Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Xi Fang Date

Improving Precision in Mediation Analysis Through Efficient Use of Case Data

By

Xi Fang

Master of Science in Public Health

Biostatistics and Bioinformatics

Robert Lyles (Thesis Advisor)

Glen Satten (Reader)

Improving Precision in Mediation Analysis Through Efficient Use of Case Data

By

Xi Fang

B.S

East China University of East China of Science and Technology

2013

Thesis Committee Chair: Robert Lyles, PhD

MS Reader: Glen Satten, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

Abstract

Improving Precision in Mediation Analysis Through Efficient Use of Case Data

By: Xi Fang

Mediation analysis refers to a set of statistical techniques designed to explore the relationship between a dependent variable and one or more independent variables of primary interest, while also accounting for one or more intermediate (or mediating) variables. A traditional method for mediation analysis is the structural equation model (SEM) which is used to test for the existence of mediators. However, this SEM model can not be used to test the significance of the indirect effect generated by the effect of exposure on the outcome through a mediator. An alternative approach is based on causal inference developments. In the method suggested by VanderWeele et al., the total effect can be decomposed to the sum of the natural indirect effect (NIE) and natural direct effect (NDE). In this thesis, a new method based on results given by Satten and Kupper[9] and Satten and Carroll[10]is introduced to estimate the indirect effects more precisely in the case-control setting by making more efficient use of data on cases. This approach is compared with the VanderWeele proposal in terms of the precision of estimated direct and indirect effects. In this new method, the multivariate delta method was used to estimate the variance of log transformed causal effect estimates based on maximum likelihood and corresponding 95% confidence intervals and assessed for coverage of the true value. After multiple simulation studies, we found that in the model without interaction between exposure and mediator, the Satten method performed well in terms of precision for estimating causal effects. If interaction between exposure and mediator exists, although new method can estimate causal effects more precisely with small sample size, the distributions of these effects were left skewed. When increasing the sample size, the distributions were closer to normal as expected, but the difference in precision of causal effect estimates between these two methods was largely decreased. The simulations suggest that the Satten proposal attains better precision when interaction was not present. When interaction was present, the methods performed similarly, with the Satten approach showing some potential benefits when sample size is relatively small.

Improving Precision in Mediation Analysis Through Efficient Use of Case Data

By

Xi Fang

B.S

East China University of East China of Science and Technology

2013

Thesis Committee Chair: Robert Lyles, PhD

MS Reader: Glen Satten, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

Acknowledgements:

I would like to express my sincere gratitude to my advisor Prof. Lyles for the continuous support of my research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. In addition, I would also thank my reader Dr. Glen Satten for providing such a great new method for me to proceed, and for this great proposed topic which gave me the incentive me to widen my outlook from various perspectives.

Table of Contents

1. Introduction

1.1 Overview of Mediation analysis

Mediation analysis refers to a set of statistical techniques designed to explore how much of the relationship between a dependent variable and one or more independent variables of primary interest is accounted for one or more intermediate (or mediating) variables. The approach is prominent in the study of epidemiology, social science and psychology, and fits within the broader realm of causal inference. In mediation analysis, the overall effect of exposure is broken down into the effect explained by mediators and the effect not explained by the same mediators. These two effects are also commonly referred to as the indirect and direct effects, respectively. Rather than considering the simple association between two variables X and Y, mediation analysis adds a third variable, M, into the mix. This third variable is distinguished from a confounder, due to the assumption that X causes the mediator, M, and that M subsequently affects Y. Hence the notion that the effect of X on Y operates through the mediating variable M.

In a simple mediation model as shown in Figure1.1, it is hypothesized that all variables are causally related, that is, X (the explanatory variable) causes changes in M (the mediator), and the effect of M on Y is then viewed as the indirect effect of X on Y mediated through M.[1,2] Figure 1.1 represents a simple example of a directed acyclic graph (DAG).

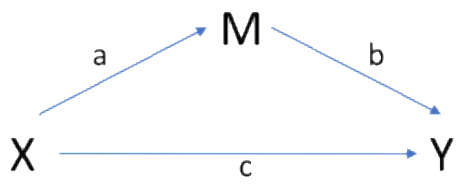


Figure 1.1 Mediation model. X is the explanatory variable(exposure), Y is the outcome, and M is a mediator. The causal change in M induced by X is represented by a, the causal change in Y induced by M is represented as b, and the direct effect of X on Y is represented as c.

1.2 Structural Equation Modeling

Traditional methods for mediation analysis were suggested by Baron and Kenny (1986), in a framework also known as structural equation modeling (SEM). This approach consists of three steps of regression models: X→Y, X→M, and X+M→Y. In the first step, Y is predicted by X, and a significant relationship is supposed. However, in modern mediation analysis, we can still move forward to the next step even if the crude association between X and Y not statistically significant. The second step is to predict M from X and check the significance of the path a. The final step is to regress X and M on Y simultaneously. In this method setup, the coefficient of M will be statistically significant but the coefficient of X will lack significance.[3] This is also a traditional method for testing for mediation. Generally, if there are significant relationships in the first and second steps, we consider full mediation to be present if X is not significant in final step when adjusting for M; if X remains significant at the final step, we will consider M to be a partial mediator.

We note, however, that Baron and Kenny's approach is not able to able to test the significance of the indirect effect, which is the pathway of X affecting Y through the mediator M

as represented by a and b in Figure 1. Another problem is that this traditional method may miss some true mediation effects due to type II errors. To solve this problem, an alternative approach is to estimate the significance of indirect effect, where the coefficient for indirect effect can be interpreted as the change in Y that is due to the change of X mediated by M. Two of the most commonly used approaches are proposed by Judd and Kenny (1981) and Sobel (1982). One involves computing the difference between two regression coefficients and the other multiplying two coefficients. In the first approach, the two basic models are shown below (in the case of linear regression):

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \epsilon$$

$$Y = \beta_0 + \beta X + \epsilon$$

Then the indirect effect can be calculated as $\beta - \beta_1$.[4] These two coefficients are both interpreted as representing an effect of X on Y, but β is the one from simple linear regression and β_1 is the partial regression coefficient. The second approach entails multiplying two coefficients from the models below:

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \epsilon$$

$$M = \beta_0 + \beta X + \epsilon$$

In this case, the indirect effect will by computed as $\beta_2 \beta$. Note that in this approach, unlike the first, the model involving the relationship between X and M is applied.[5]

After estimating the indirect effect, there are several ways to calculate the standard error which is used in test statistics or examine the existence of mediation in models. The Sobel test, Bootstrap and Monte Carlo method are the most widely used ones.[1] In Sobel Test, which is derived from delta method, the approximate estimate of the standard error of indirect effect

can be calculated by: $\beta_2^2 s_\beta^2 + \beta^2 s_{\beta_2}^2$, where β and β_2 are the coefficients in above equation, and s_β s_{β_2} are their standard errors.[11] . The Monte Carlo method can be implemented by using the parameter estimates and their associated variance and covariance to randomly draw from the joint distribution of β and β_2. After a large amount of random draws, the product of β and β_2, $\beta * \beta_2$ is used to estimate the confidence interval. If 0 falls outside of the interval, the hypothesis of no mediation will be rejected.[14]

Since mediation hypothesizes the potential causality and ordering among variables, some variables can either be a cause or an effect. Based on the linear regression model structure outlined above, SEM provides a framework for analysis and interpretation that also has the benefit of accounting for latent variables. This method can also be extended to deal with multiple independent variables and mediators as well, and the SEM modeling structure facilitates hypothesis testing.[6]

If one or both of the mediator and the outcome are binary variables, standard SEM methods are no longer available and logistic regression is a better choice. Nevertheless, traditional Baron and Kenny steps can still be used to conduct mediation testing and estimate indirect effects. As for the proper test statistics to use for indirect effects, there has been some controversy. The Sobel test is no longer available, given the failure of a necessary independence assumption. If the Y and M are binary, the coefficients in the mediation model will be on difference scales.[12]

1.3 Causal Inference Considerations

An alternative method to SEM is a formal causal inference approach, based upon the same causal structure that underlies SEM. This approach conceptualizes counterfactual and potential outcomes, which define the mechanistic process by which an exposure may causally affect the dependent variable conditional on mediators. This causal mechanism can be decomposed into direct and indirect effects as well.[7] In this framework, the following important assumptions are needed regarding confounders: 1) No unmeasurable confounders of the association between X and Y; 2) No unmeasurable confounders of the association between M and Y; 3) No unmeasurable confounders of the association between X and M. In the causal inference framework, all effects are defined using counterfactuals. Specifically, the potential outcome for exposure level x can be denoted as Y(x), for x=0,1. Then the effect of X on Y is defined as E[Y(1)]-E[Y(0)], which is the difference between the expectation of Y when X is 1 and X equals 0. Under the appropriate assumptions, this difference can be estimable even though only one of the potential outcomes can ever be observed for a given subject.

In the causal inference approach, the total effect can be decomposed to the sum of what are known as the natural indirect effect (NIE) and the natural direct effect (NDE). Here, NIE is defined as

$$NIE = E[Y(1, M_1)] - E[Y(1, M_0)]$$

where M_0 is the counterfactual value of the mediator M when X is equal to 0, and M_1 is the counterfactual value of M when X is equal to 1. And NDE is defined as:

$$NDE = E[Y(1, M_0)] - E[Y(0, M_0)]$$

Then the total effect is the sum of NIE and NDE which can be written as:

$$Total\ effect = NIE + NDE = E[Y(1, M_1)] - E[Y(0, M_0)] = E[Y(1)] - E[Y(0)]$$

For a fixed value of the mediator M, one can also define the controlled direct effect (CDE) as follows:

$$CDE(M) = E[Y(1, M)] - E[Y(0, M)]$$

Thus, the value of CDE can change for different values of M.

Valeri and VanderWeele[8] focus attention on allowing for a possible M by X interaction in regression models for an outcome Y versus exposure X and mediator M. In the case of linear regressions, the models can be written as:[8]

$$E[M|X] = \beta_m + aX$$

$$E[Y|X, M] = \beta_y + bM + cX + dXM$$

For a change in exposure from level x* to level x, the NDE and NIE will become:

$$NDE = [c + d(\beta_m + ax*)](x - x*)$$

$$NIE = a(b + dx)(x - x*)$$

and CDE can be expressed as:

$$CDE = (c + dm)(x - x*)$$

Similar counterfactual arguments are also available for binary variables (Y and M), and decomposition of the total effect of X can again be accomplished. Note that the above expressions readily simplify in the event that no M by X interaction is assumed in the model for the outcome Y. In the latter case, the CDE and the NDE definitions become equivalent.

In the case-control study, Y is a binary variable, so E[Y] can be replaced by P[Y=1]. VanderWeele and colleagues have proposed a general approach to fit the full model for E[Y|X,M,C] and mediator model for E[M|X,C], where C represents one or more additional

covariates utilized in the modeling process. Upon specifying these two models, one can define a variety of causal effects as simple functions of the parameters. However estimation of E[M|X,C] is a bit difficult for case-control data, although under the rare disease assumption it can be closely approximated by E[M|X,C,Y=0]. VanderWeele recommends fitting this model using data from controls only, in order to obtain estimates of the parameters of the E(M|X,C) model.

In this thesis, a formalism rooted in theory proposed by Satten and Kupper[9] and Satten and Carroll[10] is developed in order to allow simultaneous fitting of the response model and the mediator model to data from both cases and controls.

2. Methods

2.1 Estimation of Coefficients by Maximum Likelihood (the Satten Method)

For binary variables, A and Y, the mediation model is illustrated as shown in Figure2.1.

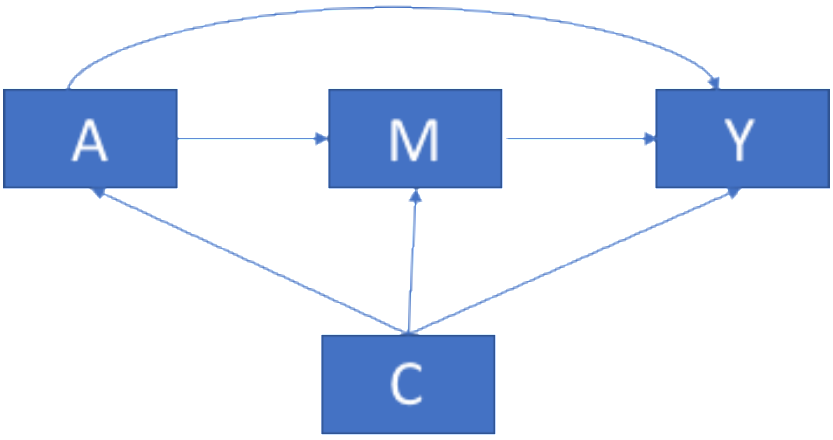


Figure2.1 Mediation model. A is exposure, Y is outcome, M is mediator, C's are confounders.

Consider two logistic regression models and a single confounder C, as follows:

$$logit[\Pr(Y = 1|A = a, M = m, C = c)] = \beta_0 + \beta_a a + \beta_m m + \beta_{am} am + \beta_c c \qquad \text{(Model 1)}$$

and

$$logit[\Pr(M = 1|A = a, C = c) = \gamma_m + \gamma_a a + \gamma_c c \qquad \text{(Model 2)}$$

Now, note that Model 1 is equivalent to

$$\theta(a, m, c) = \frac{\Pr(Y = 1|A = a, M = m, C = c)}{\Pr(Y = 0|A = a, M = m, C = c)} = e^{\beta_0 + \beta_a a + \beta_m m + \beta_{am} am + \beta_c c}$$

As before, we will refer to Model 1 as the response model, while Model 2 is known as the mediator model.

If the outcome is rare, then Model 2 holds at least approximately among the controls so that:

$$logit[\Pr(m = 1|A = a, C = c, Y = 0)] \cong \gamma_m + \gamma_a a + \gamma_c c$$

Following the Valeri and VanderWeele method[8], the odds ratios for NDE, CDE and NIE for exposure moving from a*=0 to a=1 can be defined as:

$$OR^{CDE} = e^{\beta_a + \beta_{am} m}$$

$$OR^{NDE} \cong e^{\beta_a} \frac{1 + e^{\beta_m + \beta_{am} + \gamma_m + \gamma_c c}}{1 + e^{\beta_m + \gamma_m + \gamma_c c}}$$

$$OR^{NIE} \cong \frac{[1 + e^{\gamma_m + \gamma_c c}][1 + e^{\beta_m + \beta_{am} + \gamma_m + \gamma_a + \gamma_c c}]}{[1 + e^{\gamma_m + \gamma_a + \gamma_c c}][1 + e^{\beta_m + \beta_{am} + \gamma_m + \gamma_c c}]}$$

Note that these definitions simplify easily by setting β_{am} to 0 if the interaction between A and M is assumed not to be present.

Under the assumption of rare disease and extending for the moment to the case of a polytomous mediator M, the estimation of E[M|A,C] is approximately equal to E[M|A,C,Y=0]. The mediator model can then be rewritten as:

$$\log\left(\frac{\Pr(M=m|A=a,C=c,Y=0)}{\Pr(M=1|A=a,C=c,Y=0)}\right)=\gamma_m+\gamma_a a+\gamma_c c,\qquad m>1$$

Then we have:

$$Pr(M=m|A=a,C=c,Y=0)=\frac{e^{\gamma_m+\gamma_a a+\gamma_c c}}{1+\sum_{m'}e^{\gamma_{m'}+\gamma_a a+\gamma_c c}}$$

Now following arguments analogous to those in Satten and Kupper[9] and Satten and Carroll[10], the distribution of the mediator among cases can be written as:

$$P(M=m|A=a,C=c,Y=1)=\frac{\theta(a,m,c)Pr(M=m|A=a,C=c,Y=0)}{\sum_{m'}\theta(a,m',c)Pr[M=m'|A=a,C=c,Y=0]}\qquad \text{Model 3}$$

so that

$$\Pr(M=m|A=a,C=c,Y=1)=\frac{e^{\gamma_m+\gamma_a a+\gamma_c c+\beta_0+\beta_a a+\beta_m m+\beta_{am}am+\beta_c c}}{1+\sum_{m'}e^{\gamma_m+\gamma_a a+\gamma_c c+\beta_0+\beta_a a+\beta_m m+\beta_{am}am+\beta_c c}}$$

Next we also have[9,10]:

$$\theta(a,c)=\frac{\Pr(Y=1|A=a,C=c)}{\Pr(Y=0|A=a,C=c)}=\int\theta(a,m,c)d\Pr(M=m|A=a,C=c,Y=0)\qquad \text{Model 4}$$

Given the marginal odds $\theta(a,c)$, the likelihood $\Pr[Y=y|A=a,C=c]$ will be

$$P[Y=y|A=a,C=c]=\frac{\theta(a,c)^y}{1+\theta(a,c)}$$

As a result, the case-control likelihood can be written as

$$P[Y,M|A,C]=P[M|A,C,Y]P[Y|A,C]$$

Binary mediator case:

If M is a binary variable, Model 3 can be written as:

$$\Pr(M = 1|A = a, C = c, Y = 1)$$

$$= \frac{\theta(a,1,c)Pr(M = 1|A = a, C = c, Y = 0)}{\theta(a,1,c)Pr(M = 1|A = a, C = c, Y = 0) + \theta(a,0,c)Pr(M = 0|A = a, C = c, Y = 0)}$$

Similarly, Model 4 can be rewritten as:

$$\theta(a,c) = \frac{e^{\beta_0+\beta_a a+\beta_c c} + e^{\gamma_m+\gamma_a a+\gamma_c c+\beta_0+\beta_a a+\beta_m m+\beta_{am} am+\beta_c c}}{1 + e^{\gamma_m+\gamma_a a+\gamma_c c}}$$

Since both Y and M are binary variables, we have 4 types of observations based on y and m. The likelihood construction for each type of observation is:

1) $Y = 1, M = 1$: $\Pr(M = 1|A = a, C = c, Y = 1) \times \Pr(Y = 1|A = a, C = c) = P_1$

2) $Y = 1, M = 0$: $\Pr(M = 0|A = a, C = c, Y = 1) \times \Pr(Y = 1|A = a, C = c) = P_2$

3) $Y = 0, M = 1$: $\Pr(M = 1|A = a, C = c, Y = 0) \times \Pr(Y = 0|A = a, C = c) = P_3$

4) $Y = 0, M = 0$: $\Pr(M = 0|A = a, C = c, Y = 0) \times \Pr(Y = 0|A = a, C = c) = P_4$

Then the likelihood function can be written as:

$$L = \prod_{i=1}^{n} \{{P_{1i}}^{y_i m_i} \times {P_{2i}}^{y_i(1-m_i)} \times {P_{3i}}^{(1-y_i)m_i} \times {P_{4i}}^{(1-y_i)(1-m_i)}\}$$

The estimates of the $\beta's$ and $\gamma's$ can be obtained by numerically maximizing this likelihood function.

2.2 Variance of causal effect estimates based on the delta method

The variance of the log transformed causal effects including OR_{cde}, OR_{NDE} and OR_{NIE} can be estimated by using the multivariate delta method based on the estimated parameters from Satten's method. These log transformed causal effects can be expressed as:

$$\log(OR_{CDE}) = g_1(\beta_0, \beta_a, \beta_m, \beta_{am}, \beta_c, \gamma_m, \gamma_a, \gamma_c) = \beta_a + \beta_{am}m$$

$$\log(OR_{NDE}) = g_2(\beta_0, \beta_a, \beta_m, \beta_{am}, \beta_c, \gamma_m, \gamma_a, \gamma_c)$$
$$= \beta_a + \log\left[1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}\right] - \log[1 + e^{\beta_m+\gamma_m+\gamma_c c}]$$

$$\log(OR_{NIE}) = g_3(\beta_0, \beta_a, \beta_m, \beta_{am}, \beta_c, \gamma_m, \gamma_a, \gamma_c)$$
$$= \log[1 + e^{\gamma_m+\gamma_c c}] + \log\left[1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}\right] - \log[1 + e^{\gamma_m+\gamma_a+\gamma_c c}]$$
$$- \log\left[1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}\right]$$

The partial derivatives of the log odds ratio are:

$$D'_{CDE} = \left(\frac{\partial g_1}{\partial \beta_0}, \frac{\partial g_1}{\partial \beta_a}, \frac{\partial g_1}{\partial \beta_m}, \frac{\partial g_1}{\partial \beta_{am}}, \frac{\partial g_1}{\partial \beta_c}, \frac{\partial g_1}{\partial \gamma_m}, \frac{\partial g_1}{\partial \gamma_a}, \frac{\partial g_1}{\partial \gamma_c}\right) = (0,1,0,m,0,0,0,0,0) \quad (m = 0,1)$$

$$D'_{NDE} = \left(\frac{\partial g_2}{\partial \beta_0}, \frac{\partial g_2}{\partial \beta_a}, \frac{\partial g_2}{\partial \beta_m}, \frac{\partial g_2}{\partial \beta_{am}}, \frac{\partial g_2}{\partial \beta_c}, \frac{\partial g_2}{\partial \gamma_m}, \frac{\partial g_2}{\partial \gamma_a}, \frac{\partial g_2}{\partial \gamma_c}\right)$$
$$= (0, 1, \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}$$
$$- \frac{e^{\beta_m+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\gamma_m+\gamma_c c}}, \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}, 0, \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}$$
$$- \frac{e^{\beta_m+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\gamma_m+\gamma_c c}}, 0, \frac{c e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}} - \frac{c e^{\beta_m+\gamma_m+\gamma_c c}}{1 + e^{\beta_m+\gamma_m+\gamma_c c}})$$

$$D'_{NIE} = \left(\frac{\partial g_3}{\partial \beta_0}, \frac{\partial g_3}{\partial \beta_a}, \frac{\partial g_3}{\partial \beta_m}, \frac{\partial g_3}{\partial \beta_{am}}, \frac{\partial g_3}{\partial \beta_c}, \frac{\partial g_3}{\partial \gamma_m}, \frac{\partial g_3}{\partial \gamma_a}, \frac{\partial g_3}{\partial \gamma_c}\right))$$

$$= (0, 0, \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}} - \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}},$$

$$\frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}} - \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}},$$

$$0, \frac{e^{\gamma_m+\gamma_c c}}{1+e^{\gamma_m+\gamma_c c}} + \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}} - \frac{e^{\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\gamma_m+\gamma_a+\gamma_c c}}$$

$$- \frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}},$$

$$\frac{e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}} - \frac{e^{\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\gamma_m+\gamma_a+\gamma_c c}}, \frac{ce^{\gamma_m+\gamma_c c}}{1+e^{\gamma_m+\gamma_c c}}$$

$$+ \frac{ce^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_a+\gamma_c c}} - \frac{ce^{\gamma_m+\gamma_a+\gamma_c c}}{1+e^{\gamma_m+\gamma_a+\gamma_c c}}$$

$$- \frac{ce^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}}{1+e^{\beta_m+\beta_{am}+\gamma_m+\gamma_c c}})$$

By using the above derivatives, the delta method-based approximated variance can be calculated by $Var(\log(OR)) = \boldsymbol{D'VD}$, where the elements of each **D** vector are replaced by their MLEs and **V** is the variance-covariance matrix of the vector of estimated coefficients based on maximizing the likelihood function.

2.3 Simulation Studies

All simulation processes were conducted in SAS 9[13] with the initial settings of:

$$\beta_0 = -4.25, \beta_a = 1, \beta_m = -0.5, \beta_{am} = 0.7, \beta_c = 0.5$$

where , $\beta_{am} = 0$ when excluding interaction between exposure and mediator, and

$$\gamma_m = 0.5, \gamma_a = 0.25, \gamma_c = -0.25$$

Cross-sectional data with total sample sizes of 50,000 and 2,000,000 were generated, followed by elimination of large numbers of controls at random in order to mimic case oversampling. The SAS NLMIXED procedure[13] was used for ML estimation, utilizing the 'general' log-likelihood optimization facility after specifying the likelihood function as outlined on pg. 10.

The joint ML procedure yields estimates of the coefficients in both the outcome and the mediator model, and we obtained MLEs of the causal effect measures as the appropriate functions of the coefficient estimates. The simulation data was generated via the following schemes:

C follows the normal distribution with mean=0, sd=1;

If C is greater than 0, A follows the Bernoulli distribution with p=0.2;

If C is smaller than 0 , A follows the Bernoulli distribution with p=0.4;

The probability p_m of M is calculated from the logistic regression model with the above coefficients;

M follows the Bernoulli distribution with probability p_m;

The probability p_y of Y can be calculated from the logistic regression model with the above coefficient as well;

Y follows the Bernoulli distribution with probability p_y;

To mimic oversampling of cases, all cases but only 2% of the non-cases in each original large cross-sectional sample were kept.

Bootstrap simulations:

As an alternative to the delta method, we also conducted bootstrapping[15] as an approach to estimate standard errors and confidence intervals for the causal effects of interest. The generated data were separated into 2 groups, the case group (Y=1) and control group (Y=0). For each group, the data were bootstrapped with replacement and the same size as in the original data set for that group. After bootstrapping, separately, we combined the bootstrapped case group and control group data sets to form a complete bootstrap sample. We estimate the causal effects from each bootstrap sample and repeated this procedure 50 times for each simulated dataset (bootstrap sample size=50). The standard errors of the original estimated causal effects were then derived as the empirical standard deviation of the estimated effect across the bootstrap samples.

3. Results

Sample size = 50,000 without interaction term:

To begin, the interaction between exposure and mediator was excluded from the outcome model. The mean estimated coefficients from both the VanderWeele et al. and the Satten method are listed in Table3.1. The estimated coefficients from both methods are similar to each other, and the empirical standard deviations of the estimated $\beta's$ are also close to each other. However, we note that the standard deviations of the estimated mediator model (γ) coefficients are much smaller based on the Satten method. This implies that the new method

has the potential to more precisely estimate the coefficients in the mediator model than traditional method, and has similar precision in estimating coefficients in overall model.

Table3.1 Estimated coefficients from both VanderWeele's method and Satten's method with sample size = 50,000 and number of simulations=1000 (excluding interaction term)

Method		VanderWeele's method		Satten's method	
Coefficient	True value	Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation
β_0	-4.35*	-0.338*	0.070	-0.338	0.070
β_a	1	0.999	0.101	0.999	0.101
β_m	-0.5	-0.504	0.104	-0.504	0.104
β_c	0.5	0.503	0.054	0.503	0.054
γ_m	-0.5	-0.494	0.081	-0.494	0.074
γ_a	-0.25	-0.242	0.158	-0.239	0.105
γ_c	0.5	0.507	0.074	0.506	0.054

*estimated β_0 does not matched the true β_0 because rejection smapling was used.

The relevant causal effects including direct and indirect effects were calculated from the coefficients estimated by the two methods, and the log transformation was applied to compare the empirical standard deviations of the Satten estimates with the estimated standard errors derived from the Delta method. Results are summarized in Table 3.2. Since C was simulated to follow a normal distribution with mean 0, odds ratios were evaluated with c equal to 0 as well as with c equal to its 25th and 75th percentile. Note that without interaction, the controlled

direct effects and natural direct effects are the same by definition and they do not involve the parameters of the mediation model (see expressions on pg. 8, with β_{am} set to 0). For these measures, there is no noticeable change in precision due to using full ML via the Satten's approach. But note that the natural indirect effect involves one or more parameters from the mediation model. The Satten's method yields better precision for estimating the NIE effect, due to the improvement in precision with regard to the mediation coefficients offered by the joint ML approach. That is, more efficient use of case data in estimating the mediation model parameters has yielded better precision in the NIE effect estimate. Delta method outputs the similar standard error estimations for CDE and NDE effects, and slightly different estimation for NIE effects which is still close to those from Satten's method.

The 95% confidence intervals for all causal effects based on the Satten approach were calculated by using the standard error from Delta method. We summarize the percentage coverage of these confidence intervals in Table3.3. In this table, all causal effects have almost 95% coverage rate, except the NIE effect for the 75th percentile of C, which has a somewhat conservative coverage rate. This reveals that under this no interaction condition, the standard errors estimated from the Delta method relatively reliable and associated with reasonable coverage.

The histograms in Figure3.1 show the distribution of the estimated causal effects with log transformation. Note that all display a roughly normal distribution, which is consistent with the confidence interval coverage results.

Table3.2 Estimation of causal effects based on VanderWeele's method and Satten's method. Mean estimated standard errors from the Delta Method are presented to compare with empirical SDs. Sample size = 50,000 and number of simulations=1000 (excluding interaction term).

Method		VanderWeele's method		Satten's method		Delta Method
Causal effect	True causal effect	Estimate	Standard Deviation	Estimate	Standard Deviation	Mean Standard Error
$\log(OR^{CDE})$	1.000	0.999	0.101	0.999	0.101	0.101
$\log(OR^{NDE})$	1.000	0.999	0.101	0.999	0.101	0.101
$\log(OR^{NIE}_{median})$	0.026	0.025	0.017	0.024	0.011	0.012
$\log(OR^{NIE}_{25th})$	0.022	0.021	0.014	0.021	0.010	0.012
$\log(OR^{NIE}_{75th})$	0.029	0.028	0.019	0.027	0.013	0.012

Table3.3 95% Confidence Interval coverage of causal effects based on Satten approach using the standard error from Delta method with sample size = 50,000 and number of simulations=1000 (excluding interaction term).

Causal effect	$\log(OR^{CDE})$	$\log(OR^{NDE})$	$\log(OR^{NIE}_{median})$	$\log(OR^{NIE}_{25th})$	$\log(OR^{NIE}_{75th})$
Percentage %	0.95	0.95	0.95	0.952	0.981

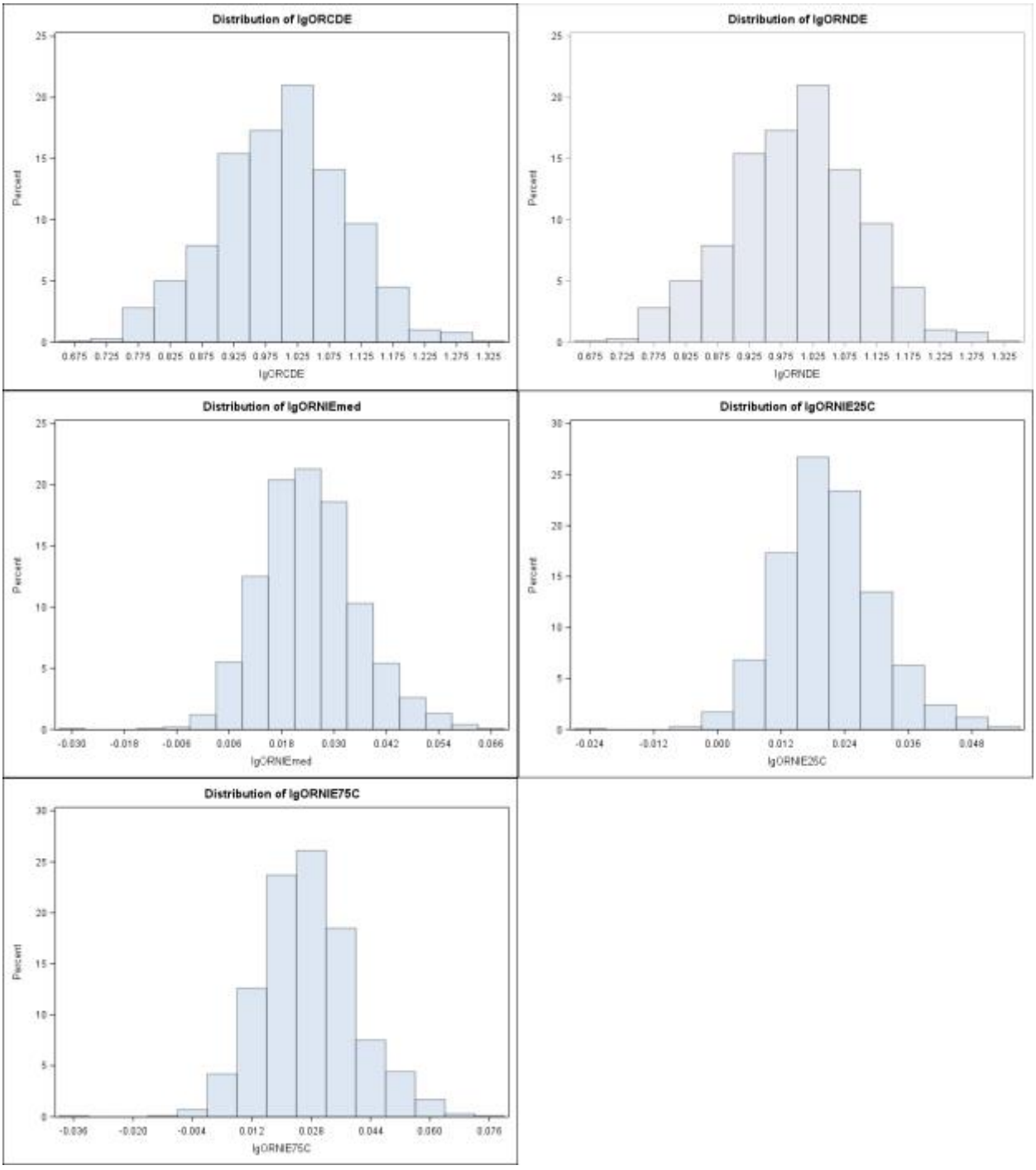


Figure3.1 The histogram of log transformed causal effects with sample size = 50,000 and number of simulations=1000 (excluding interaction term).

Sample size = 50,000 with interaction term:

To consider the case where there is interaction between exposure and mediator in the overall model, the predictor β_i was included in the model for further simulations. Again, the mean estimated coefficients and empirical standard deviations from the two methods are almost the same for the $\beta's$. However, although both methods lead to similar estimated coefficients, the Satten method can produce smaller standard deviations particularly for the mediator model coefficient estimates. This reflects the same trend that we observed in the model without interaction.

Table3.4 The estimated coefficients from VanderWeele's method and Satten's method with sample size = 50,000 and number of simulations =1000

Method		VanderWeele's method		Satten's method	
Coefficient	True value	Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation
β_0	1.000*	-0.342*	0.076	-0.342	0.075
β_a	0.500	1.008	0.123	1.009	0.122
β_m	-0.500	-0.504	0.132	-0.502	0.130
β_{am}	0.700	0.708	0.209	0.704	0.204
β_c	0.500	0.503	0.054	0.503	0.054
γ_m	-0.500	-0.496	0.081	-0.495	0.079
γ_a	-0.250	-0.263	0.166	-0.264	0.163
γ_c	0.500	0.506	0.073	0.502	0.050

*estimated β_0 does not matched the true β_0 because rejection smapling was used.

For causal effects under this condition, natural direct effects are no longer the same as controlled direct effects. However, the precision for estimating these two effects appears to remain about the same for both methods. Whereas, for natural indirect effects, Satten's method only produces a small decrease in standard deviations, with both methods producing almost the same estimated effects. Nevertheless, the Satten method is still able to improve the precision of estimating indirect natural effects at least slightly, which is consistent with our observations in the model without interaction. Under this condition, the Delta method estimates of the standard errors are noticeably different on average from the empirical SDs for either method. This indicates that the standard errors might be a biased estimate based on the Delta method. The summary of causal effects estimates is in Table3.5.

From the confidence interval coverage output which is calculated by using standard errors from the Delta method (Table 3.6), we find that all direct effects have approximately 95% coverage rates. However, for NIE effects, the coverage rates are noticeably smaller than the expected 95 percent, which likely reflects the bias in the standard error estimates of indirect effects from the Delta Method in the model with interaction.

By plotting the distributions shown in Figure3.2 of all estimated causal effects, the reason for this low coverage rate can also be explained. The distributions of all estimated direct effects follow approximate normal distributions, whereas those of the indirect effects are largely left-skewed. This means the log transformation of indirect effects cannot normalize the distribution well, and due to this, the Delta method cannot produce unbiased standard errors. We attempted other transformations (including the identity), but found that the problem of left skewness was persistent.

Table3.5 Estimation of causal effects from VanderWeele's method and Satten's method. Mean estimated standard errors from the Delta Method are presented to compare with empirical SDs. Sample size = 50,000 and number of simulations=1000 (including interaction term).

Method	True causal effect	VanderWeele's method		Satten's method		Delta Method
Causal effect		Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation	Mean Estimtated Error
$\log(OR^{CDE})$	1.000	1.008	0.123	1.009	0.122	0.120
$\log(OR^{NDE}_{median})$	1.241	1.256	0.102	1.256	0.102	0.100
$\log(OR^{NDE}_{25th})$	1.191	1.205	0.102	1.205	0.102	0.101
$\log(OR^{NDE}_{75th})$	1.296	1.312	0.106	1.311	0.106	0.100
$\log(OR^{NIE}_{median})$	-0.012	-0.017	0.020	-0.017	0.019	0.017
$\log(OR^{NIE}_{25th})$	-0.010	-0.015	0.018	-0.015	0.017	0.017
$\log(OR^{NIE}_{75th})$	-0.012	-0.018	0.021	-0.018	0.020	0.017

Table3.6 95% Confidence Interval coverage of causal effects confirmation by using the standard error from Delta method with sample size = 50,000 and number of simulations=1000

Causal effect	$\log(OR^{CDE})$	$\log(OR^{NDE}_{median})$	$\log(OR^{NDE}_{25th})$	$\log(OR^{NIE}_{median})$	$\log(OR^{NIE}_{25th})$	$\log(OR^{NIE}_{75th})$
Percentage %	94.9	94.7	95.3	93.8	83.2	85.8

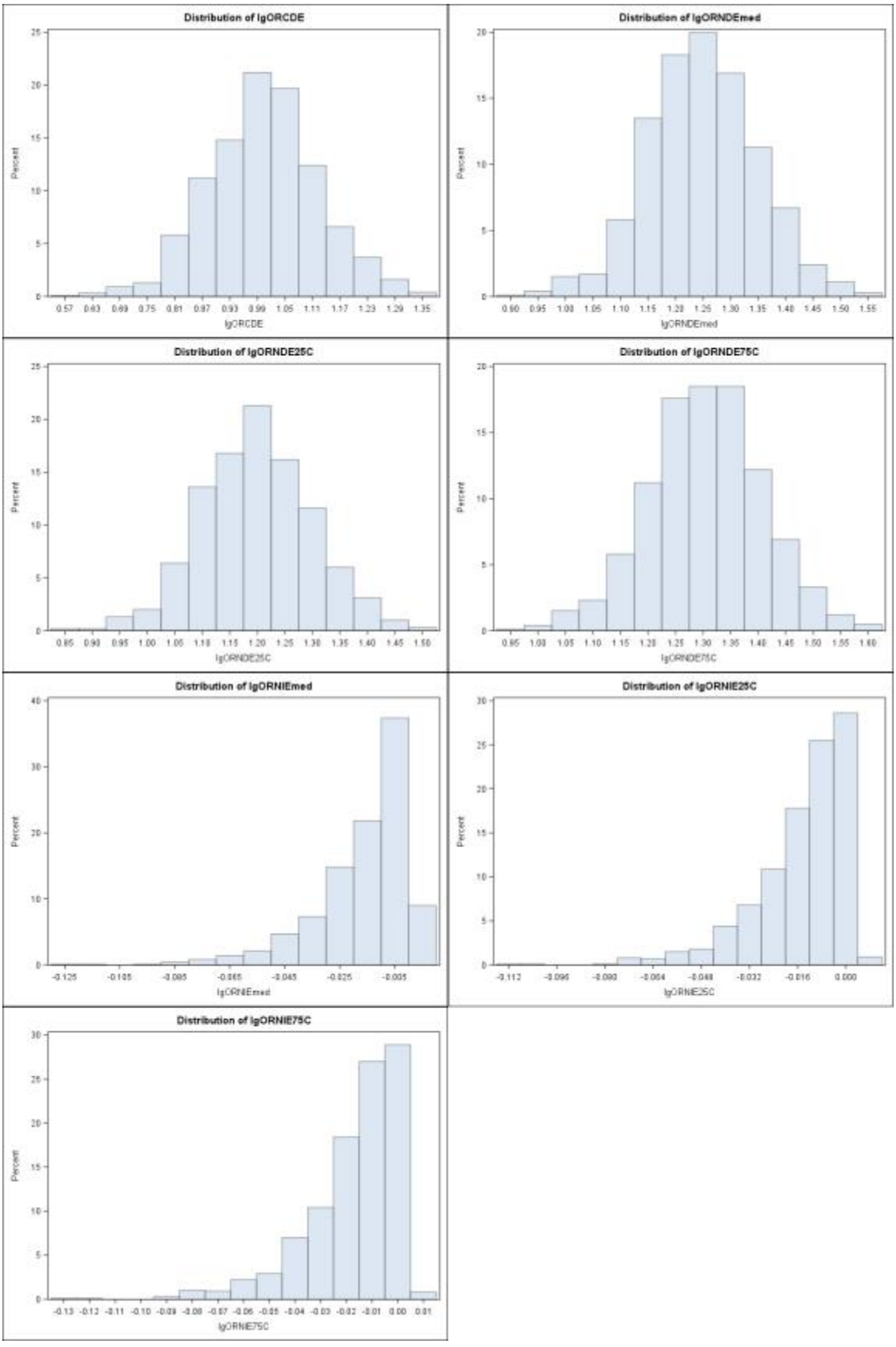

Distribution of lgORCDE
Percent
lgORCDE
Distribution of lgORNDEmed
Percent
lgORNDEmed
Distribution of lgORNDE25C
Percent
lgORNDE25C
Distribution of lgORNDE75C
Percent
lgORNDE75C
Distribution of lgORNIEmed
Percent
lgORNIEmed
Distribution of lgORNIE25C
Percent
lgORNIE25C
Distribution of lgORNIE75C
Percent
lgORNIE75C

Figure3.2 The histogram of log transformed causal effects with sample size = 50,000 and number of simulation=1000

Sample size = 2,000,000 with interaction term:

Since in the case with sample size of 50,000, the log transformed indirect causal effects displayed non-normal distributions, we performed additional simulations after increasing the sample size. After increasing the initial sample size to 2,000,000, the mean overall coefficient estimates do not change much from those with smaller sample size (Table 3.7). Notably, however, the Delta method produces mean standard errors for the indirect causal effect estimates much closer to the empirical SDs from the Satten and VanderWeele methods (Table3.8). For the standard errors of other causal effects, both methods have consistent results. The confidence interval coverage where all of the causal effects have a coverage rate close to 95% also reveals that under this condition, standard errors from the Delta method are not biased seriously (Table3.9). At the same time, the NIE effects with C in a 75^{th} percentile has a higher coverage rate which is similar to the trend in the model without interaction. From the overall distribution shown in Figure 3.3, larger sample size can make the estimated causal effects more closely follow a normal distribution, and this helps the Delta method have better performance on estimating causal effect standard errors. In this case, with a large enough initial sample, size, the precision and ability of these two methods to estimate the causal effects and to estimate coefficients in mediation model (with the exception of the coefficient for the covariate C) are similar

Table3.7 The estimated coefficients from both VanderWeele's method and the Satten method with sample size = 2,000,000 and number of simulations=1000

Method	True value	VanderWeele's method		Satten's method	
Coefficient		Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation
β_0	1.000*	-0.338*	0.012	-0.339	0.012
β_a	0.500	1.002	0.019	1.003	0.019
β_m	-0.500	-0.500	0.021	-0.498	0.021
β_{am}	0.700	0.697	0.033	0.694	0.032
β_c	0.500	0.500	0.008	0.499	0.008
γ_m	-0.500	-0.494	0.012	-0.494	0.012
γ_a	-0.250	-0.262	0.023	-0.263	0.023
γ_c	0.500	0.501	0.012	0.500	0.008

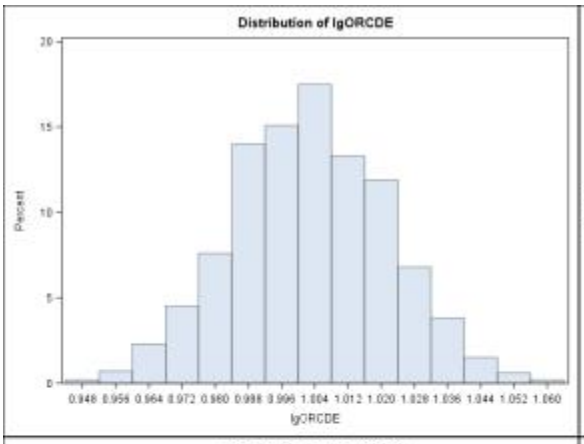
*estimated β_0 does not matched the true β_0 because rejection smapling was used.

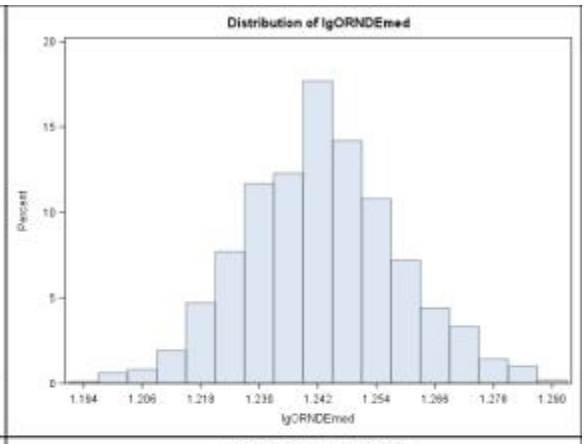
Table3.8 Estimation of causal effects based on VanderWeele and Satten method. Estimated standard errors from Delta Method with sample size = 2,000,000 and number of simulations=1000

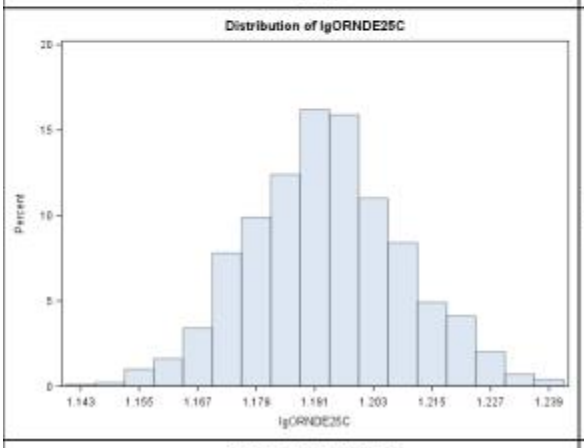
Method		VanderWeele's method		Satten's method		Delta Method
Causal effect	True causal effect	Mean Estimate	Standard Deviation	Mean estimate	Standard Deviation	Mean Standard Error
$\log(OR^{CDE})$	1.000	1.002	0.019	1.003	0.019	0.019
$\log(OR^{NDE}_{median})$	1.241	1.243	0.016	1.243	0.016	0.016
$\log(OR^{NDE}_{25th})$	1.191	1.193	0.016	1.193	0.016	0.016
$\log(OR^{NDE}_{75th})$	1.296	1.298	0.016	1.297	0.016	0.016
$\log(OR^{NIE}_{median})$	-0.012	-0.012	0.002	-0.012	0.002	0.002
$\log(OR^{NIE}_{25th})$	-0.010	-0.011	0.002	-0.011	0.002	0.002
$\log(OR^{NIE}_{75th})$	-0.012	-0.013	0.003	-0.013	0.003	0.002

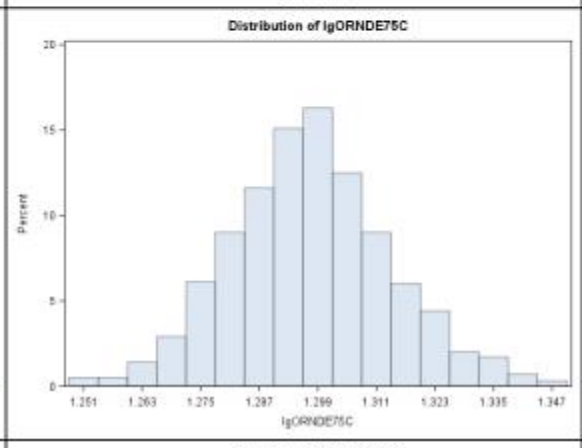
Table3.9 95% Confidence Interval coverage of causal effects confirmation by using the standard error from Delta method with sample size = 2,000,000 and number of simulations=1000

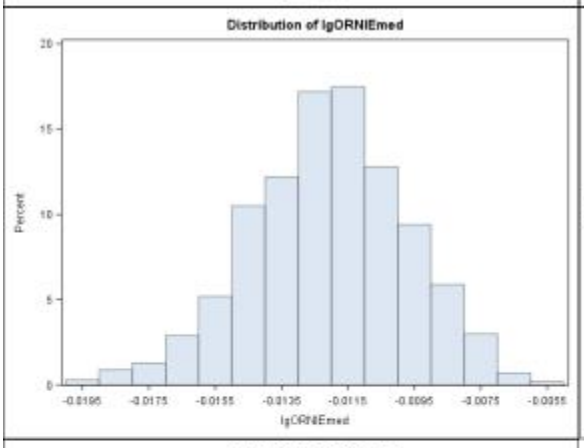
Causal effect	$\log(OR^{CDE})$	$\log(OR^{NDE}_{median})$	$\log(OR^{NDE}_{25th})$	$\log(OR^{NIE}_{median})$	$\log(OR^{NIE}_{25th})$	$\log(OR^{NIE}_{75th})$
Percentage %	95.1	93.6	94.4	93.5	95.3	97.5

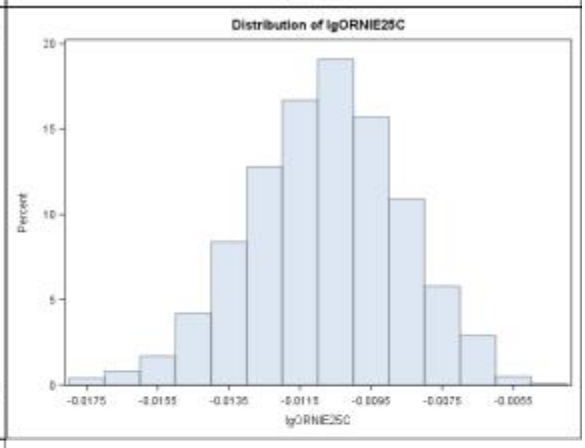

Distribution of lgORCDE
Percent
lgORCDE


Distribution of lgORNDEmed
Percent
lgORNDEmed


Distribution of lgORNDE25C
Percent
lgORNDE25C


Distribution of lgORNDE75C
Percent
lgORNDE75C


Distribution of lgORNIEmed
Percent
lgORNIEmed


Distribution of lgORNIE25C
Percent
lgORNIE25C

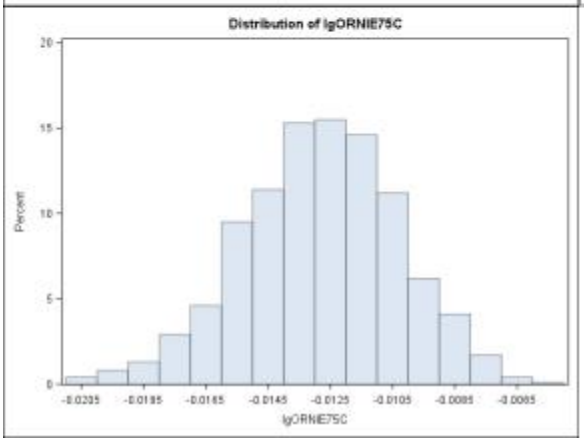

Distribution of lgORNIE75C
Percent
lgORNIE75C

Figure3.3 The histogram of log transformed causal effects with sample size = 2,000,000 and number of simulations=1000

Bootstrap Standard Error Estimation:

For estimation of standard errors, an alternative to the Delta method is the bootstrap. Table3.10 and Table3.11 summarize the results from bootstrapping separately among cases and controls (see Methods) and compares mean bootstrap standard errors vs. those based on the Delta method as well as the empirical SDs for the VanderWeele and Satten methods for estimating log transformed causal effects from the model containing interaction between exposure and mediator. The estimated standard errors of estimated natural direct effects from the bootstrap are larger on average, but in general the 95% confidence interval coverage rates are much better based on the bootstrap.

When eliminating the interaction between exposure and mediator, bootstrap outputs similar standard error estimations for CDE and NDE effects but larger estimations for NIE as shown in Table3.12 and Table3.13, which reflects similar trends as in the case of the model containing interaction. Also, the 95% confidence interval coverage rates remain favorable again when computing CIs based on bootstrap standard errors.

Table3.10 Estimation of causal effects (containing interaction between exposure and mediator) based on VanderWeele's and Satten's methods. Estimated standard errors from Delta Method and bootstrap with sample size = 50,000 and number of simulations=500, bootstrap sample size=50

Method	VanderWeele's method		Satten's method		Delta method	Bootstrap
Causal effects	Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation	Mean Standard Error	Mean Standard Error
$\log(OR^{CDE})$	1.006	0.117	1.007	0.116	0.120	0.119
$\log(OR^{NDE}_{median})$	1.248	0.097	1.247	0.097	0.100	0.100
$\log(OR^{NDE}_{25th})$	1.198	0.097	1.198	0.097	0.101	0.100
$\log(OR^{NDE}_{75th})$	1.302	0.101	1.302	0.101	0.100	0.103
$\log(OR^{NIE}_{median})$	-0.015	0.017	-0.015	0.016	0.016	0.018
$\log(OR^{NIE}_{25th})$	-0.014	0.015	-0.014	0.015	0.016	0.017
$\log(OR^{NIE}_{75th})$	-0.016	0.017	-0.016	0.017	0.016	0.019

Table3.11 95% Confidence Interval coverage of causal effects confirmation by using the standard error from bootstrap method with sample size = 50,000 and number of simulations=500, bootstrap sample size=50

Causal effect	$\log(OR^{CDE})$	$\log(OR^{NDE}_{median})$	$\log(OR^{NDE}_{25th})$	$\log(OR^{NIE}_{median})$	$\log(OR^{NIE}_{25th})$	$\log(OR^{NIE}_{75th})$
Percentage %	95.2	95.6	95.0	94.8	95.4	94.8

Table3.12 Estimation of causal effects (without interaction between exposure and mediator) based on VanderWeele's and Satten's methods. Estimated standard errors from Delta Method and bootstrap with sample size = 50,000 and number of simulations=500, bootstrap sample size=50

Method	VanderWeele's method		Satten's method		Delta method	Bootstrap
Causal effects	Mean Estimate	Standard Deviation	Mean Estimate	Standard Deviation	Mean Standard Error	Mean Standard Error
$\log(OR^{CDE})$	1.002	0.105	1.002	0.105	0.101	0.101
$\log(OR^{NDE})$	1.002	0.105	1.002	0.105	0.101	0.101
$\log(OR^{NIE}_{median})$	0.024	0.016	0.024	0.011	0.011	0.012
$\log(OR^{NIE}_{25th})$	0.021	0.014	0.020	0.009	0.011	0.010
$\log(OR^{NIE}_{75th})$	0.027	0.018	0.027	0.012	0.011	0.013

Table3.13 95% Confidence Interval coverage of causal effects confirmation by using the standard error from bootstrap method with sample size = 50,000 and number of simulations=500, bootstrap sample size=50

Causal effect	$\log(OR^{CDE})$	$\log(OR^{NDE}_{median})$	$\log(OR^{NIE}_{median})$	$\log(OR^{NIE}_{25th})$	$\log(OR^{NIE}_{75th})$
Percentage %	95.8	95.2	93.4	93.4	93.4

4. Conclusion

In this study, an alternative method to estimate causal effects in mediation analysis was conducted and compared with traditional mediation analysis method. Since this modeling

framework often allows for elimination of stratum-specific intercepts for matched data, it may also be possible to fit this model to matched data and still estimate the causal effects that VanderWeele has defined. In the case of smaller initial sample size without interaction between mediator and exposure, Satten's method can improve the precision in estimating the coefficients in mediator model and consequently in estimating natural indirect effects. In the model eliminating interaction, the Delta method can provide a good way to estimate the unbiased standard error since the estimated log transformed odds ratios follow approximately normal distribution for both direct effects and indirect effects. However, things appeared to change a lot if there exists interaction between exposure and mediators. In this case, since the distribution of log transformed indirect causal effect estimates do not follow normal distributions, the Delta method did not produce virtually unbiased standard errors anymore. The Satten method can still have better performance in estimating coefficients in the mediator model and in estimating indirect natural effects, in the sense that the empirical standard deviations were slightly decreased. Increasing the initial sample size can help to normalize the distribution of estimated causal effects. On the other hand, the efficiency advantages of Satten's method appeared muted under large sample sizes in the case where the response model included an interaction. Alternatively, the bootstrap can help to improve estimated standard errors of causal effects for Satten's method when the model contains the interaction term or not under smaller sample size. Further simulation studies under a wider variety of parameter and covariate settings will be needed to generalize these findings.

5. Discussion

In addition to the findings related to the methods for estimating causal effects in mediation analysis, this project revealed some limitations in the Delta method for standard error estimation when the distribution of the estimated function of parameters is far from normal When the overall sample size was relatively small, from the histogram of estimates based on the model containing an interaction term, the log transformed natural indirect effects was not approximately normally distributed but rather noticeably left skewed. In this case, the Delta method was not a good choice to estimate standard errors of some of the estimated causal effects. However, the bootstrap method used here is a nonparametric method that does not rely as heavily on assumptions of normal distributions. Instead, the assumption of this technique is that sampling from the empirical distribution of the data is comparable to sampling from the real distribution of the data.[16] In the model with the interaction term, the bootstrap showed a better coverage rate for the confidence interval of indirect natural effect, and can correct the biased estimator of standard error to asymptotic order. On the other hand, since bootstrapped standard error is an alternative of the asymptotic standard error, the assumption of asymptotically normal distribution of the indirect causal effects were needed, and this requirement is met when the sample size is large enough.[17] Therefore, if the sample size is not large enough, to correct the biased standard error estimators of causal effects, bootstrap performs better than the Delta method. When the sample size was larger (original N = 2,000,000 in our empirical studies), all log transformed causal effect estimates followed approximately normal distributions and the Delta method was effective in estimating the standard error.

In this study, the mediator was considered as a binary variable. However in future work, models with continuous mediators can also be built and studied. For a continuous mediator, the model with or without interaction can also studied separately to see the performance of the Satten's method for making more efficient use of case data.

References:

[1] MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation Analysis. Annual Review of Psychology, 58, 593. http://doi.org/10.1146/annurev.psych.58.110405.085542

[2] Richiardi, L., Bellocco, R., & Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. International journal of epidemiology, 42(5), 1511-1519.

[3] Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 5, 1173-1182.

[4] Judd, C.M. & Kenny, D.A. (1981). Process Analysis: Estimating mediation in treatment evaluations. Evaluation Review, 5(5), 602-619

[5] Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), Sociological Methodology 1982 (pp. 290-312).Washington DC: American Sociological Association

[6] Gunzler, D., Chen, T., Wu, P., & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. Shanghai Archives of Psychiatry, 25(6), 390–394.

[7] Ato García, M., Vallejo Seco, G., & Ato Lozano, E. (2014). Classical and causal inference approaches to statistical mediation analysis. Psicothema, 26 (2).

[8] Valeri, L., & VanderWeele, T.J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods, 18, 137-150.

[9] Satten G, Kupper L. 1993. Inferences about exposures-disease associations using probability-of-exposure information. JASA 88:200–208.Satten and Kupper (1993 JASA)

[10] Satten, G. A., & Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. Biometrics, 56(2), 384-388.

[11] Preacher, K. J., & Leonardelli, G. J. (2001). Calculation for the Sobel test. Retrieved January, 20, 2009.

[12] Herr, N. R. (2013). Mediation with dichotomous outcomes. accessed September, 12, 2014.

[13] Statistical Analysis System Institute. (2004). SAS/STAT user's guide, version 9 SAS Institute.

[14] Hayes, A. F. (2013, May). Multilevel mediation analysis. In Workshop Presented at the Association for Psychological Science (Vol. 23).

[15] Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

[16] Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. Sociological methodology, 115-140.

[17] Williams, J., & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. Structural Equation Modeling, 15(1), 23-51.