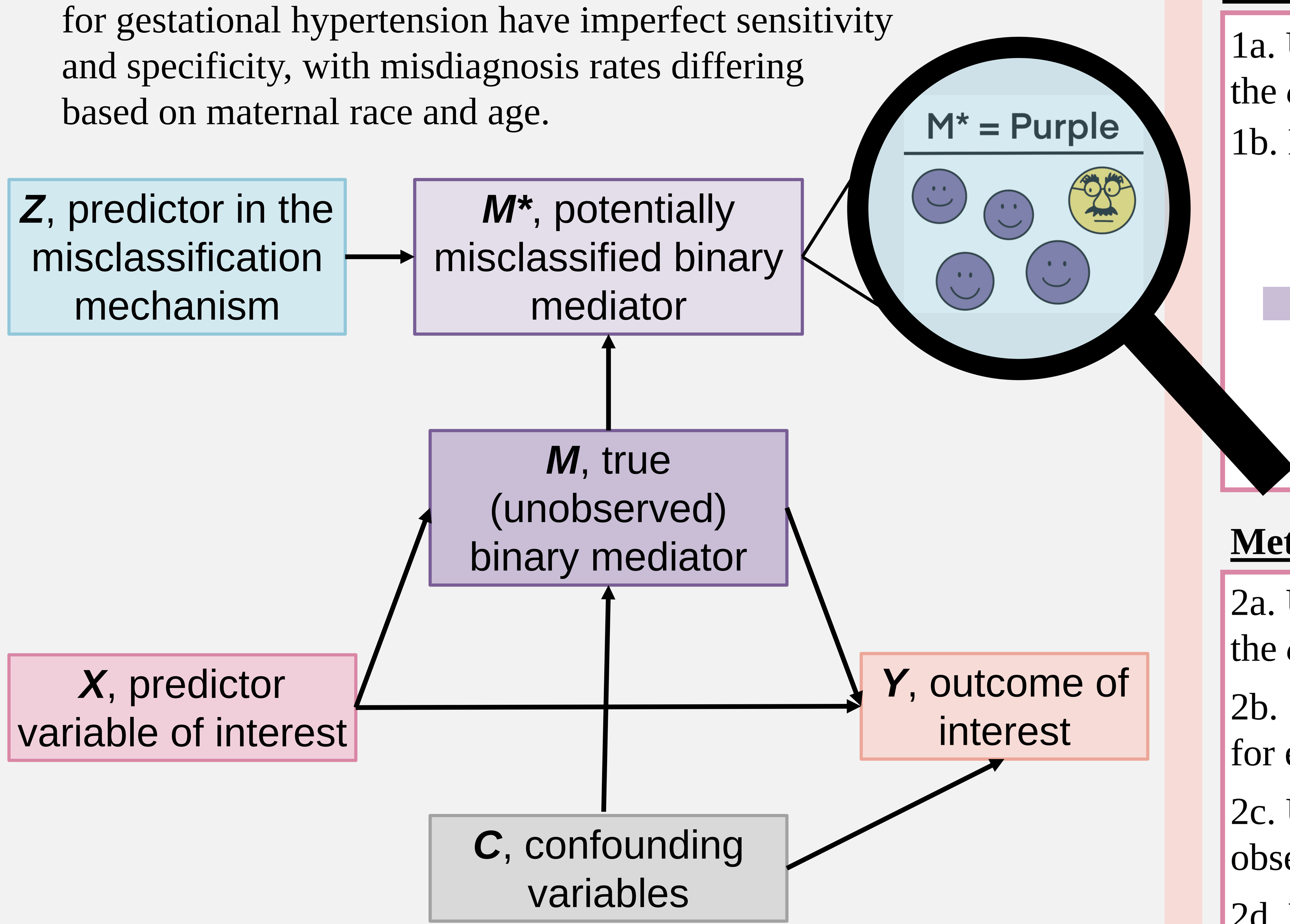
Effect estimation in the presence of a misclassified binary mediator

Kimberly A. H. Webb and Martin T. Wells
Department of Statistics and Data Science, Cornell University
Cornell Bowers CIS Statistics and Data Science

Introduction

Problem: Mediation analysis quantifies the effect of an exposure on an outcome mediated by a certain intermediate. If the ***binary mediator is misclassified***, the mediation analysis can be severely biased.

- Misclassification is especially difficult to deal with when it is ***differential*** and when there are ***no gold standard labels*** available.
- **Example:** Maternal age may be associated with gestational hypertension, which is a risk factor for preterm birth. However, tests for gestational hypertension have imperfect sensitivity and specificity, with misdiagnosis rates differing based on maternal race and age.



Research goal: To develop a suite of analysis techniques that allow researchers to estimate regression parameters in mediation models when a ***binary mediator is subject to differential misclassification, but no gold standard measures are available***.

Previous work

Webb and Wells (2023)[1] develops methods and software[2] for estimating logistic regression models with misclassified outcomes.

- **Key assumption:** Outcomes are correctly classified in at least 50% of the observations.
- **Key result:** Misclassification rates can be estimated for all subjects.

True outcome mechanism: $\text{logit}\{P(Y = j|X; \beta)\} = \beta_{j0} + \beta_{jX} X$
Observed outcome mechanism: $\text{logit}\{P(Y^* = k|Y = j, Z; \gamma)\} = \gamma_{kj0} + \gamma_{kjZ} Z$

- **COMBO** is used as a first step in ***Method #1*** and ***Method #2***.

Methods

Aim: To develop a suite of statistical methods to estimate the parameters in the following ***model specification***:

Binary mediator model: $\text{logit}\{P(M = 1|X = x, C = c)\} = \beta_0 + \beta_x x + \beta_c c$
Observed mediator model: $\text{logit}\{P(M^* = 1|M = m, Z = z)\} = \gamma_0 + \gamma_{zm} z$
Outcome model: $E(Y|X = x, C = c, M = m) = \theta_0 + \theta_x x + \theta_c c + \theta_m m$

Method #1: ***OLS Correction***[3] (only for Normal outcome models)

1a. Use the **COMBO**[1] method to estimate the ***binary mediator model***, the ***observed mediator model***, and the misclassification rates.

1b. Estimate bias adjusted parameters in the ***outcome model***

$$\begin{bmatrix}\hat\theta_m \\ \hat\theta_x\end{bmatrix} = \begin{bmatrix}(1-\zeta)S_{M^*M^*} & S_{M^*X} \\ (1+\xi)S_{XM^*} & S_{XX}\end{bmatrix}^{-1}\begin{bmatrix}S_{YM^*} \\ S_{YX}\end{bmatrix}$$

P(M* = 1 | M = 2)

P(M* = 1)

$$\hat\theta_0 = \bar{Y} - \hat\theta_m \frac{\bar{M}^* - \pi^*_{12}}{(1 - \pi^*_{12} - \pi^*_{21})} - \bar{X}^T\hat\theta_x$$

P(M* = 2 | M = 1)

$$\text{where } \zeta = 1 - \frac{(\pi^*_1 - \pi^*_{12})(1 - \pi^*_{21} - \pi^*_1)}{(1 - \pi^*_{12} - \pi^*_{21})(1 - \pi^*_1)\pi^*_1} \text{ and } \xi = \frac{(\pi^*_{21} + \pi^*_{12})}{(1 - \pi^*_{12} - \pi^*_{21})}$$

Method #2: ***Predictive Value Weighting***[4] ***(PVW)***

2a. Use the **COMBO**[1] method to estimate the ***binary mediator model***, the ***observed mediator model***, and the misclassification rates.

2b. Specify a logistic regression model to estimate *P(M* = 1 | Y, X, C)* for every subject *i*.

2c. Use the subject-specific sensitivity and specificity estimates and observed outcome probabilities to **compute the** NPV_i **and** PPV_i for all *i*.

2d. Duplicate each record in the dataset, specifying *M = 0* in the original and *M = 1* in the duplicate.

2e. Create a weight variable specified as follows:

$$M_i = 1 \cap M_i^* = 1 \implies w_i = PPV_i$$
$$M_i = 0 \cap M_i^* = 1 \implies w_i = 1 - PPV_i$$
$$M_i = 1 \cap M_i^* = 0 \implies w_i = 1 - NPV_i$$
$$M_i = 0 \cap M_i^* = 0 \implies w_i = NPV_i.$$

2f. Fit a weighted logistic regression to estimate the parameters in the ***outcome model***.

Method #3: ***An EM Algorithm***

E-Step: $P(M_i = m|x_i, m_i^*, c_i, y_i, z_i, \beta^{(t)}, \gamma^{(t)}, \theta^{(t)}, \sigma^{(t)}) =$

P(M$_i$* = ℓ | M$_i$ = m)

$$\sum_{\ell=1}^{2} \frac{m^*_{i\ell} P(y_i|x_i, m_i = m, c_i, \theta^{(t)}, \sigma^{(t)})\pi^*_{i\ell m}\pi_{im}}{\sum_{j=1}^{2} P(y_i|x_i, m_i = j, c_i, \theta^{(t)}, \sigma^{(t)})\pi^*_{i\ell j}\pi_{ij}}$$

I(M$_i$* = ℓ)

P(M$_i$ = j)

M-Step: Maximize $Q(\beta, \gamma, \theta, \sigma|\beta^{(t)}, \gamma^{(t)}, \theta^{(t)}, \sigma^{(t)}) =$

Iteration *t*

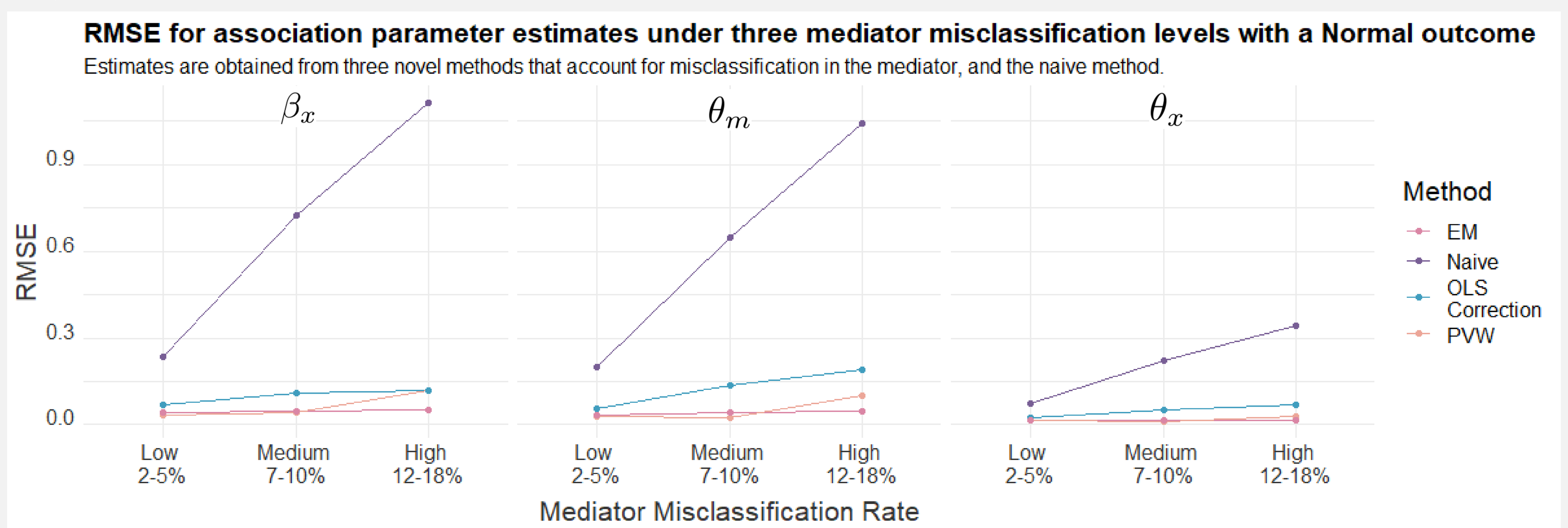
$$\sum_{i=1}^{N}\sum_{m=1}^{2} P(M_i = m|x_i, m_i^*, c_i, y_i, z_i, \beta^{(t)}, \gamma^{(t)}, \theta^{(t)}, \sigma^{(t)}) \times$$
$$[\ell_{y|x,m,c}(\theta, \sigma; x_i, m_i, c_i, y_i) + \ell_{m|x,c}(\beta; x_i, c_i, m_i) + \ell_{m^*|m,z}(\gamma; m_i, z_i, m_i^*)]$$

Results

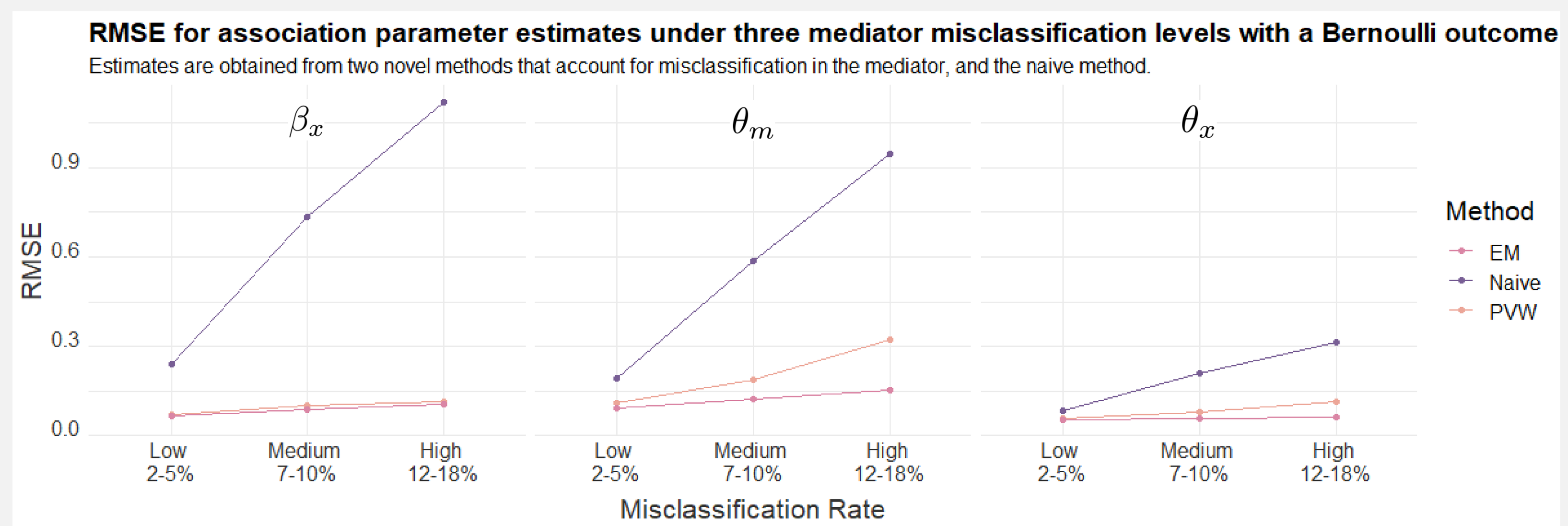
We simulate data with a misclassified binary mediator and under two conditions: ***1) a Normal outcome and 2) a Bernoulli outcome***.

- For each scenario we apply the OLS correction (if *Y* is Normal), PVW, and EM algorithm, as well as a naïve model that ignores misclassification in *M*.
- The **RMSE for three parameter estimates** are compared below for each method, at three misclassification levels.

1) Simulations with a Normal outcome



2) Simulations with a Bernoulli outcome



Conclusions

- Ignoring misclassified mediators introduces bias in association parameter estimates.
- Use of the **EM algorithm approach** to misclassified mediator correction yields more precise parameter estimates than use of the OLS correction or PVW methods.
  - The OLS correction will perform better for more uniform misclassification rates.

Primary References

1. Webb, K.A.H. and Wells, M.T. (2023). "Statistical inference for association studies in the presence of binary outcome misclassification". *arXiv preprint arXiv:2303.10215*.
2. Hochstedler, K.A. (2023). "COMBO: Correcting Misclassified Binary Outcomes in Association Studies". *R package version 1.0.0*, https://CRAN.R-project.org/package=COMBO.
3. Nguimkeu, P., Rosenman, R. and Tennekoon, V. (2020). "Regression with a misclassified binary regressor: Correcting the hidden bias".
4. Lyles, R. H. and Lin, J. (2010). "Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting". *Statistics in Medicine*, 29(22), 2297-2309.