Research Article

Yige Li* and Tyler J. VanderWeele

Direct Effects under Differential Misclassification in Outcomes, Exposures, and Mediators

<https://doi.org/10.1515/jci-2019-0020>

Received Jul 13, 2019; accepted Sep 25, 2020

Abstract: Direct effects in mediation analysis quantify the effect of an exposure on an outcome not mediated by a certain intermediate. When estimating direct effects through measured data, misclassification may occur in the outcomes, exposures, and mediators. In mediation analysis, any such misclassification may lead to biased estimates in the direct effects. Basing on the conditional dependence between the mismeasured variable and other variables given the true variable, misclassification mechanisms can be divided into non-differential misclassification and differential misclassification. In this article, several scenarios of differential misclassification will be discussed and sensitivity analysis results on direct effects will be derived for those eligible scenarios. According to our findings, the estimated direct effects are not necessarily biased in intuitively predictable directions when the misclassification is differential. The bounds of the true effects are functions of measured effects and sensitivity parameters. An example from the 2018 NCHS data will illustrate how to conduct sensitivity analyses with our results on misclassified outcomes, gestational hypertension and eclampsia, when the exposure is Hispanic women versus non-Hispanic White women and the mediator is weights gain during pregnancy.

Keywords: Controlled direct effect, Differential misclassification, Mediation analysis, Natural direct effect, Sensitivity analysis

2020 Mathematics Subject Classification: 62D20, 62P25

1 Introduction

Measurement error is one of the central threats to validity in observational studies [1–11]. Compared to unmeasured confounding, the threat of measurement error is often underestimated due to the statement that non-differential measurement error in dichotomous exposures can only bias the true odds ratio towards the null [2]. However, ignoring measurement error can lead to inverse findings, especially when the measurement error is differential [11]. In this article, we focus on sensitivity analysis for measurement error in the context of mediation analysis of binary variables. Measurement error in binary variables is also referred as “misclassification” in the literature.

Misclassification may occur in the outcomes, exposures, or mediators in mediation analysis. Previous work [12–15] on misclassification in mediation analysis have been focused on non-differential misclassification in the outcomes, exposures, or mediators respectively.

***Corresponding Author: Yige Li:** Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02215; Email: yigeli@g.harvard.edu

Tyler J. VanderWeele: Department of Epidemiology, Harvard T.H. Chan School of Public Health; Email: tvanderw@hsph.harvard.edu

Among the previous work, [12] conclude that a non-differential misclassified binary outcome biases the estimators of controlled direct effect, natural direct effect, and natural indirect effect on the risk difference scale towards the null. [13] conclude that ignoring non-differential misclassification in exposure increases type I error in mediation analysis. [14] consider non-differential misclassification in a binary mediator and conclude that the natural direct effect on the risk difference scale is overestimated. [15] study the effect of a non-differential misclassified binary mediator in the estimation of direct and indirect effects.

A recent work [11] considers the impact of differential misclassification in outcome or exposure on causal effect on risk ratio estimates. The outcome is differential misclassified when the conditional distribution of the measured outcome given the true outcome depends on the value of the exposure [11]. It happens when the sensitivity or specificity of the measured outcome varies across the two values of exposure. Differential misclassification in the exposure can occur when the subjects retrospectively report the exposure. [11] conclude the bounds of true risk ratio or odds ratio depend on the corresponding estimated ratio and ratios of some sensitivity parameters. In this article, we extend the work in [11] to provide sensitivity analysis results for controlled direct effect and natural direct effect on the risk/odds ratio scale.

In Section 2, the notation of conditional probabilities, sensitivity parameters, and causal effects on the scale of interest is introduced. In Section 3, we include the assumptions to identify controlled direct effect and natural direct effect. In Section 4, we consider several misclassification mechanisms of the outcome, exposure, or mediator respectively. The sensitivity analysis is conducted on the risk ratio scale for misclassification in the outcome and mediator and on the odds ratio scale for misclassification in the exposure.

In Section 5, we utilize data from NCHS, CDC for the year 2018 with sample size $N = 3,801,534$ and a subsample of size $n = 261,237$ to assess the conditional direct effects of ethnicity on gestational hypertension or eclampsia when the mediator is weight gain during pregnancy. According to [16], pre-eclampsia is a multifactorial disease affecting 5%-8% of all pregnancies in the United States. Pre-eclampsia is a binary variable and the diagnosis of it is subject to misclassification due to the non-specificity of signs and symptoms [15, 16]. Following the discussion in [17], we consider the situation where misclassification of gestational hypertension and eclampsia may be differential and depend on ethnicity.

2 Notation

In this article, we consider outcome Y , exposure A , and mediator M . All of those random variables are binary.

Let C denote a set of measured covariates. Let Y^* , A^* , and M^* respectively denote the measurement of Y , A , and M that may subject to misclassification.

The misclassification of the outcome Y is non-differential with respect to A and M if the distribution of measurement Y^* conditioning on Y and C is independent of A and M . On the contrary, if the distribution of measurement Y^* conditioning on Y and C is dependent on A or/and M , the misclassification is differential with respect to A or/and M .

For simplicity of notation, we assume all probabilities are conditional on covariates C and omit $C = c$ in the notation of conditional probabilities.

Definition 2.1.

Let $p_{Y|a}$ denote $P(Y = 1 | A = a)$, the probability of outcome $Y = 1$ given exposure $A = a$; let $p_{Y^*|a}$ denote $P(Y^* = 1 | A = a)$, the probability of measurement of outcome $Y^* = 1$ given exposure $A = a$; let $p_{Y|a^*}$ denote $P(Y = 1 | A^* = a)$, the probability of outcome $Y = 1$ given measurement of exposure $A^* = a$. Let $p_{A|y}$ denote $P(A = 1 | Y = y)$, the probability of exposure $A = 1$ given outcome $Y = y$; let $p_{A^*|y}$ denote $P(A^* = 1 | Y = y)$, the probability of measurement of exposure $A^* = 1$ given outcome $Y = y$.

Let $p_{M|a}$ denote $P(M = 1 | A = a)$, $p_{M^*|a}$ denote $P(M^* = 1 | A = a)$. Let $p_{A|m}$ denote $P(A = 1 | M = m)$, $p_{A^*|m}$ denote $P(A^* = 1 | M = m)$.

Let $p_{Y|am}$ denote $P(Y = 1 | A = a, M = m)$, the probability of outcome $Y = 1$ given exposure $A = a$, mediator $M = m$; let $p_{Y^*|am}$ denote $P(Y^* = 1 | A = a, M = m)$, the probability of measurement of outcome

$Y^* = 1$ given exposure $A = a$, mediator $M = m$. Let $p_{Y|a^*m}$ denote $P(Y = 1 | A^* = a, M = m)$, the probability of outcome $Y = 1$ given measurement of exposure $A^* = a$, mediator $M = m$; let $p_{Y|am^*}$ denote $P(Y = 1 | A = a, M^* = m)$, the probability of outcome $Y = 1$ given exposure $A = a$, measurement of mediator $M^* = m$.

Let $p_{M|ay}$ denote $P(M = 1 | A = a, Y = y)$, $p_{M^*|ay}$ denote $P(M^* = 1 | A = a, Y = y)$. Let $p_{A|ym}$ denote $P(A = 1 | Y = y, M = m)$, $p_{A^*|ym}$ denote $P(A^* = 1 | Y = y, M = m)$.

For the sensitivity parameters, let s_a^Y denote $P(Y^* = 1 | Y = 1, A = a)$, the sensitivity of measurement of outcome Y^* given exposure $A = a$; let s_a^M denote $P(M^* = 1 | M = 1, A = a)$; let s_y^A denote $P(A^* = 1 | A = 1, Y = y)$. Let s_{am}^Y denote $P(Y^* = 1 | Y = 1, A = a, M = m)$, the sensitivity of measurement of outcome Y^* given exposure and mediator a, m ; let s_{ay}^M denote $P(M^* = 1 | M = 1, A = a, Y = y)$; let s_{ym}^A denote $P(A^* = 1 | A = 1, Y = y, M = m)$.

For the false-positive probability parameters, let f_a^Y denote $P(Y^* = 1 | Y = 0, A = a)$, the false-positive probability of measurement of outcome Y^* given exposure $A = a$; let f_a^M denote $P(M^* = 1 | M = 0, A = a)$; let f_y^A denote $P(A^* = 1 | A = 0, Y = y)$. Let f_{am}^Y denote $P(Y^* = 1 | Y = 0, A = a, M = m)$, the false-positive probability of measurement of outcome Y^* given exposure and mediator a, m ; let f_{ay}^M denote $P(M^* = 1 | M = 0, A = a, Y = y)$; let f_{ym}^A denote $P(A^* = 1 | A = 0, Y = y, M = m)$.

Among the above definitions, $s_a^Y, s_a^M, s_y^A, s_{am}^Y, s_{ay}^M, s_{ym}^A$ are sensitivity parameters within the domain $(0, 1]$, $f_a^Y, f_a^M, f_y^A, f_{am}^Y, f_{ay}^M, f_{ym}^A$ are false-positive probability parameters within the domain $[0, 1)$.

Let Y_a and M_a denote the value of outcome and mediator that would be observed if the exposure A were set to value a , where $a \in \{0, 1\}$. Let Y_{am} denote the value of outcome that would be observed if the exposure A and the mediator M were set to value a and m respectively.

The total effect on the risk ratio scale is defined as

$$TE := \frac{P(Y_1 = 1)}{P(Y_0 = 1)}.$$

According to [18–20], the controlled direct effect on the risk ratio scale given mediator $M = m$ is defined as

$$CDE_m := \frac{P(Y_{1m} = 1)}{P(Y_{0m} = 1)}.$$

The natural direct effect on the risk ratio scale is defined as

$$NDE := \frac{P(Y_{1M_0} = 1)}{P(Y_{0M_0} = 1)}.$$

3 Identification

Let “ $A \perp B | C$ ” denote the statement that A is independent of B given C . We will review the assumptions for identifying total effect, controlled direct effect, and natural direct effect.

Assumption 3.1. (SUTVA) Stable Unit Treatment Value Assumption [21–23]. First, no interference between units. Second, there is only one version of potential outcome for each individual under each possible treatment.

Assumption 3.2. (Exchangeability) $Y_a \perp A | C$.

Assumption 3.3. (Exchangeability) $Y_{am} \perp M | A, C$.

Assumption 3.4. (Exchangeability) $M_a \perp A | C$.

Assumption 3.5. (Exchangeability) $Y_{am} \perp M_{a'} | A, C$, where $a' \neq a$.

By Assumption 3.1 and 3.2 [20], the total effect TE on the risk ratio scale is identified as

$$\text{TE} = \frac{P(Y = 1|A = 1)}{P(Y = 1|A = 0)} = \frac{p_{Y|1}}{p_{Y|0}}.$$

The total effect estimated from misclassified Y^* is denoted as TE^{Y^*} , where

$$\text{TE}^{Y^*} := \frac{P(Y^* = 1|A = 1)}{P(Y^* = 1|A = 0)} = \frac{p_{Y^*|1}}{p_{Y^*|0}}.$$

The total effect TE on the odds ratio scale is identified as

$$\text{TE}_{OR} = \frac{P(Y = 1|A = 1)/P(Y = 0|A = 1)}{P(Y = 1|A = 0)/P(Y = 0|A = 0)} = \frac{p_{Y|1}/(1 - p_{Y|1})}{p_{Y|0}/(1 - p_{Y|0})}.$$

The total effect estimated from misclassified A^* is denoted as TE^{A^*} , where

$$\text{TE}_{OR}^{A^*} := \frac{P(Y = 1|A^* = 1)/P(Y = 0|A^* = 1)}{P(Y = 1|A^* = 0)/P(Y = 0|A^* = 0)} = \frac{p_{Y|1^*}/(1 - p_{Y|1^*})}{p_{Y|0^*}/(1 - p_{Y|0^*})}.$$

By Assumption 3.1, 3.2, 3.3 [20], the controlled direct effect CDE_m is identified as

$$\text{CDE}_m = \frac{P(Y = 1|A = 1, M = m)}{P(Y = 1|A = 0, M = m)} = \frac{p_{Y|1m}}{p_{Y|0m}}.$$

The controlled direct effect estimated from misclassified Y^* is denoted as $\text{CDE}_m^{Y^*}$, where

$$\text{CDE}_m^{Y^*} := \frac{P(Y^* = 1|A = 1, M = m)}{P(Y^* = 1|A = 0, M = m)} = \frac{p_{Y^*|1m}}{p_{Y^*|0m}}.$$

The controlled direct effect estimated from misclassified M^* is denoted as $\text{CDE}_m^{M^*}$, where

$$\text{CDE}_m^{M^*} := \frac{P(Y = 1|A = 1, M^* = m)}{P(Y = 1|A = 0, M^* = m)} = \frac{p_{Y|1m^*}}{p_{Y|0m^*}}.$$

The controlled direct effect, on the odds ratio scale, estimated from misclassified A^* is denoted as $\text{CDE}_{m,OR}^{A^*}$, where

$$\text{CDE}_{m,OR}^{A^*} := \frac{P(Y = 1|A^* = 1, M = m)/P(Y = 0|A^* = 1, M = m)}{P(Y = 1|A^* = 0, M = m)/P(Y = 0|A^* = 0, M = m)} = \frac{p_{Y|1^*m}/(1 - p_{Y|1^*m})}{p_{Y|0^*m}/(1 - p_{Y|0^*m})}.$$

By Assumption 3.1, 3.2, 3.3, 3.4, and 3.5 ([20]), the natural direct effect NDE is identified as

$$\begin{aligned} \text{NDE} &= \frac{P(Y = 1|A = 1, M = 1)P(M = 1|A = 0) + P(Y = 1|A = 1, M = 0)P(M = 0|A = 0)}{P(Y = 1|A = 0, M = 1)P(M = 1|A = 0) + P(Y = 1|A = 0, M = 0)P(M = 0|A = 0)} \\ &= \frac{p_{Y|11}p_{M|0} + p_{Y|10}(1 - p_{M|0})}{p_{Y|01}p_{M|0} + p_{Y|01}(1 - p_{M|0})}. \end{aligned}$$

The natural direct effect estimated from misclassified Y^* is denoted as NDE^{Y^*} , where

$$\begin{aligned} \text{NDE}^{Y^*} &:= \frac{P(Y^* = 1|A = 1, M = 1)P(M = 1|A = 0) + P(Y^* = 1|A = 1, M = 0)P(M = 0|A = 0)}{P(Y^* = 1|A = 0, M = 1)P(M = 1|A = 0) + P(Y^* = 1|A = 0, M = 0)P(M = 0|A = 0)} \\ &= \frac{p_{Y^*|11}p_{M|0} + p_{Y^*|10}(1 - p_{M|0})}{p_{Y^*|01}p_{M|0} + p_{Y^*|01}(1 - p_{M|0})}. \end{aligned}$$

Remark 3.1. We focus on sensitivity analysis of CDE_m and NDE on the risk ratio scale. When the exposure A is misclassified, the bounds for CDE_m on the odds ratio scale are derived instead.

When estimating total effect TE, the misclassification of the outcome Y may depend on the exposure A , as illustrated in Figure 1. [11] provide sensitivity results for the causal effect on the risk ratio scale in Theorem 1. When the measurement Y^* depends on A given Y , the estimate TE^{Y^*} is not necessarily biased towards the null. The bound of true TE depends on the estimate TE^{Y^*} and sensitivity parameters $\frac{s_1^Y}{s_0^Y}$ and $\frac{f_1^Y}{f_0^Y}$. If the misclassification of Y is non-differential, i.e., sensitivity parameters are not dependent on A , then $s_1^Y = s_0^Y$ and $f_1^Y = f_0^Y$, the estimate TE^{Y^*} is biased towards the null.

Figure 1: Differential misclassification in the measurement of outcome Y^*

When the measurement of exposure A^* depends on Y conditioning on A , as illustrated in Figure 2, the sensitivity analysis is in Theorem 2 in [11]. The estimate $TE_{OR}^{A^*}$ on the odds ratio scale is not necessarily biased towards the null. It depends on true TE and four parameters $\frac{s_1^A/(1-s_1^A)}{s_0^A/(1-s_0^A)}$, $\frac{f_1^A/(1-f_1^A)}{f_0^A/(1-f_0^A)}$, $\frac{s_1^A/(1-f_1^A)}{s_0^A/(1-f_0^A)}$, $\frac{f_1^A/(1-s_1^A)}{f_0^A/(1-s_0^A)}$. If the misclassification is non-differential, then $s_1^A = s_0^A$ and $f_1^A = f_0^A$, the estimate $TE_{OR}^{A^*}$ on the odds ratio scale is biased towards the null.

Figure 2: Differential misclassification in the measurement of exposure A^*

We extend the Theorem 1 and 2 in [11] to more misclassification mechanisms in mediation analysis in Section 4.

4 Misclassification scenarios

In this section, we consider several misclassification scenarios in the outcome, the mediator, or the exposure.

4.1 Misclassification in the outcome

The misclassification of outcome Y may depend on both exposure A and mediator M . The misclassification mechanism of Y is illustrated in Figure 3. We derive the following two theorems respectively for controlled direct effect and natural direct effect.

Figure 3: Differential misclassification in the measurement of outcome Y^* in mediation analysis

Theorem 4.1.

If $p_{Y|10} \geq p_{Y|00}$, i.e. $\text{CDE}_{m=0} \geq 1$,

$$\text{CDE}_{m=0} \geq \text{CDE}_{m=0}^{Y^*} / \max \left(\frac{s_{10}^Y}{s_{00}^Y}, \frac{f_{10}^Y}{f_{00}^Y} \right). \quad (1)$$

If $p_{Y|11} \geq p_{Y|01}$, i.e. $\text{CDE}_{m=1} \geq 1$,

$$\text{CDE}_{m=1} \geq \text{CDE}_{m=1}^{Y^*} / \max \left(\frac{s_{11}^Y}{s_{01}^Y}, \frac{f_{11}^Y}{f_{01}^Y} \right). \quad (2)$$

If $p_{Y|10} \leq p_{Y|00}$, i.e. $\text{CDE}_{m=0} \leq 1$,

$$\text{CDE}_{m=0} \leq \text{CDE}_{m=0}^{Y^*} / \min \left(\frac{s_{10}^Y}{s_{00}^Y}, \frac{f_{10}^Y}{f_{00}^Y} \right). \quad (3)$$

If $p_{Y|11} \leq p_{Y|01}$, i.e. $\text{CDE}_{m=1} \leq 1$,

$$\text{CDE}_{m=1} \leq \text{CDE}_{m=1}^{Y^*} / \min \left(\frac{s_{11}^Y}{s_{01}^Y}, \frac{f_{11}^Y}{f_{01}^Y} \right). \quad (4)$$

Proof. By definition and identification of controlled direct effect,

$$\text{CDE}_{m=0}^{Y^*} = \frac{P(Y^* = 1 | A = 1, M = 0)}{P(Y^* = 1 | A = 0, M = 0)} = \frac{s_{10}^Y p_{Y|10} + f_{10}^Y (1 - p_{Y|10})}{s_{00}^Y p_{Y|00} + f_{00}^Y (1 - p_{Y|00})}.$$

The rest of proof follows immediately from the analytic result in Theorem 1 in [11] by replacing $(C = c)$ with $(C = c, M = m)$. $\square$

Remark 4.1. When the misclassification of Y is differential, the estimate CDE^{Y^*} is not necessarily biased towards the null. When the misclassification is non-differential, i.e. $s_{10}^Y = s_{00}^Y, f_{10}^Y = f_{00}^Y, s_{11}^Y = s_{01}^Y, f_{11}^Y = f_{01}^Y$, the estimate CDE^{Y^*} is biased towards the null.

Theorem 4.2.

If $p_{Y|10} \geq p_{Y|00}$ and $p_{Y|11} \geq p_{Y|01}$, i.e. $\text{CDE}_{m=0} \geq 1$ and $\text{CDE}_{m=1} \geq 1$,

$$\text{NDE} \geq \text{NDE}^{Y^*} / \max \left(\frac{\max(s_{11}^Y, s_{10}^Y)}{\min(s_{01}^Y, s_{00}^Y)}, \frac{\max(f_{11}^Y, f_{10}^Y)}{\min(f_{01}^Y, f_{00}^Y)} \right). \quad (5)$$

If $p_{Y|10} \leq p_{Y|00}$ and $p_{Y|11} \leq p_{Y|01}$, i.e. $\text{CDE}_{m=0} \leq 1$ and $\text{CDE}_{m=1} \leq 1$,

$$\text{NDE} \leq \text{NDE}^{Y^*} / \min \left(\frac{\min(s_{11}^Y, s_{10}^Y)}{\max(s_{01}^Y, s_{00}^Y)}, \frac{\min(f_{11}^Y, f_{10}^Y)}{\max(f_{01}^Y, f_{00}^Y)} \right). \quad (6)$$

Proof. By the definition of NDE,

$$\begin{aligned} \text{NDE}^{Y^*} &= \frac{P(Y^* = 1|A = 1, M = 1)P(M = 1|A = 0) + P(Y^* = 1|A = 1, M = 0)P(M = 0|A = 0)}{P(Y^* = 1|A = 0, M = 1)P(M = 1|A = 0) + P(Y^* = 1|A = 0, M = 0)P(M = 0|A = 0)} \\ &= \frac{[s_{11}^Y p_{Y|11} + f_{11}^Y(1 - p_{Y|11})]p_{M|0} + [s_{10}^Y p_{Y|10} + f_{10}^Y(1 - p_{Y|10})](1 - p_{M|0})}{[s_{01}^Y p_{Y|01} + f_{01}^Y(1 - p_{Y|01})]p_{M|0} + [s_{00}^Y p_{Y|00} + f_{00}^Y(1 - p_{Y|00})](1 - p_{M|0})}. \end{aligned}$$

If $p_{Y|10} \geq p_{Y|00}$, $p_{Y|11} \geq p_{Y|01}$,

$$\begin{aligned} \text{NDE}^{Y^*} &\leq \max \left(\frac{s_{11}^Y p_{M|0} p_{Y|11} + s_{10}^Y (1 - p_{M|0}) p_{Y|10}}{s_{01}^Y p_{M|0} p_{Y|01} + s_{00}^Y (1 - p_{M|0}) p_{Y|00}}, \frac{f_{11}^Y p_{M|0} (1 - p_{Y|11}) + f_{10}^Y (1 - p_{M|0}) (1 - p_{Y|10})}{f_{01}^Y p_{M|0} (1 - p_{Y|01}) + f_{00}^Y (1 - p_{M|0}) (1 - p_{Y|00})} \right) \\ &\leq \frac{p_{Y|11} p_{M|0} + p_{Y|10} (1 - p_{M|0})}{p_{Y|01} p_{M|0} + p_{Y|00} (1 - p_{M|0})} \max \left(\frac{\max(s_{11}^Y, s_{10}^Y)}{\min(s_{01}^Y, s_{00}^Y)}, \frac{\max(f_{11}^Y, f_{10}^Y)}{\min(f_{01}^Y, f_{00}^Y)} \right) \\ &= \text{NDE} \max \left(\frac{\max(s_{11}^Y, s_{10}^Y)}{\min(s_{01}^Y, s_{00}^Y)}, \frac{\max(f_{11}^Y, f_{10}^Y)}{\min(f_{01}^Y, f_{00}^Y)} \right). \end{aligned}$$

The second inequality is because when $\frac{p_{Y|10}}{p_{Y|00}} \geq 1$, $\frac{p_{Y|11}}{p_{Y|01}} \geq 1$, we have $\frac{f_{11}^Y p_{M|0} (1 - p_{Y|11}) + f_{10}^Y (1 - p_{M|0}) (1 - p_{Y|10})}{f_{01}^Y p_{M|0} (1 - p_{Y|01}) + f_{00}^Y (1 - p_{M|0}) (1 - p_{Y|00})} \leq \frac{f_{11}^Y p_{M|0} p_{Y|11} + f_{10}^Y (1 - p_{M|0}) p_{Y|10}}{f_{01}^Y p_{M|0} p_{Y|01} + f_{00}^Y (1 - p_{M|0}) p_{Y|00}}$.

If $p_{Y|10} \leq p_{Y|00}$, $p_{Y|11} \leq p_{Y|01}$,

$$\begin{aligned} \text{NDE}^{Y^*} &\geq \min \left(\frac{s_{11}^Y p_{M|0} p_{Y|11} + s_{10}^Y (1 - p_{M|0}) p_{Y|10}}{s_{01}^Y p_{M|0} p_{Y|01} + s_{00}^Y (1 - p_{M|0}) p_{Y|00}}, \frac{f_{11}^Y p_{M|0} (1 - p_{Y|11}) + f_{10}^Y (1 - p_{M|0}) (1 - p_{Y|10})}{f_{01}^Y p_{M|0} (1 - p_{Y|01}) + f_{00}^Y (1 - p_{M|0}) (1 - p_{Y|00})} \right) \\ &\geq \frac{p_{Y|11} p_{M|0} + p_{Y|10} (1 - p_{M|0})}{p_{Y|01} p_{M|0} + p_{Y|00} (1 - p_{M|0})} \min \left(\frac{\min(s_{11}^Y, s_{10}^Y)}{\max(s_{01}^Y, s_{00}^Y)}, \frac{\min(f_{11}^Y, f_{10}^Y)}{\max(f_{01}^Y, f_{00}^Y)} \right) \\ &= \text{NDE} \min \left(\frac{\min(s_{11}^Y, s_{10}^Y)}{\max(s_{01}^Y, s_{00}^Y)}, \frac{\min(f_{11}^Y, f_{10}^Y)}{\max(f_{01}^Y, f_{00}^Y)} \right). \end{aligned}$$

The second inequality is because when $\frac{p_{Y|10}}{p_{Y|00}} \leq 1$, $\frac{p_{Y|11}}{p_{Y|01}} \leq 1$, we have $\frac{f_{11}^Y p_{M|0} (1 - p_{Y|11}) + f_{10}^Y (1 - p_{M|0}) (1 - p_{Y|10})}{f_{01}^Y p_{M|0} (1 - p_{Y|01}) + f_{00}^Y (1 - p_{M|0}) (1 - p_{Y|00})} \geq \frac{f_{11}^Y p_{M|0} p_{Y|11} + f_{10}^Y (1 - p_{M|0}) p_{Y|10}}{f_{01}^Y p_{M|0} p_{Y|01} + f_{00}^Y (1 - p_{M|0}) p_{Y|00}}$. □

Remark 4.2. When the misclassification of Y is differential, the estimate NDE^{Y^*} is not necessarily biased towards the null. When the misclassification is non-differential, where $s_{10}^Y = s_{00}^Y = s_{11}^Y = s_{01}^Y$, $f_{10}^Y = f_{00}^Y = f_{11}^Y = f_{01}^Y$, the estimate NDE^{Y^*} is biased towards the null.

Corollary 4.1.

If we add additional constraints $s_{11}^Y = s_{10}^Y = s_{11}^Y$, $s_{01}^Y = s_{00}^Y = s_{01}^Y$, $f_{11}^Y = f_{10}^Y = f_{11}^Y$, $f_{01}^Y = f_{00}^Y = f_{01}^Y$ to Theorem 4.2, which means there is no direct path from M to Y^* . The misclassification mechanism is illustrated by Figure 4.

Figure 4: Differential misclassification in the measurement of outcome Y^* , no direct path from M to Y^*

If $p_{Y|10} \geq p_{Y|00}$, $p_{Y|11} \geq p_{Y|01}$,

$$\text{NDE} \geq \text{NDE}^{Y^*} / \max \left(\frac{s_1^Y}{s_0^Y}, \frac{f_1^Y}{f_0^Y} \right). \quad (7)$$

If $p_{Y|10} \leq p_{Y|00}$, $p_{Y|11} \leq p_{Y|01}$,

$$\text{NDE} \leq \text{NDE}^{Y^*} / \min \left(\frac{s_1^Y}{s_0^Y}, \frac{f_1^Y}{f_0^Y} \right). \quad (8)$$

4.2 Misclassification in the mediator

A misclassification mechanism of the mediator M is shown in Figure 5.

Figure 5: Differential misclassification in the measurement of mediator M^* in mediation analysis

Theorem 4.3.

If $p_{M|1} \geq p_{M|0}$,

$$\begin{aligned} & \frac{p_{M|1}/(1-p_{M|1})}{p_{M|0}/(1-p_{M|0})} \max(\text{CDE}_{m=1}, \text{CDE}_{m=0}) \\ & \geq \text{CDE}_{m=0}^{M^*} \min \left(\frac{1-s_1^M}{1-s_0^M}, \frac{1-f_1^M}{1-f_0^M} \right) / \max \left(\frac{1-s_1^M}{1-s_0^M}, \frac{1-f_1^M}{1-f_0^M} \right). \end{aligned} \quad (9)$$

If $p_{M|1} \leq p_{M|0}$,

$$\begin{aligned} & \frac{p_{M|1}/(1-p_{M|1})}{p_{M|0}/(1-p_{M|0})} \min(\text{CDE}_{m=1}, \text{CDE}_{m=0}) \\ & \leq \text{CDE}_{m=0}^{M^*} \max \left(\frac{1-s_1^M}{1-s_0^M}, \frac{1-f_1^M}{1-f_0^M} \right) / \min \left(\frac{1-s_1^M}{1-s_0^M}, \frac{1-f_1^M}{1-f_0^M} \right). \end{aligned} \quad (10)$$

Proof.

$$\begin{aligned} \text{CDE}_{m=0}^{M^*} &= \frac{P(Y=1|A=1, M^*=0)}{P(Y=1|A=0, M^*=0)} \\ &= \frac{P(Y=1|A=1, M=1)P(M=1|A=1, M^*=0) + P(Y=1|A=1, M=0)P(M=0|A=1, M^*=0)}{P(Y=1|A=0, M=1)P(M=1|A=0, M^*=0) + P(Y=1|A=0, M=0)P(M=0|A=0, M^*=0)} \\ &= \frac{p_{Y|11}(1-s_1^M)p_{M|1}/(1-p_{M^*|1}) + p_{Y|10}(1-f_1^M)(1-p_{M|1})/(1-p_{M^*|1})}{p_{Y|01}(1-s_0^M)p_{M|0}/(1-p_{M^*|0}) + p_{Y|00}(1-f_0^M)(1-p_{M|0})/(1-p_{M^*|0})} \end{aligned}$$

If $p_{M|1} \geq p_{M|0}$,

$$\frac{p_{M|1}/(1-p_{M|1})}{p_{M|0}/(1-p_{M|0})} \max \left(\frac{1-s_1^M}{1-s_0^M} \text{CDE}_{m=1}, \frac{1-f_1^M}{1-f_0^M} \text{CDE}_{m=0} \right)$$

$$\geq \text{CDE}_{m=0}^{M^*} \frac{1 - p_{M^*|1}}{1 - p_{M^*|0}} \bigg/ \frac{1 - p_{M|1}}{1 - p_{M|0}} \geq \text{CDE}_{m=0}^{M^*} \min \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right).$$

Thus,

$$\begin{aligned} & \frac{p_{M|1}/(1 - p_{M|1})}{p_{M|0}/(1 - p_{M|0})} \max(\text{CDE}_{m=1}, \text{CDE}_{m=0}) \\ & \geq \text{CDE}_{m=0}^{M^*} \min \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right) \bigg/ \max \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right). \end{aligned}$$

If $p_{M|1} \leq p_{M|0}$,

$$\begin{aligned} & \frac{p_{M|1}/(1 - p_{M|1})}{p_{M|0}/(1 - p_{M|0})} \min \left(\frac{1 - s_1^M}{1 - s_0^M} \text{CDE}_{m=1}, \frac{1 - f_1^M}{1 - f_0^M} \text{CDE}_{m=0} \right) \\ & \leq \text{CDE}_{m=0}^{M^*} \frac{1 - p_{M^*|1}}{1 - p_{M^*|0}} \bigg/ \frac{1 - p_{M|1}}{1 - p_{M|0}} \leq \text{CDE}_{m=0}^{M^*} \max \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right). \end{aligned}$$

Thus,

$$\begin{aligned} & \frac{p_{M|1}/(1 - p_{M|1})}{p_{M|0}/(1 - p_{M|0})} \min(\text{CDE}_{m=1}, \text{CDE}_{m=0}) \\ & \leq \text{CDE}_{m=0}^{M^*} \max \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right) \bigg/ \min \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right). \end{aligned}$$

□

Remark 4.3. When the misclassification of M is differential, the estimate $\text{CDE}_m^{M^*}$ is not necessarily biased towards the null.

Corollary 4.2. If applying an additional constraint $\text{CDE}_{m=1} = \text{CDE}_{m=0}$ to Theorem 4.3, we can conclude:

when $p_{M|1} \geq p_{M|0}$,

$$\frac{p_{M|1}/(1 - p_{M|1})}{p_{M|0}/(1 - p_{M|0})} \text{CDE}_{m=0} \geq \text{CDE}_{m=0}^{M^*} \min \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right) \bigg/ \max \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right),$$

when $p_{M|1} \leq p_{M|0}$,

$$\frac{p_{M|1}/(1 - p_{M|1})}{p_{M|0}/(1 - p_{M|0})} \text{CDE}_{m=0} \leq \text{CDE}_{m=0}^{M^*} \max \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right) \bigg/ \min \left(\frac{1 - s_1^M}{1 - s_0^M}, \frac{1 - f_1^M}{1 - f_0^M} \right).$$

4.3 Misclassification in the exposure

For the misclassification mechanism in Figure 6, we derive the following theorem for CDE on the odds ratio scale.

Figure 6: Differential misclassification in the measurement of exposure A^* in mediation analysis

$$\text{CDE}_{m,OR}^{A^*} = \frac{P(Y=1|A^*=1, M=m)/(1-P(Y=1|A^*=1, M=m))}{P(Y=1|A^*=0, M=m)/(1-P(Y=1|A^*=0, M=m))}.$$

Theorem 4.4.

If $p_{Y|10} \geq p_{Y|00}$, then

$$\text{CDE}_{m=0,OR} \geq \text{CDE}_{m=0,OR}^{A^*} / \max \left(\frac{s_{10}^A/(1-s_{10}^A)}{s_{00}^A/(1-s_{00}^A)}, \frac{f_{10}^A/(1-f_{10}^A)}{f_{00}^A/(1-f_{00}^A)}, \frac{s_{10}^A/(1-f_{10}^A)}{s_{00}^A/(1-f_{00}^A)}, \frac{f_{10}^A/(1-s_{10}^A)}{f_{00}^A/(1-s_{00}^A)} \right).$$

If $p_{Y|11} \geq p_{Y|01}$, then

$$\text{CDE}_{m=1,OR} \geq \text{CDE}_{m=1,OR}^{A^*} / \max \left(\frac{s_{11}^A/(1-s_{11}^A)}{s_{01}^A/(1-s_{01}^A)}, \frac{f_{11}^A/(1-f_{11}^A)}{f_{01}^A/(1-f_{01}^A)}, \frac{s_{11}^A/(1-f_{11}^A)}{s_{01}^A/(1-f_{01}^A)}, \frac{f_{11}^A/(1-s_{11}^A)}{f_{01}^A/(1-s_{01}^A)} \right).$$

If $p_{Y|10} \leq p_{Y|00}$, then

$$\text{CDE}_{m=0,OR} \leq \text{CDE}_{m=0,OR}^{A^*} / \min \left(\frac{s_{10}^A/(1-s_{10}^A)}{s_{00}^A/(1-s_{00}^A)}, \frac{f_{10}^A/(1-f_{10}^A)}{f_{00}^A/(1-f_{00}^A)}, \frac{s_{10}^A/(1-f_{10}^A)}{s_{00}^A/(1-f_{00}^A)}, \frac{f_{10}^A/(1-s_{10}^A)}{f_{00}^A/(1-s_{00}^A)} \right).$$

If $p_{Y|11} \leq p_{Y|01}$, then

$$\text{CDE}_{m=1,OR} \leq \text{CDE}_{m=1,OR}^{A^*} / \min \left(\frac{s_{11}^A/(1-s_{11}^A)}{s_{01}^A/(1-s_{01}^A)}, \frac{f_{11}^A/(1-f_{11}^A)}{f_{01}^A/(1-f_{01}^A)}, \frac{s_{11}^A/(1-f_{11}^A)}{s_{01}^A/(1-f_{01}^A)}, \frac{f_{11}^A/(1-s_{11}^A)}{f_{01}^A/(1-s_{01}^A)} \right).$$

If

$$\begin{aligned} & \min \left(\frac{P(Y=1|A^*=1, A=1, M=0)}{P(Y=0|A^*=1, A=1, M=0)}, \frac{P(Y=1|A^*=0, A=1, M=0)}{P(Y=0|A^*=0, A=1, M=0)}, \right. \\ & \quad \left. \frac{P(Y=1|A^*=1, A=0, M=0)}{P(Y=0|A^*=1, A=0, M=0)}, \frac{P(Y=1|A^*=0, A=0, M=0)}{P(Y=0|A^*=0, A=0, M=0)} \right) \leq \text{CDE}_{m=0,OR}^{A^*} \\ & \leq \max \left(\frac{P(Y=1|A^*=1, A=1, M=0)}{P(Y=0|A^*=1, A=1, M=0)}, \frac{P(Y=1|A^*=0, A=1, M=0)}{P(Y=0|A^*=0, A=1, M=0)}, \right. \\ & \quad \left. \frac{P(Y=1|A^*=1, A=0, M=0)}{P(Y=0|A^*=1, A=0, M=0)}, \frac{P(Y=1|A^*=0, A=0, M=0)}{P(Y=0|A^*=0, A=0, M=0)} \right), \end{aligned}$$

when $p_{Y|10} \geq p_{Y|00}$,

$$\text{CDE}_{m=0,OR} \geq \text{CDE}_{m=0,OR}^{A^*} / \max \left(\frac{s_{10}^A/(1-s_{10}^A)}{s_{00}^A/(1-s_{00}^A)}, \frac{f_{10}^A/(1-f_{10}^A)}{f_{00}^A/(1-f_{00}^A)} \right).$$

when $p_{Y|10} \leq p_{Y|00}$,

$$\text{CDE}_{m=0,OR} \leq \text{CDE}_{m=0,OR}^{A^*} / \min \left(\frac{s_{10}^A/(1-s_{10}^A)}{s_{00}^A/(1-s_{00}^A)}, \frac{f_{10}^A/(1-f_{10}^A)}{f_{00}^A/(1-f_{00}^A)} \right).$$

If

$$\begin{aligned} & \min \left(\frac{P(Y=1|A^*=1, A=1, M=1)}{P(Y=0|A^*=1, A=1, M=1)}, \frac{P(Y=1|A^*=0, A=1, M=1)}{P(Y=0|A^*=0, A=1, M=1)}, \right. \\ & \quad \left. \frac{P(Y=1|A^*=1, A=0, M=1)}{P(Y=0|A^*=1, A=0, M=1)}, \frac{P(Y=1|A^*=0, A=0, M=1)}{P(Y=0|A^*=0, A=0, M=1)} \right) \leq \text{CDE}_{m=1,OR}^{A^*} \\ & \leq \max \left(\frac{P(Y=1|A^*=1, A=1, M=1)}{P(Y=0|A^*=1, A=1, M=1)}, \frac{P(Y=1|A^*=0, A=1, M=1)}{P(Y=0|A^*=0, A=1, M=1)}, \right. \\ & \quad \left. \frac{P(Y=1|A^*=1, A=0, M=1)}{P(Y=0|A^*=1, A=0, M=1)}, \frac{P(Y=1|A^*=0, A=0, M=1)}{P(Y=0|A^*=0, A=0, M=1)} \right), \end{aligned}$$

when $p_{Y|11} \geq p_{Y|01}$,

$$\text{CDE}_{m=1,OR} \geq \text{CDE}_{m=1,OR}^{A^*} / \max \left(\frac{s_{11}^A/(1-s_{11}^A)}{s_{01}^A/(1-s_{01}^A)}, \frac{f_{11}^A/(1-f_{11}^A)}{f_{01}^A/(1-f_{01}^A)} \right).$$

when $p_{Y|11} \leq p_{Y|01}$,

$$\text{CDE}_{m=1,OR} \leq \text{CDE}_{m=1,OR}^{A^*} / \min \left(\frac{s_{11}^A/(1-s_{11}^A)}{s_{01}^A/(1-s_{01}^A)}, \frac{f_{11}^A/(1-f_{11}^A)}{f_{01}^A/(1-f_{01}^A)} \right).$$

Proof. The proof is a direct extension of the Theorem for misclassified exposure in [11]. $\square$

Remark 4.4. Theorem 4.2 is a direct application of Theorem 2 in [11] on conditional direct effect.

When the misclassification of A is differential, the estimate CDE^{A^*} is not necessarily biased towards the null. When the misclassification is nondifferential, where $s_{10}^A = s_{00}^A, f_{10}^A = f_{00}^A, s_{11}^A = s_{01}^A, f_{11}^A = f_{01}^A$, the estimate CDE^{A^*} on the odds ratio scale is biased towards the null.

5 Case study

In our case study, we first conduct sensitivity analyses for the differential misclassification in the outcomes, gestational hypertension and eclampsia, respectively. The data is from the 2018 national birth cohort. Gestational hypertension is defined as isolated elevated blood pressure during pregnancy [17, 24]. Eclampsia is the final stage of preeclampsia that causes convulsions and/or coma [25, 26]. Studies find Hispanic women, compared to non-Hispanic White women, have a lower risk of developing gestational hypertension [17, 24]. For preeclampsia, the results of that comparison vary [17, 24, 27].

[17] mention that women with preeclampsia being falsely labeled as having gestational hypertension is more often among Hispanic women than non-Hispanic White women. Therefore, the estimate of the risk ratio for preeclampsia among Hispanic women versus non-Hispanic white women can be biased, as well as the risk ratio of gestational hypertension.

Following the work in [17], we carry out sensitivity analysis for the controlled direct effect and natural direct effect of the risk factor ethnicity on gestational hypertension and eclampsia with US National Natality Data 2018 ($N = 3,801,534$). We study eclampsia instead of preeclampsia because the latter is not measured in the cohort. In our case study, the exposure A is ethnicity (1: Hispanic vs. 0: non-Hispanic White), which is acquired from Mother's Race/Hispanic Origin in the record. The outcomes are Y_1 : gestational hypertension and Y_2 : eclampsia, which are acquired from Risk Factors in the record. Excess gestational weight gain is a possible risk factor for gestational hypertension [26]. Therefore, the mediator M is weights gain during pregnancy (1: greater than 40 pounds vs. 0: less than or equal to 40 pounds), dichotomized from the Weight Gain in the data. According to [28], 25-35 pounds gestational weights gain is recommended for women with a prepregnancy BMI of 19.8 to 26.0 kg/m², so we choose 40 pounds as the threshold.

The covariates to adjust for both outcomes Y_1 and Y_2 include age, prepregnancy hypertension status, BMI, smoking, gestational age when prenatal care begins, plurality, prepregnancy diabetes, and gestational diabetes [17]. 2,706,066 observations are kept after the removal of observations with missing values. For an illustration of our theorems, we are interested in a conditional controlled direct effect and natural direct effect in the sub-population ($n = 261,237$) with age greater than or equal to 35, no prepregnancy hypertension, BMI less than 30 kg/m², and no smoking during pregnancy, prenatal care begin before 7th month, singleton birth, and no diabetes.

The estimated conditional total effect of ethnicity on gestational hypertension $\text{TE}^{Y_1^*}$ on the risk ratio scale is 0.89. The conditional control direct effect $\text{CDE}_{m=1}^{Y_1^*}$ in weight gain > 40 pounds group is 0.85; the $\text{CDE}_{m=0}^{Y_1^*}$ in weight gain $\leq$ 40 pounds group is 0.96; the conditional natural direct effect $\text{NDE}^{Y_1^*}$ is 0.93. For eclampsia, total effect is 1.09; $\text{CDE}_{m=1}^{Y_2^*}$ in weight gain > 40 pounds group is 1.13; $\text{CDE}_{m=0}^{Y_2^*}$ in weight gain $\leq$ 40 pounds group is 1.13; $\text{NDE}^{Y_2^*}$ is 1.13.

The evidence in [17] suggests that mislabeling preeclampsia as gestational hypertension is more common among Hispanic women. The authors conclude that such misclassification bias the estimate of the risk ratio of gestational hypertension and that of preeclampsia towards each another.

Since gestational hypertension Y_1 cannot be mislabeled as preeclampsia [17], we assume the sensitivity of gestational hypertension is the same for Hispanic women and non-Hispanic White women, i.e. $P(Y_1^* = 1|Y_1 = 1, A = 1, M = m)/P(Y_1^* = 1|Y_1 = 1, A = 0, M = m) = 1$. If eclampsia, similar to preeclampsia, is more often mislabeled as gestational hypertension in Hispanic women, then $P(Y_1^* = 1|Y_1 = 0, A = 1, M = m)/P(Y_1^* = 1|Y_1 = 0, A = 0, M = m) \geq 1$, thus $\min(\frac{s_1^Y}{s_0^Y}, \frac{f_1^Y}{f_0^Y}) = 1$. By Theorem 4.1 and Corollary 4.1, if we assume the true

CDE's and NDE of Y_1 on the risk ratio scale is less than or equal to 1, then the true $CDE_{m=1}$ in weight gain > 40 pounds group is less than or equal to 0.85 and the true $CDE_{m=0}$ in weight gain ≤ 40 pounds is less than or equal to 0.96. The true NDE is less than or equal to 0.93.

The second outcome Y_2 is eclampsia. Eclampsia may be misclassified as gestational hypertension, thus $P(Y_2^* = 1|Y_2 = 1, A = 1, M = m)/P(Y_2^* = 1|Y_2 = 1, A = 0, M = m) \leq 1$. We assume that the false-positive probability is the same for Hispanic women and non-Hispanic White women. Thus $\max(\frac{s_1^Y}{s_0^Y}, \frac{f_1^Y}{f_0^Y}) = 1$. If we assume the true CDE's and NDE on the risk ratio scale are greater than or equal to 1, then the true $CDE_{m=1}$ in weight gain > 40 pounds group is greater than or equal to 1.13 and the true $CDE_{m=0}$ in weight gain ≤ 40 pounds is greater than or equal to 1.13. The true NDE is greater than or equal to 1.13.

The misclassification mechanism in the above example always biases the estimated risk ratio of gestational hypertension and that of eclampsia towards each other, which verify the statement in the discussion in [17]. The results are summarized in Table 1.

Table 1: The estimated conditional direct effects and true conditional direct effects of ethnicity, assuming the true effects on Y_1 is less than or equal to 1, the true effects on Y_2 is greater than or equal to 1. Mislabeling eclampsia as gestational hypertension (GHP) is more common among Hispanic women. By Corollary 4.1, the bounds of true risk ratios are derived as follows

Outcome	$CDE_{m=1}^{Y^*}$	$CDE_{m=0}^{Y^*}$	NDE^{Y^*}	$CDE_{m=1}$	$CDE_{m=0}$	NDE
Y_1 (GHP)	0.85	0.96	0.93	≤ 0.85	≤ 0.96	≤ 0.93
Y_2 (Eclampsia)	1.13	1.13	1.13	≥ 1.13	≥ 1.13	≥ 1.13

However, in another example, misclassification does not necessarily bias the estimated risk ratio of gestational hypertension and that of eclampsia towards each other, which is different from the statement in [17].

If we are interested in conditional CDE's and NDE in the subpopulation ($n = 1,162,740$) with age between 20 and 34, no prepregnancy hypertension, BMI less than 30 kg/m^2 , and no smoking during pregnancy, prenatal care begin before 7th month, singleton birth, and no diabetes, we may not able to extend the statement in [17]. The estimated conditional TE of ethnicity on gestational hypertension on the risk ratio scale is 0.69; the estimated conditional TE on eclampsia is 0.89. The estimated CDE of ethnicity on gestational hypertension on the risk ratio scale in weight gain > 40 pounds group is 0.74; the estimated CDE of ethnicity on eclampsia in the same group is 0.88. The estimated CDE of ethnicity on gestational hypertension on the risk ratio scale in weight gain ≤ 40 pounds group is 0.72; the estimated CDE of ethnicity on eclampsia in the same group is 0.96. The NDE of gestational hypertension is 0.73; the NDE of eclampsia is 0.93.

If for gestational hypertension, we assume the true CDE and is less than or equal to 1, then the true CDE in weight gain > 40 pounds group is less than or equal to 0.74 and the CDE in weight gain ≤ 40 pounds is less than or equal to 0.72. The NDE is less than or equal to 0.72. When the outcome is eclampsia, let $P(Y_2^* = 1|Y_2 = 1, A = 1, M = m)/P(Y_2^* = 1|Y_2 = 1, A = 0, M = m) = 0.95$, then $\min(\frac{s_1^Y}{s_0^Y}, \frac{f_1^Y}{f_0^Y}) = 0.95$. If we assume the true CDE is less than or equal to 1, then the true CDE in weight gain > 40 pounds group is less than or equal to 0.92 and the CDE in weight gain ≤ 40 pounds is less than or equal to 1.01. The actual NDE is less than or equal to 0.98. The results are summarized in Table 2. This result is not enough to verify the statement in [17].

Table 2: The estimated conditional direct effects and true conditional direct effects of ethnicity, assuming the true effects on Y_1 is less than or equal to 1, the true effects on Y_2 is less than or equal to 1. Mislabeling eclampsia as gestational hypertension (GHP) is more common among Hispanic women. Letting $s_1^Y/s_0^Y = 0.95$, by Corollary 4.1, the bounds of true risk ratios are derived as follows

Outcome	$CDE_{m=1}^{Y^*}$	$CDE_{m=0}^{Y^*}$	NDE^{Y^*}	$CDE_{m=1}$	$CDE_{m=0}$	NDE
Y_1 (GHP)	0.74	0.72	0.73	≤ 0.74	≤ 0.72	≤ 0.73
Y_2 (Eclampsia)	0.88	0.96	0.93	≤ 0.92	≤ 1.01	≤ 0.98

Then, we consider misclassification in the mediator. In this example, the exposure is ethnicity. The outcome is low birthweight (1: less than 2500 grams, 0: greater than or equal to 2500 grams). The mediator is eclampsia. For illustration, we are interested in conditional CDE in the population ($n = 261,237$) with age greater than or equal to 35, no prepregnancy hypertension, BMI less than 30, and no smoking during pregnancy, prenatal care begin before 7th month, singleton birth, and no diabetes. Conditional TE on the risk ratio scale is 1.44. Conditional CDE in women with eclampsia is 0.96, in women without eclampsia is 1.44. By Theorem 4.4, maximum of the true conditional CDE's in the two groups is greater than or equal to

$$1.44 \frac{\min(\frac{1-s_1^M}{1-s_0^M}, 1)}{\max(\frac{1-s_1^M}{1-s_0^M}, 1)} \bigg/ \frac{p_{M|1}/(1-p_{M|1})}{p_{M|0}/(1-p_{M|0})} \text{ if the risk ratio of ethnicity on eclampsia is greater than or equal to 1. } s_1^M < s_0^M.$$

Let $\frac{1-s_1^M}{1-s_0^M} = 0.9$, then the true CDE is greater than $1.30 \bigg/ \frac{p_{M|1}/(1-p_{M|1})}{p_{M|0}/(1-p_{M|0})}$. The true odds ratio of the exposure on the mediator cannot be acquired from the same data because of misclassification of M . But a range of the odds ratio from other studies can help getting a bound.

6 Discussion

In this article, we provide sensitivity analysis results for controlled direct effect and natural direct effect in mediation analysis. We discuss misclassification in the outcome, exposure, and mediator.

Our results do not require internal validation data. If the true exposure and outcome are measured by a gold standard diagnostic technique in a subsample, differential misclassification errors can be corrected by regression analysis [29, 30].

For misclassification in the outcome, we derive bounds for the controlled direct effect and the natural direct effect on the risk ratio scale. When putting extra constraints on the sensitivity parameters to enforce non-differential misclassification, the results of non-differential misclassification can be acquired from differential misclassification. For differential misclassification in the mediator and exposure, sensitivity results for the controlled direct effects are derived. The bounds for the controlled direct effect in the presence of misclassified exposure are derived on the odds ratio scale instead of the risk ratio scale, because the latter one is not currently available.

In our case study, we verify the statement in [17] that the differential misclassification in the outcomes, gestational hypertension and preeclampsia, bias the estimated risk ratio of ethnicity on them towards each other. The facts are: mislabeling preeclampsia as gestational hypertension is more common among Hispanic women compared to non-Hispanic White women; gestational hypertension cannot be mislabeled as preeclampsia [17]. According to our derivation, the additional conditions sufficient for this statement are that the sensitivities of gestational hypertension are the same for the two groups; the false-positive probabilities of preeclampsia are the same for the two groups; the true risk ratio of ethnicity on gestational hypertension is less than or equal to 1; the true risk ratio of ethnicity on eclampsia is greater than or equal to 1. When an above condition is violated, the validity of the statement may be threatened or extra results on the bounds are required.

One limitation we mentioned is that when the misclassification occurs in the exposure, the bounds on the odds ratio scale are derived instead of the risk ratio scale. The bounds of the natural direct effect are only derived for misclassification in the outcome. Another limitation is the bounds of marginal risk ratio in the presence of confounding are not acquired because of the complexity caused by the non-collapsibility nature of conditional risk ratio estimates.

References

- [1] I. Bross, "Misclassification in 2 x 2 tables," *Biometrics*, vol. 10, no. 4, pp. 478–486, 1954.
- [2] M. Dosemeci, S. Wacholder, and J. H. Lubin, "Does nondifferential misclassification of exposure always bias a true effect toward the null value?," *Am. J. Epidemiol.*, vol. 132, pp. 746–748, 10 1990.

- [3] C. Weinberg, D. Umbach, and S. Greenland, "When will nondifferential misclassification of an exposure preserve the direction of a trend?," *Am. J. Epidemiol.*, vol. 140, pp. 565–571, September 1994.
- [4] J. M. Neuhaus, "Bias and efficiency loss due to misclassified responses in binary regression," *Biometrika*, vol. 86, no. 4, pp. 843–855, 1999.
- [5] M. A. Hernán and S. R. Cole, "Invited Commentary: Causal Diagrams and Measurement Bias," *Am. J. Epidemiol.*, vol. 170, pp. 959–962, 09 2009.
- [6] R. H. Lyles and J. Lin, "Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting," *Stat. Med.*, vol. 29, no. 22, pp. 2297–2309, 2010.
- [7] S. le Cessie, J. Debeij, F. Rosendaal, S. Cannegieter, and J. Vandenbroucke, "Quantification of bias in direct effects estimates due to different types of measurement error in the mediator," *Epidemiology*, vol. 23, pp. 551–560, 07 2012.
- [8] T. J. VanderWeele and M. A. Hernán, "Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs," *Am. J. Epidemiol.*, vol. 175, pp. 1303–1310, 05 2012.
- [9] P. Ding and T. J. VanderWeele, "Sensitivity Analysis Without Assumptions," *Epidemiology*, vol. 27, no. 3, pp. 368–377, 2016.
- [10] T. J. VanderWeele and P. Ding, "Sensitivity Analysis in Observational Research: Introducing the E-Value," *Ann. Intern. Med.*, vol. 167, pp. 268–274, 08 2017.
- [11] T. J. VanderWeele and Y. Li, "Simple sensitivity analysis for differential measurement error," *Am. J. Epidemiol.*, 05 2019.
- [12] Z. Jiang and T. J. VanderWeele, "Causal mediation analysis in the presence of a mismeasured outcome," *Epidemiology*, vol. 26, no. 1, 2015.
- [13] L. Valeri, S. L. Reese, S. Zhao, C. M. Page, W. Nystad, B. A. Coull, and S. J. London, "Misclassified exposure in epigenetic mediation analyses. does dna methylation mediate effects of smoking on birthweight?," *Epigenomics*, vol. 9, no. 3, pp. 253–265, 2017.
- [14] E. L. Ogburn and T. J. VanderWeele, "Analytic results on the bias due to nondifferential misclassification of a binary mediator," *Am. J. Epidemiol.*, vol. 176, pp. 555–561, 9 2012.
- [15] L. Valeri and T. J. VanderWeele, "The estimation of direct and indirect causal effects in the presence of misclassified binary mediator," *Biostatistics*, vol. 15, no. 3, pp. 498–512, 2014.
- [16] J. A. Turner, "Diagnosis and management of pre-eclampsia: an update," *Int. J. Womens Health*, vol. 2, pp. 327–337, 2010.
- [17] M. Wolf, A. Shah, R. Jimenez-Kimble, J. Sauk, J. L. Ecker, and R. Thadhani, "Differential risk of hypertensive disorders of pregnancy among hispanic women," *J. Am. Soc. Nephrol.*, vol. 15, no. 5, pp. 1330–1338, 2004.
- [18] J. M. Robins and S. Greenland, "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, vol. 3, no. 2, pp. 143–155, 1992.
- [19] J. Pearl, "Direct and indirect effects," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, (San Francisco, CA, USA), pp. 411–420, Morgan Kaufmann Publishers Inc., 2001.
- [20] T. J. VanderWeele, *Explanation in causal inference : methods for mediation and interaction*. Oxford University Press, 2015.
- [21] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *Ann. Stat.*, vol. 6, no. 1, pp. 34–58, 1978.
- [22] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *J. Am. Stat. Assoc.*, vol. 75, no. 371, pp. 591–593, 1980.
- [23] P. W. Holland, "Statistics and causal inference," *J. Am. Stat. Assoc.*, vol. 81, no. 396, pp. 945–960, 1986.
- [24] G. Ghosh, J. Grewal, T. Männistö, P. Mendola, Z. Chen, Y. Xie, and S. K. Laughon, "Racial/ethnic differences in pregnancy-related hypertensive disease in nulliparous women," *Int. J. Hypertens.*, vol. 24, no. 3, pp. 283–289, 2014.
- [25] J. J. Chang, J. F. Strauss, III, J. P. Deshazo, F. B. Rigby, D. P. Chelmow, and G. A. Macones, "Reassessing the impact of smoking on preeclampsia/eclampsia: Are there age and racial differences?," *PLOS ONE*, vol. 9, pp. 1–8, 10 2014.
- [26] F. Veglio, G. K. Singh, M. Siahpush, L. Liu, and M. Allender, "Racial/ethnic, nativity, and sociodemographic disparities in maternal hypertension in the united states, 2014-2015," *Int. J. Hypertens.*, vol. 2018, p. 7897189, 2018.
- [27] J. Gong, D. A. Savitz, C. R. Stein, and S. M. Engel, "Maternal ethnicity and pre-eclampsia in new york city, 1995-2003.," *Paediatr. Perinat. Ep.*, vol. 26, no. 1, pp. 45–52, 2012.
- [28] R. T. Fortner, P. Pekow, C. G. Solomon, G. Markenson, and L. Chasan-Taber, "Prepregnancy body mass index, gestational weight gain, and risk of hypertensive pregnancy among latina women," *Am. J. Obstet. Gynecol.*, vol. 200, no. 2, pp. 167.e1–167.e7, 2009.
- [29] L. Tang, R. H. Lyles, C. C. King, D. D. Celentano, and Y. Lo, "Binary regression with differentially misclassified response and exposure variables," *Stat. Med.*, vol. 34, no. 9, pp. 1605–1620, 2015.
- [30] L. Tang, R. H. Lyles, C. C. King, J. W. Hogan, and Y. Lo, "Regression analysis for differentially misclassified correlated binary outcomes," *J. Roy. Stat. Soc. C-App.*, vol. 64, no. 3, pp. 433–449, 2015.