.028)	-0.004 (0.021)
SP=SN=0.80	NDE	-0.016 (0.210)	-0.004 (0.089)	0.001 (0.040)	0.002 (0.028)
	NIE	-0.022 (0.160)	-0.009 (0.071)	-0.008 (0.032)	-0.006 (0.023)
Naive estimator when SP=SN=0.80	NDE	-0.500 (0.526)	-0.509 (0.514)	-0.506 (0.507)	-0.506 (0.507)
	NIE	-0.073 (0.126)	-0.073 (0.086)	-0.072 (0.075)	-0.072 (0.073)
Bin M & bin Y					
NDE=4.04, NIE=1.15, TE=4.64(OR)					
SP=SN=0.90	NDE	0.281 (4.485)	0.331 (1.430)	-0.002 (0.491)	-0.015 (0.346)
	NIE	-0.008 (0.127)	0.002 (0.056)	0.002 (0.025)	0.000 (0.018)
	TE	0.280 (6.406)	0.402 (1.706)	0.008 (0.590)	-0.015 (0.417)
SP=SN=0.85	NDE	0.117 (4.721)	0.411 (1.846)	0.003 (0.579)	-0.002 (0.411)
	NIE	0.004 (0.170)	0.001 (0.064)	0.002 (0.028)	0.001 (0.019)
	TE	0.173 (7.303)	0.497 (2.205)	0.015 (0.703)	0.002 (0.494)
SP=SN=0.80	NDE	-0.155 (5.206)	0.515 (2.408)	0.047 (0.706)	-0.015 (0.462)
	NIE	-0.007 (0.198)	0.000 (0.074)	0.002 (0.033)	0.001 (0.023)
	TE	-0.185 (7.883)	0.612 (2.879)	0.066 (0.850)	-0.011 (0.555)
Naive estimator when SP=SN=0.80	NDE	-1.392 (3.737)	-1.730 (1.796)	-1.771 (1.783)	-1.786 (1.792)
	NIE	-0.069 (0.114)	-0.068 (0.078)	-0.067 (0.069)	-0.067 (0.068)
	TE	-1.772 (5.385)	-2.142 (2.210)	-2.186 (2.198)	-2.202 (2.208)

We then compared our methods with two widely used approaches, method of moments and regression calibration (Carroll et al. 2012; Fuller 2009). The method of moments is based on the asymptotic biases of the estimated coefficients from the regression models using the observed exposure. By solving the system of equations that arises from the asymptotic biases, we can obtain the method of moments estimators of β and θ . To obtain the regression calibration estimators, we substituted the observed exposure in the naive mediator and outcome regression with the best linear predictor of the latent true exposure given A^* , M and C . After obtaining the estimators of β and θ , we then calculated the *NDE* and the *NIE*. Table 3 presents the simulation results of our approach and regression calibration for cases with binary mediators and outcomes. Because we cannot obtain the asymptotic biases for the coefficients in logistic regression models under misclassification, we do not give the result for the method of moments in Table 3. In Table S3 of the appendix, we show the comparison of our approach with the method of moments for cases with continuous mediators and outcomes. From these results, we can see that our approach is better than the other two approaches in all the cases.

Table 3: Biases and RMSEs (in parentheses) of EM algorithm and regression calibration for binary (bin) mediators and outcomes with and without interaction (int) when $SN = SP = 0.8$.

No int & bin M & bin Y		NDE=1.65, NIE=1.06(OR)			
		$N = 200$	$N = 1000$	$N = 5000$	$N = 10000$
EM correction	NDE	0.317 (1.617)	0.061 (0.402)	0.012 (0.174)	0.014 (0.122)
	NIE	-0.012 (0.128)	-0.001 (0.042)	0.002 (0.018)	0.001 (0.013)
Naive estimator	NDE	-0.233 (0.514)	-0.289 (0.347)	-0.300 (0.310)	-0.302 (0.307)
	NIE	-0.022 (0.052)	-0.022 (0.029)	-0.022 (0.023)	-0.021 (0.022)
Regression calibration	NDE	0.175 (1.040)	-0.003 (0.387)	-0.055 (0.177)	-0.062 (0.130)
	NIE	0.006 (0.159)	0.021 (0.058)	0.029 (0.036)	0.028 (0.032)
Int & bin M & bin Y		NDE=4.04, NIE=1.15(OR)			
EM correction	NDE	0.155 (5.206)	0.515 (2.408)	0.047 (0.706)	-0.015 (0.462)
	NIE	-0.007 (0.198)	0.000 (0.074)	0.002 (0.033)	0.001 (0.023)
Naive estimator	NDE	-1.392 (3.737)	-1.730 (1.796)	-1.771 (1.783)	-1.786 (1.792)
	NIE	-0.069 (0.114)	-0.068 (0.078)	-0.067 (0.069)	-0.067 (0.068)
Regression calibration	NDE	0.120 (5.166)	0.503 (3.451)	0.617 (1.763)	0.425 (0.984)
	NIE	-0.378 (0.585)	-0.247 (0.382)	-0.173 (0.220)	-0.163 (0.192)

6 Example

We apply our correction approach to a perinatal epidemiological study using National Center for Health Statistics (NCHS) birth certificate data of all US births in 2003. The causal mechanisms by which the smoking status is related to preterm birth are not entirely known and a potential intermediate variable is pre-eclampsia. However, because the smoking status of a person is self-reported, it is typically subject to misclassification.

We carry out mediation analysis to calculate the natural indirect effect of the smoking status on the preterm birth, mediated by pre-eclampsia, and also the natural direct effect through other pathways independent of pre-eclampsia. The total sample size is 3, 918, 542. Because calculating the weights in the E-step is computationally intensive, we use a random sample of 500,000 individuals and adjust for the variables which may confound smoking and preterm birth relationship, smoking and pre-eclampsia relationship, and pre-eclampsia and preterm birth relationship. We assume that there are no confounders of the pre-eclampsia and preterm birth relationship that are affected by smoking. As potential confounders C , we include mother's ethnicity, age, marital status, drinking status and education level. In addition to the smoking status, misclassification of other variables such as drinking status and pre-eclampsia is also possible in our data. Some methods in the previous literature can be applied to deal with the misclassification of other variables (VanderWeele et al. 2012; Carroll et al. 2012). However, because we consider exposure misclassification in our theory, we focus only on the misclassification of smoking status in our example. The data analysis is intended only for illustrative purposes to give an example of the exposure misclassification results. We therefore do not consider misclassification of other variables.

The traditional non-differential misclassification assumption is not plausible in the data because the misclassification probability may depend on the covariates. For example, the education level of a mother is probably

related to whether she would misreport her smoking status. However, after controlling for the covariates, it is plausible that the misclassification probability is independent of the mediator and the outcome.

We use the following models to estimate the NDE and NIE,

$$\begin{aligned}\text{logit}\{P(A = 1 | \mathbf{C} = \mathbf{c})\} &= \boldsymbol{\gamma}'\mathbf{c}, \\ \text{logit}\{P(M = 1 | A = a, \mathbf{C} = \mathbf{c})\} &= \beta_0 + \beta_1 a + \boldsymbol{\beta}'_2 \mathbf{c}, \\ \text{logit}\{P(Y = 1 | A = a, M = m, \mathbf{C} = \mathbf{c})\} &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}'_4 \mathbf{c}.\end{aligned}$$

The model of Y allows for the interaction of A and M . We calculate the NDE and NIE by first conditioning on $\mathbf{c}$ and then averaging over $\mathbf{c}$. If we ignore the misclassification and treat the observed exposure as the true exposure, the estimated natural direct effect on the odds ratio scale is 1.206 (95 % CI: (1.176, 1.251)), and the estimated natural indirect effect is 0.988 (95 % CI: (0.985, 0.992)). The naive estimators suggest a small protective effect through pre-eclampsia but a much larger detrimental direct effect through other pathways.

Next, we use our correction approach to estimating the natural direct and indirect effects under different values of sensitivity and specificity parameters. Generally, mothers are more likely to misreport their smoking status when they are smoking and are less likely to misreport when not smoking (Boyd et al. 1998) (i. e. high specificity but low sensitivity). In addition, it is reasonable to assume $\text{SN}_{\mathbf{c}} \geq 1 - \text{SP}_{\mathbf{c}}$, which means that the probability of reporting smoking when a mother is actually smoking is no less than that when a mother is not smoking. In our analysis, for simplicity, we set the values of $\text{SN}_{\mathbf{c}}$ and $\text{SP}_{\mathbf{c}}$ to be independent of $\mathbf{c}$, i. e. $\text{SN}_{\mathbf{c}} = \text{SN}$ and $\text{SP}_{\mathbf{c}} = \text{SP}$. We can also choose the values of $\text{SN}_{\mathbf{c}}$ and $\text{SP}_{\mathbf{c}}$ to depend on $\mathbf{c}$, but for the purposes of this analysis, since we do not have prior knowledge about the relations of the missing probabilities in different subpopulations, we thus do not consider this in our analysis. From the conditional non-differential misclassification assumption and the law of total probability, we have

$$\begin{aligned}P(A^* = 1 | M, Y, \mathbf{C} = \mathbf{c}) &= P(A^* = 1 | A = 1, \mathbf{C} = \mathbf{c})P(A = 1 | M, Y, \mathbf{C} = \mathbf{c}) \\ &\quad + P(A^* = 1 | A = 0, \mathbf{C} = \mathbf{c})P(A = 0 | M, Y, \mathbf{C} = \mathbf{c}) \\ &= \text{SN} \cdot P(A = 1 | M, Y, \mathbf{C} = \mathbf{c}) + (1 - \text{SP}) \cdot \{1 - P(A = 1 | M, Y, \mathbf{C} = \mathbf{c})\} \\ &= (\text{SN} + \text{SP} - 1)P(A = 1 | M, Y, \mathbf{C} = \mathbf{c}) + 1 - \text{SP} \\ &\geq 1 - \text{SP},\end{aligned}$$

which leads to $\text{SP} \geq 1 - P(A^* = 1 | M, Y, \mathbf{C} = \mathbf{c})$. From the observed data, we find that $P(A^* = 1 | M = m, Y = y, \mathbf{C} = \mathbf{c}) \leq 0.03$ in some subpopulations. Thus, the specificity should be larger than 0.97 and we choose two levels for SP , 0.97, 0.99 and vary SN from 0.7 to 1. Note that we may have different lower bounds on the sensitivity parameter if we allow for different specificity parameters for different subpopulations.

Figure 2 shows the natural direct and indirect effects on the odds ratio scale when $\text{SP} = 0.97$ and $\text{SP} = 0.99$. As SN increases from 0.7 to 1, the estimate of the natural direct effect decreases and the estimate of the natural indirect effect increases. However, in both plots, the corrected estimates do not change substantially. Due to the large sample size of 500,000, it will be very time consuming to implement the full sample bootstrap to obtain the confidence intervals. We implement the subsample bootstrap (Dimitris and Joseph 1994) to obtain the confidence intervals for the following four cases: (1) $\text{SN} = 0.7, \text{SP} = 0.97$; (2) $\text{SN} = 1, \text{SP} = 0.97$; (3) $\text{SN} = 0.7, \text{SP} = 0.99$; (4) $\text{SN} = 1, \text{SP} = 0.99$. In Table 4, we display the results of the subsample bootstrap. The results depend on the values of the sensitivity and specificity parameters considered. However, for all the four cases, the natural direct effect is detrimental and the natural indirect effect on the odds ratio scale is protective but small. The effect of smoking status on preterm birth is primarily through pathways other than pre-eclampsia status. The conclusion seems relatively robust to misclassification.

Figure 2: Sensitivity analysis for the NCHS birth certificate data. The effects are on the odds ratio scale. The solid lines are the effects when $SP = 0.97$ and the dotted lines are the effects when $SP = 0.99$.

Table 4: Estimates and confidence intervals (in parentheses) of the natural direct and indirect effects (on the odds ratio scale) for the NCHS birth certificate data under different values of sensitivity and specificity parameters.

	NDE	NIE
SN = 0.7, SP = 0.97	1.252 (1.202, 1.334)	0.981 (0.975, 0.986)
SN = 1.0, SP = 0.97	1.237 (1.197, 1.303)	0.982 (0.976, 0.986)
SN = 0.7, SP = 0.99	1.230 (1.191, 1.278)	0.996 (0.989, 0.997)
SN = 1.0, SP = 0.99	1.218 (1.183, 1.272)	0.986 (0.981, 0.990)

7 Discussion

In this paper, we study the problem of misclassification of a binary exposure in the context of causal mediation analysis with a binary mediator, allowing for the presence of exposure-mediator interaction. We show that when conditional non-differential misclassification of a binary exposure is ignored, in the absence of additive exposure-mediator interaction, the estimator estimates the natural direct effect towards the null. We deduce formulas for the biases of the naive estimators of the natural direct and indirect effects when the mediator is binary.

Our results require that there are no post-exposure mediator-outcome confounders. However, this assumption is often violated in applied mediation settings. Without this assumption, the randomized interventional direct and indirect effects can be identified by (1) and (2) under the confounding assumptions (i)–(iii) in Section 2 (Didelez et al. 2006; VanderWeele 2015). Therefore, our approach can be generalized to these randomized interventional effects. However, if there are mediator-outcome confounders affected by the exposure, then identification formulas for the interventional effects are different and thus this would require additional technical development when assessing the consequences of misclassification.

We propose a correction approach based on the EM algorithm as a possible strategy of correction for misclassification. The correction strategy works well for moderately large sample sizes except for the direct effect when both the mediator and outcome are binary and there is exposure-mediator interaction. In this setting, much larger sample sizes are needed. Applying the approach to the NCHS birth certificate data, we find that the smoking status affects the preterm birth principally through pathways other than pre-eclampsia but there may be a small protective effect mediated by pre-eclampsia. Our conclusion seems fairly robust for the presence of misclassification in the exposure.

In our bias correction approach, we treat the true binary exposure as missing data and deal with the mixture model likelihood. A general technique for finding maximum likelihood estimators in these models is the EM algorithm. Because we are dealing with a low-dimensional problem, in the sense that we have only two latent clusters, the convergence of the EM algorithm is not very slow. In the simulation study, we compared

our correction approach with two alternative approaches, the method of moments and the regression calibration, which do not require iteration. Our approach can get more accurate and precise estimates but is more computationally intensive than the two other approaches. In addition, the method of moments estimator does not require the model for $P(A = 1 | C = c)$, which is more robust.

Although our theory allows for the misclassification probabilities to depend on the covariates, we may have some practical issues when applying our correction approach. First, we will often not have subject knowledge about how the misclassification probabilities depend on the covariates. Second, we may have too many sensitivity and specificity parameters to easily interpret when we allow for the misclassification probabilities to depend on the covariates. Therefore, when there is no subject knowledge about the misclassification, it may be easier to use the same value for the misclassification probabilities to gain at least some understanding of the potential role of misclassification.

Acknowledgment

We thank the Associate Editor and two reviewers for helpful comments.

Funding

This work was supported by the National Institute of Environmental Health Sciences (Funder Id: <http://dx.doi.org/10.13039/1000000066>, Grant Number: ES017876).

References

- Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1:49–59.
- Baron, R. M., and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.
- Boyd, N. R., Windsor, R. A., Perkins, L. L., and Lowe, J. B. (1998). Quality of measurement of smoking status by self-report and saliva cotinine among pregnant women. *Maternal and Child Health Journal*, 2:77–83.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2012). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, Florida: Chapman and Hall/CRC Press.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39:1–38.
- Didelez, V., Dawid, P., and Geneletti, S. (2006). Direct and indirect effects of sequential decisions. In: *Proceedings of the 22nd Annual Conference for Uncertainty in Artificial Intelligence*, 138–146. Arlington, VA: AUAI Press.
- Dimitris, N. P., and Joseph, P. R. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22:2031–2050.
- Freeman, R. B. (1983). Longitudinal analyses of the effects of trade unions. *Journal of Labor Economics*, 2:1–26.
- Fuller, W. A. (2009). *Measurement Error Models*, volume 305. New York: John Wiley & Sons.
- Hoyle, R. H., and Kenny, D. A. A. (1999). Sample size, reliability, and tests of statistical mediation. *Statistical Strategies for Small Sample Research*, 1:195–222.
- Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15:309–334.
- Jiang, Z., Ding, P., and Geng, Z. (2015). Qualitative evaluation of associations by the transitivity of the association signs. *Statistica Sinica*, 25:1065–1079.
- le Cessie, S., Debeij, J., Rosendaal, F. R., Cannegieter, S. C., and Vandenbroucke, J. P. (2012). Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, 23:551–560.
- Pearl, J. (2001). Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 411–420. San Francisco, CA: Morgan Kaufmann.
- Robins, J. M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155.
- Savoca, E. (2000). Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings. *Health Services and Outcomes Research Methodology*, 1:149–164.
- Valeri, L., Lin, X., and VanderWeele, T. J. (2014). Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Statistics in Medicine*, 33:4875–4890.
- Valeri, L., and VanderWeele, T. J. (2014). The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics*, 15: 498–512.
- VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.
- VanderWeele, T. J., Valeri, L., and Ogburn, E. L. (2012). The role of measurement error and misclassification in mediation analysis. *Epidemiology*, 23:561–564.
- VanderWeele, T. J., and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

VanderWeele, T. J., and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172:1339–1348.

Supplementary Material: The online version of this article offers supplementary material (DOI:<https://doi.org/10.1515/em-2016-0006>).