Statistical Methods for Causal Mediation Analysis

Statistical Methods for Causal Mediation Analysis

A dissertation presented

by

Linda Valeri

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 2012

Dissertation Advisor: Professor Tyler J. VanderWeele Linda Valeri

Statistical Methods for Causal Mediation Analysis

Abstract

Mediation analysis is a popular approach in the social an biomedical sciences to examine the extent to which the effect of an exposure on an outcome is through an intermediate variable (mediator) and the extent to which the effect is direct. We first develop statistical methods and software for the estimation of direct and indirect causal effects in generalized linear models when exposure-mediator interaction may be present. We then study the bias of direct and indirect effects estimators that arise in this context when a continuous mediator is measured with error or a binary mediator is misclassified. We develop methods of correction for measurement error and misclassification coupled with sensitivity analyses for which no auxiliary information on the mediator measured with error is needed. The proposed methods are applied to a lung cancer study to evaluate the effect of genetic variants mediated through smoking on lung cancer risk and to a perinatal epidemiological study on the determinants of preterm birth.

Contents

Title page i

Abstract iii

Table of Contents iv

List of Figures viii

List of Tables x

Acknowledgments xi

1 Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros 1

1.1 Introduction 3

1.2 Classic Regression Approach to Mediation Analysis 5

1.3 Counterfactual Approach to Mediation Analysis 7

1.4 Identification 9

1.5 Binary Outcome 12

1.6 Mediation analysis for models with non-linearities - a comparison of approaches 14

1.6.1 Traditional approaches to mediation analysis 15

1.6.2 Comparison of traditional approaches with the counterfactual approach when there are interactions and non-linearities 17

1.7 Description of the SAS macro 20

1.7.1 Basic SAS Macro 20

1.7.2 Other options in the SAS Macro 21

1.7.3 Comparison with other macros 24

1.8 Description of the SPSS macro . . . 27
1.9 Example . . . 29
1.10 Discussion . . . 31
1.A Definition of causal effects and Identifiability conditions . . . 33
1.B Continuous Mediator and Outcome . . . 35
1.C Continuous Mediator and Binary Outcome . . . 38
1.D Binary Mediator and Continuous Outcome . . . 43
1.E Binary Mediator and Binary Outcome . . . 48

2 Mediation analysis in generalized linear models when the mediator is measured with error 54

2.1 Introduction . . . 56
2.2 Mediation analysis within the counterfactual framework in the absence of measurement error . . . 57
2.3 Asymptotic bias of direct and indirect effects when the mediator is measured with error . . . 61
2.3.1 GLM with mis-measured mediator . . . 61
2.3.2 Asymptotic Limit of parameters of the outcome regression in the presence of exposure-mediator interaction . . . 62
2.3.3 Asymptotic bias of direct and indirect causal effects . . . 64
2.4 Correction strategy for direct and indirect effects estimators . . . 68
2.4.1 Method of moments estimators . . . 69
2.4.2 Regression calibration estimators . . . 71
2.4.3 SIMEX . . . 72
2.4.4 Simulations . . . 73
2.5 Example . . . 76
2.6 Discussion . . . 79
2.A Definition of causal effects and identifiability conditions . . . 81
2.B Continuous Mediator and Outcome . . . 82
2.C Continuous Mediator and Binary-Logistic, Binary-Log-linear, or Count Outcome . . . 95

3 The estimation of direct and indirect causal effects in the presence of misclassified binary mediator 115

3.1 Introduction . . . 117

3.2 Mediation analysis within the counterfactual framework in the absence of misclassification . . . 119

3.3 Results on direct and indirect effects naive estimators when the mediator is misclassified . . . 123

3.3.1 Mediator and outcome regressions when mediator is misclassified . 123

3.3.2 Asymptotic limit of parameters of the mediator regression . . . 124

3.3.3 Asymptotic limit of parameters of the outcome regression . . . 125

3.3.4 Asymptotic bias of the direct and indirect causal effects . . . 126

3.3.5 Additional results on the behavior of direct and indirect causal effects naive estimators in the absence of exposure-mediator interaction 128

3.3.6 Numerical bias analysis . . . 130

3.4 Correction strategy for direct and indirect effects estimators . . . 134

3.4.1 Iteratively Re-weighted Least Squares estimators for mediator regression . . . 134

3.4.2 Predictive Value Weighting estimators for outcome regression . . . 135

3.4.3 Likelihood-based approach for outcome and covariate misclassification . . . 136

3.4.4 Corrected estimators for direct and indirect effects . . . 138

3.4.5 Simulations . . . 138

3.5 Example . . . 141

3.5.1 Mother's age above 35, pre-eclampsia and preterm birth: background and data description . . . 141

3.5.2 Mother's age above 35, pre-eclampsia and preterm birth: naive analyses . . . 143

3.5.3 Mother's age above 35, pre-eclampsia and preterm birth: mediation analysis corrected for misclassification . . . 144

3.6 Discussion . . . 147

3.A Description of measurement error mechanism . . . 148

3.B Probability Limit of MLE of Continuous Outcome Regression . . . 150

3.C Asymptotic limit of naive outcome and mediator regression parameters in the absence of exposure-mediator interaction (θ^*,β^*) in terms of misclassification probabilities and true prevalence of the mediator 152

3.D Asymptotic bias of direct and indirect effects naive estimators 153

3.E Other theoretical results for direct effect naive estimators 153

3.F When the mediator is misclassified will the sum of the biased direct and indirect effects estimator still give an unbiased estimate of the total effect? . 154

3.G Standard errors of the PVW/IRLS estimators for direct and indirect causal effects 157

3.H Direction of asymptotic bias of naive direct and indirect effects estimators in the absence of exposure-mediator interaction 162

3.I Predictive Value Weighting approach with correctly specified model for $M^*|Y, A, C$ 165

3.J Maximum Likelihood (″Direct Method″) approach to misclassification correction in the outcome regression 166

References **168**

List of Figures

1.1 Mediation model in Baron and Kenny 1986 paper. 5

1.2 Mediation DAG 10

2.1 Numerical analysis of relative bias of direct (NDE) and indirect (NIE) effect naive estimators. Simulations run for continuous outcome modeled using linear regression and binary outcome modeled using logistic regression; exposure-mediator interaction both present or absent. Sample size $n = 10,000$. Measurement error variance, $\sigma_U^2 \in (0, 1)$, corresponding to a reliability ratio, $\lambda \in (0, 1)$. 67

2.2 Sensitivity analyses for direct (OR^{NDE}), indirect (OR^{NIE}), total (OR^{TE}) effects and proportion mediated ($PM = OR^{NDE} \times (OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1)$) for variant rs1051730. Absent to severe measurement error ($\sigma_U^2 \in (0, 2.5)$) which corresponds to a reliability ratio, $\lambda \in (0, 1)$. 78

2.3 Density of causal effect estimators: small error, linear Y model and $\theta_3 = 0$. . 112

2.4 Density of causal effect estimators: moderate error, linear Y model and $\theta_3 = 0$. 112

2.5 Density of causal effect estimators: small error, linear Y model and $\theta_3 \neq 0$. . 112

2.6 Density of causal effect estimators: moderate error, linear Y model and $\theta_3 \neq 0$. 113

2.7 Density of causal effect estimators: small error, logistic Y model and $\theta_3 = 0$. 113

2.8 Density of causal effect estimators: moderate error, logistic Y model and $\theta_3 = 0$. 113

2.9 Density of causal effect estimators: small error, logistic Y model and $\theta_3 \neq 0$. 114

2.10 Density of causal effect estimators: moderate error, logistic Y model and $\theta_3 \neq 0$. 114

3.1 Mediation Directed Acyclic Graph (DAG) 120

3.2 Relative bias of direct (NDE) and indirect (NIE) naive causal effects for the grid (SN,SP)=$(0.1, 0.9) \times (0.1, 0.9)$ for continuous outcome modeled using linear regression (BLUE-No exposure-mediator interaction/PINK-Exposure-mediator interaction) 132

3.3 Relative bias of direct (NDE) and indirect (NIE) naive causal effects for the grid (SN,SP)=$(0.1, 0.9) \times (0.1, 0.9)$ for binary outcome modeled using logistic regression (GREEN-No exposure-mediator interaction/YELLOW-Exposure-mediator interaction) 133

List of Tables

1.1 Macro Comparison (*Hayes and Preacher **Imai et al. ***Valeri and VanderWeele † Muthen) 26

2.1 Simulations for naive, method of moments (MoM), regression calibation (RC) and SIMEX estimators of direct, indirect and total effects with continuous (linear link) outcome. 74

2.2 Simulations for naive, method of moments (MoM), regression calibation (RC) and SIMEX estimators of direct, indirect and total effects with binary (logistic link) outcome. 75

2.3 Sensitivity analysis results for direct (OR^{NDE}), indirect (OR^{NIE}) effects and proportion mediated ($PM = OR^{NDE} \times (OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1)$) for variants rs1051730 and rs8034191 allowing for exposure-mediator interaction and attenuation factor λ up to 0.25. 77

2.4 Simulations results for continuous outcome 110

2.5 Simulation results for binary (logistic link) outcome 111

3.1 Asymptotic bias of naive, predictive value weighting (IRLS/PVW, IRLS/**s**PVW, IRLS/**t**PVW) and direct maximum likelihood (ML) of controlled direct effect (CDE) natural direct (NDE), natural indirect effect (NIE) and total effect (TE) when sample size is $n = 10,000$, marginal probability of the true mediator is 50%, and the outcome is continuous or binary. 140

3.2 Naive and misclassification-corrected mediation analysis allowing for exposure-mediator interaction ($SP = 0.99$, $SN = (0.8, 0.9, 0.95, 0.99)$, CI obtained from delta method standard errors) 145

3.3 Naive and misclassification-corrected mediation analysis assuming no exposure-mediator interaction ($SP = 0.99$, $SN = (0.8, 0.9, 0.95, 0.99)$, CI obtained from delta method standard errors) 146

Acknowledgments

I am deeply grateful to my advisor, Tyler VanderWeele, for supporting, teaching, and guiding me in these years.
My sincere thanks goes to Xihong Lin and Judith Lok, my dissertation committee members, for their valuable help in improving my dissertation.
This intense life as a PhD student, that now draws to an end, would have never started without the encouragement and the support of my undergraduate and master thesis advisor Marco Bonetti. Thank You!
The beauty of research comes along with the fight for being a better researcher every day.
I thank my family and my friends for walking with me in this wonderful journey.

Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros

[1]Linda Valeri and [1,2] Tyler J. VanderWeele

[1]Department of Biostatistics, Harvard School of Public Health and
[2]Department of Epidemiology, Harvard School of Public Health

Abstract

Mediation analysis is a useful and widely employed approach to studies in the field of psychology and in the social and biomedical sciences. The contributions of this paper are several-fold. First we seek to bring the developments in mediation analysis for non linear models within the counterfactual framework to the psychology audience in an accessible format and compare the sorts of inferences about mediation that are possible in the presence of exposure-mediator interaction when using a counterfactual versus a purely statistical approach. Second, the work by VanderWeele and Vansteelandt (2009, 2010) is extended here to allow for dichotomous mediators and count outcomes. Third, we provide SAS and SPSS macros to implement all of these mediation analysis techniques automatically and we compare the types of inferences about mediation that are allowed by a variety of software macros.

Keywords: Causal Inference, Direct and Indirect Effects, Mediation Analysis, Interaction, Software Macro.

1.1 Introduction

Mediation analysis investigates the mechanisms that underlie an observed relationship between an exposure variable and an outcome variable and examines how they relate to a third intermediate variable, the mediator. Rather than hypothesizing only a direct causal relationship between the independent variable and the dependent variable, a mediational model hypothesizes that the exposure variable causes the mediator variable, which in turn causes the outcome variable. The mediator variable then serves to clarify the nature of the relationship between the exposure and outcome variable (MacKinnon, 2008). For example, it might be of interest to understand whether a rehabilitation program for drug-addicted individuals, with methadone as treatment, leads to increased work activity and whether drug use may mediate some of this effect. In this example, drug use may be a potential mediator of the relationship between the methadone treatment and the work activity outcome since the level of methadone may affect drug use, which may in turn affect work activity.

The use of mediation analysis in psychology and in the social sciences is widespread and has been strongly influenced by the seminal paper of Baron and Kenny (1986). More recently, new advances in mediation analysis have been made by using the counterfactual framework (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010ab). Using the counterfactual framework has allowed for definitions of direct and indirect effects and for decomposition of a total effect into direct and indirect effects even in models with interactions and non-linearities. In many contexts investigators are interested in assessing whether most of the effect is mediated through a particular intermediate or the extent to which it is through other pathways. Decomposition of a total effect into direct and indirect effects accomplishes this goal.

It is then possible to use this counterfactual framework to extend formulae from Baron and Kenny (1986) to allow for mediation analysis even in the presence of exposure mediator interactions. Special cases for mediated effects in the presence of interaction have appeared previously in the literature (e.g. Preacher et al., 2007) but do not give definitions of

direct effects such that the total effect decomposes into a direct and indirect effect. In particular, VanderWeele and Vansteelandt (2009, 2010) derived results for direct and indirect effects for linear and logistic regressions when exposure-mediator interaction is present. In many studies it is unrealistic to assume that the exposure and mediator do not interact in their effects on the outcome. Carrying out mediation analysis incorrectly assuming no interaction may result in invalid inferences. The present paper makes a number of important contributions to mediation analysis from both methodological and implementation perspectives. First, we extend work on causal mediation analysis for parametric models with interactions (VanderWeele and Vansteelandt, 2009, 2010) to allow for dichotomous mediators, and not simply continuous mediators as were previously considered. This is done using Pearl's mediation formula (Pearl, 2001), also described outside the context of counterfactuals elsewhere (Huang et al., 2004). Second, we moreover extend the results to count data. Third, we provide SAS and SPSS macros, which give estimates and confidence intervals for direct and indirect effects when interactions between the mediator of interest and the exposure are present and we compare the types of inference about mediation that are available in a variety of software packages. Finally, we will compare and contrast the inferences that are possible about direct and indirect effects in the presence of exposure-mediator interaction, when using the counterfactual framework versus a purely statistical approach. We consider both continuous and dichotomous variables as outcomes and mediators and allow for general treatment variables. The approach here enriches the contributions of Baron and Kenny and expands the previous software developed by Preacher and Hayes (2004) and Preacher et al. (2007) to allow for effect decomposition of a total effect into direct and indirect effects in the presence of exposure-mediator interaction and other non-linearities.

The paper is organized as follows. The first section discusses the approach to mediation analysis sometimes referred to as the "product method" and made popular by Baron and Kenny. The second section introduces the reader to the counterfactual approach which gives rise to broader definitions of direct and indirect effects and allows one to carry out mediation analysis when interaction between exposure and mediator is present. In the following section, conditions are given for the identifiability of direct and indirect effects

in mediation analysis; these are the conditions needed for the results of statistical procedures to have a causal interpretation. The next section clarifies the relationship between the results on mediation analysis that arise within the counterfactual framework with other popular approaches to mediation analysis. The paper continues with instructions for using the software developed (SAS and SPSS) and a description of the output is provided. We conclude by providing an example of mediation analysis performed using the mediation macros.

1.2 Classic Regression Approach to Mediation Analysis

The practice of mediation analysis in the field of psychology has been highly influenced by the work of Baron and Kenny (1986). The causal diagram in Figure 1.1 captures how they conceptualized the role of a mediator variable. In this graph, which represents a simple mediation model, A denotes an exposure (or treatment) variable, M denotes the mediator and Y denotes the outcome variable. According to Baron and Kenny the follow-

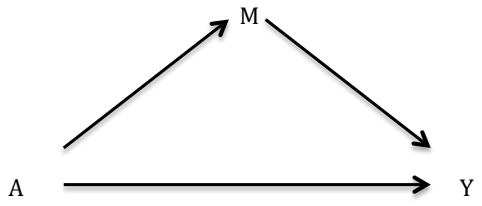

Figure 1.1: Mediation model in Baron and Kenny 1986 paper.

ing criteria need to be satisfied for a variable to be defined as mediator: (i) a change in levels of the exposure variable significantly affects the changes in the mediator (i.e., Path from A to M), (ii) there is a significant relationship between the mediator and the outcome

(i.e., Path from M to Y), (iii) a change in levels of the exposure variable significantly affects the changes in the outcome (i.e., total effect of A on Y is significant), and (iv) when the previously defined paths are controlled, a previously significant relation between the exposure and outcome is no longer significant, with the strongest demonstration of mediation occurring when the path from the independent variable to the outcome variable is zero.

While requirements (i) and (ii) have been accepted as correct criteria to identify a potential mediator, requirement (iii) has been critiqued by many scholars (MacKinnon, 2008). Consensus has now been reached that the relationship between A and Y need not be statistically significant for M to be a mediator. The reason is that the effect of A on Y may not be significant when direct and mediated effects have opposite sign. This phenomenon is commonly known as inconsistent mediation. Requirement (iv) is also not necessary because mediation can be partial or complete. When mediation is complete, after controlling for M, the direct path from A to Y would be zero. When mediation is partial, the path from A to Y can still be significant, but the effect should be reduced if mediation is indeed present. In the present work we allow for both partial and complete mediation.

In 1986, Baron and Kenny also proposed a parametric approach to estimate and test for mediation. The approach is often simply referred to as the "Baron and Kenny approach", however others had proposed it previously (Hyman, 1955; Alwin and Hausen, 1975; Judd and Kenny, 1981; Sobel 1982) and is also more generally referred to as the "product method". Let A be the treatment, Y the outcome, M the mediator and C additional covariates. For the case of continuous mediator and outcome, consider the following regression models:

$$E(M|A = a, C = c) = \beta_0 + \beta_1 a + \beta_2^{'} c \tag{1.1}$$

$$E(Y|A = a, M = m, C = c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_4^{'} c \tag{1.2}$$

The original Baron and Kenny approach did not have covariates, but the same general approach applies with covariates (i.e. $\beta_2^{'} c$ and $\theta_4^{'} c$ were not included in the original models by the authors; here c is considered a vector and may contain multiple confounders). Establishing mediation entails estimating the parameters of these regression models. In

particular, Baron and Kenny proposed that the direct effect be assessed by estimating θ_1 and that the indirect effect be assessed by estimating $\theta_2\beta_1$. The direct effect can be conceived of as the treatment effect on the outcome at a fixed level of the mediator variable, which is different from the total effect, which represents simply the overall effect of exposure or treatment on the outcome. The indirect effect can be conceived of as the effect on the outcome of changes of the exposure which operate through mediator levels.

1.3 Counterfactual Approach to Mediation Analysis

While the concept of mediation, as defined within psychology and the social sciences, is theoretically appealing, the methods traditionally used to study mediation empirically have important limitations concerning their applicability in models with interactions or non-linearities (Robins and Greenland, 1992; Pearl, 2001).

Recent contributions in mediation analysis have emphasized the importance of articulating identifiability conditions for a causal interpretation and have extended definitions and results for direct and indirect effect to settings in which non-linearities and interactions are present (Robins and Greenland, 1992; Pearl, 2001). This is relevant especially when mediation analysis is implemented in social science contexts where, for example, the exposure of interest might interact in its effect on the outcome with the mediator.

The approach advocated by Baron and Kenny is widely applied for mediation analysis and software is available to implement it (Preacher and Hayes, 2004, 2008). However, this method does not fully accommodate settings in which the exposure and the mediator interact in their effects on the outcome. Although special cases for mediated effects in the presence of interaction are available (e.g. Preacher et al., 2007), these do not give definitions of direct effects such that the total effect decomposes into a direct and indirect effect. VanderWeele and Vansteelandt (2009, 2010) show how the notions of direct and indirect causal effects from causal inference in the counterfactual framework (Greenland and Robins, 1992; Pearl, 2001) can extend the Baron and Kenny formulae for direct and indirect effects to settings in which there is an interaction term between exposure and

mediator in the outcome regression.

Suppose we have a continuous outcome and mediator and the mediator regression remains as in model (1.1) while the outcome regression is reformulated as

$$E(Y|A=a, M=m, C=c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c \tag{1.3}$$

The use of the causal inference approach to mediation analysis gives rise to counterfactual definitions of direct and indirect effects, which were formulated by Pearl (2001) and Greenland and Robins (1992). These effects can be estimated from the regression parameters in models (1.1) and (1.3), provided certain identifiability assumptions (no confounding), described below, hold and models are correctly specified (VanderWeele and Vansteelandt, 2009, 2010). In particular, from models (1.1) and (1.3) what can be defined as the controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) for change in exposure from level a^* to level a, are given by

$$CDE = (\theta_1 + \theta_3 m)(a - a^*)$$

$$NDE = (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'} c)(a - a^*)$$

$$NIE = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

These expressions generalize those of Baron and Kenny to allow for interactions between the exposure and the mediator. We describe these effects below. Note that if interaction is not present, so that $\theta_3 = 0$, the controlled direct effect and the natural direct effect are equal to the direct effect obtained using Baron and Kenny approach θ_1 times $(a - a^*)$ and the natural indirect effect is equal to the indirect effect of the Baron and Kenny approach $\theta_2\beta_1$ times $(a - a^*)$.

The controlled direct effect (CDE) expresses how much the outcome would change on average if the mediator were controlled at level m uniformly in the population but the treatment were changed from level $a^* = 0$ to level $a = 1$. The natural direct effect (NDE) expresses how much the outcome would change if the exposure were set at level $a = 1$ versus level $a^* = 0$ but for each individual the mediator were kept at the level it would have taken in the absence of the exposure. The natural indirect effect (NIE) expresses

how much the outcome would change on average if the exposure were controlled at level $a = 1$, but the mediator were changed from the level it would take if $a^* = 0$ to the level it would take if $a = 1$. The total effect (TE) can be defined as how much the outcome would change overall for a change in the exposure from level $a^* = 0$ to level $a = 1$. More formal definitions of these effects explicitly in terms of counterfactuals are given in the appendix. An important property of the natural indirect effect and the natural direct effect is that the total effect decomposes into the sum of these two effects; this holds even in models with interactions or non-linearities (Pearl, 2001). The expressions given above involving the coefficients of models (1.1) and (1.3) will be equal to the effects we have just discussed under certain identifiability assumptions given in the next section. These identifiability assumptions allow for a causal interpretation of the direct and indirect effects. These effects are conditional on the level of the covariates C. For continuous outcomes, if C were set at its average level we would obtain marginal effects on the entire population.
While controlled direct effects are often of greater interest in policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms (Robins, 2003; Joffe et al., 2007).

1.4 Identification

The conditions for a causal interpretation of the direct and indirect effects defined in the previous section can be usefully characterized via causal diagrams. Consider the relation between the variables in Figure 1.2, which might encompass a wide range of scenarios in mediation analysis. A careful study of this graph will be useful in clearly formulating the identifiability assumptions for the direct and indirect causal effects of interest: The variables in the graph are: exposure (A), mediator (M), outcome (Y), covariates ($C = (C_1, C_2)$), which include exposure-outcome confounders (C_1) and mediator-outcome confounders (C_2). All the comments below will still hold if C_1 affects C_2 or if C_2 affects C_1.
Consider the example of working activity of a drug addicted individual as the outcome of

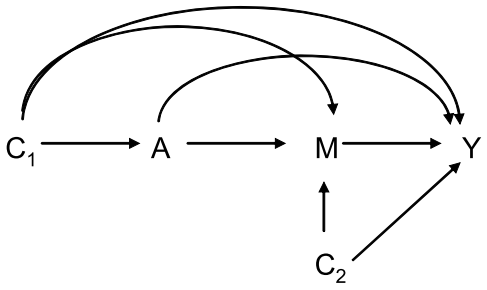

Figure 1.2: Mediation DAG

interest (Y). Let the treatment be methadone (A), and the potential mediator be the level of drug use (M). Under this scenario, the investigator may be interested in studying how the effect of the treatment A on the outcome Y is mediated by the level of drug use of an individual (M). In addressing this question of interest, the investigator must think carefully about and try to control for variables that may be exposure-outcome confounders (C_1) or mediator-outcome confounders (C_2). For example, there might be social and biological factors, such as income and hypertension status (C_1), that affect the decision of the level of treatment (A) and the working activity outcome (Y), or other factors, such as neighborhood of residence or alcohol consumption (C_2), which affect both the level of drug use (M) and the working activity outcome (Y).

In order for the effects to have a causal interpretation, control must be made for the confounding variables. In order to ensure identifiability of controlled direct effect, two assumptions are needed: namely those of (i) no unmeasured confounding of the treatment-outcome relationship and (ii) no unmeasured confounding of mediator-outcome relationship. The first of these assumptions would be automatically satisfied if treatment were randomized, but even with randomized treatment the second assumption might not be satisfied. If we refer to the example above, to control for (i) confounding of the treatment-outcome relationship the investigator must adjust for common causes of the treatment and the outcome e.g. information on income and hypertension status and any

other treatment-outcome confounding variable (C_1) in the analysis. To control for (ii) mediator-outcome confounding the investigator must adjust for common causes of the mediator and the outcome e.g. alcohol consumption and neighborhood of residence or any other mediator-outcome confounding variable (C_2). In practice, both sets of covariates would simply be included in the overall set C for which adjustment is made; the investigator does not need to distinguish in this regression approach the treatment-outcome and the mediator-outcome confounding variables but the collection of covariates must include both sets for estimates to have a causal interpretation.

The assumptions we have described are for controlled direct effects; the identification of natural direct and indirect effects uses these two assumptions above along with two additional assumptions. In particular, for natural direct and indirect effect there must also be (iii) no unmeasured confounding of the treatment-mediator relationship. Control must be made for variables that cause both the level of treatment and the level of the mediator. In the context of our example, hypertension may be a factor which influences the use of treatment as well as the level of drug addiction, and it would need to be controlled for in the analysis. This third assumption, like the first, would also be satisfied automatically if the treatment were randomized. Finally, for the natural direct effect and indirect effects to be identified it also needs to be the case that (iv) there is no mediator-outcome confounder that is affected by the treatment (i.e. no arrow from A to C_2 in Figure 1.2).

It should be noted that assumptions (i), (ii), and (iii) also require an assumption of temporal ordering. This assumption of temporal ordering is implicitly or explicitly present in various approaches to mediation analysis (Cole and Maxwell, 2003). In particular, the assumption of no unmeasured confounding of the treatment-outcome relationship implicitly assumes that the treatment temporally precedes the outcome. The assumption of no unmeasured confounding of the mediator-outcome relationship implicitly assumes that the mediator precedes temporally the outcome. Finally, the assumption of no unmeasured treatment-mediator confounding implicitly assumes that the exposure must precede the mediator. Formally the no unmeasured confounding assumptions require that associations reflect causal effects; if the temporal ordering assumptions were not satisfied then neither would the no unmeasured confounding assumptions since associations

would not represent causal effects.

In summary, controlled direct effects require (i) no unmeasured treatment-outcome confounding and (ii) no unmeasured mediator-outcome confounding. Natural direct and indirect effects require these assumptions and also (iii) no unmeasured treatment-mediator confounding and (iv) no mediator-outcome confounder affected by treatment. It is important to note that randomizing the treatment is not enough to rule out confounding issues in mediation analysis. This is because randomization of the treatment rules out the problem of treatment-outcome and treatment-mediator confounding but does not guarantee that the assumption of no confounding of mediator-outcome relationship holds. This is because even if the treatment is randomized, the mediator generally will not be. This was pointed out by Judd and Kenny (1981), James et al. (1984), MacKinnon (2008), but unfortunately not mentioned in the popular paper by Baron and Kenny (1986). If there are confounders of the mediator-outcome relationship for which control has not been made, then direct and indirect effect estimates will not have a causal interpretation; they will be biased. This is true for the controlled direct effect and natural direct and indirect effects described above and also for the effects described by Baron and Kenny. Investigators should think more carefully about and collect data on and control for such mediator-outcome confounding variables when mediation analysis is of interest. If the investigator is aware that unmeasured confounding may be an issue in his or her study, sensitivity analyses (VanderWeele, 2010; Imai et al. 2010a) should be implemented.

1.5 Binary Outcome

We have thus far considered only the case in which both outcome and mediator are continuous. The results can be extended to cases in which one or both of the mediator and outcome variables are binary. For example, when the outcome is binary and mediator is continuous the model for the mediator is represented by (1.1) and the outcome can be modeled via a logistic regression

$$logit[P(Y=1|A=a,M=m,C=c)] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \tag{1.4}$$

For this case, provided the outcome is relatively rare and assumptions $(i) - (iv)$ hold, we can derive controlled direct effects, and natural direct and indirect effects on the odds ratio scale (VanderWeele and Vansteelandt, 2010a) as:

$$log\{OR^{CDE}\} = (\theta_1 + \theta_3 m)(a - a^*)$$

$$log\{OR^{NDE}\} \approx (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' C + \theta_3\theta_2\sigma^2)(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})$$

$$log\{OR^{NIE}\} \approx (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

where σ^2 is the variance of the error term in the regression for the mediator, M, and where the approximations hold to the extent that the outcome Y is rare.

With these odds ratios, the total effect is equal to the product of the natural direct and indirect effects (rather than the sum).

When the outcome is not rare, the odds ratio does not approximate the risk ratio anymore. Therefore, the causal effects previously defined will be biased if logistic regression is used to model the outcome. In this case the investigator can estimate the causal effect by running a generalized linear model regression with a binomial distribution and a log link and the causal effects will have a risk ratio interpretation and the formulas hold exactly.

When the outcome is rare then the direct and indirect effects can be estimated even in case-control designs. The formulas for the effects remain the same, however the mediator regression is run only for controls, to take into account the case-control design (VanderWeele and Vansteelandt, 2010). This approach works because with a rare outcome Y, the distribution of M among the controls will approximate the distribution in the population. We also extend the previous results to the cases in which the mediator is a dichotomous variable. The identifiability assumptions do not change but now we would use a logistic

model for the mediator:

$$logit[P(M = 1|A = a, C = c)] = \beta_0 + \beta_1 a + \beta_2^{'} c \tag{1.5}$$

Formulas for controlled direct effects and natural direct and indirect effects when the mediator is dichotomous are given in the appendix. Finally, in the online appendix we show that these formulas for causal effects for binary outcome, along with their standard errors, extend to count variables when modeled with a log link.

The total effect is computed as the sum of the natural direct effect and the natural indirect effect when the outcome is continuous and as the product of the natural direct and indirect effect odds ratios when the outcome is binary. Another measure that has been popular in mediation analysis is the proportion mediated. The proportion mediated can be defined as the ratio of the natural indirect effect to the total effect when the outcome is continuous; the proportion mediated on risk difference scale can also be calculated when the outcome is binary using a transformation of the odds ratios (VanderWeele and Vansteelandt, 2010a). Several authors have, however, issued cautions on its use. Kenny (1998) warns about the instability of such measure, especially when the association between the exposure and the outcome is weak. Consequently, we have not implemented this measure in the macro.

Estimates described later in the paper of the direct and indirect effects of interest are obtained by plugging in the estimated coefficient values while the standard errors can be obtained using the delta method or by bootstrapping techniques. The reader can refer to the online supplement for derivations of the direct and indirect effects and delta method standard errors. The macro we provide will calculate these automatically.

1.6 Mediation analysis for models with non-linearities - a comparison of approaches

The counterfactual approach to mediation analysis displays all its power and flexibility when the causal relationships under study are complex and the investigator needs to

depart from simple linear models and allow for non-linearities and interactions. In this section we describe some of the advantages of employing the counterfactual framework to causal mediation that we presented in the previous sections by comparing it to other popular methods to address mediation questions. In this comparison we will focus on the so-called product method, the difference method, and the MacArthur approach, and address also some developments with regard to "moderated mediation". We first describe traditional statistical approaches and we then discuss what the counterfactual approach contributes over and above them and comment on the relation between the two.

1.6.1 Traditional approaches to mediation analysis

Modern approaches to mediation have been inspired by the pioneering work of the geneticist Sewall Wright (1920) who developed the path analysis method. Path analysis is now viewed as a special case of structural equation modeling (SEM). Structural equations methods allow for the estimation of direct and indirect effects by modeling covariance and correlation matrices. Most mediation analyses in psychological studies have been conducted using the structural equation modeling (SEM) approach (Baron and Kenny, 1986; Judd and Kenny, 1981; MacKinnon, 2008). Methods to improve estimation and inferential procedures for SEM-based mediation analyses have continued to develop (e.g. MacKinnon, 2008; Sobel, 1982). Structural equation models are often criticized for not adequately addressing issues of confounding/endogeneity in inferring causal relationships. However, if such issues of confounding are adequately addressed by including all relevant confounders (as described in detail above) in the structural equation model then the SEM approach can be a useful tool. The counterfactual approach has placed strong emphasis on identifiability assumptions and conceptual definitions of causal effects, and recently, a number of authors have been using the counterfactual framework to translate the SEM approach within the counterfactual framework [1] (e.g. Jo, 2008; Sobel, 2008; Van-

[1]Note that a different way to think about inference with regard to an intermediate within the counterfactual approach framework is to use the concept of "principal strata" (Frangakis and Rubin, 2002; Jo, 2008; Rubin, 2004; VanderWeele, 2008; Chiba, 2010). For a discussion on the use of principal stratification in mediation analysis the interested reader can refer to the commentaries in the International Journal of

derWeele and Vansteelandt, 2009; Imai et al., 2010a; Pearl, 2011). Among traditional SEM methods, we describe the product method and the difference method. Assume a simple mediation model with no exposure-mediator interaction. The rationale behind the product method is that mediation depends on the extent to which the exposure A changes the mediator M, β_1 from equation (1.1), and the extent to which the mediator affects the outcome Y, θ_2 from equation (1.2). The product method estimator of the indirect effect is then simply $\theta_2\beta_1$. Sobel (1982) proposed a test for a mediated effect from the product method estimator. The difference method approach is implemented by fitting an outcome model with the mediator as in equation (1.2) and also an outcome model with no mediator:

$$E(Y|A = a, C = c) = \theta_0^\dagger + \theta_1^\dagger a + \theta_4^\dagger c \tag{1.6}$$

The value of the mediated or indirect effect is then estimated by taking the difference in the coefficients from equations (1.6) and (1.2), $\theta_1^\dagger - \theta_1$ this corresponds to the reduction in the independent variable effect on the dependent variable when adjustment is made for the mediator. The algebraic equivalence of the indirect effect using the product method, $\theta_2\beta_1$, and the difference method, $\theta_1^\dagger - \theta_1$ was shown by MacKinnon et al. (1995) for ordinary least squares in linear models with continuous outcomes and discussed also in Alwin and Hauser (1975). The product method and difference method diverge however when using a binary outcome and logistic regression (MacKinnon and Dwyer, 1993), a point to which we return below. When mediation models include an exposure-mediator interaction term in the outcome regression, this is a particular case or a variant of what is sometimes referred to as "moderated mediation" (James and Brett, 1984; Preacher et al., 2007). Moderated mediation considers the case in which a covariate moderates the mediated effect (cf. MacKinnon, 2007) i.e. when the mediated effect varies by the level of a covariate. Such moderated mediation by a covariate was also analyzed by Yzerbyt, Muller, and Judd (2004) and Muller, Yzerbyt and Judd (2008). When the treatment itself is the moderator for the mediator (as considered in Preacher et al. 2007), the effect of the mediator is allowed to vary by treatment status; or, conceived of another way, the effect of treatment is allowed to vary with (i.e. it interacts with) the mediator. In this setting,

Preacher et al. (2007) derived an indirect effect estimator in the context of moderated mediation using the product method.

The MacArthur approach (Kraemer et al., 2008) gives criteria somewhat different than that of Baron and Kenny in assessing mediation and allows also for assessing exposure-mediator interactions. This approach to mediation analysis is based on the assumption that temporal antecedence and association are necessary (but not sufficient) for a causal relationship. The approach allows for non-linear relations among variables to qualify as mediation as long as there is a relationship between the exposure A, and the mediator M. In particular, it is proposed, first, that if there is no association between A and M, and if M precedes A, and if the $A \times M$ interaction is significant, then the variable M is to be considered as a moderator rather than a mediator. Second, for M to be a mediator for the effect of A on outcome Y, A should precede M and M should precede Y, the variables A and M should be correlated, and either the main effect of M on the outcome or the $A \times M$ interaction should be significant.

1.6.2 Comparison of traditional approaches with the counterfactual approach when there are interactions and non-linearities

One of the chief advantages of the counterfactual approach to mediation analysis is that it allows for the decomposition of a total effect into a direct effect and an indirect effect even when there are interactions and non-linearities. As noted above, some of the statistical approaches, such as that of Preacher et al. (2007) or Kraemer et al. (2008) allow one to assess mediation even when there is exposure-mediator interaction. In fact, the indirect effect of Preacher et al. (2007) for continuous outcome when there is an exposure-mediator interaction is equivalent to the one given here. However, neither Preacher et al. (2007) nor Kraemer et al. (2008) give a definition of a direct effect in the presence of exposure-mediator interaction such that the sum of the direct and indirect effects equals a total effect. The counterfactual approach provides a general approach to do effect de-

composition irrespective of the statistical model and irrespective of possible interactions. The counterfactual approach coincides with the criteria for mediation of the MacArthur approach (Kraemer et al., 2008) but provides actual direct and indirect effect estimates that combine to a total effect and makes clear the no-unmeasured-confounding assumptions needed for a causal interpretation. The counterfactual approach also helps in understanding mediation with binary outcomes and binary mediators. As noted above, with a binary outcome and logistic regression, the product method and difference method give different results (MacKinnon and Dwyer, 1993). In fact, neither in general will be equal to an estimate of an indirect effect with a causal interpretation (VanderWeele and Vansteelandt, 2010). VanderWeele and Vansteelandt (2010) did, however, show that when there is no exposure-mediator interaction, the product method and difference method will be approximately equivalent when the outcome is rare; and both will then be approximately equal to the natural indirect effect when all the no confounding assumptions hold. The problem with dichotomous outcomes arises when the outcome is common and has to do with the fact that logistic regression uses the odds ratio, which is a measure that is "non-collapsible". Viewed intuitively, the problem occurs because when the outcome is common, the odds ratio does not approximate the risk ratio, and the extent of this lack of approximation can vary with the other covariates in the models. With a common outcome, the odds ratios with the mediator in the model versus without the mediator in the model are thus not directly comparable, and so the difference method essentially breaks down. The risk ratio does not suffer this problem and it is for this reason that we propose using a log-linear model in this paper when the outcome is common. Moreover, this approach also allows us to define and estimate direct and indirect effects when the outcome is binary and an exposure-mediator interaction is present. We have moreover, using the counterfactual approach in this paper, derived analytic expressions for cases when the mediator itself is binary. The counterfactual approach provides a versatile framework to derive direct and indirect effects and to do effect decomposition even with binary variables and non-linear models.

As is perhaps now clear from this discussion, the traditional statistical approach and the counterfactual approach to mediation will in some settings coincide. For linear models

and log-linear models, they will coincide when there is no exposure-mediator interaction; for logistic models, they will coincide when there is no exposure-mediator interaction and when the outcome is rare (VanderWeele and Vansteelandt, 2009, 2010). Thus, before an investigator proceeds with one of the traditional approaches (the product method or difference method) he or she should: (i) consider whether control has been made for exposure-outcome confounders, mediator-outcome confounders, and exposure-mediator confounders, (ii) check whether there is exposure-mediator interaction, and (iii) if the outcome is binary and logistic regression is used, check whether the outcome is rare. If the no-unmeasured-confounding conditions are satisfied, there is no interaction, and the outcome is rare if logistic regression is used, then proceeding with the traditional statistical approaches is fine. If there are exposure-mediator interactions then the approach described in this paper, or another counterfactual-based approach, should be used. If the outcome is common, a log-linear model can be used. If there are confounders of the exposure-outcome, mediator-outcome, or exposure-mediator relationship then, to the extent possible, these should be controlled for in the models; otherwise sensitivity analysis techniques (VanderWeele, 2010; Imai et al., 2010a) can be used.

As a final point of discussion, we note that even in the presence of interaction and non-linearities, the product method may be useful to test for mediation even if the estimates are not themselves interpretable as estimates of an indirect effect. In other words, to test for mediation we can test for whether the product of the coefficients is non-zero even if this product is not equal to a causal indirect effect measure. For example, with logistic model with common outcome, the product method estimates will not in general have a causal interpretation as a natural indirect effect. It is nonetheless the case that although the product-method estimator is not itself a measure of an indirect effect, the product method still gives a valid test for the presence of a mediated effect, provided that the identification assumptions hold and that the models are correctly specified (a formal proof of this is given in the e-Appendix of VanderWeele, 2011). The intuition is that even if the product of the coefficients is not equal to a causal indirect effect, if the product is non-zero then there must be an effect of the exposure on the mediator and an effect of the mediator on the outcome, and under the identification assumptions, this

would also imply the presence of a natural indirect effect. Thus, the product-method approach can still be useful in testing for mediation even when there are interactions and non-linearities. For estimation and for decomposing a total effect into a direct and indirect effect (arguably the chief advantages of the counterfactual approach), rather than just testing, methods such as those described in this paper can be employed.

1.7 Description of the SAS macro

The present macro is designed to enable the investigator to easily implement mediation analysis in the presence of exposure-mediator interaction accounting for different types of outcomes (normal, dichotomous-logistic or dichotomous log-linear, poisson, negative binomial) and mediators of interest (normal or dichotomous with logit link). The logit link for dichotomous outcomes should only be used if the outcome is rare. If the outcome is not rare the log link can be used (though the outcome model may not always converge). In the case of using the log link the direct and indirect effects are on the risk ratio scale. In particular, these macros for SAS and SPSS provide estimates, and confidence intervals for the direct and indirect effects previously defined. The estimates assume the model assumptions are correct and the identifiability assumptions discussed in the previous section hold.

1.7.1 Basic SAS Macro

The macro has been developed using the 9.2 version of SAS software. In order to implement mediation analysis via the *mediation macro* in SAS the investigator first opens a new SAS session and inputs the data, which has to include the outcome, treatment and mediator variables as well as the covariates to be adjusted for in the model. Macro activation requires then the investigator to save the macro script and input information in the statement

```
% mediation(data=,yvar=,avar=,mvar=,cvar=,a0=,a1=,m=,nc=,yreg=,mreg=,
interaction=)
run;
```

First one inputs the name of the dataset ($data =$), then the name of the outcome variable ($yvar =$), the treatment variable ($avar =$), the mediator variable ($mvar =$), the other covariates, ($cvar =$). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro $dumvar$ from MCHP SAS Macros, for example, can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* ($a0 =$), the new exposure level a ($a1 =$), the level of mediator m ($m =$) at which the controlled direct effect is to be estimated and the number of covariates to be used ($nc =$). When no covariates are entered, then the user still needs to write the commands $cvar =$ and $nc =$ even though both are left blank. The user must also specify which types of regression have to be implemented. In particular, linear, logistic, loglinear, poisson or negbin can be specified ($yreg =$). For the mediator either linear or logistic regressions are allowed ($mreg =$). Finally, the analyst needs to specify whether an exposure-mediator interaction is present ($interaction = true$ or $false$).
The macro provides the following output: first the regression output for outcome and mediator models is provided. The output in the SAS macro is derived from the procedures of proc reg when the variable is continuous, proc logistic when the variable is binary. When the outcome is specified as poisson, negative binomial or log-linear the procedure proc genmod is employed. If the dataset contains missing data the macro implements a complete case only analysis. A table with direct and indirect effects together with total effects follows. The effects are reported for the mean level of the covariates C. The table contains standard errors, and confidence intervals for each effect.

1.7.2 Other options in the SAS Macro

The reduced output is the default option. The table will just display controlled direct effect, natural direct effect, natural indirect effect and total effect described above. When the option $output = full$ is used, both conditional effects and effects evaluated

at the mean covariate levels are shown. When the $output = full$ option is chosen, the investigator must enter fixed values for the covariates C at which compute conditional effects. The macro statement is as follows:

```
%mediation(data=,yvar=,avar=,mvar=,cvar=,a0=,a1=,m=,nc=,yreg=,mreg=,
interaction=,output=,c=)
run;
```

When $output = full$ is added, then, in addition to the controlled direct effect, and the natural direct and indirect effect described above, two other effects are displayed. The natural direct and indirect effects we have been considering are sometimes called the "pure" natural direct effect and the "total" natural indirect effect (Robins and Greenland, 1992). We can also consider the "total" natural direct effect and the "pure" natural indirect effect. For binary exposure the total natural direct effect expresses how much the outcome would change on average if the exposure changed from level $a^* = 0$ to level $a = 1$, but the mediator for each individual was fixed at the natural level which would have taken at exposure level $a = 1$. The pure natural indirect effect expresses how much the outcome would change on average if the exposure were controlled at level $a^* = 0$ but the mediator were changed from the natural level it would take if $a^* = 0$ to the level that would have taken at exposure level $a = 1$. These effects are also reported if the user selects $output = full$. If there is no exposure-mediator interaction, the "pure" and "total" natural direct effects will coincide and the "pure" and "total" natural indirect effects will coincide.

The investigator also has the option of implementing mediation analysis when data arise from a case-control design, provided the outcome in the population is rare. To do so the option $casecontrol = true$ can be used. In this case the macro statement changes to:

```
%mediation(data=,yvar=,avar=,mvar=,cvar=,a0=,a1=,m=,nc=,yreg=,mreg=,
interaction=,casecontrol=)
run;
```

Finally, the investigator can choose whether to obtain standard errors and confidence intervals via the delta method or a bootstrapping technique. The default is the delta method. To use bootstrapping the option $boot = true$ can be given. In this case the macro will compute 1,000 bootstrap samples from which causal effects are obtained along with their standard errors ($s.e.$) and percentile confidence intervals ($p_{95_{CIlower}}, p_{95_{CIupper}}$). If the investigator wishes to use a higher number of bootstrap samples, instead of "true" he or she inputs the number of bootstrap samples desired (e.g. boot=5000 would estimate standard errors and confidence intervals using 5000 bootstrap samples). The use of bootstrap for standard errors is generally to be preferred if the sample size of the original sample is small as it will lead to more accurate inferences than the delta method (MacKinnon, 2008). However, these issues are less important if the original sample is large and if this is the case the use of delta method standard errors may be preferred because of computational efficiency. (For example, Ananth and VanderWeele (2011) conducted a mediation analysis using a sample of 26,000,000 individuals and bootstrapping would have been completely infeasible). When using the bootstrap the macro statement changes to:

```
%mediation(data=,yvar=,avar=,mvar=,cvar=,a0=,a1=,m=,nc=,yreg=,mreg=,
interaction=,boot=)
run;
```

As noted above, if the investigator wants to add a categorical variable as covariate, this must be recoded as a series of indicator variables. For example, if a covariate, named catvar, takes four levels (1,2,3,4) we could construct three "dummy" or "indicator" variables, named, for example, ivar2, ivar3, and ivar4, leaving the first value as the reference. The variable ivar2 would take the value 1 for all observations which had catvar=2, and 0 for all other observations. The variable ivar3 would take the value 1 for all observations that had catvar=3 and 0 for all other observations, etc. The macro dumvar mentioned previously requires the user to list the dataset (data=), the categorical variable (e.g. catvar) that needs to be transformed in the input (dvar=). The user needs also to input the prefix of

the name of the dummy variables (e.g. ivar) that will be generated (prefix=) and the reference category (drop=). Categorical variables can be both character and numerical using $dumvar$. For example we can run the following:

```
dumvar data=dat dvar="catvar" prefix="ivar" drop="ivar1"
```

Running this command will generate three indicator variables: ”ivar2”, ”ivar3”, ”ivar4”. For more examples: http://mchpappserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1048.

1.7.3 Comparison with other macros

Before concluding the section we would like the reader to be aware that a rich set of alternative programs is also available to implement mediation analyses in certain settings. We believe that our macro provides unique features that may be useful to investigators. At the end of this section table 1.1 compares our macro to some of the existing and popular software tools. Preacher, Hayes et al. developed several macros for mediation mainly implementable in SAS, SPSS and Mplus (indirect, mediate, modmed, medcurve); Imai et al. also developed a macro in R (mediate). We also compare the macros to recent procedures that have been developed in Mplus (Muthn, 2012) in part based on the work we present in this paper. We compare the macros on the basis of certain features. We check whether they provide both direct and indirect effects and if they allow for non-linearities such as interactions, and binary or count variables. We also consider whether they accommodate case-control designs and in which software packages they can be implemented. Our macro, in contrast with Preacher and Hayes', (i) allows for effect decomposition into direct and indirect effects even in the presence of exposure-mediator interaction, (ii) allows for dichotomous mediators and count outcomes, (iii) allows for case-control designs, and (iv) gives estimates with a clear interpretation within the counterfactual framework. In contrast with that of Imai et al. (2010), our macro (i) provides direct and indirect effects on a ratio scale for dichotomous outcomes, (ii) allows for case-control sampling designs,

(iii) is implemented in SAS and SPSS which are more commonly employed in the social sciences. Our macro provides similar features to Mplus which is in part because recent developments in Mplus (Muthn, 2012) were implemented following the results of our paper. Our macro, in contrast to Mplus, allows for case-control designs; Mplus, in contrast to our macro, allows for the flexibility to handle ordinal outcomes.

Table 1.1: Macro Comparison (*Hayes and Preacher **Imai et al. ***Valeri and VanderWeele † Muthen)

	mediation***	mediation**	modmed*	mediate*	Sobel*	Indirect*	medcurve*	Mplus†
Causal Effects								
direct effects	√	√	×	√	×	√	×	√
indirect effects	√	√	√	√	√	√	√	√
Interaction								
M-A	√	√	√	×	×	×	×	√
M-C	×	×	√	×	×	×	×	√
Type of variables								
continuous M	√	√	√	√ (+ M & A)	√	√ (+ M)	√	√
binary M	√	√	×	×	×	×	×	√
continuous Y	√	√	√	√	√	√	√	√
binary Y	√	√	×	√	√	√	×	√
count Y	√	√	×	×	×	×	×	√
ordinal Y	×	√	×	×	×	×	×	√
Additional covariate	√	√	√	√	×	√	√	√
Design								
Cross-Sectional	√	√	√	√	√	√	√	√
Cohort	√	√	√	√	√	√	√	√
Case-Control	√	√	×	×	√	√	×	×
Standard Errors								
delta method	√	×	×	×	×	×	×	√
bootstrap	√	√	√	√	√	√	√	√
Software								
SAS	√	×	×	√	√	√	√	×
SPSS	√	×	√	√	√	√	√	×
R	×	√	×	×	×	×	×	×
MPLUS	×	×	√	×	×	×	×	√

† A number of the recent developments in Mplus were motivated by the results of the present paper

** The Imai et al. (2009, 2010b) macros contain a sensitivity analysis option, Mplus is adding these features in keeping up with the literature and our macros will eventually have these features as well.

1.8 Description of the SPSS macro

The SPSS macro that we provide, which was developed under the version 19.0, performs exactly the same tasks described in the previous section for the SAS macro. However, we point out some small differences that the investigator has to take into account when running mediation analysis using SPSS software.
Before invoking the mediation macro the user has to open a new SPSS session and needs to specify the path in which he or she wants to save relevant estimates from the mediator and outcome regressions. This is simply done by running this command:

```
DEFINE !path()"C:\ "!ENDDEFINE.
```

In between the quotation marks the path is defined, here for example the path "C:\" has been entered. For SPSS users, macro activation requires that the macro script is then saved as a syntax file (the syntax file should be called from the session that has just been opened) and information is input in the following statement:

```
mediation data= / yvar= /avar= /mvar= /cvar= /NC= /a0= /a1=
/m= /yreg= /mreg= /interaction=
[/casecontrol= /Output= /c=]
```

First one inputs the name of the dataset (including the path, e.g. $data =" C : .sav"$), then the name of the outcome variable ($yvar =$), the treatment variable (avar=), the mediator variable ($mvar =$), and the other covariates ($cvar =$). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro $dummit$ can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* ($a0 =$), the new exposure level a ($a1 =$), the level of mediator m at which the controlled direct effect is to be estimated and the number of covariates to be used ($nc =$). When no covariates are entered, then the user still needs to write the command $cvar =$ and needs to specify that $nc = 0$. The user must also specify which types of

regression have to be implemented. In particular, $LINEAR$, $LOGISTIC$, $LOGLINEAR$, $POISSON$ or $NEGBIN$ can be specified in the option $yreg$. Logistic links for $yreg$ can be used for rare dichotomous outcomes; otherwise for dichotomous outcomes that are not rare, log links should be used for the outcome regression and the effects are then given on the risk ratio scale. For the option mreg either $LINEAR$ or $LOGISTIC$ regressions are allowed. If the dataset contains missing data the macro implements a complete case only analysis.

Finally, the analyst needs to specify whether an exposure-mediator interaction is present ($TRUE$ or $FALSE$). As optional inputs, the investigator can use the option $casecontrol = TRUE$, when the data arise from a case-control study and the outcome is rare. More complete output (described in the previous section) can be obtained using the option $Output = FULL$ and entering the values for the covariates at which to compute causal effects conditional on those covariate values ($c =$). In order to enter the covariate values the investigator needs to create a separate dataset that contains those values. For example, if two covariates C are present in the model and the value at which the investigator wants to fix the first is 4 and the value at which the investigator wants to fix the second is 10, at the beginning of the script the following commands need to be run:

```
Matrix.
compute c=make(1,2,0).
compute c(1,1)=4.
compute c(1,2)=10.
SAVE {c(1,:)} /OUTFILE="C:\c.sav".
end matrix.
```

After having created dataset for the covariate values, the user can specify the option $Output = FULL/c =$ "C : ¸sav" to obtain the more complete output. If the investigator wishes to obtain bootstrap standard errors, he or she can use the option $boot = true$ followed by the number of observations in the dataset ($nobs =$) to compute causal effects and standard errors with 1,000 bootstrap replications (or "$boot = n$", where n is the

desired number of bootstrap samples). Otherwise delta method standard errors is the default option. As we mentioned in the previous section, if the investigator needs to add a categorical variable as covariate, a series of indicator variables needs to be generated. The SPSS macro dummit works very similarly to the SAS macro. In particular the investigator needs to call the macro followed by three parentheses. In the first parenthesis the number of levels is entered, in the second parenthesis the name of the variable needs to be specified. Finally, in the third parenthesis, the prefix for the new variables is entered. For example if the variable we need to recode is ”smoking” which takes levels ”never”, ”past”, ”current”. Then we can run the following macro:

```
dummit (3) (smoking) (smoke)
```

This macro would generate the following variables: ”smokedum2”, ”smokedum3”. The category ”never” is automatically taken as a reference. More examples can be found following the link: http://www.glennlthompson.com/?p=92.

1.9 Example

We present in this section an example of using the mediation macro. We implement the analyses on a modified version of the fictitious dataset used by Preacher and Hayes (2004) to explain their Sobel macro. The interest lies in the effects of a new cognitive therapy on life satisfaction after retirement. Residents of a retirement home diagnosed as clinically depressed are randomly assigned to receive 10 sessions of a new cognitive therapy ($A = 1$) or 10 sessions of an alternative therapeutic method ($A = 0$). After Session 8, the positivity of the evaluation the residents make for a recent failure experience is assessed (M). Finally, at the end of Session 10, the residents are given a questionnaire to measure life satisfaction (Y). The question is whether the cognitive therapy's effect on life satisfaction is mediated by the positivity of their attributions of negative experiences.
The new dataset that we employ differs with respect to Preacher and Hayes' one only

in the way in which the outcome is simulated. In particular, the exposure and mediator variables are the same but now the outcome is simulated as a normally distributed variable with mean equal to the linear regression estimated with the original data (the coefficients given in the outcome regression in Preacher and Hayes, 2004) plus a new term, the exposure-mediator interaction term, with coefficient equal to $\theta_3 = 0.5$ indicating a weak positive interaction, and standard deviation equal to the standard error of the residuals obtained from the outcome regression using Preacher and Hayes data (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html).

We first consider the case in which the interaction between the therapy and the attributions of negative experiences is omitted by the investigator. After having saved the dataset and inserted macro script we run the following command:

```
%mediation(data=dat,yvar=satis,avar=therapy,mvar=attrib,cvar=,a0=0,
a1=1,m=0,nc=,yreg=linear,mreg=linear,interaction=false)
run;
```

The first output provided is the results of the outcome and mediator regressions:

Dependent Variable: satis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.71479	0.20449	-3.50	0.0017
therapy	1	0.66788	0.30147	2.22	0.0354
attrib	1	0.67186	0.16923	3.97	0.0005

Dependent Variable: attrib

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.35357	0.21837	-1.62	0.1166
therapy	1	0.81857	0.29902	2.74	0.0106

Then the direct effects and indirect effects follow. We give the reduced output, which provides estimates for the controlled direct effect, the natural indirect effect, and the total effect:

Obs	Effect	Estimate	s_e_	p_value	lower	upper
1	cde=nde	0.66788	0.30147	0.026733	0.07700	1.25877
2	nie	0.54997	0.24403	0.024215	0.07167	1.02827
3	total effect	1.21785	0.33475	0.000275	0.56174	1.87396

We then run the mediation macro with the correctly specified outcome regression model that includes the exposure-mediator interaction term. We type the following command:

```
%mediation(data=dat,yvar=satis,avar=therapy,mvar=attrib,cvar=,a0=0,
a1=1,m=0,nc= ,yreg=linear,mreg=linear,interaction=true)
run;
```

The output from the outcome regression is the following (the mediator regression will be the same):

Dependent Variable: satis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.84424	0.19646	-4.30	0.0002
therapy	1	0.62132	0.27901	2.23	0.0348
attrib	1	0.30575	0.21913	1.40	0.1747
int	1	0.74464	0.31251	2.38	0.0248

We obtain the following estimates for the effects:

Obs	Effect	Estimate	s_e_	p_value	_95__CI_ lower	_95__CI_ upper
1	cde	0.62132	0.27901	0.02596	0.07446	1.16818
2	nde	0.35804	0.34759	0.30298	-0.32323	1.03931
3	nie	0.85981	0.28782	0.00281	0.29568	1.42395
4	marginal total effect	1.21785	0.33407	0.00027	0.56307	1.87263

We can see how the estimate of the indirect effect is downward biased and is less significant if the interaction term is omitted. Moreover, when the interaction term is correctly added in the model, controlled direct effects and natural direct effects differ.

1.10 Discussion

With the present work we have provided several contributions that will likely be important for research in psychology and in the social and biomedical sciences. First, by using a counterfactual approach for the definition of the causal effects of interest, along with their identifiability conditions, we give the reader some intuitive rules allowing for causal interpretation in mediation analysis. Issues of identification and causal

interpretation have often been neglected when using the Baron and Kenny approach and other traditional approaches; the overview here will hopefully guide researchers in thinking about these questions. Second, we have described how progress in mediation analysis can be made in the case in which exposure-mediator interaction is present and we have derived new formulas in the appendix for settings with a binary mediator allowing for exposure-mediator interactions. We have also extended this approach to count outcomes. Third, the investigator who wishes to pursue mediation analysis using regression models will find useful resources in the SAS and SPSS macro that we developed. These macros target the implementation of mediation analysis allowing for the presence of exposure-mediator interaction. The macro was created by applying and extending the work on identification and estimation of direct and indirect causal effects of VanderWeele and Vansteelandt (2009, 2010). We provided a table that summarizes the features of the most popular existing macros for mediation. The current macro also allows for binary and count data as outcomes and provides valid estimation under case-control designs provided the outcome is rare.

Mediation analysis from a counterfactual perspective with exposure-mediator interaction can also be performed in R and STATA using the macro provided by Imai et al. (2010a, 2010b). Their approach to mediation analysis relies on Monte Carlo methods. However, the connections to product method and other popular methods in mediation analysis are clearer with the regression-based approach we have presented in that we have provided analytic formulae for the direct and indirect effects and these formulae coincide with the product method when there are no interactions.

The reader should note that if interactions between exposure or mediator and additional covariates (C) are present, these might need to be included in order to have a correctly specified model. However, the identifiability conditions that we described above under the counterfactual framework are applicable also to these more complex models. An investigator can still pursue mediation analysis with these different models, but new formulas for the direct and indirect effects defined above would have to be derived. The

derivations in the online appendix provide a template that could be used to derive these new formulas for the direct and indirect effects and their standard errors in other types of models that may include interactions between covariates and treatment or mediator or quadratic terms.

Finally we emphasize that the investigator needs to take particular care in controlling for mediator-outcome confounding. The estimates from the product method or difference method or our approach will be biased if control is not made for these variables. Mediator-outcome confounding can be present even if the exposure is randomized (since the mediator is not randomized). Unfortunately, this point was not made in the popular Baron and Kenny (1986) paper, though it was made by Judd and Kenny (1981) five years earlier and it has now been emphasized and clarified in the causal inference literature and is being emphasized again in psychology. Psychologists, social scientists, and biomedical researchers need to take this assumption seriously if they hope to obtain valid conclusions about direct and indirect effects. If the investigator thinks that unmeasured confounding may be present, sensitivity analysis should be used (VanderWeele, 2010b; Imai et al. 2010a). We hope to automate sensitivity analysis in the macro in future work.

1.A Definition of causal effects and Identifiability conditions

We let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a. We let Y_{am} denote the value of the outcome that would have been observed had the exposure, A, and mediator, M, been set to levels a and m, respectively.

The average controlled direct effect comparing exposure level a to a^* and fixing the mediator to level m is defined by $CDE_{a,a^*}(m) = E[Y_{am} - Y_{a^*m}]$. The average natural direct effect is then defined by $NDE_{a,a^*}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$. The average natural indirect effect can be defined as $NIE_{a,a^*}(a) = E[Y_{aM_a} - Y_{aM_{a^*}}]$, which compares the effect of

the mediator at levels M_a and M_{a^*} on the outcome when exposure A is set to a. Controlled direct effects and natural direct and indirect effects within strata of $C = c$ are then defined by: $CDE_{a,a^*|c}(m) = E[Y_{am} - Y_{a^*m}|c]$, $NDE_{a,a^*|c}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]$ and $NIE_{a,a^*|c}(a) = E[Y_{aM_a} - Y_{aM_{a^*}}|c]$ respectively.

For a dichotomous outcome the total effect on the odds ratio scale conditional on $C = c$ is given by $OR^{TE}_{a,a^*|c} = \frac{P(Y_a=1|c)/\{1-P(Y_a=1|c)\}}{P(Y_{a^*}=1|c)/\{1-P(Y_{a^*}=1|c)\}}$. The controlled direct effect on the odds ratio scale is given by $OR^{CDE}_{a,a^*|c}(m) = \frac{P(Y_{am}=1|c)/\{1-P(Y_{am}=1|c)\}}{P(Y_{a^*m}=1|c)/\{1-P(Y_{a^*m}=1|c)\}}$. The natural direct effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}{P(Y_{a^*M_{a^*}}=1|c)/\{1-P(Y_{a^*M_{a^*}}=1|c)\}}$. The natural indirect effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NIE}_{a,a^*|c}(a) = \frac{P(Y_{aM_a}=1|c)/\{1-P(Y_{aM_a}=1|c)\}}{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}$.

As discussed in the text, identification assumptions (i)-(iv) will suffice to identify these direct and indirect effects. If we let $X \perp Y|Z$ denote that X is independent of Y conditional on Z then these four identification assumptions can be expressed formally in terms of counterfactual independence as (i) $Y_{am} \perp A|C$, (ii) $Y_{am} \perp M|\{A, C\}$, (iii) $M_a \perp A|C$, and (iv) $Y_{am} \perp M_{a^*}|C$. Assumptions (i) and (ii) suffice to identify controlled direct effects; assumptions (i)-(iv) suffice to identify natural direct and indirect effects (Pearl, 2001; VanderWeele and Vansteelandt, 2009). The intuitive interpretation of these assuptions as described in the text follows from the theory of causal diagrams (Pearl, 2001). Alternative identification assumptions have also been proposed (Imai 2010a; Hafeman and VanderWeele, 2011). However, it has been shown that the intuitive graphical interpretation of these alternative assumptions are in fact equivalent (Shpitser and VanderWeele, 2011). Technical examples can be constructed where one set of identifiation assumptions holds and another does not, but on a causal diagram corresponding to a set of non-parametric structural equations, whenever one set of the assumptions among those in VanderWeele and Vansteelandt (2009), Imai (2010a), and Hafeman and VanderWeele (2011) holds, the others will also.

1.B Continuous Mediator and Outcome

Effects using regression

Suppose that both the mediator and the outcome are continuous and that the following models fit the observed data:

$$E(M|A = a, C = c) = \beta_0 + \beta_1 a + \beta_2^{'} c \tag{1.7}$$

$$E(Y|A = a, M = m, C = c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c \tag{1.8}$$

If the covariates C satisfied the no-unmeasured confounding assumptions (i)-(iv) above, then the average controlled direct effect and the average natural direct and indirect effects were derived by VanderWeele and Vansteelandt, 2009.

In particular, if the regression models (1.7) and (1.8) are correctly specified and assumptions of no unmeasured confounding of exposure-outcome relationship (i) and no unmeasured confounding of the mediator-outcome relationship (ii) hold, then we could compute the controlled direct effect as follows:

$$\begin{aligned}
&CDE = E[Y_{am} - Y_{a^*m}|C = c] \\
&= E[Y|C = c, A = a, M = m] - E[Y|C = c, A = a^*, M = m] \\
&= (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c) - (\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4^{'} c) \\
&= (\theta_1 a + \theta_3 am - \theta_1 a^* - \theta_3 a^* m) \\
&= \theta_1(a - a^*) + \theta_3 m(a - a^*).
\end{aligned}$$

If the regression models (1.7) and (1.8) are correctly specified and assumptions (i) and (ii) together with two additional assumptions of (iii) no unmeasured confounding of the exposure-mediator relationship and (iv) that there is no mediator-outcome confounder that is affected by the exposure hold, then we could compute the natural direct effects by:

$NDE = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|C = c]$
$= \sum_m \{E[Y|C = c, A = a, M = m] - E[Y|C = c, A = a^*, M = m]\} \times P(M = m|C = c, A = a^*)$
$= \sum_m \{(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) - (\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4' c)\} \times P(M = m|C = c, A = a^*)$
$= \sum_m \{(\theta_1 a + \theta_2 m + \theta_3 am) - (\theta_1 a^* + \theta_2 m + \theta_3 a^* m)\} \times P(M = m|C = c, A = a^*)$
$= \{(\theta_1 a + \theta_2 E[M|A = a^*, C = c] + \theta_3 a E[M|A = a^*, C = c]) - (\theta_1 a^* + \theta_2 E[M|A = a^*, C = c] + \theta_3 a^* E[M|A = a^*, C = c])\}$
$= \{(\theta_1 a + \theta_2(\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_3 a(\beta_0 + \beta_1 a^* + \beta_2' c) - (\theta_1 a^* + \theta_2(\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_3 a^*(\beta_0 + \beta_1 a^* + \beta_2' c))\}$
$= \{\theta_1 a + \theta_3 a(\beta_0 + \beta_1 a^* + \beta_2' c) - (\theta_1 a^* + \theta_3 a^*(\beta_0 + \beta_1 a^* + \beta_2' c))\}$
$= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' c)(a - a^*).$

Moreover under the same assumptions we can compute the natural indirect effects by:

$NIE = E[Y_{aM_a} - Y_{aM_{a^*}}|C = c]$
$= \sum_m E[Y|C = c, A = a, M = m] \times P(M = m|C = c, A = a) - \sum_m E[Y|C = c, A = a, M = m] \times P(M = m|C = c, A = a^*)$
$= \sum_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) \times P(M = m|C = c, A = a) - \sum_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) \times P(M = m|C = c, A = a^*)$
$= (\theta_0 + \theta_1 a + \theta_2 E[M|A = a, C = c] + \theta_3 a E[M|A = a, C = c] + \theta_4' c) - (\theta_0 + \theta_1 a + \theta_2 E[M|A = a^*, C = c] + \theta_3 a^* E[M|A = a^*, C = c] + \theta_4' c)$
$= (\theta_1 a + \theta_2(\beta_0 + \beta_1 a + \beta_2' c) + \theta_3 a(\beta_0 + \beta_1 a + \beta_2' c)) - (\theta_1 a^* + \theta_2(\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_3 a^*(\beta_0 + \beta_1 a^* + \beta_2' c))$
$= (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*).$

If the regression models (1.7) and (1.8) are correctly specified and assumptions (i) and (ii) hold, then we could compute the total effect by:

$$TE = E[Y_a - Y_{a^*}|C = c]$$
$$= E[Y_{a,M(a^*)} - Y_{a^*,M(a^*)}|C = c] + E[Y_{a,M(a)} - Y_{a^*,M(a^*)}|C = c]$$
$$= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*).$$

Finally if the regression models (1.7) and (1.8) are correctly specified and assumptions (i)-(iv) hold then we could compute the proportion mediated by:

$$PM = \frac{E[Y_{aM_a} - Y_{a^*M_{a^*}}|C=c]}{E[Y_a - Y_{a^*}|C=c]}$$
$$= \frac{\theta_2\beta_1 + \theta_3\beta_1 a}{\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a}.$$

Standard errors

Suppose that model (1.7) and (1.8) have been fit using standard linear regression software and that the resulting estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2^{'})^{'}$ and $\hat{\theta}$ of $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4^{'})^{'}$ have covariance matrices Σ_β and Σ_θ. Then the covariance matrix of $(\hat{\beta}^{'}, \hat{\theta}^{'})$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{bmatrix}$$

Standard errors of the controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma\Sigma\Gamma'}|a - a^*|$$

with $\Gamma = (0, 0, 0^{'}, 0, 1, 0, m, 0^{'})$ for the controlled direct effect, $\Gamma = (\theta_3, \theta_3 a^*, \theta_3 c^{'}, 0, 1, \beta_0 + \beta_1 a^* + \beta_2^{'}c, 0^{'})$ for the pure natural direct effect (same expression holds for the total natural direct effect upon substituting a and a^*), $\Gamma = (0, \theta_2 + \theta_3 a, 0^{'}, 0, 0, \beta_1, \beta_1 a, 0^{'})$ for the total natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*), $\Gamma = (\theta_3, \theta_3(a + a^*) + \theta_2, \theta_3 c^{'}, 0, 1, \beta_1, \beta_0 + \beta_1(a + a^*) + \beta_2^{'}c, 0^{'})$ for the total effect and for the proportion mediated $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ where

$$d_1 = -\theta_3 \frac{\theta_2\beta_1 + \theta_3\beta_1 a}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_2 = \frac{(\theta_2 + \theta_3 a)(-(\theta_2\beta_1 + \theta_3\beta_1 a) + (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)) - \theta_3 a^*}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_3 = -\frac{\theta_3 c^{'}(\theta_2\beta_1 + \theta_3\beta_1 a)}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_4 = 0$$

$$d_5 = -\frac{\theta_2\beta_1 + \theta_3\beta_1 a}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_6 = \frac{\beta_1(-(\theta_2\beta_1 + \theta_3\beta_1 a) + (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a))}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_7 = \frac{\beta_1 a(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a) - (\beta_0 + \beta_1(a + a^*) + \beta_2^{'}c)(\theta_2\beta_1 + \theta_3\beta_1 a)}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'}c + \theta_2\beta_1 + \theta_3\beta_1 a)^2}$$

$$d_8 = 0^{'}.$$

1.C Continuous Mediator and Binary Outcome

Effects using regression

Suppose that the mediator is continuous and the outcome is binary and is rare. Suppose that the following models fit the observed data:

$$E(M|A = a, C = c) = \beta_0 + \beta_1 a + \beta_2^{'}c \tag{1.9}$$

$$logit\{P(Y=1|A=a,M=m,C=c)\}=\theta_0+\theta_1 a+\theta_2 m+\theta_3 a*m+\theta_4^{'}c \tag{1.10}$$

and that the error term in the regression model for M is normally distributed with mean 0 and variance σ^2. If the regression models (1.9) and (1.10) are correctly specified and assumptions (i) and (ii) hold then the conditional controlled direct effect on the odds ratio scale would be given by (VanderWeele and Vansteelandt, 2010):

$OR^{CDE}=\frac{P(Y_{am}=1|c)/(1-P(Y_{am}=1|c))}{P(Y_{a^*m}=1|c)/(1-P(Y_{a^*m}=1|c))}$
$=\frac{P(Y=1|a,m,c)/(1-P(Y=1|a,m,c))}{P(Y=1|a^*,m,c)/(1-P(Y=1|a^*m,c))}$

$=\frac{exp[\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4^{'}c]}{exp[\theta_0+\theta_1 a^*+\theta_2 m+\theta_3 a^* m+\theta_4^{'}c]}$
$=exp[(\theta_1+\theta_3 m)(a-a^*)].$

If the regression models (1.9) and (1.10) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, and the error term for linear regression model ((2.12)) is normally distributed and has constant variance σ^2, then we could compute the natural direct effects by:

$OR^{NDE}=exp[log\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}]$
$=exp[logit\{P(Y_{aM_{a^*}}=1|c)\}-logit\{P(Y_{a^*M_{a^*}}=1|c)\}]$
$\sim exp[\theta_0+\theta_1 a+\theta_4^{'}c+(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2^{'}c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2-\{\theta_0+\theta_1 a^*+\theta_4^{'}c+(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2^{'}c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2\}]$
$=exp[\{\theta_1+\theta_3(\beta_0+\beta_1 a^*+\beta_2^{'}c+\theta_2\sigma^2)\}(a-a^*)+0.5\theta_3^2\sigma^2(a^2-a^{*2})].$

If the regression models (1.9) and (1.10) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, and the error term for linear regression model (1.9) is normally distributed and has constant variance σ^2, then we could compute the natural indirect effects by:

$$OR^{NIE} = exp[log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}]$$
$$= exp[logit\{P(Y_{aM_a}=1|c)\} - logit\{P(Y_{aM_{a^*}}=1|c)\}]$$
$$\sim exp[\theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2 - \{\theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2\}]$$
$$= exp[(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)].$$

If the regression models (1.9) and (1.10) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, and the error term for linear regression model (1.9) is normally distributed and has constant variance σ^2, then we could compute the total effects by:

$$OR^{TE} = exp[log\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}] \times exp[log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}]$$

$$= E[Y_{a,M_{a^*}} - Y_{a^*,M_{a^*}}|C=c] \times E[Y_{a,M_a} - Y_{a^*,M_{a^*}}|C=c]$$
$$= exp[(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' c + \theta_2\beta_1 + \theta_3\beta_1 a + \theta_3\theta_2\sigma^2)(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})].$$

If the regression models (1.9) and (1.10) are correctly specified and assumptions (i)-(iv) hold then we can compute the proportion mediated by:

$$PM = \frac{log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}}{log\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\} + log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}}$$
$$= \frac{(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)}{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' c + \theta_2\beta_1 + \theta_3\beta_1 a + \theta_3\theta_2\sigma^2)(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})}.$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the outcome model; the direct and indirect effect will have now an interpretation on the risk ratio scale rather than on the odds ratio scale.

These expressions apply also if the outcome is a count variable. In particular if $Y \sim Poi(\lambda)$ for $\lambda = \exp\{\theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c\})$ the outcome regression can be defined as:

$$log\{E(Y|A = a, M = m, C = c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c$$

The natural direct effect for binary outcome on the risk ratio scale coincides with the natural direct effect for poisson count outcome since:

$$RR^{NDE} = exp[log\{\frac{E(Y_{aM_{a^*}}|c)}{E(Y_{a^*M_{a^*}}|c)}\}]$$

The same argument holds for the natural indirect effect. Finally, the argument can be extended to the case in which the count outcome is modeled with a negative binomial distribution. This is the case since the negative binomial distribution can be represented as an over-dispersed poisson and the mean of the two models coincide.

Standard errors

We now consider standard errors for the controlled direct effect and natural direct and indirect effect odds ratios. Suppose that model (1.10) has been fit using standard logistic regression software and that model (1.9) has been fit using standard linear regression software. Suppose furthermore that the resulting estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2^{'})^{'}$, $\hat{\theta}$ of $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4^{'})^{'}$ and $\hat{\sigma}^2$ of σ have covariance matrices Σ_β and Σ_θ. Then the covariance matrix of $(\hat{\beta}^{'}, \hat{\theta}^{'}, \hat{\sigma}^2)$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta & 0 \\ 0 & 0 & \Sigma_{\sigma^2} \end{bmatrix}$$

Standard errors of the controlled and natural direct and indirect effects can be obtained

(using the delta method) as

$$\sqrt{\Gamma\Sigma\Gamma'}|a-a^*|$$

with $\Gamma = (0, 0, 0^{'}, 0, 1, 0, m, 0^{'}, 0)$ for the log of controlled direct effect odds ratio, $\Gamma = (\theta_3, \theta_3 a^*, \theta_3 c^{'}, 0, 1, \theta_3\sigma^2, \beta_0 + \beta_1 a^* + \beta_2^{'}c + \theta_2\sigma^2 + \theta_3\sigma^2(a+a^*), 0^{'}, \theta_2\theta_3 + 0.5\theta_3^2(a+a^*))$ for the log pure natural direct effect odds ratio (same expression holds for the total natural direct effect upon substituting a and a^*), $\Gamma = (0, \theta_2 + \theta_3 a, 0^{'}, 0, 0, \beta_1, \beta_1 a, 0^{'}, 0)$ for the log of total natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*), $\Gamma = (\theta_3, \theta_3(a+a^*) + \theta_2, \theta_3 c^{'}, 0, 1, \theta_3\sigma^2 + \beta_1, \beta_0 + \beta_1(a+a^*) + \beta_2^{'}c + \theta_2\sigma^2 + \theta_3\sigma^2(a^2 - a^{*2}), 0^{'}, 0.5\theta_3^2(a^2 - a^{*2}))$ for the logarithm of the total effect. Standard errors for the proportion mediated can be obtained (using the delta method) as

$$\sqrt{\Gamma\Sigma\Gamma'}$$

where $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9)$.

Let

$$A = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

$$B = [\{\theta_1 + \theta_3(\beta_0 + \beta_1(a+a^*) + \beta_2^{'}c + \theta_2\sigma^2) + \beta_1\theta_2\}(a-a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})]$$

$$d_1 = -\frac{\theta_3(a-a^*)A}{B^2}$$

$$d_2 = \frac{(\theta_2 + \theta_3 a)(a-a^*)B - (\theta_3(a+a^*) + \theta_2)(a-a^*)A}{B^2}$$

$$d_3 = -\frac{\theta_3^{'}c(a-a^*)A}{B^2}$$

$$d_4 = 0$$

$$d_5 = -\frac{A(a-a^*)}{B^2}$$

$$d_6 = \frac{\beta_1(a-a^*)B - (\theta_2\sigma^2 + \beta_1)(a-a^*)A}{B^2}$$

$$d_7 = \frac{\beta_1 a(a-a^*)B + (\beta_0 + \beta_1(a+a^*) + \beta_2^{'}c + \theta_2\sigma^2)(a-a^*) - (\theta_2\sigma^2)(a-a^*)A}{B^2}$$

$$d_8 = 0^{'}$$

$$d_9 = -\frac{[\theta_3\theta_2(a-a^*) + 0.5\theta_3^2(a^2 - a^{*2})]A}{B^2}$$

1.D Binary Mediator and Continuous Outcome

Effects using regression

Suppose that the outcome is continuous, the mediator is binary and that the following models fit the observed data:

$$logit\{P(M=1|A=a, C=c)\} = \beta_0 + \beta_1 a + \beta_2^{'}c \tag{1.11}$$

$$E(Y|A=a, M=m, C=c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'}c \tag{1.12}$$

In particular, if the regression models (1.11) and (1.12) are correctly specified and assumptions (i) and (ii) hold then we could compute the average controlled direct effect as in section 1

If the regression models (1.11) and (1.12) are correctly specified and assumptions (i)-(iv) hold then we could compute the average natural direct effects by:

$NDE = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|C=c]$

$= \sum_m \{E[Y|C=c, A=a, M=m] - E[Y|C=c, A=a^*, M=m]\} \times P(M=m|C=c, A=$

$a^*)$
$= \sum_m \{(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c) - (\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4^{'} c)\} \times P(M = m|C = c, A = a^*)$
$= \sum_m \{(\theta_1 a + \theta_2 m + \theta_3 am) - (\theta_1 a^* + \theta_2 m + \theta_3 a^* m)\} \times P(M = m|C = c, A = a^*)$
$= \{\theta_1(a - a^*)\} + \{\theta_3(a - a^*)\} \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}.$

If the regression models (1.11) and (1.12) are correctly specified and assumptions (i)-(iv) hold then we could compute the average natural indirect effects by:

$NIE = E[Y_{aM_a} - Y_{aM_{a^*}}|C = c]$
$= \sum_m E[Y|C = c, A = a, M = m] \times P(M = m|C = c, A = a) - \sum_m E[Y|C = c, A = a, M = m] \times P(M = m|C = c, A = a^*)$
$= \sum_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c) \times P(M = m|C = c, A = a) - \sum_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c) \times P(M = m|C = c, A = a^*)$
$= (\theta_2 + \theta_3 a)\{E[M|A = a, C = c] - E[M|A = a^*, C = c]\}$
$=(\theta_2 + \theta_3 a)\{\frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]} - \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}\}.$

If the regression models (1.11) and (1.12) are correctly specified and assumptions (i)-(iv) hold then we could compute the total effect by:

$TE = E[Y_a - Y_{a^*}|C = c]$
$= E[Y_{aM_{a^*}} - Y_{a^* M_{a^*}}|C = c] + E[Y_{aM_a} - Y_{a^* M_{a^*}}|C = c]$
$= \{\theta_1(a - a^*)\} + \{\theta_3(a - a^*)\} \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]} + (\theta_2 + \theta_3 a)\{\frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]} - \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}\}.$

If the regression models (1.11) and (1.12) are correctly specified and assumptions (i)-(iv) hold then we could compute the proportion mediated by:

$$PM = \frac{E[Y_{aM_a} - Y_{a^*M_{a^*}}|C=c]}{E[Y_a - Y_{a^*}|C=c]}$$

$$= \frac{(\theta_2+\theta_3 a)\{\frac{exp[\beta_0+\beta_1 a+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]} - \frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}}{(\theta_2+\theta_3 a)\{\frac{exp[\beta_0+\beta_1 a+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]} - \frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\} + \{\theta_1(a-a^*)\} + \{\theta_3(a-a^*)\}\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}}.$$

Standard errors

Suppose that model (1.12) have been fit using standard linear regression software and that model (1.11) have been fit using standard logistic regression. The resulting estimates are $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2^{'})^{'}$and $\hat{\theta}$ of $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4^{'})^{'}$ have covariance matrices Σ_β and Σ_θ. Then the covariance matrix of $(\hat{\beta}^{'}, \hat{\theta}^{'})$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{bmatrix}$$

Standard errors of the controlled and natural direct can be obtained (using the delta method) as

$$\sqrt{\Gamma\Sigma\Gamma^{'}}|a - a^*|$$

with $\Gamma = (0, 0, 0^{'}, 0, 1, 0, m, 0^{'})$ for the controlled direct effect, $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the pure natural direct effect (same expression holds for the total natural direct effect upon substituting a and a^*), where

$$d_1 = \frac{\theta_3 exp[[\beta_0 + \beta_1 a^* + \beta_2^{'}c](1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]) - \theta_3\{exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]\}^2}{(1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c])^2}$$

$$d_2 = \frac{\theta_3 a^* exp[[\beta_0 + \beta_1 a^* + \beta_2^{'}c](1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]) - \{exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]\}^2}{(1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c])^2}$$

$$d_3 = \frac{\theta_3 c^{'} exp[[\beta_0 + \beta_1 a^* + \beta_2^{'}c](1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]) - \{exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c]\}^2}{(1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'}c])^2}$$

$$d_4 = 0$$

$$d_5 = 1$$

$$d_6 = 0$$

$$d_7 = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}$$

$$d_8 = 0^{'}$$

Standard errors of the natural indirect can be obtained (using the delta method) as

$$\sqrt{\Gamma \Sigma \Gamma^{'}}$$

For the natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*) let

$$A = \frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\} - \{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}{\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}$$

$$B = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\} - \{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}{\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}$$

$$K = \frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$D = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

and

$\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$, where

$$d_1 = \{\theta_2 + \theta_3 a\}[A - B]$$

$$d_2 = \{\theta_2 + \theta_3 a\}[aA - a^* B]$$

$$d_3 = \{\theta_2 + \theta_3 a\} c^{'} [A - B]$$

$$d_4 = 0$$

$$d_5 = 0$$

$$d_6 = K - D$$

$$d_7 = a[K - D]$$

$$d_8 = 0^{'}$$

Standard errors of the controlled and total effect and percentage mediated can be obtained (using the delta method) as

$$\sqrt{\Gamma \Sigma \Gamma'}$$

let

$$A = \frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\} - \{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}{\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}$$

$$B = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\} - \{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}{\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}$$

$$K = \frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$D = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

for the total effect $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$, where

$$d_1 = \theta_3 (a - a^*) B + (\theta_2 + \theta_3 a)(A - B)$$

$$d_2 = a^* \theta_3 (a - a^*) B + (\theta_2 + \theta_3 a)(aA - a^* B)$$

$$d_3 = c^{'} \theta_3 (a - a^*) B + (\theta_2 + \theta_3 a)(A - B)$$

$$d_4 = 0$$

$$d_5 = a - a^*$$

$$d_6 = K - D$$

$$d_7 = (a - a^*)D + a[K - D]$$

$$d_8 = 0^{'}$$

and for the proportion mediated $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ where

$$d_1 = \frac{[(\theta_2+\theta_3 a)(A-B)]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}-\{[(\theta_2+\theta_3 a)(A-B)]+(a-a^*)\theta_3 B\}(\theta_2+\theta_3 a)[K-D]}{\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}^2}$$

$$d_2 = \frac{[(\theta_2+\theta_3 a)(aA-a^*B)]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}-\{[(\theta_2+\theta_3 a)(aA-a^*B)]+a^*(a-a^*)\theta_3 B\}(\theta_2+\theta_3 a)[K-D]}{\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}^2}$$

$$d_3 = \frac{[(\theta_2+\theta_3 a)c^{'}(A-B)]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}-c^{'}\{[(\theta_2+\theta_3 a)(A-B)]+(a-a^*)\theta_3 B\}(\theta_2+\theta_3 a)[K-D]}{\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}^2}$$

$$d_4 = 0$$

$$d_5 = \frac{(a - a^*)(\theta_2 + \theta_3 a)[K - D]}{\{(\theta_2 + \theta_3 a)(K - D) + (a - a^*)[\theta_1 + \theta_3 D]\}^2}$$

$$d_6 = \frac{a[K-D]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}-[K-D]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}}{\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}^2}$$

$$d_7 = \frac{[K-D]\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}-\{a[K-D]+(a-a^*D)\}\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}}{\{(\theta_2+\theta_3 a)(K-D)+(a-a^*)[\theta_1+\theta_3 D]\}^2}$$

$$d_8 = 0^{'}.$$

1.E Binary Mediator and Binary Outcome

Effects using regression

Suppose that both the outcome and the mediator are binary and that the following models fit the observed data:

$$logit\{P(M = 1|A = a, C = c)\} = \beta_0 + \beta_1 a + \beta_2^{'} c \tag{1.13}$$

$$logit\{P(Y = 1|A = a, M = m, C = c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c \tag{1.14}$$

If the regression models (1.13) and (1.14) are correctly specified and assumptions (i) and (ii) hold then we can compute the controlled direct effect odds ratio as the case in which

the mediator is continuous and the outcome is binary.

If the regression models (1.13) and (1.14) are correctly specified and assumptions (i)-(iv) hold and the outcome Y is rare, then we could compute the average natural direct effects by:

$$
\begin{aligned}
&OR^{NDE} = exp[log\{\tfrac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}] \\
&= exp[logit\{P(Y_{aM_{a^*}}=1|c)\} - logit\{P(Y_{a^*M_{a^*}}=1|c)\}] \\
&\sim exp[log\{\tfrac{exp(\theta_0+\theta_1 a+\theta_4^{'}c)+exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}C]}\} - log\{\tfrac{exp(\theta_0+\theta_1 a^*+\theta_4^{'}c)+exp(\theta_0+\theta_1 a^*+\theta_4^{'}c+\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}] \\
&= \{\tfrac{exp[\theta_0+\theta_1 a+\theta_4^{'}c]+exp[\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c]}{exp[\theta_0+\theta_1 a^*+\theta_4^{'}c]+exp(\theta_0+\theta_1 a^*+\theta_4^{'}c+\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c)}\}
\end{aligned}
$$

$$
= \{\tfrac{exp[\theta_1 a](1+exp[\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c])}{exp[\theta_1 a^*](1+exp[\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c])}\}.
$$

If the regression models (1.13) and (1.14) are correctly specified and assumptions (i)-(iv) hold and the outcome Y is rare, then we could compute the average natural indirect effects by:

$$
\begin{aligned}
&OR^{NIE} = exp[log\{\tfrac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}] \\
&= exp[logit\{P(Y_{aM_a}=1|c)\} - logit\{P(Y_{aM_{a^*}}=1|c)\}] \\
&\sim exp[log\{\tfrac{exp(\theta_0+\theta_1 a+\theta_4^{'}c)+exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]}\} - log\{\tfrac{exp(\theta_0+\theta_1 a+\theta_4^{'}c)+exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}] \\
&= \tfrac{[1+exp(\beta_0+\beta_1 a^*+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)]}{[1+exp(\beta_0+\beta_1 a+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)]}.
\end{aligned}
$$

If the regression models (1.13) and (1.14) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, then we could compute the total effects by:

$$
OR^{TE} = exp[log\{\tfrac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}] \times exp[log\{\tfrac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}]
$$

$$= \{\frac{exp[\theta_1 a](1+exp[\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c])}{exp[\theta_1 a^*](1+exp[\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c])}\} \times \{\frac{[1+exp(\beta_0+\beta_1 a^*+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)]}{[1+exp(\beta_0+\beta_1 a+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)]}\}.$$

If the regression models (1.13) and (1.14) are correctly specified and assumptions (i)-(iv) hold then we can compute the proportion mediated by:

$$PM = \frac{log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}}{log\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}+log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}}$$
$$= \frac{log[\frac{[1+exp(\beta_0+\beta_1 a^*+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)]}{[1+exp(\beta_0+\beta_1 a+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)]}]}{log[\{\frac{exp[\theta_1 a](1+exp[\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c])}{exp[\theta_1 a^*](1+exp[\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c])}\}\times\{\frac{[1+exp(\beta_0+\beta_1 a^*+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)]}{[1+exp(\beta_0+\beta_1 a+\beta_2^{'}c)][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)]}\}]}.$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the outcome model; the direct and indirect effect will have now an interpretation on the risk ratio scale rather than on the odds ratio scale.
These expressions apply also if the outcome is a count variable. In particular if $Y \sim Poi(\lambda)$ for $\lambda = \exp\{\theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c\}$) the outcome regression can be defined as:

$$log\{E(Y|A=a, M=m, C=c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c$$

The natural direct effect for binary outcome on the risk ratio scale coincides with the natural direct effect for poisson count outcome since:

$$RR^{NDE} = exp[log\{\frac{E(Y_{aM_{a^*}}|c)}{E(Y_{a^*M_{a^*}}|c)}\}]$$

The same argument holds for the natural indirect effect. Finally, the argument can be extended to the case in which the count outcome is modeled with a negative binomial distribution. This is the case since the negative binomial distribution can be represented as an over-dispersed poisson and the mean of the two models coincide.

Standard Errors:

Suppose that model (1.13) and (1.14) have been fit using standard logistic regression software and that the resulting estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}$ of $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$ have covariance matrices Σ_β and Σ_θ. Then the covariance matrix of $(\hat{\beta}', \hat{\theta}')$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{bmatrix}$$

Standard errors of the controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma \Sigma \Gamma'}$$

with $\Gamma = (0, 0, 0', 0, (a - a^*), 0, m(a - a^*), 0')$ for the controlled direct effect, $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the logarithm of the pure natural direct effect (same expression holds for the logarithm of the total natural direct effect upon substituting a and a^*), where let

$$A = \frac{exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c]}{\{1 + exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c]\}}$$

$$B = \frac{exp[\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c]}{\{1 + exp[\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c]\}}$$

and

$$d_1 = A - B$$

$$d_2 = a^*(A - B)$$

$$d_3 = c'(A - B)$$

$$d_4 = 0$$

$$d_5 = (a - a^*)$$

$$d_6 = A - B$$

$$d_7 = aA - a^*B$$

$$d_8 = 0^{'}$$

for the logarithm of the natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*) let

$$A = \frac{exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$B = \frac{exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

$$K = \frac{exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$D = \frac{exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

and

$\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ where

$$d_1 = (D + A) - (K + B)$$

$$d_2 = a^*[D - B] + a[A - K]$$

$$d_3 = c^{'}[(D + A) - (K + B)]$$

$$d_4 = 0$$

$$d_5 = 0$$

$$d_6 = A - B$$

$$d_7 = a[A - B]$$

$$d_8 = 0^{'}$$

Standard errors of the logarithm of the total effect and percentage mediated can be obtained (using the delta method) as

$$\sqrt{\Gamma \Sigma \Gamma^{'}}$$

Let $d_i(log(pnde))$ and $d_i(log(tnie))$ for $i = 1, ..., 8$, the gamma elements derived for the logarithm of the pure natural direct effect and the total natural indirect effect respectively.

For the total effect $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$,where

$$d_1 = d_1(log(pnde)) + d_1(log(tnie))$$

$$d_2 = d_2(log(pnde)) + d_2(log(tnie))$$

$$d_3 = d_3(log(pnde)) + d_3(log(tnie))$$

$$d_4 = d_4(log(pnde)) + d_4(log(tnie))$$

$$d_5 = d_5(log(pnde)) + d_5(log(tnie))$$

$$d_6 = d_6(log(pnde)) + d_6(log(tnie))$$

$$d_7 = d_7(log(pnde)) + d_7(log(tnie))$$

$$d_8 = d_8(log(pnde)) + d_8(log(tnie))$$

and for the proportion mediated $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$. Let

$$d_1 = \frac{d_1(log(tnie)) * log(te) - d_1(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_2 = \frac{d_2(log(tnie)) * log(te) - d_2(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_3 = \frac{d_3(log(tnie)) * log(te) - d_3(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_4 = \frac{d_4(log(tnie)) * log(te) - d_4(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_5 = \frac{d_5(log(tnie)) * log(te) - d_5(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_6 = \frac{d_6(log(tnie)) * log(te) - d_6(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_7 = \frac{d_7(log(tnie)) * log(te) - d_7(log(te)) * log(tnie)}{[log(te)]^2}$$
$$d_8 = \frac{d_8(log(tnie)) * log(te) - d_8(log(te)) * log(tnie)}{[log(te)]^2}.$$

Mediation analysis in generalized linear models when a continuous mediator is measured with error

[1]Linda Valeri, [1]Xihong Lin, [1,2] Tyler J. VanderWeele

[1]Department of Biostatistics, Harvard School of Public Health and
[2]Department of Epidemiology, Harvard School of Public Health

Abstract

Mediation analysis is a popular approach to examine the extent to which the effect of an exposure on an outcome is through an intermediate variable (mediator) and the extent to which the effect is direct. When the mediator is mis-measured the validity of mediation analysis can be severely undermined. In this paper we first study the bias of classical, non-differential measurement error on a continuous mediator in the estimation of direct and indirect causal effects in generalized linear models when the outcome is either continuous or discrete and exposure-mediator interaction may be present. We then develop methods to correct for measurement error. Three correction approaches using method of moments, regression calibration and SIMEX are compared. We apply the proposed method to the Massachusetts General Hospital lung cancer study to evaluate the effect of genetic variants mediated through smoking on lung cancer risk.

Keywords: Asymptotic bias; Measurement error; Mediation analysis; Method of moments; Regression calibration; SIMEX.

2.1 Introduction

Mediation analysis investigates the role of intermediate variables (mediators) in governing an observed relationship between an exposure variable and an outcome variable. Rather than hypothesizing only a direct causal relationship between the independent variable and the dependent variable, a mediational model hypothesizes that the exposure variable causes the mediator variable, which in turn causes the outcome variable (MacKinnon, 2008). The use of mediation analysis in biomedical and social sciences is widespread and has been strongly influenced by the seminal paper of Baron and Kenny (1986). More recently, new advances in mediation analysis have been made by applying the counterfactual framework in this field (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010).

A recent epidemiological study by VanderWeele et al. (2012a) on the etiology of lung cancer motivates the present work. VanderWeele et al. (2012a) investigated the extent to which the effect of genetic variants rs8034191 and rs1051730 on chromosome 15q25.1 on the risk of lung cancer is direct and to what extent that association is mediated by pathways related to smoking behavior. The question was addressed using a case-control study at Massachusetts General Hospital. A potential concern about the validity of their findings arises from the fact that the mediator, measured as self-reported average cigarettes smoked per day, was likely subject to measurement error. It is of interest to understand how sensitive the results of their study are with respect to measurement error in the intermediate variable (smoking), while allowing for gene-environment interaction.

The literature on measurement error in generalized linear models is rich and rapidly evolving. In this study, we extensively use results that have been derived about the consequences of measurement error on parameter estimators in parametric regression models when a covariate is mis-measured (Cochran, 1968; McCallum, 1972; Carroll et al., 2006; Fuller, 2006).

The present work makes two main contributions. First, we study the implications of classical non-differential measurement error in the mediator variable on the validity of

mediation analysis. We derive the asymptotic bias of direct and indirect causal effects estimators in closed form when interaction between exposure and mediator may be present in the outcome model, which follows a generalized linear model (GLM). We demonstrate that even if the error is assumed to be non-differential, regression coefficient estimators obtained in mediation analysis ignoring measurement error can sometimes be severely biased and therefore induce bias in estimation of causal direct and indirect effects.

The second contribution is to propose strategies for measurement error correction that yield consistent or approximately consistent estimators of the direct and indirect causal effects under classical non-differential measurement error model. We propose three different correction approaches coupled with sensitivity analyses when no gold standard or validation samples for the mis-measured mediator are available. In particular, we compare the performance of measurement-error-corrected estimators for direct and indirect causal effects using method of moments (Fuller, 2006; Murad and Freedman, 2007), regression calibration (Spiegelman et al., 1997) and SIMEX (Carroll and Stefansky, 1995; Wang et al., 1997) estimators.

The paper is organized as follows. Section 2.2 discusses some results from mediation analysis and reviews the direct and indirect causal effects. Section 2.3 introduces mediation measurement error models, and studies the asymptotic bias in direct and indirect causal effects when the mediator is measured with error. In Section 2.4 we propose three approaches for measurement error correction and compare their performance in estimating direct and indirect causal effects via a simulation study. In Section 2.5 we apply the proposed methods to the Massachusetts General Hospital (MGH) lung cancer genetic epidemiological study, followed by discussion in Section 2.6.

2.2 Mediation analysis within the counterfactual framework in the absence of measurement error

Let A be an exposure or treatment, Y an outcome, M a mediator and $\boldsymbol{C}$, a k-dimensional vector of covariates. Baron and Kenny (1986) defined, for the case of a continuous medi-

ator and outcome, the following regression models:

$$E(Y|A=a, M=m, \boldsymbol{C}=\boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \boldsymbol{\theta_4'}\boldsymbol{c} \tag{2.1}$$

$$E(M|A=a, \boldsymbol{C}=\boldsymbol{c}) = \beta_0 + \beta_1 a + \boldsymbol{\beta_2'}\boldsymbol{c}. \tag{2.2}$$

They proposed that the causal direct effect of the exposure can be assessed by estimating θ_1 and that the indirect causal effect of the exposure can be assessed by estimating $\beta_1\theta_2$. Using counterfactual definitions of direct and indirect causal effects of the exposure, the approach of Baron and Kenny can be extended to non-linear models and to allow for the presence of exposure-mediator interaction. Let A and C be continuous or categorical and assume M continuous. Assume that the conditional mean, μ, of the outcome Y given the exposure A, the mediator M, and the covariates $\boldsymbol{C}$ follows a generalized linear model (GLM) (McCullagh and Nelder, 1989)

$$g(\mu) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4'}\boldsymbol{c},$$

where $g(\cdot)$ is a monotone link function.

When we have a continuous outcome and mediator, and both are modeled using the linear link, the mediator regression remains as in model (2.2), but the outcome regression, allowing for exposure-mediator interaction, is as follows:

$$E(Y|A=a, M=m, \boldsymbol{C}=\boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4'}\boldsymbol{c}. \tag{2.3}$$

The use of the causal inference approach to mediation analysis gives rise to the counterfactual definition of direct and indirect effects of the exposure which was formulated by Pearl (2001) and Greenland and Robins (1992). These effects can be estimated from the regression parameters in models (2.2) and (2.3), provided certain identifiability assumptions (no confounding) hold (VanderWeele and Vansteelandt, 2009, 2010). In particular, from models (2.2) and (2.3) what can be defined as the controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) for a change in exposure from

level $\tilde{a}$ to level a are given by (VanderWeele and Vansteelandt, 2009):

$$\begin{aligned} CDE &= (\theta_1 + \theta_3 m)(a - \tilde{a}) \\ NDE &= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1\tilde{a} + \theta_3\boldsymbol{\beta_2'c})(a - \tilde{a}) \\ NIE &= (\theta_2\beta_1 + \theta_3\beta_1 a)(a - \tilde{a}). \end{aligned}$$

These expressions generalize those of Baron and Kenny (1986) to allow for interaction between the exposure and the mediator. While controlled direct effects are often of greater interest in policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms (Robins, 2003; Joffe et al., 2007).

Let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a. Let Y_{am} denote the value of the outcome that would have been observed had the exposure, A, and mediator, M, been set to levels a and m, respectively. The controlled direct effect (CDE), defined by $E[Y_{am} - Y_{\tilde{a}m}|\boldsymbol{C}]$, measures how much the mean of the outcome would change if the mediator were controlled at level m uniformly in the population but the treatment were changed from level $\tilde{a}$ to level a. The natural direct effect (NDE), defined by $E[Y_{aM_{\tilde{a}}} - Y_{\tilde{a}M_{\tilde{a}}}|\boldsymbol{C}]$, measures how much the mean of the outcome would change if the exposure were set at level a versus level $\tilde{a}$ but the mediator were kept at the level it would have taken under $\tilde{a}$. The natural indirect effect (NIE), defined by $E[Y_{aM_a} - Y_{aM_{\tilde{a}}}|\boldsymbol{C}]$, measures how much the mean of the outcome would change if the exposures were controlled at level a, but the mediator were changed from the level it would take under $\tilde{a}$ to the level it would take under a.

The expressions above in terms of regression coefficients will be equal to the counterfactual direct and indirect effects provided that conditional on covariates $\boldsymbol{C}$ there is no unmeasured confounding of (i) the exposure-outcome relationship, (ii) the mediator outcome relationship, (iii) the exposure-mediator relationship, and (iv) that there is no variable affected by the exposure that confounds the mediator outcome relationship. In the counterfactual notation this is: (i) $Y_{am} \amalg A|\boldsymbol{C}$, (ii) $Y_{am} \amalg M|\boldsymbol{C}$, (iii) $M_a \amalg A|\boldsymbol{C}$, (iv) $Y_{am} \amalg M_{\tilde{a}}|\boldsymbol{C}$ (See Pearl (2001) and Robins and Richardson (2010) for further discussion of these assumptions).

When the outcome is binary modeled with a logit link, equation (3) can be replaced by

$$logit\{P(Y=1|A=a,M=m,\boldsymbol{C}=\boldsymbol{c})\}=\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\boldsymbol{\theta_4'c}. \qquad (2.4)$$

If the outcome is case and rare, then from models (2.2) and (2.4) the average controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) for a change in exposure from level $\tilde{a}$ to level a are given in terms of odds ratios by (VanderWeele and Vansteelandt, 2010):

$$\begin{aligned} OR^{CDE} &= exp[(\theta_1+\theta_3 m)(a-\tilde{a})] \\ OR^{NDE} &\approx exp[\{\theta_1+\theta_3(\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}+\theta_2\sigma^2)\}(a-\tilde{a})+0.5\theta_3^2\sigma^2(a^2-\tilde{a}^2)] \\ OR^{NIE} &\approx exp[(\theta_2\beta_1+\theta_3\beta_1 a)(a-\tilde{a})]. \end{aligned}$$

The same formulas apply exactly if the logit link in (2.4) is replaced by a logarithmic link for log-linear, binary or count outcome. In this case the average controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) have a risk ratio or rate ratio interpretation (Valeri and VanderWeele, 2012). If we replace (β_0, β_1, $\boldsymbol{\beta_2'}$) and (θ_0, θ_1, θ_2, θ_3, $\boldsymbol{\theta_4'}$) with their maximum likelihood estimators, we will, by the continuous mapping theorem, have consistent estimates for the direct and indirect effects.

Note that in presence of multiple exposure variables the vector of exposures $\boldsymbol{A}$ can be defined. Replacing θ_1 and θ_3 with the vectors of parameters $\boldsymbol{\theta_1'}$ and $\boldsymbol{\theta_3'}$ in the outcome regression, and replacing β_1 with the vector of parameters $\boldsymbol{\beta_1'}$ in the mediator regression, direct and indirect effects for a joint intervention on the vector of exposures $\boldsymbol{A}$ take the same form as described above. In what follows we consider the exposure as a scalar, but all the results directly extend to the case in which multiple exposures are of interest.

2.3 Asymptotic bias of direct and indirect effects when the mediator is measured with error

2.3.1 GLM with mis-measured mediator

Using the notation in section 2.2, assume that both A and $\boldsymbol{C}$, as well as the outcome Y, are correctly measured. Let M be the continuous mediator at its true level and M^* the version of M measured with error. Let the error, u, be additive with mean zero and constant variance σ_u^2,

$$M^* \quad = \quad M + u.$$

When the mediator is mis-measured, an investigator operates with an observed version of the generalized linear model for the outcome where the true intermediate M is replaced by the observed intermediate M^*

$$g^*(\mu) = \theta_0^* + \theta_1^* a + \theta_2^* m^* + \theta_3^* a m^* + \boldsymbol{\theta_4^{*\prime}} \boldsymbol{c},$$

where θ^* is the asymptotic limit of the estimators of the outcome regression parameters, $\hat{\theta}^*$, when M is replaced by M^*.

In the following we assume that the measurement error is characterized by the property of $Cov(M, u) = 0$ and $Cov(M^*, u) \neq 0$, usually referred as classical measurement error. Moreover, we assume that the measurement error mechanism is independent of the outcome, the exposure, and the covariates (i.e. non-differential).

When the mediator is continuous and measurement error follows the classical measurement error model, it has been shown (McCallum, 1972) that ordinary least squares (OLS) estimators of the coefficients of the mediator regression (2.2) are asymptotically unbiased. However, the assumption that $Cov(M^*, u) \neq 0$ typically causes parameter estimates of the outcome regression to be asymptotically biased. We proceed by deriving the asymptotic limit for the coefficients' estimators of the outcome equation assuming that mediator-exposure interaction may be present. We will present the results for the outcome regression parameters that are involved in the estimation of direct and indirect effects, that is, θ_1, θ_2, θ_3.

2.3.2 Asymptotic Limit of parameters of the outcome regression in the presence of exposure-mediator interaction

Suppose that M is subject to classical measurement error and measured as M^* and that we fit the outcome regression model with either a linear (2.3), a logit (2.4), a probit or a logarithmic link using M^* rather than M. Let ($\hat{\theta}_1^*$, $\hat{\theta}_2^*$, $\hat{\theta}_3^*$) be the naive maximum likelihood estimators of the outcome regressors if M is replaced by M^*. Let (θ_1, θ_2, θ_3) be the true parameters of the regressors. We study the asymptotic limit of the naive estimators of the exposure, mediator and the exposure-mediator interaction coefficients and we denote them by θ_1^*, θ_2^* and θ_3^* respectively.

Let σ_m^2 and σ_u^2 denote the variance of the true mediator given the exposure and the additional covariates $\boldsymbol{C}$ in the outcome model, and of the measurement error respectively. Set $\lambda = \sigma_m^2/(\sigma_m^2 + \sigma_u^2)$, which takes values from 0 to 1 and encodes of the reliability of the measure for the observed mediator. Recall that β_0, β_1, and β_2' are the coefficients of the mediator regression from equation (2.2). Let $X = (1, A, M, AM, \boldsymbol{C})$ and $X^* = (1, A, M^*, AM^*, \boldsymbol{C})$ denote the matrix of the true and observed covariates respectively. Let $Cov(X^*, A\boldsymbol{C})$ and $Cov(X^*, A^2)$ denote the vectors of covariances between the variables in X^* with $A\boldsymbol{C}$ and A^2 respectively. Note that by the assumption of independence between the error, u, and the covariates A and $\boldsymbol{C}$, the covariances just defined do not depend on the moments of u and therefore $Cov(X^*, A\boldsymbol{C}) = Cov(X, A\boldsymbol{C})$ and $Cov(X^*, A^2) = Cov(X, A^2)$. Define δ_A, δ_{M^*}, δ_{AM^*} to be the row vectors of the matrix $E(X^{*T}X^*)^{-1}$. For a continuous outcome modeled using the linear link,

$$
\begin{aligned}
\theta_1^* &= \theta_1 + (1-\lambda)[\theta_2\beta_1 + \theta_3\{\beta_0 + \delta_A Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_A Cov(X^*, A^2)\}] \\
\theta_2^* &= \theta_2\lambda + (1-\lambda)\theta_3\{\delta_{M^*} Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_{M^*} Cov(X^*, A^2)\} \\
\theta_3^* &= \theta_3[\lambda + (1-\lambda)\{\delta_{AM^*} Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_{AM^*} Cov(X^*, A^2)\}].
\end{aligned}
$$

The same result holds when Y is either continuous, binary or count and modeled using the log link. For binary outcome modeled either with logit or probit link we can obtain

only an approximation of the asymptotic limit

$$\begin{aligned}
\theta_1^* &\approx \{\theta_1 + (1-\lambda)[\theta_2\beta_1 + \theta_3\{\beta_0 + \delta_A Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_A Cov(X^*, A^2)\}]\} * H_A(0) \\
\theta_2^* &\approx [\theta_2\lambda + (1-\lambda)\theta_3\{\delta_{M^*} Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_{M^*} Cov(X^*, A^2)\}] * H_{M^*}(0) \\
\theta_3^* &\approx [\theta_3[\lambda + (1-\lambda)\{\delta_{AM^*} Cov(X^*, A\boldsymbol{C})\boldsymbol{\beta_2} + \beta_1\delta_{AM^*} Cov(X^*, A^2)\}]] * H_{AM^*}(0),
\end{aligned}$$

where $H_Z(0)$ is a function of the joint conditional distribution of $A\boldsymbol{C} - E(A\boldsymbol{C}), A^2 - E(A^2)|Z$ with Z equal to either A, M^*, or AM^*. In general this functional is not recoverable in closed form (Neuhaus and Jewell, 1993). However, a numerical bias analysis can still be carried out (Wang et al., 1998). Finally, if we assume a binary exposure for which $A = A^2$, then the two terms can be incorporated and if additionally the true model included the exposure-covariates interaction terms then the asymptotic limit of the estimators of the regression coefficients could be easily derived in closed form as

$$\begin{aligned}
\theta_1^* &= [\theta_1 + \theta_2(1-\lambda)\beta_1 + \theta_3(1-\lambda)(\beta_0 + \beta_1)]/\tau \\
\theta_2^* &= \theta_2\lambda/\tau \\
\theta_3^* &= \theta_3\lambda/\tau,
\end{aligned}$$

where $\tau = (1 + \theta_2^2\lambda\sigma_u^2/S^2)^{\frac{1}{2}}$ with $S = \frac{15\pi}{16\sqrt{3}} \sim 1.7$ when logit link is used and S=1 for probit link (note that allowing for exposure-covariates interaction would change the form of the direct and indirect causal effects estimators).

We note that when an exposure-mediator interaction is present, the asymptotic bias has a complex structure. The bias induced by measurement error is coupled with an omitted-variable type of bias induced by the interaction between a variable measured with error, the mediator, and another covariate in the model, the exposure. The above calculations show that when the true model has interaction terms and the outcome is binary, in general no closed form solutions of the asymptotic bias are available. The magnitude of the

distortion is related to the magnitude of the measurement error expressed by the reliability factor, λ, and to the magnitude of the parameters θ_2 and β_1. We also observe that the magnitude of the distortion is related to the magnitude of the interaction term, θ_3, and the covariance between the variables in the outcome model has impact on how bad the bias could be.

Finally, note that under the non-differential and classical measurement error model and in the absence of exposure-mediator interaction, measurement error typically induces a dilution of the effect of the mediator on the outcome and an over-estimation or an under-estimation of the effect of the exposure on the outcome depending on the sign of the effect of the mediator on the outcome and the sign of the effect of the exposure on the mediator (Carroll et al., 2006; Wang et al., 1998).

2.3.3 Asymptotic bias of direct and indirect causal effects

Given the asymptotic convergence of the outcome regression parameters, the asymptotic bias of the estimators of direct and indirect effects when the mediator is measured with error can be straightforwardly obtained. Let $\gamma_1 = \delta_A Cov(X^*, A\boldsymbol{C})$, $\gamma_2 = \delta_{M^*} Cov(X^*, A\boldsymbol{C})$, $\gamma_3 = \delta_{AM^*} Cov(X^*, A\boldsymbol{C})$, $\gamma_4 = \delta_A Cov(X^*, A^2)$, $\gamma_5 = \delta_{M^*} Cov(X^*, A^2)$, and $\gamma_6 = \delta_{AM^*} Cov(X^*, A^2)$. The asymptotic bias for controlled direct effects, natural direct effects and natural indirect effects when the continuous outcome is modeled using a linear link and exposure-mediator interaction is present is derived is the appendix as:

$$
\begin{aligned}
ABIAS(\widehat{CDE}) &= (1-\lambda)[\theta_2\beta_1 + \theta_3\{\beta_0 + \gamma_1\boldsymbol{\beta_2} + \beta_1\gamma_4 + m(\gamma_3\boldsymbol{\beta_2} + \beta_1\gamma_6 - 1)\}](a-\tilde{a}) \\
ABIAS(\widehat{NDE}) &= (1-\lambda)[\theta_2\beta_1 + \theta_3\{\beta_0 + \gamma_1\boldsymbol{\beta_2} + \beta_1\gamma_4 + (\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c})(\gamma_3\boldsymbol{\beta_2} + \beta_1\gamma_6 - 1)\}] \times \\
&\quad (a-\tilde{a}) \\
ABIAS(\widehat{NIE}) &= (1-\lambda)[\theta_3\{\gamma_2\boldsymbol{\beta_2} + \beta_1\gamma_5 + a(\gamma_3\boldsymbol{\beta_2} + \beta_1\gamma_6 - 1)\} - \theta_2]\beta_1(a-\tilde{a}).
\end{aligned}
$$

When exposure-mediator interaction is absent the formulas can be simplified and we note that measurement error typically induces an under-estimation of the indirect effect and an over-estimation of the direct effect

$$
\begin{aligned}
ABIAS(\widehat{NDE}) &= ABIAS(\widehat{CDE}) = [\theta_2(1-\lambda)\beta_1](a-\tilde{a}) \\
ABIAS(\widehat{NIE}) &= [\theta_2(\lambda-1)]\beta_1(a-\tilde{a}).
\end{aligned}
$$

In the absence of exposure-mediator interaction, if a count, log-linear, or binary outcome is modeled using the log link, the asymptotic bias of the estimators of direct and indirect effects on the log-risk ratio scale takes the same form as that given above. If the binary outcome is modeled using logit or probit link, the asymptotic bias for the indirect effect on the log-odds ratio scale is similar to the one derived above (with λ replaced by $\frac{\lambda}{\tau}$) while the asymptotic bias of the natural direct effect on the log-odds ratio scale is given by

$$
ABIAS(log(\widehat{OR^{NDE}})) = ABIAS(log(\widehat{OR^{CDE}})) = \left[\theta_1(\frac{1}{\tau}-1) + \frac{\theta_2(1-\lambda)\beta_1}{\tau}\right](a-\tilde{a}),
$$

which depends additionally on the magnitude of the effect of the exposure on the outcome, θ_1, and the term τ. Therefore, the choice of link function shapes the impact that measurement error can have on the estimation of direct and indirect causal effect.

Results on asymptotic bias of direct and indirect effects for outcome modeled using either logit, probit, or log link are more complex and less intuitive in presence of exposure-mediator interaction and are provided in the supplementary materials.

For continuous and binary outcome, we carried out a simulation study for large sample size to investigate the change of asymptotic relative bias for the naive direct and indirect causal effect estimators as a function of the magnitude of σ_u^2. We generate samples of dimension $n = 10,000$ with $r = 100$ runs. We define a binary exposure $A_i \sim Ber(p_a)$ with $p_a = 0.4$ and a continuous covariate $C \sim N(0,1)$. The true mediator conditional on A and C is defined as $M|A,C \sim N(\mu_M, \sigma_M^2)$, where $\mu_M = \beta_0 + \beta_1 A + \beta_2 C$ and $\sigma_M^2 = 1$, with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$. The observed mediator is defined as $M^* = M + u$ and $u \sim N(0, \sigma_u^2)$, with σ_u^2 taking values in the range $(0,1)$ which correspond to $\lambda = (0.5, 1)$. The outcome is either normal or binary and in particular we generate $Y|A,M,C \sim N(\mu_Y, \sigma_Y^2)$, where $\mu_Y = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C$ and $\sigma_Y^2 = 1$, with $\theta_0 = 0$, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 1$, $\theta_4 = 1$ or $Y|A,M,C \sim Ber(p_Y)$ with $p_Y = F(\mu_y)$ where $F(u) = exp(u)/(1+exp(u))$ and

$\mu_y = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C$ with $\theta_0 = -2$, $\theta_1 = 0.25$, $\theta_2 = 1$, $\theta_3 = 0.25$, $\theta_4 = 0.25$. The naive outcome and mediator regression models are run simply substituting M with the observed mediator M^*.

Figure 2.1 summarizes the findings for naive direct and indirect effect estimators under the particular setting just described. A figure describing the asymptotic relative bias study for naive estimators $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, $\hat{\theta}_3^*$ is presented in the online appendix. As expected, the naive estimator for the exposure regression parameter ignoring measurement error, $\hat{\theta}_1^*$, is biased upward and the bias increases with the presence of exposure-mediator interaction and for binary outcome modeled using logistic regression. The naive estimator of the mediator, $\hat{\theta}_2^*$, is instead biased downward with a less dramatic change in bias when the outcome is binary and/or in the presence of exposure-mediator interaction. The asymptotic relative bias of the naive estimator of the interaction parameter, $\hat{\theta}_3^*$, is positive in this setting and increases when the outcome is binary relative to continuous. Simulation results are found to be consistent with the theoretical results.

The study of asymptotic relative bias for naive direct and indirect effects estimators reveals that measurement error might exert a stronger impact when the outcome is binary rather than normal, and when the interaction is present rather than absent in the estimation of direct effects. For the particular setting just described, the estimated direct effect is biased upward both in the presence and in the absence of exposure mediator interaction and the estimated indirect effect is biased downward in the absence of interaction and when the outcome is binary. However, when the outcome is linear and exposure-mediator interaction is present, the indirect effect is found to be over-estimated. We note that measurement error could induce either downward or upward biases of both direct and indirect effect estimators in the presence of interaction, depending on the sign and the magnitude of the vector of parameters θ and β. This contrasts with the result obtained in the context of simple mediation models with no interaction for which it is known that measurement error will bias the direct effect upward and the indirect effect downward. The counterintuitive result occurs because covariate measurement error in non-linear models induces additionally an omitted variable problem. We showed that the asymptotic relative bias of the outcome regression parameters estimators in the pres-

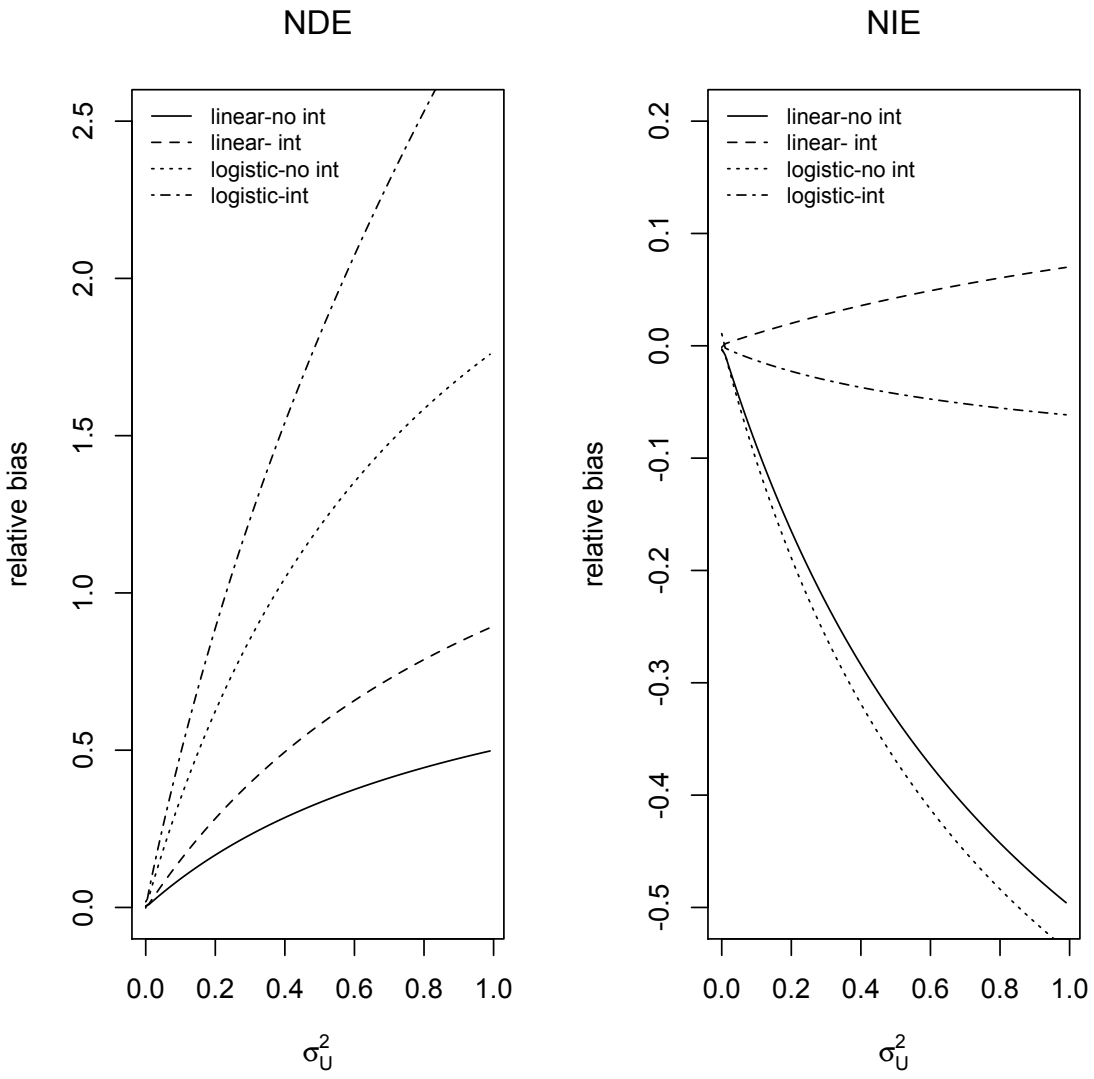

Figure 2.1: Numerical analysis of relative bias of direct (NDE) and indirect (NIE) effect naive estimators. Simulations run for continuous outcome modeled using linear regression and binary outcome modeled using logistic regression; exposure-mediator interaction both present or absent. Sample size $n = 10,000$. Measurement error variance, $\sigma_U^2 \in (0, 1)$, corresponding to a reliability ratio, $\lambda \in (0, 1)$.

ence of exposure mediator interaction contains covariances involving the terms A^2 and $A\boldsymbol{C}$, which are not included the naive outcome regression (5), which consequently is misspecified. VanderWeele et al. (2012b) show that although measurement error in the mediator induces biased direct and indirect effects, the combination of these biased effects is in fact unbiased for the total effect. However, this statement is true only if the mediator and outcome models with M^* replacing M are correctly specified.

Finally, we note that the change in relative bias for the indirect effect naive estimator is different from the one that we observe for the direct effect under the simulation scenarios considered. The naive estimator of the indirect effect for normal outcome in the presence of exposure-mediator interaction is found to be less biased than in the absence of the interaction, the opposite is found for direct effect estimation. This result is consistent with the theoretical results.

2.4 Correction strategy for direct and indirect effects estimators

In what follows we consider three different approaches to measurement error correction of the outcome regression models, namely method of moments, regression calibration, and SIMEX (Fuller, 2006; Carroll et al., 2006; Spiegelman, Rosner and McDermott, 1997; Cook and Stefanski, 1995). These methods are among the most popular and widely used in statistics and epidemiology but they have not been applied to mediation problems and their performance in this context has not been compared when models have interactions and non-linearities, as in our case. All three methods are appealing for several reasons. First, they require assumptions on the moments of the error, rather than assumptions on its complete distribution, which is typically assumed in structural measurement error models. Second, they can be implemented even when auxiliary data on the mediator are not available but the investigator is willing to implement sensitivity analyses on the measurement error magnitude. Finally, their rationale is very intuitive. We will first describe the proposed methods and we will then illustrate their salient properties via a

simulation study. We will compare their performance considering continuous and binary outcomes, and allowing for exposure-mediator interaction.

2.4.1 Method of moments estimators

The most intuitive way to recover consistent estimators for the outcome regression parameters is by solving the system of equations that arises from the study of the limit of the naive estimators with respect to the true parameters. The limit of the naive estimators depends not only on the true parameters but also on population moments and the measurement error variance. Method of moments estimators arise when the system is solved with respect to the true parameters and the population moments are replaced by sample moments.

If the assumptions on the measurement error mechanism and the modeling assumptions hold, and if we assume that the variance of the measurement error, σ_u^2, is known or can be specified in a sensitivity analysis, and we assume that there is no exposure-mediator interaction, then estimators that consistently estimate θ_1 and θ_2 are easily derived from the results given in the previous sections. When the outcome is continuous the method of moments estimators are given by:

$$\begin{aligned} \hat{\theta_1}^{MoM} &= \hat{\theta}_1^* - \hat{\theta}_2^*(1-\lambda)\hat{\beta}_1/\lambda \\ \hat{\theta}_2^{MoM} &= \hat{\theta}_2^*/\lambda, \end{aligned}$$

where $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ are the naive estimators of θ_1 and θ_2.

For a binary outcome, the system that arises from the approximate limit of the naive estimators can again be solved and the method of moments estimators are given by:

$$\begin{aligned}\hat{\theta}_1^{MoM} &= \hat{\theta}_1^*(1+\hat{\theta}_2^{MoM2}\sigma_u^2\lambda) - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_1 \\ \hat{\theta}_2^{MoM} &= \hat{\theta}_2^*/(\lambda^2 - \hat{\theta}_2^*\lambda\sigma_u^2/S^2)^{\frac{1}{2}},\end{aligned}$$

where $S = \frac{15\pi}{16\sqrt{3}} \sim 1.7$ when logit link is used and $S = 1$ when probit link is used.
When exposure-mediator interaction is present in the true model, if again the assumptions on the measurement error mechanism and the modeling assumptions hold, if we assume that the variance of the measurement error, σ_u^2, is known, then estimators that consistently estimate θ_1, θ_2, and θ_3 are derived from the results given in the previous sections. For continuous, binary and count outcomes modeled using linear and log-linear links the method of moment estimators for θ_1, θ_2 and θ_3 are given by:

$$\begin{aligned}\hat{\theta}_1^{MoM} &= \hat{\theta}_1^* - (1-\lambda)\{\hat{\theta}_2^{MoM}\hat{\beta}_1 - \hat{\theta}_3^{MoM}(\hat{\beta}_0 + \gamma_1\widehat{\boldsymbol{\beta}}_{\mathbf{2}} + \hat{\beta}_1\gamma_4)\} \\ \hat{\theta}_2^{MoM} &= [\hat{\theta}_2^* - (1-\lambda)\hat{\theta}_3^{MoM}\{\gamma_2\widehat{\boldsymbol{\beta}}_{\mathbf{2}} + \hat{\beta}_1\gamma_5\}]/\lambda \\ \hat{\theta}_3^{MoM} &= \hat{\theta}_3^*/[(1-\lambda)(\gamma_3\widehat{\boldsymbol{\beta}}_{\mathbf{2}} + \hat{\beta}_1\gamma_6) + \lambda]\end{aligned}$$

When the binary outcome follows a logistic or a probit model the estimators described above are an approximation of the method of moments estimators, given in the online supplement. Finally, consistent estimators for direct and indirect causal effects are easily obtained by substituting the naive estimators $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, and in the presence of exposure-mediator interaction, $\hat{\theta}_3^*$ with the method of moments estimators. For example, we can define method of moments estimators of direct and indirect effects when Y is continuous as follows:

$$\begin{aligned}\widehat{CDE}_{MoM} &= (\hat{\theta}_1^{MoM} + \hat{\theta}_3^{MoM}m)(a-\tilde{a}) \\ \widehat{NDE}_{MoM} &= (\hat{\theta}_1^{MoM} + \hat{\theta}_3^{MoM}\hat{\beta}_0 + \hat{\theta}_3^{MoM}\hat{\beta}_1\tilde{a} + \hat{\theta}_3^{MoM}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}'\boldsymbol{c})(a-\tilde{a}) \\ \widehat{NIE}_{MoM} &= (\hat{\theta}_2^{MoM}\hat{\beta}_1 + \hat{\theta}_3^{MoM}\hat{\beta}_1 a)(a-\tilde{a}).\end{aligned}$$

The implementation of method of moments estimators is straightforward in the absence of exposure-mediator interaction in the true outcome regression model.

In the appendix we discuss the method of moments estimator for binary-logistic and binary-probit outcome in the presence of interaction. In general, for this case the method of moments estimator is not of practical use since it involves the previously described functions $H_Z(0)$ with Z equal to either A, M^*, AM^*, or $\boldsymbol{C}$ which depends on the conditional distribution of $(A\boldsymbol{C} - E(A\boldsymbol{C}), A_i^2 - E(A^2))$ given the covariates in the outcome regression model, an object that is usually hard to derive.

When the function $H_Z(0)$ cannot be recovered, an approximate method of moments estimator or other approximately consistent estimators such as regression calibration and SIMEX estimators should be considered.

2.4.2 Regression calibration estimators

The use of regression calibration to obtain consistent estimators for linear regression coefficients and approximately consistent estimators in the case of logistic regression is based on the assumption of non-differential measurement error (Armstrong, 1985; Spiegelman, Rosner and McDermott, 1997; Carroll et al. 2006).

Regression calibration estimators $\hat{\theta}_1^{rc}$, $\hat{\theta}_2^{rc}$, $\hat{\theta}_3^{rc}$ can be recovered in a rather simple way. First, a calibration model for the regression of the unknown covariate M on the observed mediator M^*, exposure A and the covariates C is developed and fitted. This can be accomplished using replication, validation, or instrumental data. When auxiliary data are not available and the variance of the measurement error is unknown, the value of the measurement error variance σ_u^2 is set as a sensitivity analysis parameter. The unobserved M is then replaced by its predicted values $\hat{M}$ from the calibration model in a standard analysis. Finally, the standard errors are adjusted to account for the estimation of the unknown covariates.

Regression calibration estimators $\hat{\theta}_1^{rc}$, $\hat{\theta}_2^{rc}$, $\hat{\theta}_3^{rc}$ can be recovered in a similar way in the case of logistic regression. Armstrong (1985) showed that for binary outcome regression calibration estimators will yield approximately consistent estimators, provided measurement

error is small and the effect of the mediator on the outcome is not too large in absolute value.

Note that regression calibration estimators and method of moments estimators for parameters of linear regression, under the assumption that σ_u^2 is known, coincide if there is no exposure-mediator interaction (Carroll et al., 2006). However, method of moments estimators and regression calibration estimators won't coincide when the outcome model is non-linear or in the presence of exposure-mediator interaction. In particular, for binary outcome in the absence of exposure-mediator interaction, the method of moments estimator might be preferred to the regression calibration estimator since it provides a better approximation to the consistent estimators and is expected to perform better when measurement error is large. In the presence of exposure-mediator interaction instead regression calibration might be preferred since the method of moments estimator is hard to recover.

2.4.3 SIMEX

SIMEX is a simulation-based approach for measurement error correction, a full description is given by Carroll et al. (2006) and Cook and Stefanski (1995). The SIMEX-method exploits the following relationship between the measurement error variance, σ_u^2, and the limit of the naive estimators, θ^*

$$\sigma_u^2 \to \theta^*(\sigma_u^2) = \mathcal{G}(\sigma_u^2).$$

A consistent estimator of θ when there is no measurement error is such that $\mathcal{G}(0) = \theta$. SIMEX approximates the function $\mathcal{G}(\sigma_u^2)$ by a parametric approach $\mathcal{G}(\sigma_u^2, \Gamma)$, for example with a quadratic approximation $\mathcal{G}_{quadratic}(\sigma_u^2, \Gamma) = \gamma_0 + \gamma_1\sigma_u^2 + \gamma_2(\sigma_u^2)^2$.

Given σ_u^2 either known or specified in a sensitivity analysis, the SIMEX approach consists of two steps. To estimate Γ a simulation step is carried out that adds measurement error with variance $\lambda\sigma_u^2$ to the contaminated variable. The resulting measurement error variance is then $(1 + \lambda)\sigma_u^2$. The naive estimator for this increased measurement error is calculated and repeated B times. The average over B converges to $\mathcal{G}((1 + \lambda)\sigma_u^2)$. Re-

peating this simulations for a fixed grid of λ, leads to an estimator $\hat{\Gamma}$ of the parameters $\mathcal{G}_{quadratic}(\sigma_u^2, \Gamma)$, for example by least squares. In a second step, the extrapolation step, the approximated function $\mathcal{G}_{quadratic}(\sigma_u^2, \hat{\Gamma})$ is extrapolated back to the case of no measurement error and so the SIMEX estimator is defined by

$$\theta_{SIMEX}(\sigma_u^2) = \mathcal{G}(0, \hat{\Gamma})$$

which corresponds to $\lambda = -1$.

Some drawbacks of this method should be mentioned. The SIMEX method is almost always only approximately consistent due to the fact that we generally don't know the true extrapolation function. When the magnitude of the measurement error is substantial the method might not perform well if the extrapolation function is far from the truth. Moreover, SIMEX is computationally less efficient than the regression calibration estimator.

Even if this method is only approximately consistent, we consider implementing it for several reasons. First of all, we have seen in the previous sections that, in general, intuitive analytical formulae for asymptotic bias for binary outcome regression parameters in the presence of exposure-mediator interaction, cannot be recovered. Therefore, the first step of the SIMEX approach can be useful in visualizing the effect of measurement error on the parameter estimates for a given or estimated value of σ_u^2. Second, this method is particularly robust against modeling the structure of the unobservable mediator since it does not require any assumptions on the latent mediator nor on the moments of the measurement error. Finally, this approach has been widely used in the context of generalized linear models.

When the outcome is linear and there is no exposure-mediator interaction SIMEX approach and regression calibration will yield very similar estimators but in general they will differ.

2.4.4 Simulations

We now evaluate the performance in estimating the outcome regression parameters and the direct and indirect effects of interest of the three methods proposed for measurement

($\sigma_u^2 = 0.1$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.093	-0.006	0.006	0.006	0.003	0.003	0.003	0.003	0.01	0.003	0.003	0.003
NIE	-0.087	0.004	0.004	0.000	0.003	0.004	0.004	0.004	0.01	0.004	0.004	0.004
TE	0.0015	0.0015	0.0015	0.0015	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
($\sigma_u^2 = 0.5$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.332	-0.004	-0.004	0.099	0.004	0.005	0.005	0.005	0.11	0.005	0.005	0.014
NIE	-0.333	0.008	0.008	0.096	0.002	0.007	0.007	0.005	0.1	0.007	0.007	0.014
TE	0.0015	0.0015	0.0015	0.0015	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
($\sigma_u^2 = 0.1$)	**Relative Bias**				**Variance**				**MSE**			
Effect($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.157	-0.012	-0.000	0.005	0.004	0.004	0.005	0.005	0.029	0.004	0.005	0.005
NDE	0.164	0.004	0.006	0.012	0.008	0.009	0.008	0.005	0.035	0.009	0.008	0.005
NIE	0.0145	0.099	0.005	0.005	0.01	0.01	0.01	0.01	0.015	0.056	0.014	0.014
TE	0.064	0.064	0.005	0.006	0.01	0.01	0.01	0.005	0.055	0.055	0.016	0.017
($\sigma_u^2 = 0.5$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.588	-0.24	-0.01	0.156	0.007	0.01	0.01	0.01	0.35	0.072	0.012	0.034
NDE	0.595	-0.23	-0.003	0.16	0.01	0.02	0.01	0.01	0.36	0.079	0.01	0.04
NIE	0.046	0.46	0.005	0.0125	0.02	0.04	0.02	0.02	0.029	0.89	0.020	0.02
TE	0.226	0.226	0.002	0.062	0.02	0.02	0.02	0.02	0.499	0.496	0.019	0.05

Table 2.1: Simulations for naive, method of moments (MoM), regression calibation (RC) and SIMEX estimators of direct, indirect and total effects with continuous (linear link) outcome.

error correction, namely method of moments, regression calibration and SIMEX (with quadratic extrapolation function). To compare the methodologies for each estimator we estimate their relative bias, variance and mean squared error. In particular we are interested in comparing their behavior in the presence of non-linearities, which in our study arise when an exposure-mediator interaction is present and if the link is non-linear.

The simulation setting is the same as the one used for the numerical bias analysis in section 2.3.3 which implies a scenario under which the indirect effect of A on Y through M as well as the exposure-mediator interaction are particularly strong. The simulations are now run using a sample size of $n = 1,500$, which mimics a more realistic study sample size, and $r = 100$ runs. Tables 2.1 and 2.2 present the simulations results for $\sigma_u^2 = (0.1, 0.5)$, considering cases of small and moderate measurement error.

($\sigma_u^2 = 0.1$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.258	-0.059	-0.107	-0.06	0.022	0.019	0.019	0.019	0.022	0.019	0.019	0.019
NIE	-0.09	0.009	0.001	0.010	0.008	0.011	0.010	0.011	0.016	0.011	0.010	0.011
TE	-0.02	-0.004	-0.02	-0.004	0.018	0.019	0.018	0.019	0.019	0.019	0.019	0.019
($\sigma_u^2 = 0.5$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	1.141	-0.090	-0.145	0.35	0.016	0.019	0.02	0.018	0.097	0.021	0.021	0.026
NIE	-0.36	0.01	-0.038	-0.118	0.004	0.011	0.013	0.010	0.135	0.014	0.013	0.02
TE	-0.06	-0.006	-0.06	-0.024	0.017	0.019	0.017	0.018	0.022	0.019	0.022	0.019
($\sigma_u^2 = 0.1$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.41	-0.009	-0.034	-0.039	0.04	0.039	0.04	0.05	0.051	0.039	0.044	0.045
NDE	0.399	0.308	-0.008	0.023	0.07	0.07	0.08	0.08	0.121	0.107	0.08	0.09
NIE	-0.001	0.08	-0.006	0.01	0.017	0.02	0.017	0.018	0.017	0.03	0.017	0.018
TE	0.119	0.15	-0.006	0.014	0.12	0.13	0.13	0.148	0.17	0.21	0.13	0.149
($\sigma_u^2 = 0.5$)	**Relative Bias**				**Variance**				**MSE**			
Effect ($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	1.598	0.194	0.047	0.527	0.03	0.02	0.04	0.05	0.191	0.031	0.042	0.05
NDE	1.684	1.85	0.16	0.78	0.07	0.11	0.09	0.111	0.900	1.10	0.107	0.29
NIE	-0.031	0.24	-0.04	-0.007	0.016	0.027	0.021	0.02	0.018	0.12	0.02	0.02
TE	0.485	0.732	0.014	0.231	0.133	0.194	0.15	0.187	0.887	1.9	0.15	0.35

Table 2.2: Simulations for naive, method of moments (MoM), regression calibation (RC) and SIMEX estimators of direct, indirect and total effects with binary (logistic link) outcome.

Comparing the three proposed methods of correction we note that regression calibration has consistently good performance over all the scenarios considered. We notice that method of moments estimator and SIMEX do substantially worse than regression calibration in the presence of exposure-mediator interaction and moderate to severe measurement error. These results are mainly driven by the fact that both method of moments and SIMEX estimators for the interaction term have a poor performance (results shown in the online appendix). However, when the outcome is binary and in the absence of exposure-mediator interaction, as expected, regression calibration, being an approximately consistent estimator, does slightly worse than method of moments in terms of relative bias. Method of moments in this case gives a better approximation to the consistent estimator. The simulation results for SIMEX are similar to the ones of regression calibration when measurement error is small. However, when measurement error is moderate, regression calibration is found to outperform SIMEX under the cases considered.

2.5 Example

We applied the proposed methods to a recent study on the etiology of lung cancer. VanderWeele et al. (2012a) investigated the extent to which the effect of genetic variants rs8034191 and rs1051730 on chromosome 15q25.1 on lung cancer is direct and to what extent that association is mediated by cigarette smoking. Mediation analysis allowing for gene-environment interaction, as described in the second section, was applied to a case-control study of Massachusetts General Hospital (MGH) where 1836 cases and 1452 controls were sampled. Eligible cases included any person over the age of 18 years, with a diagnosis of primary lung cancer that was further confirmed by an MGH lung pathologist. The controls (with no previous history of cancer) were recruited from among the friends or spouses of cancer patients or the friends or spouses of other surgery patients in the same hospital.

rs1051730	1 (naive)	.75	.50	.25
OR^{NDE}	1.26 (1.19-1.33)	1.278 (1.13 -1.46)	1.271 (1.12 - 1.45)	1.307 (1.16 - 1.49)
OR^{NIE}	1.0 (1-1.01)	1.014 (0.99 - 1.03)	1.021 (0.99 - 1.04)	1.045 (0.99 - 1.10)
PM^*	0.023	0.063	0.095	0.159
rs8034191	1 (naive)	.75	.50	.25
OR^{NDE}	1.26 (1.19-1.33)	1.299 (1.14 -1.47)	1.292 (1.14 -1.47)	1.330 (1.17 -1.49)
OR^{NIE}	1.01 (1-1.01)	1.014 (0.99 - 1.03)	1.021 (0.99 - 1.05)	1.044 (0.99 - 1.11)
PM^*	0.032	0.059	0.088	0.152

Table 2.3: Sensitivity analysis results for direct (OR^{NDE}), indirect (OR^{NIE}) effects and proportion mediated ($PM = OR^{NDE} \times (OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1)$) for variants rs1051730 and rs8034191 allowing for exposure-mediator interaction and attenuation factor λ up to 0.25.

VanderWeele et al. (2012a) reported statistically significant additive interaction (P=2×10^{-10} and P=1×10^{-9}) and multiplicative interaction (P=0.01 and P=0.01) between the genetic variants and smoking behavior, measured in terms of square-root average cigarettes per-day. The authors implemented the methodology for mediation analysis in the presence of exposure-mediator interaction adjusting for race, sex, and college education (results in Table 2.3).

We now present the results of the adjustment for measurement error allowing for the presence of exposure-mediator interaction by means of a sensitivity analysis using regression calibration, which was the method that performed best in the simulation study. Setting the attenuation factor λ equal to 0.75, 0.5, and 0.25 (which correspond, given our data, to a variance of the measurement error variable u, σ_u^2, equal to 0.65, 1.3, and 2 respectively), we obtain the corrected direct and indirect effects and percentile confidence intervals from 1,000 bootstrap replications and proportion mediated presented in Table 2.3. The analysis reveals that measurement error induces an underestimate of the indirect effect of the genetic variants on lung cancer mediated by smoking behavior. The direct effect is also found to be slightly underestimated when measurement error is severe. Figure 2.2 depicts the sensitivity of the estimates of direct and indirect effects to the increase of measurement error had we assumed exposure-mediator interaction either present or

absent. We note that ignoring the presence of gene-environment interaction, for moderate

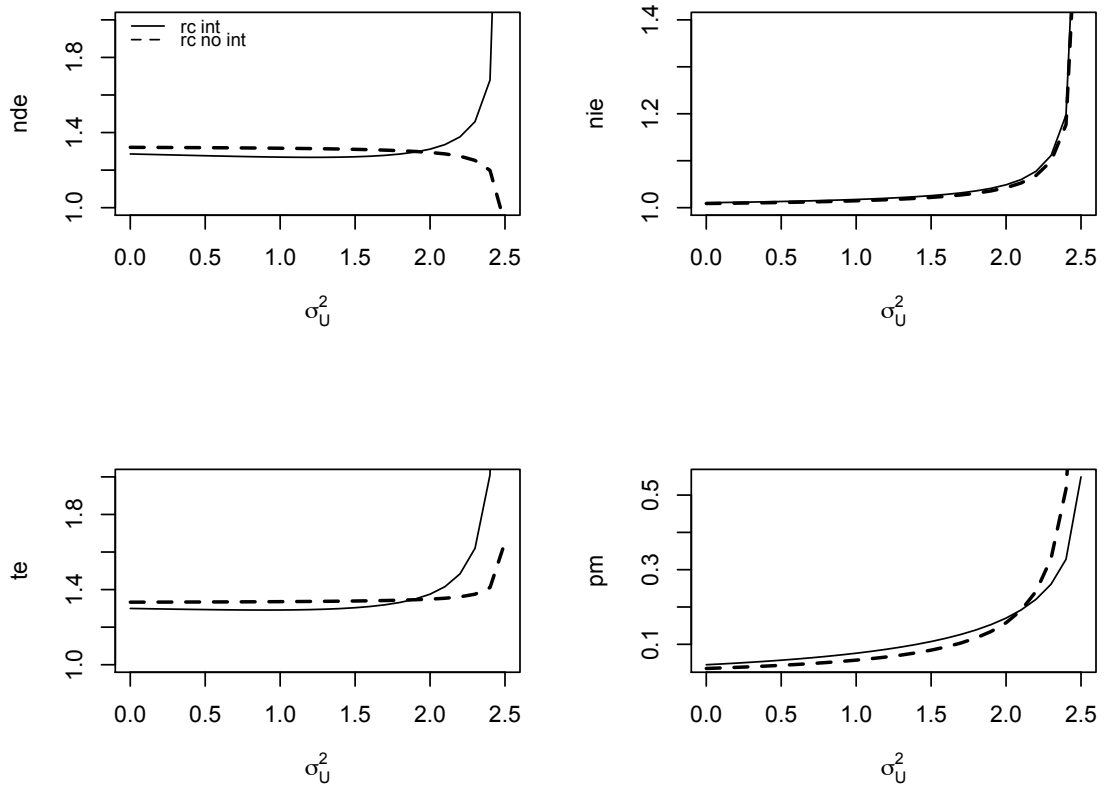

Figure 2.2: Sensitivity analyses for direct (OR^{NDE}), indirect (OR^{NIE}), total (OR^{TE}) effects and proportion mediated ($PM = OR^{NDE} \times (OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1)$) for variant rs1051730. Absent to severe measurement error ($\sigma^2_U \in (0, 2.5)$) which corresponds to a reliability ratio, $\lambda \in (0, 1)$.

to large measurement error, the sensitivity analysis would show an upward bias for the direct effect, which is the opposite from what the analysis taking into account the interaction reveals. Although correcting for measurement error we obtain slightly different results from the naive analysis, we can still conclude that for all the values considered in the sensitivity analysis ($\lambda = 0.75, 0.5, 0.25$) the association of the variants with lung cancer is primarily through pathways other than cigarettes per day (with proportion of the effect of the genetic variants on lung cancer mediated through smoking taking up to the values of 16% in case of lager measurement error).

2.6 Discussion

We have studied the problem of measurement error in the context of causal mediation analysis in GLMs, where exposure-mediator interaction can be present. We have demonstrated that classical and non-differential measurement error on a continuous mediator can undermine the validity of the estimators of direct and indirect causal effects that have been employed. The theoretical results and a numerical study illustrate that when exposure-mediator interaction is present or the outcome is not continuous, the impact of measurement error might be severe. We showed that the bias of the causal effects estimators that ignore measurement error can take unintuitive directions in the presence of non-linearities.

VanderWeele et al. (2012b) show that although measurement error in the mediator induces biased direct and indirect effects, the combination of these biased effects is in fact unbiased for the total effect. However, this statement is true only if the mediator and outcome models with M^* replacing M are correctly specified. In both the simulations and the example above, when exposure-mediator interaction is present or the link function of the outcome model is non-linear, the total effect of the exposure on the outcome (computed as either the sum or the product of direct and indirect effects) was also biased. This phenomenon occurs because covariate measurement error in non-linear models additionally induces model mis-specification, which is what gives rise to the bias in the estimates of total effects as well.

We propose a solution to the problem of measurement error that does not require distributional assumptions on the latent mediator observed with error. We considered regression calibration, SIMEX procedure, and a moment method as possible strategies of correction for measurement error. We compared the performance of corrected estimators for direct and indirect effects in a simulation study. Regression calibration has been found to perform well over all the scenarios considered. Method of moments estimators outperformed the other two approaches only in the case of binary outcome in the absence of interaction. The SIMEX approach performed poorly when the magnitude of measure-

ment error was moderate to severe.

In many instances auxiliary information on the mis-measured intermediate is not available in mediation studies. We illustrated in a real data example the correction strategy coupled with sensitivity analysis for the unknown variance of the measurement error, σ_u^2, for which no validation data or replicates for the mis-measured mediator is needed. Although the correction strategy using sensitivity analysis does not require validation data or replicates for the mis-measured mediator, the corrected estimators could be recovered making use of this information, if available.

Throughout the paper we took a functional approach rather than a structural approach to measurement error. The former makes no assumptions about the distribution of the unobservables, the latter typically makes distributional assumptions. The appeal of functional modeling is model robustness. Alternatively, we could have taken a structural approach as in MacKinnon (2008). In this paper, we assumed classical measurement error for which $Cov(M, u) = 0$ and $Cov(M^*, u) \neq 0$. Alternatively we could have assumed $Cov(M, u) \neq 0$ and $Cov(M^*, u) = 0$, also called Berkson measurement error model. Note that under the Berkson model, for the case of continuous mediator and outcome, it follows from Carroll et al. (2006) that the estimators of direct and indirect effect result in unbiased estimates, even if measurement error was ignored.

Some possible extensions of our study should be mentioned. While leaving the distribution of the latent mediator unspecified, we make the strong assumptions of independence between the measurement error mechanism and all the other variables measured without error. In particular, the assumption about the independence between the measurement error variable, u, and the outcome, Y, is critical for the validity of our asymptotic bias calculations as well as the proposed methods of correction. Care should be given in evaluating the plausibility of the assumption of non-differential measurement error. The effect of a misclassified binary mediator on the validity of mediation analysis is also of interest and will be object of future work.

2.A Definition of causal effects and identifiability conditions

We let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a. We let Y_{am} denote the value of the outcome that would have been observed had the exposure, A, and mediator, M, been set to levels a and m, respectively.

The average controlled direct effect comparing exposure level a to a^* and fixing the mediator to level m is defined by $CDE_{a,a^*}(m) = E[Y_{am} - Y_{a^*m}]$. The average natural direct effect is then defined by $NDE_{a,a^*}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$. The average natural indirect effect can be defined as $NIE_{a,a^*}(a) = E[Y_{aM_a} - Y_{aM_{a^*}}]$, which compares the effect of the mediator at levels M_a and M_{a^*} on the outcome when exposure A is set to a. Controlled direct effects and natural direct and indirect effects within strata of $C = c$ are then defined by: $CDE_{a,a^*|c}(m) = E[Y_{am} - Y_{a^*m}|c]$, $NDE_{a,a^*|c}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]$ and $NIE_{a,a^*|c}(a) = E[Y_{aM_a} - Y_{aM_{a^*}}|c]$ respectively.

For a dichotomous outcome the total effect on the odds ratio scale conditional on $C = c$ is given by $OR^{TE}_{a,a^*|c} = \frac{P(Y_a=1|c)/\{1-P(Y_a=1|c)\}}{P(Y_{a^*}=1|c)/\{1-P(Y_{a^*}=1|c)\}}$. The controlled direct effect on the odds ratio scale is given by $OR^{CDE}_{a,a^*|c}(m) = \frac{P(Y_{am}=1|c)/\{1-P(Y_{am}=1|c)\}}{P(Y_{a^*m}=1|c)/\{1-P(Y_{a^*m}=1|c)\}}$. The natural direct effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}{P(Y_{a^*M_{a^*}}=1|c)/\{1-P(Y_{a^*M_{a^*}}=1|c)\}}$. The natural indirect effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NIE}_{a,a^*|c}(a) = \frac{P(Y_{aM_a}=1|c)/\{1-P(Y_{aM_a}=1|c)\}}{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}$.

If we let $X \perp Y|Z$ denote that X is independent of Y conditional on Z then the identification assumptions for the causal effects previously defined can be expressed formally in terms of counterfactual independence as (i) $Y_{am} \perp A|C$, (ii) $Y_{am} \perp M|\{A, C\}$, (iii) $M_a \perp A|C$, and (iv) $Y_{am} \perp M_{a^*}|C$. Assumptions (i) and (ii) suffice to identify controlled direct effects; assumptions (i)-(iv) suffice to identify natural direct and indirect effects (Pearl, 2001; VanderWeele and Vansteelandt, 2009). The intuitive interpretation of these assuptions follows from the theory of causal diagrams (Pearl, 2001). Alternative

identification assumptions have also been proposed (Imai 2010a; Hafeman and VanderWeele, 2011). However, it has been shown that the intuitive graphical interpretation of these alternative assumptions are in fact equivalent (Shpitser and VanderWeele, 2011). Technical examples can be constructed where one set of identifiation assumptions holds and another does not, but on a causal diagram corresponding to a set of non-parametric structural equations, whenever one set of the assumptions among those in VanderWeele and Vansteelandt (2009), Imai (2010a), and Hafeman and VanderWeele (2011) holds, the others will also.

2.B Continuous Mediator and Outcome

Effects using regression when the mediator is perfectly measured

Suppose that both the mediator and the outcome are continuous and that the following models fit the observed data:

$$M_i = \beta_0 + \beta_1 A_i + \beta_2^{'} C_i + \epsilon_{2i} \tag{2.5}$$

$$Y_i = \theta_0 + \theta_1 A_i + \theta_2 M_i + \theta_3 A_i * M_i + \theta_4^{'} C_i + \epsilon_{1i} \tag{2.6}$$

If the covariates C satisfied the no-unmeasured confounding assumptions (i)-(iv) above, then the average controlled direct effect and the average natural direct and indirect effects were derived by VanderWeele and Vansteelandt (2009).

In particular, if the regression models (2.5) and (2.6) are correctly specified and assumptions of no unmeasured confounding of exposure-outcome relationship (i) and no unmeasured confounding of the mediator-outcome relationship (ii) hold, then we could compute the controlled direct effect as follows:

$CDE = E[Y_{am} - Y_{a^*m}|C = c]$
$= \theta_1(a - a^*) + \theta_3 m(a - a^*).$

If the regression models (2.5) and (2.6) are correctly specified and assumptions (i) and (ii) together with two additional assumptions of (iii) no unmeasured confounding of the exposure-mediator relationship and (iv) that there is no mediator-outcome confounder that is affected by the exposure hold, then we could compute the natural direct effects by:

$NDE = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|C = c]$
$= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' c)(a - a^*).$

Moreover under the same assumptions we can compute the natural indirect effects by:

$NIE = E[Y_{aM_a} - Y_{a^*M_{a^*}}|C = c]$
$= (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*).$

Standard errors for these estimators can ben obtained either via bootstrap procedure or by the delta method (VanderWeele and Vansteelandt, 2009).

Asymptotic bias of outcome regression parameters when the mediator is measured with error

Suppose now that we do not observe the true mediator M, rather a mis-measured version of it $M^* = M + u$. We assume that the measurement error is non-differential, additive, mean zero and with constant variance σ_u^2. In this case we would fit the following *observed* mediator regression and *observed* outcome regression

$$M_i^* = \beta_0^* + \beta_1^* A_i + \beta_2^{*'} C_i + \epsilon_{2i}^* \tag{2.7}$$

$$Y_i = \theta_0^* + \theta_1^* A_i + \theta_2^* M_i^* + \theta_3^* A * M_i^* + \theta_4^{*'} C_i + \epsilon_{1i}^* \tag{2.8}$$

Estimators of the observed mediator regression parameters have been shown to consistently estimate the coefficients of the true mediator regression (Fuller, 2006).

We proceed by studying the asymptotic limit in probability of the estimators of the outcome regression coefficients. We exploit the assumptions on the measurement error mechanism and on the relationship between the true mediator and the other covariates. In particular it is useful to re-write the *true* outcome regression in terms of the *observed* mediator, M^*, in such a way that the observed mis-measured mediator is uncorrelated with a new error term. We then compare the *mis-specified* outcome model with the *true* outcome model.

We note that $M_i = (1-\lambda)M_i + \lambda M_i$. Moreover we have previously defined M as a linear function of A and C plus an error ϵ_2 and the same M can be written as the sum of the observed mediator M^* plus another error u. Therefore, we can write $(1-\lambda)M_i + \lambda M_i = (1-\lambda)(\beta_0 + \beta_1 A_i + \beta_2' C_i + \epsilon_{2i}) + \lambda(M_i^* - u_i)$, with $\lambda = \frac{\sigma^2_{m|a,c}}{\sigma^2_{m|a,c}+\sigma^2_u}$ (Carroll et al., 2006). The terms in this equivalence can be rearranged separating the error terms from the observed variables as $M_i = (1-\lambda)(\beta_0 + \beta_1 A_i + \beta_2' C_i) + \lambda(M_i^*) + (1-\lambda)\epsilon_{2i} - \lambda u_i$ where we recognize that $(1-\lambda)(\beta_0 + \beta_1 A_i + \beta_2' C_i) + \lambda(M_i^*) = E[M|M^*, A, C]$ and $(1-\lambda)\epsilon_{2i} - \lambda u_i = M_i - E[M|M^*, A, C]$ and these two terms are uncorrelated. That is, the true mediator can be defined as its best linear prediction given the observed covariates plus an error. Introducing $E[M|M^*, A, C] + M_i - E[M|M^*, A, C]$ in lieu of M in equation (2.6) helps us understanding how measurement error on M can induce bias in the parameter estimates. Thus, we can write the *mis-specified* outcome model as

$$\begin{aligned}
Y_i &= \theta_0 + \theta_1 A_i + \theta_2\{E[M_i|M_i^*, A_i, C_i] + M_i - E[M_i|M_i^*, A_i, C_i]\} + \\
&\quad \theta_3 A * \{E[M_i|M_i^*, A_i, C_i] + M_i - E[M_i|M_i^*, A_i, C_i]\} + \theta_4' C_i + \epsilon_{1i} \\
Y_i &= \theta_0 + \theta_1 A_i + \theta_2\{(1-\lambda)(\beta_0 + \beta_1 A_i + \beta_2' C_i) + \lambda M_i^* + (1-\lambda)\epsilon_{2i} - \lambda u_i\} + \\
&\quad \theta_3 A * \{(1-\lambda)(\beta_0 + \beta_1 A_i + \beta_2' C_i) + \lambda M_i^* + (1-\lambda)\epsilon_{2i} - \lambda u_i\} + \theta_4' C_i + \epsilon_{1i}
\end{aligned}$$

And we finally obtain

$$\begin{aligned} Y_i &= (\theta_0 + \theta_2(1-\lambda)\beta_0) + (\theta_1 + \theta_2(1-\lambda)\beta_1 + \theta_3(1-\lambda)\beta_0)A_i + \theta_2\lambda M_i^* + \\ &\quad \theta_3\lambda A * M^* + (\theta_4' + \theta_2(1-\lambda)\beta_2')C_i + \epsilon_{1i} - \lambda u_i(\theta_2 + \theta_3 A_i) + (1-\lambda)\epsilon_{2i}(\theta_2 + \theta_3 A_i) + \\ &\quad + \theta_3(1-\lambda)\beta_2' A_i C_i + \theta_3(1-\lambda)\beta_1 A_i^2 \end{aligned} \tag{2.9}$$

When the outcome is continuous and there is no exposure mediator interaction the probability limit of the naive estimator is easily derived by direct comparison of equations (2.8) and (2.9). We first consider the case of no interaction $\theta_3 = 0$. Since $Cov(M^*, M_i - E[M_i|M^*, A_i, C_i]) = Cov(A_i, M_i - E[M_i|M^*, A_i, C_i]) = Cov(C_i, M_i - E[M_i|M^*, A_i, C_i]) = 0$, it is straightforward to derive the probability limit of the parameter estimates from the *observed* regression.

When the outcome is continuous fitting outcome model with the ordinary least squares estimator (OLS) is equivalent to using the generalized linear model estimator with linear link. The OLS estimators of the vector of parameters in the *observed* regression can be written as

$$\hat{\theta}_n^* = (X^{*T}X^*)^{-1}X^{*T}Y = (X^{*T}X^*)^{-1}X^{*T}(X^*\theta^{mis} + \epsilon_1 + \theta_2[(1-\lambda)\epsilon_2 - \lambda u])$$

where $X^* = (1, A, M^*, C)$ is the matrix of observed covariates, and $\theta^{misT} = (\theta_0 + \theta_2(1-\lambda)\beta_0, \theta_1 + (1-\lambda)\beta_1\theta_2, \theta_2\lambda, \theta_2(1-\lambda)\beta_2' + \theta_4')$.

The estimators of the parameters from the *observed* regression will converge to

$$\begin{aligned} plim_{n\to\infty}\hat{\theta}_{0n}^* &= \theta_0 + \theta_2(1-\lambda)\beta_0 \\ plim_{n\to\infty}\hat{\theta}_{1n}^* &= \theta_1 + \theta_2(1-\lambda)\beta_1 \\ plim_{n\to\infty}\hat{\theta}_{2n}^* &= \theta_2\lambda \\ plim_{n\to\infty}\hat{\theta}_{4n}^{*'} &= \theta_4' + \theta_2(1-\lambda)\beta_2' \end{aligned}$$

Since $plim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(x_i^{*T}\theta_2((1-\lambda)\epsilon_{2i}-\lambda u_i))=0$.

Note again that this result is obtained by directly comparing equations (2.8) and (2.9).

When the exposure-mediator interaction term is present the derivation of the asymptotic limit is more complex. In order to facilitate the bias analysis we reparametrize the misspecified outcome model (2.9) so that all the new terms that appear after ϵ_{1i} have mean zero. By adding and subtracting at the rhs of (2.8) the term $(\theta_3(1-\lambda)\beta_2' E(AC)+\theta_3(1-\lambda)\beta_1 E(A^2))$ we obtain:

$$\begin{aligned} Y_i &= (\theta_0+\theta_2(1-\lambda)\beta_0+\theta_3(1-\lambda)\beta_2' E(AC)+\theta_3(1-\lambda)\beta_1 E(A^2))+(\theta_1+\theta_2(1-\lambda)\beta_1+ \\ &\quad +\theta_3(1-\lambda)\beta_0)A_i+\theta_2\lambda M_i^*+\theta_3\lambda A*M^*+(\theta_4'+\theta_2(1-\lambda)\beta_2')C_i+\epsilon_{1i}+ \\ &\quad -\lambda u_i(\theta_2+\theta_3 A_i)+(1-\lambda)\epsilon_{2i}(\theta_2+\theta_3 A_i)+\theta_3(1-\lambda)\beta_2'(A_iC_i-E(AC))+ \\ &\quad +\theta_3(1-\lambda)\beta_1(A_i^2-E(A^2)) \end{aligned} \tag{2.10}$$

A direct comparison of equations (2.8) and (2.10) does not help in this case to study the probability limit of the naive estimators. The OLS estimators of the vector of parameters in the *observed* regression can be written as

$$\begin{aligned} \hat{\theta}_n^* &= (X^{*T}X^*)^{-1}X^{*T}Y \\ &= (X^{*T}X^*)^{-1}X^{*T}(X^*\theta^*+\epsilon_1^*) \\ &= (X^{*T}X^*)^{-1}X^{*T}(X^*\theta^{mis}+\epsilon_1+\theta_2[(1-\lambda)\epsilon_2-\lambda u]+\theta_3 A[(1-\lambda)\epsilon_2-u]+ \\ &\quad +\theta_3(1-\lambda)(AC-E(AC))\beta_2+\theta_3(1-\lambda)\beta_1(A^2-E(A^2)) \end{aligned}$$

where $X^* = (1, A, M^*, AM^*, C)$ is the matrix of observed covariates, and $\theta^{misT}=(\theta_0+\theta_2(1-\lambda)\beta_0, \theta_1+(1-\lambda)\beta_1\theta_2+(1-\lambda)\beta_0\theta_3, \theta_2\lambda, \theta_3\lambda, \theta_2(1-\lambda)\beta_2+\theta_4')$.

The asymptotic limit of the naive estimators can be derived as:

$$\begin{aligned} plim_{n\to\infty}(\hat{\theta}_n^*) &= \theta^{mis}+\theta_3(1-\lambda)E[(X^{*T}X^*)]^{-1}E[X^{*T}(AC-E(AC))]\beta_2 \\ &\quad +\theta_3(1-\lambda)\beta_1 E[(X^{*T}X^*)]^{-1}E[X^{*T}(A^2-E(A^2))]. \end{aligned}$$

Define the matrix $E[(X^{*T}X^*)]^{-1}$ in terms of its row vectors, and let $E[(X^{*T}X^*)]^{-1} = (\delta_I, \delta_A, \delta_{M^*}, \delta_{AM^*}, \delta_C)^T$.

We then obtain the following probability limits for the naive estimators of the outcome regression:

$$
\begin{aligned}
plim_{n\to\infty}\hat{\theta}^{*}_{0n} &= \theta_0 + \theta_2(1-\lambda)\beta_0 + \theta_3(1-\lambda)\beta_2' E(AC) + +\theta_3(1-\lambda)\beta_1 E(A^2) + \\
&\quad +\theta_3(1-\lambda)\delta_I E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad +\theta_3(1-\lambda)\beta_1\delta_I E[X^{*T} * (A^2 - E(A^2))] \\
plim_{n\to\infty}\hat{\theta}^{*}_{1n} &= \theta_1 + \theta_2(1-\lambda)\beta_1 + \theta_3(1-\lambda)\beta_0 + \theta_3(1-\lambda)\delta_A E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad +\theta_3(1-\lambda)\beta_1\delta_A E[X^{*T} * (A^2 - E(A^2))] \\
plim_{n\to\infty}\hat{\theta}^{*}_{2n} &= \theta_2\lambda + \theta_3(1-\lambda)\{\delta_{M^*} E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad +\beta_1\delta_{M^*} E[X^{*T} * (A^2 - E(A^2))]\} \\
plim_{n\to\infty}\hat{\theta}^{*}_{3n} &= \theta_3[\lambda + (1-\lambda)\{\delta_{AM^*} E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad +\beta_1\delta_{AM^*} E[X^{*T} * (A^2 - E(A^2))]\}] \\
plim_{n\to\infty}\hat{\theta}^{*'}_{4n} &= (\theta_4' + \theta_2(1-\lambda)\beta_2' + \theta_3(1-\lambda)\delta_C E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad +\theta_3(1-\lambda)\beta_1\delta_C E[X^{*T} * (A^2 - E(A^2))])
\end{aligned}
$$

Since $plim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(x_i^{*T}\theta_2((1-\lambda)\epsilon_{2i} - \lambda u_i))$ =plim$_{n\to\infty} n^{-1}\sum_{i=1}^{n}(x_i^{*T}\theta_3 A_i((1-\lambda)\epsilon_{2i} - \lambda u_i)) = 0$.

Note that in absence of exposure-mediator interaction we obtain $plim_{n\to\infty}(\hat{\theta}^{*}_{n}) = \theta^{mis}$, which is identical to the probability limit formulae given in the previous section.

Asymptotic bias of causal effects when the mediator is measured with error

Bias formulae for the effects of interest follow directly from the results derived in the previous section.

Let, $\gamma_1 = \delta_A E[X^{*T} * (AC - E(AC))]$, $\gamma_2 = \delta_{M^*} E[X^{*T} * (AC - E(AC))]$, $\gamma_3 = \delta_{AM^*} E[X^{*T} * (AC - E(AC))]$, $\gamma_4 = \delta_A E[X^{*T} * (A^2 - E(A^2))]$, $\gamma_5 = \delta_{M^*} E[X^{*T} * (A^2 - E(A^2))]$, and $\gamma_6 = \delta_{AM^*} E[X^{*T} * (A^2 - E(A^2))]$ and assume exposure-mediator interaction is present.

The asymptotic bias for controlled direct effects, natural direct effects and natural indirect effects when the outcome is continuous and exposure-mediator interaction is present can be derived as:

$$\begin{aligned}
ABIAS(\widehat{CDE}) &= [\theta_2(1-\lambda)\beta_1 + \theta_3\{(1-\lambda)\beta_0 + (1-\lambda)\gamma_1\beta_2 + (1-\lambda)\beta_1\gamma_4 + \\
&\quad + m(\lambda + (1-\lambda)\gamma_3\beta_2 + (1-\lambda)\beta_1\gamma_6 - 1)\}](a-a^*) \\
ABIAS(\widehat{NDE}) &= [\theta_2(1-\lambda)\beta_1 + \theta_3\{(1-\lambda)\beta_0 + (1-\lambda)\gamma_1\beta_2 + (1-\lambda)\beta_1\gamma_4 + \\
&\quad + (\beta_0 + \beta_1 a^* + \beta_2' c)(\lambda + (1-\lambda)\gamma_3\beta_2 + (1-\lambda)\beta_1\gamma_6 - 1)\}](a-a^*) \\
ABIAS(\widehat{NIE}) &= [\theta_2(\lambda - 1) + \theta_3((1-\lambda)\gamma_2\beta_2 + (1-\lambda)\beta_1\gamma_5 + a\lambda + a(1-\lambda)\gamma_3\beta_2 + \\
&\quad + a(1-\lambda)\beta_1\gamma_6 - a)]\beta_1(a-a^*)
\end{aligned}$$

In absence of exposure-mediator interaction all terms involving the parameter θ_3 drop and we obtain

$$\begin{aligned}
ABIAS(\widehat{CDE}) &= ABIAS(\hat{NDE}) = \theta_2(1-\lambda)\beta_1(a-a^*) \\
ABIAS(\widehat{NIE}) &= [\theta_2\beta_1(\lambda-1)](a-a^*)
\end{aligned}$$

Method of Moments Estimators for regression parameters and causal effects

Method of moments estimators for the parameters involved in mediation analysis can be obtained by solving the system of equations that arises from the previous results on the probability limit of naive estimators when the mediator is measured with error.
When exposure-mediator is absent in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2. The method of moments estimators of the outcome regression parameters are given by

$$\begin{aligned}
\hat{\theta}_0^{MoM} &= \hat{\theta}_0^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_0 \\
\hat{\theta}_1^{MoM} &= \hat{\theta}_1^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_1 \\
\hat{\theta}_2^{MoM} &= \hat{\theta}_2^*/\lambda \\
\hat{\theta}_4^{MoM'} &= \hat{\theta}_4^{*'} - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_2'
\end{aligned}$$

When exposure-mediator is present in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2 and θ_3. The method of moments estimators for outcome regression parameters is given by

$$\begin{aligned}
\hat{\theta}_0^{MoM} &= \hat{\theta}_0^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_0 - \hat{\theta}_3^{MoM}(1-\lambda)\{\hat{\beta}_2' E(AC) + \hat{\beta}_1 E(A^2) + \\
&\quad \delta_I E[X^{*T}(AC - E(AC))]\hat{\beta}_2 + \hat{\beta}_1 \delta_I E[X^{*T}(A^2 - E(A^2))]\} \\
\hat{\theta}_1^{MoM} &= \hat{\theta}_1^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_1 - \hat{\theta}_3^{MoM}(1-\lambda)(\hat{\beta}_0 + \gamma_1\hat{\beta}_2 + \hat{\beta}_1\gamma_4) \\
\hat{\theta}_2^{MoM} &= \frac{\hat{\theta}_2^*}{\lambda} - \frac{\hat{\theta}_3^{MoM}(1-\lambda)}{\lambda}(\gamma_2\hat{\beta}_2 + \hat{\beta}_1\gamma_5) \\
\hat{\theta}_3^{MoM} &= \frac{\hat{\theta}_3^*}{(1-\lambda)\gamma_3\hat{\beta}_2 + (1-\lambda)\hat{\beta}_1\gamma_6 + \lambda} \\
\hat{\theta}_4^{MoM'} &= \hat{\theta}_4^{*'} - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_2' - \hat{\theta}_3^{MoM}(1-\lambda)[\delta_C E[X^{*T}(AC - E(AC))]\hat{\beta}_2 + \\
&\quad \hat{\beta}_1\delta_C E[X^{*T}(A^2 - E(A^2))]].
\end{aligned}$$

Regression Calibration Estimators for regression parameters and causal effects

Regression calibration estimators for the parameters involved in mediation analysis can be obtained by substituting the observed mediator in the naive outcome regression with the best linear predictor of the latent true mediator given A, C and M^* .

When exposure-mediator interaction is absent in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2. The regression calibration estimators for these two regression parameters coincide with the method of moments estimators given in the previous section (Carroll et al., 2006).

The use of regression calibration to obtain consistent estimators for linear regression coefficients is based on the assumption of non differential measurement error.

For simple linear regression we can show that under non differential measurement error

$$\begin{aligned}
E(Y|M) &= E_m\{E(Y|M, M^*)|M^*\} \\
&= E_m\{E(Y|M)|M^*\} \\
&= E_m(\theta_0 + \theta_1 M|M^*) \\
&= \theta_0 + \theta_1 E_m(M|M^*)
\end{aligned}$$

We can show that regression approximation is exact in the exposure-mediator interaction model.

We claim that

$$E(Y|A, M^*, C) \;=\; E_m\{E(Y|A, M, C, M^*)|A, M^*, C\} = E_m\{E(Y|A, M, C)|A, M^*, C\}$$

The proof is as follows:

$$\begin{aligned} E_m\{E(Y|A, M, C, M^*)|A, M^*, C\} &= \int_m \int_y y f(y|a, m, c, m^*) dy f(a, m, c|a, m^*, c) dm \\ &= \int_y y \int_m \frac{f(y|a, m, c, m^*) f(a, m, c, m^*)}{f(a, m^*, c)} dm dy \\ &= \int_y \frac{y}{f(a, m^*, c)} \int_m f(y, a, m, c, m^*) dm dy \\ &= \int_y y f(y|a, m^*, c) dy = E(Y|A, M^*, C) \end{aligned}$$

and we note that

$$E(Y|A, M^*, C) \;=\; \theta_0 + \theta_1 A + \theta_2 E_m(M|A, M^*, C) + \theta_3 A E_m(M|A, M^*, C) + \theta_4' C'$$

Therefore, estimating the latent mediator as a function of the observed covariates and running the usual regression of Y on the exposure, the covariates C and the calibration function of M will yield consistent estimators θ^{rc} for θ.

Standard errors of the method of moments estimators for direct and indirect causal effects

We now derive the standard errors of method of moments estimators for controlled direct, natural direct and natural indirect effects assuming that exposure-mediator interaction may be present.

Define the corrected method of moments estimators of the causal effects of interest as

$$\widehat{CDE} = E[Y_{am} - Y_{a^*m}|C = c]$$
$$= \hat{\theta}_1^{MoM}(a - a^*) + \hat{\theta}_3^{MoM} m(a - a^*).$$

$$\widehat{NDE} = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|C = c]$$
$$= (\hat{\theta}_1^{MoM} + \hat{\theta}_3^{MoM}\hat{\beta}_0 + \hat{\theta}_3^{MoM}\hat{\beta}_1 a^* + \hat{\theta}_3^{MoM}\hat{\beta}_2' c)(a - a^*).$$

$$\widehat{NIE} = E[Y_{aM_a} - Y_{a^*M_{a^*}}|C = c]$$
$$= (\hat{\theta}_2^{MoM}\hat{\beta}_1 + \hat{\theta}_3^{MoM}\hat{\beta}_1 a)(a - a^*).$$

$$\widehat{TE} = E[Y_a - Y_{a^*}|C = c]$$
$$= \widehat{NDE} + \widehat{NIE}.$$

Suppose that model (2.7) and (2.8) have been fit using standard linear regression software and that the resulting estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}^*$ of $\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^{*\prime})'$ have covariance matrices Σ_β and Σ_{θ^*}. Then the covariance matrix of $(\hat{\beta}', \hat{\theta}^{*\prime})$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_{\theta^*} \end{bmatrix}$$

Suppose further that method of moments estimators of the outcome regression parameters are obtained as described in the previous section $\hat{\theta}^{MoM} = (\hat{\theta}_0^{MoM}, \hat{\theta}_1^{MoM}, \hat{\theta}_2^{MoM}, \hat{\theta}_3^{MoM}, \hat{\theta}_4^{MoM\prime})'$. Then, the covariance matrix of $(\hat{\beta}', \hat{\theta}^{MoM\prime})$ is

$$\Sigma^{MoM} = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_{\theta^{MoM}} \end{bmatrix}$$

$\Sigma_{\theta^{MoM}}$ is found using the multivariate delta method.

$$\Sigma_{\theta^{MoM}} = D^{MoM}\Sigma D^{MoM^T}$$

where,

$$D^{MoM} = \begin{bmatrix} \frac{\partial\theta_0^{MoM}}{\partial\beta_0} & \frac{\partial\theta_0^{MoM}}{\partial\beta_1} & \frac{\partial\theta_0^{MoM}}{\partial\beta_2'} & \frac{\partial\theta_0^{MoM}}{\partial\theta_0^*} & \frac{\partial\theta_0^{MoM}}{\partial\theta_1^*} & \frac{\partial\theta_0^{MoM}}{\partial\theta_2^*} & \frac{\partial\theta_0^{MoM}}{\partial\theta_3^*} & \frac{\partial\theta_0^{MoM}}{\partial\theta_4^{*\prime}} \\ \frac{\partial\theta_1^{MoM}}{\partial\beta_0} & \frac{\partial\theta_1^{MoM}}{\partial\beta_1} & \frac{\partial\theta_1^{MoM}}{\partial\beta_2'} & \frac{\partial\theta_1^{MoM}}{\partial\theta_0^*} & \frac{\partial\theta_1^{MoM}}{\partial\theta_1^*} & \frac{\partial\theta_1^{MoM}}{\partial\theta_2^*} & \frac{\partial\theta_1^{MoM}}{\partial\theta_3^*} & \frac{\partial\theta_1^{MoM}}{\partial\theta_4^{*\prime}} \\ \frac{\partial\theta_2^{MoM}}{\partial\beta_0} & \frac{\partial\theta_2^{MoM}}{\partial\beta_1} & \frac{\partial\theta_2^{MoM}}{\partial\beta_2'} & \frac{\partial\theta_2^{MoM}}{\partial\theta_0^*} & \frac{\partial\theta_2^{MoM}}{\partial\theta_1^*} & \frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} & \frac{\partial\theta_2^{MoM}}{\partial\theta_3^*} & \frac{\partial\theta_2^{MoM}}{\partial\theta_4^{*\prime}} \\ \frac{\partial\theta_3^{MoM}}{\partial\beta_0} & \frac{\partial\theta_3^{MoM}}{\partial\beta_1} & \frac{\partial\theta_3^{MoM}}{\partial\beta_2'} & \frac{\partial\theta_3^{MoM}}{\partial\theta_0^*} & \frac{\partial\theta_3^{MoM}}{\partial\theta_1^*} & \frac{\partial\theta_3^{MoM}}{\partial\theta_2^*} & \frac{\partial\theta_3^{MoM}}{\partial\theta_3^*} & \frac{\partial\theta_3^{MoM}}{\partial\theta_4^{*\prime}} \\ \frac{\partial\theta_4^{MoM\prime}}{\partial\beta_0} & \frac{\partial\theta_4^{MoM\prime}}{\partial\beta_1} & \frac{\partial\theta_4^{MoM\prime}}{\partial\beta_2'} & \frac{\partial\theta_4^{MoM\prime}}{\partial\theta_0^*} & \frac{\partial\theta_4^{MoM\prime}}{\partial\theta_1^*} & \frac{\partial\theta_4^{MoM\prime}}{\partial\theta_2^*} & \frac{\partial\theta_4^{MoM\prime}}{\partial\theta_3^*} & \frac{\partial\theta_4^{MoM\prime}}{\partial\theta_4^{*\prime}} \end{bmatrix}$$

with the partial derivatives derived as,

$$\frac{\partial\theta_0^{MoM}}{\partial\beta_0} = -(1-\lambda)\theta_2^{MoM}$$

$$\frac{\partial\theta_0^{MoM}}{\partial\beta_1} = -(1-\lambda)\beta_0\frac{\partial\theta_2^{MoM}}{\partial\beta_1} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_1}(1-\lambda)(\beta_2'(E(AC) + \delta_I E[X^{*T}(AC - E(AC))]) + \beta_1(E(A^2) + \delta_I E[X^{*T}(A^2 - E(A^2))])) + \theta_3^{MoM}(1-\lambda)(E(A^2) + \delta_I E[X^{*T}(A^2 - E(A^2))])]$$

$$\frac{\partial\theta_0^{MoM}}{\partial\beta_2'} = -(1-\lambda)\beta_0\frac{\partial\theta_2^{MoM}}{\partial\beta_2'} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_2'}(1-\lambda)(\beta_2'(E(AC) + \delta_I E[X^{*T}(AC - E(AC))]) + \beta_1(E(A^2) + \delta_I E[X^{*T}(A^2 - E(A^2))])) + \theta_3^{MoM}(1-\lambda)(E(AC) + \delta_I E[X^{*T}(AC - E(AC))])]$$

$$\frac{\partial\theta_1^{MoM}}{\partial\beta_0} = -(1-\lambda)\theta_3^{MoM}$$

$$\frac{\partial\theta_1^{MoM}}{\partial\beta_1} = -(1-\lambda)\theta_2^{MoM}\frac{\partial\theta_2^{MoM}}{\partial\beta_1} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_1}(1-\lambda)(\beta_0 + \beta_2'\gamma_1 + \beta_1\gamma_4) + \theta_3^{MoM}(1-\lambda)\gamma_4]$$

$$\frac{\partial\theta_1^{MoM}}{\partial\beta_2'} = -(1-\lambda)\beta_1\frac{\partial\theta_2^{MoM}}{\partial\beta_2'} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_2'}(1-\lambda)(\beta_0 + \beta_2'\gamma_1 + \beta_1\gamma_4) + \theta_3^{MoM}(1-\lambda)\gamma_1]$$

$$\frac{\partial\theta_2^{MoM}}{\partial\beta_0} = \frac{\partial\theta_3^{MoM}}{\partial\beta_0} = \frac{\partial\theta_4^{MoM\prime}}{\partial\beta_0} = 0$$

$$\frac{\partial\theta_2^{MoM}}{\partial\beta_1} = \frac{\theta_3(1-\lambda)\gamma_5[\lambda\{(1-\lambda)(\beta_2'\gamma_3 + \beta_2\gamma_6) + \lambda\}] - \lambda(1-\lambda)\gamma_6[\theta_3^*(1-\lambda)(\beta_2'\gamma_2 + \beta_1\gamma_5)]}{[\lambda\{(1-\lambda)(\beta_2'\gamma_3 + \beta_1\gamma_6) + \lambda\}]^2}$$

$$\frac{\partial\theta_2^{MoM}}{\partial\beta_2'} = \frac{\theta_3(1-\lambda)\gamma_2[\lambda\{(1-\lambda)(\beta_2'\gamma_3+\beta_2\gamma_6)+\lambda\}] - \lambda(1-\lambda)\gamma_3[\theta_3^*(1-\lambda)(\beta_2'\gamma_2+\beta_1\gamma_5)]}{[\lambda\{(1-\lambda)(\beta_2'\gamma_3+\beta_1\gamma_6)+\lambda\}]^2}$$

$$\frac{\partial\theta_3^{MoM}}{\partial\beta_1} = -\frac{(1-\lambda)\gamma_6}{[\lambda+(1-\lambda)(\beta_2'\gamma_3+\beta_1\gamma_6)]^2}$$

$$\frac{\partial\theta_3^{MoM}}{\partial\beta_2'} = -\frac{(1-\lambda)\gamma_3}{[\lambda+(1-\lambda)(\beta_2'\gamma_3+\beta_1\gamma_6)]^2}$$

$\frac{\partial\theta_4^{MoM'}}{\partial\beta_1} = -(1-\lambda)\beta_2'\frac{\partial\theta_2^{MoM}}{\partial\beta_1} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_1}(1-\lambda)(\beta_2'\delta_C E[X^{*T}(AC-E(AC))] + \beta_1\delta_C E[X^{*T}(A^2 - E(A^2))]) + \theta_3^{MoM}(1-\lambda)\delta_C E[X^{*T}(A^2-E(A^2))]]$

$\frac{\partial\theta_4^{MoM'}}{\partial\beta_2'} = -(1-\lambda)\theta_2^{MoM}\frac{\partial\theta_2^{MoM}}{\partial\beta_2'} - [\frac{\partial\theta_3^{MoM}}{\partial\beta_2'}(1-\lambda)(\beta_2'\delta_C E[X^{*T}(AC-E(AC))] + \beta_1\delta_C E[X^{*T}(A^2 - E(A^2))]) + \theta_3^{MoM}(1-\lambda)\delta_C E[X^{*T}(AC-E(AC))]]$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_0^*} = \frac{\partial\theta_1^{MoM}}{\partial\theta_1^*} = 1$$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_1^*} = \frac{\partial\theta_1^{MoM}}{\partial\theta_0^*} = \frac{\partial\theta_2^{MoM}}{\partial\theta_0^*} = \frac{\partial\theta_2^{MoM}}{\partial\theta_1^*} = \frac{\partial\theta_3^{MoM}}{\partial\theta_0^*} = \frac{\partial\theta_3^{MoM}}{\partial\theta_1^*} = \frac{\partial\theta_3^{MoM}}{\partial\theta_2^*} = 0$$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_4^{*\prime}} = \frac{\partial\theta_1}{\partial\theta_4^{*\prime}} = \frac{\partial\theta_2^{MoM}}{\partial\theta_4^{*\prime}} = \frac{\partial\theta_3^{MoM}}{\partial\theta_4^{*\prime}} = \frac{\partial\theta_4^{MoM'}}{\partial\theta_0^*} = \frac{\partial\theta_4^{MoM'}}{\partial\theta_1^*} = 0'$$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_2^*} = -\frac{(1-\lambda)\beta_0}{\lambda}$$

$\frac{\partial\theta_0^{MoM}}{\partial\theta_3^*} = \frac{(1-\lambda)\beta_0[(1-\lambda)(\beta_2'\gamma_2+\beta_1\gamma_5)-\lambda(1-\lambda)(\beta_2'E(AC)+\beta_1E(A^2)+\beta_2'\delta_I E[X^{*T}(AC-E(AC))]+\beta_1\delta_I E[X^{*T}(A^2-E(A^2))])]}{\lambda\{(1-\lambda)[\beta_2'\gamma_3+\beta_1\gamma_6]+\lambda\}}$

$$\frac{\partial\theta_1^{MoM}}{\partial\theta_2^*} = -\frac{(1-\lambda)\beta_1}{\lambda}$$

$$\frac{\partial\theta_1^{MoM}}{\partial\theta_3^*} = \frac{(1-\lambda)\beta_1[(1-\lambda)(\beta_2'\gamma_2+\beta_1\gamma_5) - \lambda(1-\lambda)(\beta_0+\beta_2'\gamma_1+\beta_1\gamma_4)]}{\lambda\{(1-\lambda)[\beta_2'\gamma_3+\beta_1\gamma_6]+\lambda\}}$$

$$\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} = \frac{1}{\lambda}$$

$$\frac{\partial\theta_2^{MoM}}{\partial\theta_3^*} = -\frac{(1-\lambda)(\beta_2'\gamma_2+\beta_1\gamma_5)}{\lambda[(1-\lambda)(\beta_2'\gamma_3+\beta_1\gamma_6)+\lambda]}$$

$$\frac{\partial \theta_3^{MoM}}{\partial \theta_3^*} = \frac{1}{(1-\lambda)(\beta_2'\gamma_3 + \beta_1\gamma_6) + \lambda}$$
$$\frac{\partial \theta_4'}{\partial \theta_2^*} = -\frac{(1-\lambda\beta_2')}{\lambda}$$

$$\frac{\partial \theta_4^{MoM'}}{\partial \theta_3^*} = \frac{(1-\lambda)\beta_2'[(1-\lambda)(\beta_2'\gamma_2+\beta_1\gamma_5)-\lambda(1-\lambda)(\beta_2'\delta_C E[X^{*T}(AC-E(AC))]+\beta_1\delta_C E[X^{*T}(A^2-E(A^2))])]}{\lambda\{(1-\lambda)[\beta_2'\gamma_3+\beta_1\gamma_6]+\lambda\}}$$

$$\frac{\partial \theta_4^{MoM'}}{\partial \theta_4^{*'}} = I.$$

Standard errors of the method of moments controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma^{MoM}\Sigma^{MoM}\Gamma^{MoM\prime}}|a-a^*|$$

with $\Gamma^{MoM} = (0, 0, 0^{'}, 0, 1, 0, m, 0^{'})$ for the controlled direct effect, $\Gamma^{MoM} = (\theta_3^{MoM}, \theta_3^{MoM}a^*, \theta_3^{MoM}c^{'}, 0, 1, \beta_0 + \beta_1 a^* + \beta_2' c, 0^{'})$ for the pure natural direct effect (same expression holds for the total natural direct effect upon substituting a and a^*), $\Gamma^{MoM} = (0, \theta_2^{MoM} + \theta_3^{MoM}a, 0^{'}, 0, 0, \beta_1, \beta_1 a, 0^{'})$ for the total natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*), $\Gamma^{MoM} = (\theta_3^{MoM}, \theta_3^{MoM}(a + a^*) + \theta_2^{MoM}, \theta_3^{MoM}c^{'}, 0, 1, \beta_1, \beta_0 + \beta_1(a + a^*) + \beta_2' c, 0^{'})$ for the total effect.

Standard errors of the method of moments estimators of the causal effects of interest can be obtained in absence of exposure-mediator interaction in a similar way by setting $\theta_3 = 0$. In absence of exposure mediator interaction the standard errors of the controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma^{MoM}\Sigma^{MoM}\Gamma^{MoM\prime}}|a-a^*|$$

with $\Gamma^{MoM} = (0, 0, 0^{'}, 0, 1, 0, 0^{'})$ for the controlled direct effect and for the pure natural direct effect, $\Gamma^{MoM} = (0, \theta_2^{MoM}, 0^{'}, 0, \beta_1, 0^{'})$ for the total natural indirect effect, $\Gamma^{MoM} = (0, \theta_2^{MoM}, 0\prime, 0, 1, \beta_1, 0^{'})$ for the total effect.

$$\Sigma^{MoM} = \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta^{MoM} \end{bmatrix}$$

where,

$$\Sigma_{\theta}^{MoM} = D^{MoM}\Sigma D^{MoM^T}$$

with,

$$D^{MoM} = \begin{bmatrix} -\theta_2^{MoM}(1-\lambda) & 0 & 0 & 1 & 0 & -\frac{(1-\lambda)\beta_0}{\lambda} & 0 \\ 0 & -\theta_2^{MoM}(1-\lambda) & 0 & 0 & 1 & -\frac{(1-\lambda)\beta_1}{\lambda} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\lambda} & 0 \\ 0 & 0 & -\theta_2^{MoM}(1-\lambda) & 0 & 0 & -\frac{(1-\lambda)\beta_2'}{\lambda} & I \end{bmatrix}$$

2.C Continuous Mediator and Binary-Logistic, Binary-Log-linear, or Count Outcome

Effects using regression when the mediator is prefectly measured

Suppose that the mediator is continuous and the outcome is binary and is rare. Suppose that the following models fit the observed data:

$$E(M|A=a, C=c) = \beta_0 + \beta_1 a + \beta_2^{'} c \tag{2.11}$$

$$logit\{P(Y=1|A=a, M=m, C=c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'} c \tag{2.12}$$

and that the error term in the regression model for M is normally distributed with mean 0 and variance $\sigma^2 = \sigma^2_{m|a,c}$. If the regression models (2.11) and (2.12) are correctly specified and assumptions (i) and (ii) hold then the conditional controlled direct effect on the odds ratio scale would be given by (VanderWeele and Vansteelandt, 2010):

$OR^{CDE} = \frac{P(Y_{am}=1|c)/(1-P(Y_{am}=1|c))}{P(Y_{a^*m}=1|c)/(1-P(Y_{a^*m}=1|c))}$
$= exp[(\theta_1 + \theta_3 m)(a - a^*)]$.

If the regression models (2.11) and (2.12) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, and the error term for linear regression model (2.11) is normally distributed and has constant variance σ^2, then we could compute the natural direct effects by:

$$
\begin{aligned}
&OR^{NDE} = exp[log\{\tfrac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}] \\
&= exp[logit\{P(Y_{aM_{a^*}}=1|c)\} - logit\{P(Y_{a^*M_{a^*}}=1|c)\}] \\
&= exp[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2^{'}c + \theta_2\sigma^2)\}(a-a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})].
\end{aligned}
$$

If the regression models (2.11) and (2.12) are correctly specified and assumptions (i)-(iv) hold, the outcome Y is rare, and the error term for linear regression model (2.11) is normally distributed and has constant variance $\sigma^2 = \sigma^2_{m|a,c}$, then we could compute the natural indirect effects by:

$$
\begin{aligned}
&OR^{NIE} = exp[log\{\tfrac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}] \\
&= exp[logit\{P(Y_{aM_a}=1|c)\} - logit\{P(Y_{aM_{a^*}}=1|c)\}] \\
&= exp[(\theta_2\beta_1 + \theta_3\beta_1 a)(a-a^*)].
\end{aligned}
$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the outcome model; the direct and indirect effect will have now an interpretation on the risk ratio scale rather than on the odds ratio scale.
These expressions apply also if the outcome is a count variable. In particular if $Y \sim Poi(\lambda)$ for $\lambda = \exp\{\theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'}c\})$ the outcome regression can be defined as:

$$log\{E(Y|A=a, M=m, C=c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a * m + \theta_4^{'}c$$

The natural direct effect for binary outcome on the risk ratio scale coincides with the natural direct effect for poisson count outcome since:

$$RR^{NDE} = exp[log\{\frac{E(Y_{aM_{a^*}}|c)}{E(Y_{a^*M_{a^*}}|c)}\}]$$

The same argument holds for the natural indirect effect. Finally, the argument can be extended to the case in which the count outcome is modeled with a negative binomial distribution. This is the case since the negative binomial distribution can be represented as an over-dispersed poisson and the mean of the two models coincide.

Standard errors for these estimators can ben obtained either via bootstrap procedure or by delta method (VanderWeele and Vansteelandt, 2010; Valeri and VanderWeele, 2012).

Asymptotic bias of outcome regression parameters when the mediator is measured with error

If we did not observe the true mediator, but a mis-measured version of it instead, we would fit the *observed* mediator regression (2.7) that, as we have discussed above, will yield valid estimates of the parameters. When the outcome is binary we assume that Y_i arises from a continuous latent variable Y_i^{latent} that we would model as (2.6), had the true mediator been observed and as (2.8) had M^* been observed. Then, $Y_i = 1$ iff $Y_i^{latent} > 0$, $Y_i = 0$ otherwise.

For some symmetrical distribution function F_{ϵ_1}, when the true mediator is observed we have $P(Y_i = 1) = F_{\epsilon_1}^{-1}(\theta_0+\theta_1 A_i+\theta_2 M_i+\theta_3 A*M_i+\theta_4^{'}C_i)$, when we observe a mis-measured version of the mediator instead we have $P(Y_i = 1) = F_{\epsilon_1^*}^{-1}(\theta_0^* + \theta_1^* A_i + \theta_2^* M_i^* + \theta_3^* A * M_i^* + \theta_4^{*'} C_i)$. For logistic and probit regression F_{ϵ_1} and $F_{\epsilon_1^*}$ are the cumulative distribution functions of the logistic and the normal distribution respectively.

Both in ordinary linear regression and in the discrete model, identification of the parameters requires further assumptions about the disturbances. In both models their mean must be specified or the intercept is not identified. In the discrete model, the variance of the disturbances, $\sigma_{\epsilon_1}^2$ and $\sigma_{\epsilon_1^*}^2$, must be specified too, since $Y_i^{latent} > 0$ is invariant to scaling of Y_i^{latent}, and hence to scaling of ϵ_{1i} and ϵ_{1i}^* and of the vector θ and θ^*, so that neither the variance nor the parameters are identified. This indeterminacy is resolved by imposing a

set value S on $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_1^*}$. Both sides of the latent variable models (2.6) and (2.8) are multiplied by $S/\sigma^2_{\epsilon_1}$ and $S/\sigma^2_{\epsilon_1^*}$ respectively, and thus the latent variable models are replaced by

$$Y_i^\dagger = \theta_0^\dagger + \theta_1^\dagger A_i + \theta_2^\dagger M_i + \theta_3^\dagger A * M_i + \theta_4^{\dagger'} C_i + \epsilon_{1i}^\dagger \tag{2.13}$$

with $Y_i^\dagger = Y_i \frac{S}{\sigma^2_{\epsilon_1}}$, $\theta^\dagger = \theta \frac{S}{\sigma^2_{\epsilon_1}}$, $\epsilon_{1i}^\dagger = \epsilon_{1i} \frac{S}{\sigma^2_{\epsilon_1}}$, when the true mediator is observed, and

$$Y_i^\dagger = \theta_0^{*\dagger} + \theta_1^{*\dagger} A_i + \theta_2^{*\dagger} M_i^* + \theta_3^{*\dagger} A * M_i^* + \theta_4^{*\dagger'} C_i + \epsilon_{1i}^{*\dagger} \tag{2.14}$$

when the mis-measured version of the mediator is observed, with $Y_i^\dagger = Y_i \frac{S}{\sigma^2_{\epsilon_1^*}}$, $\theta^{*\dagger} = \theta^* \frac{S}{\sigma^2_{\epsilon_1^*}}$, $\epsilon_{1i}^{*\dagger} = \epsilon_{1i}^* \frac{S}{\sigma^2_{\epsilon_1^*}}$. The observed outcome is now defined as $Y_i = 1$ iff $Y_i^\dagger > 0$ and $Y_i = 0$ otherwise.

In the probit model $\epsilon_{1i}^\dagger$ and $\epsilon_{1i}^{\dagger *}$ have a standard normal distribution and $S = 1$; in the logit model $\epsilon_{1i}^\dagger$ and $\epsilon_{1i}^{\dagger *}$ have a logistic distribution and $S = \frac{15\pi}{16\sqrt{3}}$.

Consider now equation (2.10) described in the previous section. For binary outcome equation (2.10) corresponds to the *mis-specified* latent outcome model. Let $\epsilon_i^{mis} = \epsilon_{1i} - \lambda u_i(\theta_2 + \theta_3 A_i) + (1-\lambda)\epsilon_{2i}(\theta_2 + \theta_3 A_i) + \theta_3(1-\lambda)\beta_2'(A_i C_i - E(AC)) + \theta_3(1-\lambda)\beta_1(A_i^2 - E(A)^2)$, and denote with $\sigma^2_{\epsilon^{mis}}$ the variance of ϵ_i^{mis}, model (2.10) can be rewritten, multiplying both sides by $S/\sigma^2_{\epsilon_1^{mis}}$, as

$$\begin{aligned} Y_i^\dagger \;=\; & (\theta_0^\dagger + \theta_2^\dagger(1-\lambda)\beta_0 + \theta_3^\dagger(1-\lambda)\beta_2' E(AC) + \theta_3^\dagger(1-\lambda)\beta_1 E(A^2)) + (\theta_1^\dagger + \theta_2^\dagger(1-\lambda)\beta_1 + \\ & + \theta_3^\dagger(1-\lambda)\beta_0) A_i + \theta_2^\dagger \lambda M_i^* + \theta_3^\dagger \lambda A * M^* + (\theta_4^{\dagger'} + \theta_2^\dagger(1-\lambda)\beta_2') C_i + \epsilon_{1i}^\dagger - \lambda u_i(\theta_2^\dagger + \\ & + \theta_3^\dagger A_i) + (1-\lambda)\epsilon_{2i}(\theta_2^\dagger + \theta_3^\dagger A_i) + \theta_3^\dagger(1-\lambda)\beta_2'(A_i C_i - E(AC)) + \\ & + \theta_3^\dagger(1-\lambda)\beta_1(A_i^2 - E(A)^2) \end{aligned} \tag{2.15}$$

Since the estimation of the θ parameters now involves also the variance of ϵ_{1i}, which is mis-specified too, the bias formulae are more complex.

When there is no exposure-mediator interaction the probability limit can be derived by direct comparison of equations (2.13), (2.14), and (2.15) dropping all θ_3 terms. Since

$Cov(M^*, M_i - E[M_i|M^*, A_i, C_i]) = Cov(A_i, M_i - E[M_i|M^*, A_i, C_i]) = Cov(C_i, M_i - E[M_i|M^*, A_i, C_i]) = 0$, by comparing (2.14) and (2.15) we note that the estimators of the *observed* regression parameters will converge to $\theta_0^{*\dagger} = \theta_0^\dagger + \theta_2^\dagger(1-\lambda)\beta_0$, $\theta_1^{*\dagger} = \theta_1^\dagger + \theta_2^\dagger(1-\lambda)\beta_1$, $\theta_2^{*\dagger} = \theta_2^\dagger\lambda$, and $\theta_4^{'*\dagger} = \theta_4^{\dagger'} + \theta_2^\dagger(1-\lambda)\beta_2'$.

Rewrite equation (2.15) as

$$Y_i^\dagger = \theta_0^{mis\dagger} + \theta_1^{mis\dagger}A_i + \theta_2^{mis\dagger}M_i^* + \theta_3^{mis\dagger}A * M^* + \theta_4^{mis\dagger'}C_i + \epsilon_{1i}^{mis\dagger} \tag{2.16}$$

By comparing equations (2.16) and (2.13) we further note that in absence of exposure-mediator interaction for $i = 0, 1, 2, 4$, $\frac{\theta_i^{mis\dagger}}{\theta_i^\dagger} = \frac{\theta^{mis}(S/\sigma^2_{\epsilon_1^{mis}})}{\theta(S/\sigma^2_{\epsilon_1})} = \frac{1}{(1+\theta_2^2\lambda\sigma_u^2/S^2)^{\frac{1}{2}}}$. Therefore, we can redefine the probability limit for the outcome regression parameters as

$$\begin{aligned}
plim_{n\to\infty}\hat{\theta}_{0n}^* &= \frac{\theta_0 + \theta_2(1-\lambda)\beta_0}{\tau} \\
plim_{n\to\infty}\hat{\theta}_{1n}^* &= \frac{\theta_1 + \theta_2(1-\lambda)\beta_1}{\tau} \\
plim_{n\to\infty}\hat{\theta}_{2n}^* &= \frac{\theta_2\lambda}{\tau} \\
plim_{n\to\infty}\hat{\theta}_{4n}^{*'} &= \frac{\theta_4' + \theta_2(1-\lambda)\beta_2'}{\tau},
\end{aligned}$$

where $\tau = (1+\theta_2^2\lambda\sigma_u^2/S^2)^{\frac{1}{2}}$ with $S = \frac{15\pi}{16\sqrt{3}} \sim 1.7$ when logit link is used and S=1 for probit link.

When exposure-mediator interaction is present, again $Cov(M^*, M_i - E[M_i|M^*, A_i, C_i]) = Cov(A_i, M_i - E[M_i|M^*, A_i, C_i]) = Cov(C_i, M_i - E[M_i|M^*, A_i, C_i]) = 0$, but the terms $(A_iC_i - E(AC))$ and $(A_i^2 - E(A^2))$ that appear in equation (2.15), are embedded in the error term of the *observed* outcome model and are correlated with the variables specified in the observed model. This introduces an important difference from the previous section. We can see that the mis-specified model and the observed data model won't induce the same type of GLM in general. This prevents us to directly compare the observed and mis-specified outcome models as we did before.

In this case, measurement error causes bias in two ways, one way is through the mis-specification of the coefficients and the other one is by omitted correlated variables. We

have noted in fact that two new terms that are correlated with the covariates in the model, arise, namely $A_iC_i - E(AC)$ and $A_i^2 - E(A^2)$. Therefore, the bias will depend on the mis-specification that is directly visible by comparing (2.14) and (2.15), on the particular link function that is used, and on the conditional distributions of the omitted variables given each variable in the *observed* outcome model. The latter feature is the most problematic and induces a modification of the link function in an almost unpredictable way.
Following the reasoning of Neuhaus and Jewell (1993), we note that the parameter estimates from the *observed* regression will approximately converge to

$$\begin{aligned}
plim_{n\to\infty}\hat{\theta}^*_{0n} &= \{\theta_0 + \theta_2(1-\lambda)\beta_0 + \theta_3(1-\lambda)[\beta_2' E(AC) + \beta_1 E(A^2) + \\
&\quad + \delta_I E[X^{*T} * (AC - E(AC))]\beta_2 + \beta_1\delta_I E[X^{*T} * (A^2 - E(A^2))]]\} H_I(0) \\
plim_{n\to\infty}\hat{\theta}^*_{1n} &= \{\theta_1 + \theta_2(1-\lambda)\beta_1 + \theta_3(1-\lambda)[\beta_0 + \delta_A E[X^{*T} * (AC - E(AC))]\beta_2 + \\
&\quad + \beta_1\delta_A E[X^{*T} * (A^2 - E(A^2))]]\} * H_A(0) \\
plim_{n\to\infty}\hat{\theta}^*_{2n} &= \{\theta_2\lambda + \theta_3(1-\lambda)\delta_{M^*}[E[X^{*T} * (AC - E(AC))]\beta_2 + \beta_1 E[X^{*T} * (A^2 + \\
&\quad - E(A^2))]]\} * H_{M^*}(0) \\
plim_{n\to\infty}\hat{\theta}^*_{3n} &= \theta_3[\lambda + (1-\lambda)\delta_{AM^*}\{E[X^{*T} * (AC - E(AC))]\beta_2 + \beta_1 E[X^{*T} * (A^2 + \\
&\quad - E(A^2))]]\}] * H_{AM^*}(0) \\
plim_{n\to\infty}\hat{\theta}^{*'}_{4n} &= \{(\theta_4' + \theta_2(1-\lambda)\beta_2' + \theta_3(1-\lambda)\delta_C E[X^{*T} * (AC - E(AC))]\beta_2 + \theta_3(1 + \\
&\quad - \lambda)\beta_1\delta_C E[X^{*T} * (A^2 - E(A^2))])\} * H_C(0)
\end{aligned}$$

Where, $H_X(0)$ is a function that depends on the variance of ϵ_1^{mis}, the error term in the mis-specified outcome model, previously defined, as well as on the type of link function chosen, and the joint conditional distribution of $A_iC_i - E(AC)$ and $A_i^2 - E(A^2)$ given X. Let $\mu_0 = E[Y|X, \alpha]$ where X is the variable whose coefficient we want to estimate and α is particular value of the omitted covariates effect. Intuitively, $H_X(0)$ expresses the sensitivity of the mean of Y given X to a change in the effect of the omitted variables on Y, which depends of course on the type of link function is used. The function $H_X(0)$ is equal to one in the case of linear link, and therefore we get the same result as the one obtained assuming an OLS model for the continuous outcome. The function $H_X(0)$ is rather complex for logit and probit links. For the logit link $H_X(0) = (1 - \frac{var\{\mu_0\}}{E\{\mu_0\}(1-E\{\mu_0\})})$,

for the probit link $H_X(0) = \frac{E\phi\{\Phi^{-1}(\mu_0)\}}{\phi[\Phi^{-1}\{E(\mu_0)\}]}$. In general this functional is not recoverable in closed form. However, a numerical bias analysis can still be carried out.

In order to implement the numerical bias analysis one needs to write the score equation under the naive model and take its expectation under the true model and solve for the naive parameters as a function of the true parameters (Wang et al., 1998). Finally, if we assume a binary exposure so that $A_i = A_i^2$, then the two terms can be incorporated and if additionally the true model included the exposure-covariates interaction terms then the asymptotic limit of the estimators of the regression coefficients could be easily derived in closed form as was shown in the case of no exposure-mediator interaction because the observed and mis-specified model would be directly comparable since no omitted covariates would appear in the error term anymore.

When a logarithmic link is used for binary, log-linear or count outcomes $H_X(0) = 1$, therefore the probability limit of the parameter estimates θ_1, θ_2, and, in presence of interaction, θ_3, that are involved in the direct and indirect effect estimation, is equal to the one derived for the linear-link.

Asymptotic bias of causal effects when the mediator is measured with error

Let, $\gamma_1 = \delta_A E[X^{*T} * (AC - E(AC))]$, $\gamma_2 = \delta_{M^*} E[X^{*T} * (AC - E(AC))]$, $\gamma_3 = \delta_{AM^*} E[X^{*T} * (AC - E(AC))]$, $\gamma_4 = \delta_A E[X^{*T} * (A^2 - E(A^2))]$, $\gamma_5 = \delta_{M^*} E[X^{*T} * (A^2 - E(A^2))]$, and $\gamma_6 = \delta_{AM^*} E[X^{*T} * (A^2 - E(A^2))]$. Note that because the natural direct effect involves $\sigma^2 = \sigma^2_{m|a,c}$, the conditional variance of the mediator given the exposure and additional covariates, we need to consider also the bias of the estimate $\hat{\sigma}^{*2} = \sigma^2_{m*|a,c}$ from the observed mediator regression which converges in probability to $\sigma^{*2} = \sigma^2 + \sigma^2_u$. Assuming exposure-mediator interaction is present, when the outcome is binary and a logit link is used the asymptotic

bias for controlled direct effect, natural direct effect and natural indirect effect is given by

$$
\begin{aligned}
ABIAS(log(\widehat{OR}^{CDE})) &= [\theta_1(H_A(0)-1)+\theta_2(1-\lambda)\beta_1 H_A(0)+\theta_3\{H_A(0)((1-\lambda)\beta_0+ \\
&\quad +(1-\lambda)\gamma_1\beta_2+(1-\lambda)\beta_1\gamma_4)+H_{AM^*}(0)(\lambda+(1-\lambda)\gamma_3\beta_2+ \\
&\quad +(1-\lambda)\beta_1\gamma_6)-m\}](a-a^*) \\
ABIAS(log(\widehat{OR}^{NDE})) &= \{H_A(0)[\theta_1+\theta_2(1-\lambda)\beta_1+\theta_3((1-\lambda)\beta_0+(1-\lambda)\gamma_1\beta_2+ \\
&\quad +(1-\lambda)\beta_1\gamma_4)]+H_{AM^*}(0)\theta_3[\beta_0+\beta_1 a^*+\beta_2' c+ \\
&\quad +(\theta_2\lambda+\theta_3((1-\lambda)\gamma_2\beta_2+(1-\lambda)\beta_1\gamma_5)\sigma^{*2})]\times \\
&\quad \times(\lambda+(1-\lambda)\gamma_3\beta_2+(1-\lambda)\beta_1\gamma_6)\}(a-a^*)+(a^2+a^{*2})\times \\
&\quad \times[H_{AM^*(0)}\theta_3\sigma^*(\lambda+(1-\lambda)\gamma_3\beta_2+(1-\lambda)\beta_1\gamma_6)]^2-OR^{NDE} \\
ABIAS(log(\widehat{OR}^{NIE})) &= \{\theta_2[\lambda H_{M^*}(0)-1]+\theta_3[H_{M^*}(0)((1-\lambda)\gamma_2\beta_2+(1-\lambda)\beta_1\gamma_5)+ \\
&\quad +aH_{AM^*}(0)(\lambda+(1-\lambda)\gamma_3\beta_2+(1-\lambda)\beta_1\gamma_6)-1)]\}\beta_1(a-a^*)
\end{aligned}
$$

In absence of exposure-mediator interaction the formulas simplify substantially and the asymptotic bias is given by

$$
\begin{aligned}
ABIAS(log(\widehat{OR}^{CDE})) &= ABIAS(log(\widehat{OR}^{NDE})) = [\theta_1(\frac{1}{\tau}-1)+\frac{\theta_2(1-\lambda)\beta_1}{\tau}](a-a^*) \\
ABIAS(log(\widehat{OR}^{NIE})) &= [\theta_2\beta_1(\frac{\lambda}{\tau}-1)](a-a^*)
\end{aligned}
$$

We turn now to consider the asymptotic bias of the causal effects estimators when a probit link is used to model the binary outcome. For this purpose we define the causal effect in the risk difference scale:

$$
\begin{aligned}
RD^{CDE} &= \Phi\{\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c\}-\Phi\{\theta_0+\theta_1 a^*+\theta_2 m+\theta_3 a^* m+\theta_4' c\} \\
RD^{NDE} &= \int_m [\Phi\{\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c\}-\Phi\{\theta_0+\theta_1 a^*+\theta_2 m+\theta_3 a^* m+\theta_4' c\}]\times \\
&\quad \times f_{M|AC}(m|a^*,c)dm \\
RR^{NIE} &= \int_m \Phi\{\theta_0+\theta_1 a^*+\theta_2 m+\theta_3 a^* m+\theta_4' c\}[f_{M|AC}(m|a,c)-f_{M|AC}(m|a^*,c)]dm
\end{aligned}
$$

We see that when the probit link is specified, the causal effect estimators are function potentially of all the parameters of the outcome regression. This implies that the estimator

will be much more affected by the bias due to measurement error. Moreover, the bias formulae are complex and they don't have a clear interpretation in terms of specific parameters or functions. Therefore, we omit the results on asymptotic bias for direct and indirect causal effects estimators under probit outcome regression.
Finally, the asymptotic biases of direct and indirect effects estimators when the outcome is modeled using a logarithmic link are given by

$$\begin{aligned}
ABIAS(log(\widehat{RR^{CDE}})) &= [\theta_2(1-\lambda)\beta_1 + \theta_3\{(1-\lambda)\beta_0 + (1-\lambda)\gamma_1\beta_2 + (1-\lambda)\beta_1\gamma_4 + \\
&\quad + m(\lambda + (1-\lambda)\gamma_3\beta_2 + (1-\lambda)\beta_1\gamma_6 - 1)\}](a-a^*) \\
ABIAS(log(\widehat{RR^{NDE}})) &= [(1-\lambda)\beta_1\theta_2 + \theta_3((1-\lambda)\beta_0 + (1-\lambda)\gamma_1\beta_2 + (1-\lambda)\beta_1\gamma_4)] \times \\
&\quad \times (a-a^*) + [\theta_3\{((\lambda + (1-\lambda)\gamma_3\beta_2 + (1-\lambda)\beta_1\gamma_6(\beta_0 + \\
&\quad + \beta_1 a^* + \beta_2' c + \sigma^{2*}(\theta_2\lambda + \theta_3((1-\lambda)\gamma_2\beta_2 + (1-\lambda)\beta_1\gamma_5)))) + \\
&\quad - (\beta_0 + \beta_1 a^* + \beta_2' c + \sigma^2\theta_2)\}](a-a^*) + [\theta_3^2\{\sigma^{2*}(\lambda + (1-\lambda)\gamma_3\beta_2 + \\
&\quad + (1-\lambda)\beta_1\gamma_6) - \sigma^2\}]0.5(a^2 - a^{*2}) \\
ABIAS(log(\widehat{RR^{NIE}})) &= [\theta_2(\lambda - 1) + \theta_3((1-\lambda)\gamma_2\beta_2 + (1-\lambda)\beta_1\gamma_5 + a\lambda + a(1-\lambda)\gamma_3\beta_2 + \\
&\quad + a(1-\lambda)\beta_1\gamma_6 - a)]\beta_1(a-a^*).
\end{aligned}$$

Method of Moments Estimators for regression parameters and causal effects

Method of moments estimators for the parameters involved in mediation analysis can be obtained by solving the system of equations that arises from the previous results on the probability limit of naive estimators when the mediator is measured with error.
When exposure-mediator is absent in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2. The method of moments estimators for the outcome regression parameters, when the outcome is modelled using

a log-link, are given by

$$\begin{aligned}
\hat{\theta}_0^{MoM} &= \hat{\theta}_0^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_0 \\
\hat{\theta}_1^{MoM} &= \hat{\theta}_1^* - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_1 \\
\hat{\theta}_2^{MoM} &= \hat{\theta}_2^*/\lambda \\
\hat{\theta}_4^{MoM'} &= \hat{\theta}_4^{*'} - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_2'
\end{aligned}$$

We note that the estimators are the same as the ones we derived for the continuous outcome.

When the outcome is binary the method of moments estimators can be defined as described above, but a better approximation is given by

$$\begin{aligned}
\hat{\theta}_0^{MoM} &= \hat{\theta}_0^*(1+\hat{\theta}_2^{MoM2}\sigma_u^2\lambda/S^2)^{\frac{1}{2}} - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_0 \\
\hat{\theta}_1^{MoM} &= \hat{\theta}_1^*(1+\hat{\theta}_2^{MoM2}\sigma_u^2\lambda/S^2)^{\frac{1}{2}} - \hat{\theta}_2^{MoM}(1-\lambda)\hat{\beta}_1 \\
\hat{\theta}_2^{MoM} &= \frac{\hat{\theta}_2^*}{(\lambda^2 - \hat{\theta}_2^*\lambda\sigma_u^2/S^2)^{\frac{1}{2}}} \\
\hat{\theta}_4^{MoM'} &= \hat{\theta}_4^{*'}(1+\hat{\theta}_2^{MoM2}\sigma_u^2\lambda/S^2)^{\frac{1}{2}} - \hat{\theta}_2^{MoM}(1-\lambda)\beta_2'
\end{aligned}$$

When exposure-mediator is present in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2 and θ_3. The method of moments estimators for these three regression parameters when the outcome is binary is not of operational use since the function $H_X(0)$ previously defined cannot be recovered in general. For binary-log-linear or count outcomes modeled with a glm with log-link the method of moments estimators again coincide with the ones derived for the continuous outcome case. The same estimator could be used as approximate method of moments estimator for binary-logistic and binary-probit outcomes.

Regression Calibration Estimators for regression parameters and causal effects

When exposure-mediator is absent in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2. The regression calibration

estimators for these two regression parameters, when the outcome is binary or count, are given by

$$\begin{aligned}\theta_1^{RC} &= \theta_1^* - \theta_2^* \frac{(1-\lambda)\beta_1}{\lambda} \\ \theta_2^{RC} &= \theta_2^*/\lambda\end{aligned}$$

When exposure-mediator is present in order to consistently estimate direct and indirect effects we need to consistently estimate the parameters θ_1 and θ_2 and θ_3. For both binary and log-linear or count outcomes modelled with a glm with log-link the regression calibration estimators again can be recovered as described in the section on continuous outcome.

Armstrong (1985) showed that for logistic regression regression calibration estimators will yield approximately consistent estimators.

$$P(Y_i|M_i) = expit(\theta_0 + \theta_1 A_i + \theta_2 M_i + \theta_3 A_i M_i + \theta_4' C_i)$$

$$P(Y_i|M_i) \sim expit(\theta_0 + \theta_1 A_i + \theta_2 E[M_i|M_i^*, A_i, C_i] + \theta_3 A_i E[M_i|M_i^*, A_i, C_i] + \theta_4' C_i) +$$

$$+ \frac{\partial}{\partial M_i} expit(\theta_0 + \theta_1 A_i + \theta_2 M_i + \theta_3 A_i M_i + \theta_4' C_i)|_{M_i = E[M_i|M_i^*, A_i, C_i]} (M_i - E[M_i|M_i^*, A_i, C_i]) +$$

$$+ \frac{\partial^2}{\partial M_i^2} expit(\theta_0 + \theta_1 A_i + \theta_2 M_i + \theta_3 A_i M_i + \theta_4' C_i)|_{M_i = E[M_i|M_i^*, A_i, C_i]} (M_i - E[M_i|M_i^*, A_i, C_i])^2.$$

Again, regression calibration estimators for $\hat{\theta}_1^{rc}$, $\hat{\theta}_2^{rc}$, and $\hat{\theta}_3^{rc}$ can be derived by running a logistic regression where M^* is replaced by $E[M_i|M_i^*, A_i, C_i]$. From a careful analysis of the approximation just derived, we can see that when there is no exposure interaction the regression calibration estimators should perform reasonably well if the measurement error is small (i.e. σ_u^2 is small) and if the effect of the mediator on the outcome is not too large in absolute value. When an exposure-mediator interaction is

present the validity of the approximation might be undermined. We note that the term $(M_i - E[M_i|M_i^*, A_i, C_i])^2$ will now depend on the distribution of the exposure variable as well and in general we might not be able to assume that this term takes on small values.

Standard errors of the method of moments estimators for direct and indirect causal effects

We now derive the standard errors of method of moments estimators for controlled direct, natural direct and natural indirect effects assuming that exposure-mediator interaction may be present.
Define the corrected method of moments estimators of the causal effects of interest as

$$\widehat{OR}^{CDE} = \frac{P(Y_{am}=1|c)/(1-P(Y_{am}=1|c))}{P(Y_{a^*m}=1|c)/(1-P(Y_{a^*m}=1|c))}$$
$$= exp[(\hat{\theta}_1^{MoM} + \hat{\theta}_3^{MoM} m)(a - a^*)].$$

$$\widehat{OR}^{NDE} = exp[log\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\}]$$
$$= exp[\{\hat{\theta}_1^{MoM} + \hat{\theta}_3^{MoM}(\hat{\beta}_0 + \hat{\beta}_1 a^* + \hat{\beta}_2' c + \hat{\theta}_2^{MoM}\sigma^{MoM2})\}(a - a^*) + 0.5\hat{\theta}_3^{MoM2}\sigma^{MoM2}(a^2 - a^{*2})].$$

$$\widehat{OR}^{NIE} = exp[log\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\}]$$
$$= exp[(\hat{\theta}_2^{MoM}\hat{\beta}_1 + \hat{\theta}_3^{MoM}\hat{\beta}_1 a)(a - a^*)].$$

$$\widehat{OR}^{TE} = E[Y_a - Y_{a^*}|C = c]$$
$$= \widehat{OR}^{NDE} \times \widehat{OR}^{NIE}.$$

Suppose that model (2.7) has been fit using standard linear regression software, the observed outcome regression model has been fit using a logistic regression model and that the resulting estimates $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}^*$ of $\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^{*\prime})'$ and $\hat{\sigma}^{*2}$ of σ^{*2}

have covariance matrices Σ_β, Σ_{θ^*} and $\Sigma_{\sigma^{*2}}$. Then the covariance matrix of $(\hat{\beta}', \hat{\theta}^{*\prime}, \hat{\sigma}^{*2})$ is

$$\Sigma = \begin{bmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_{\theta^*} & 0 \\ 0 & 0 & \Sigma_{\sigma^{*2}} \end{bmatrix}$$

Suppose further that method of moments estimators of the outcome regression parameters are obtained as described in the previous section $\hat{\theta}^{MoM} = (\hat{\theta}_0^{MoM}, \hat{\theta}_1^{MoM}, \hat{\theta}_2^{MoM}, \hat{\theta}_3^{MoM}, \hat{\theta}_4^{MoM\prime})'$ and a consistent estimator for σ^2 is given by $\hat{\sigma}^{MoM2} = \hat{\sigma}^{*2}\hat{\lambda}$, with $\hat{\lambda} = \frac{\hat{\sigma}^{*2} - \sigma_u^2}{\hat{\sigma}^{*2}}$. Recall that in the previous section we noted that the method of moments estimator in presence of interaction is not of operational use since it's typically hard to recover the function $H_X(0)$. The method of moments estimators described for the linear outcome case are equivalent to the method of moments estimators of the regression parameters when the outcome is modeled using a logarithmic link and can be considered an approximation to the method of moments estimators for binary outcome regression parameters.
Then the covariance matrix of $(\hat{\beta}', \hat{\theta}^{MoM\prime}, \hat{\sigma^{MoM2}})$ is

$$\Sigma^{MoM} = \begin{bmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_{\theta^{MoM}} & 0 \\ 0 & 0 & \Sigma_{\sigma^{MoM2}} \end{bmatrix}$$

Where $\Sigma_{\sigma^{MoM2}} = (\frac{\partial \hat{\sigma}^{MoM2}}{\partial \hat{\sigma}^{*2}})^2 \Sigma_{\sigma^{*2}} = \Sigma_{\sigma^{*2}}$ and $\Sigma_{\theta^{MoM}}$ is found using the multivariate delta method.

$$\Sigma_{\theta^{MoM}} = D^{MoM} \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_{\theta^*} \end{bmatrix} D^{MoM^T}$$

where D^{MoM} is recovered as we described in the continuous outcome section.

Standard errors of the method of moments controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma^{MoM} \Sigma^{MoM} \Gamma^{MoM\prime}} |a - a^*|$$

with $\Gamma^{MoM} = (0, 0, 0^{'}, 0, 1, 0, m, 0^{'}, 0)$ for the log of controlled direct effect odds ratio, $\Gamma^{MoM} = (\theta_3^{MoM}, \theta_3^{MoM}a^*, \theta_3^{MoM}c^{'}, 0, 1, \theta_3^{MoM}\sigma^{MoM2}, \beta_0 + \beta_1 a^* + \beta_2^{'}c + \theta_2^{MoM}\sigma^{MoM2} + \theta_3^{MoM}\sigma^{MoM2}(a+a^*), 0^{'}, \theta_2^{MoM}\theta_3^{MoM} + 0.5\theta_3^{MoM2}(a+a^*))$ for the log pure natural direct effect odds ratio (same expression holds for the total natural direct effect upon substituting a and a^*), $\Gamma^{MoM} = (0, \theta_2^{MoM} + \theta_3^{MoM}a, 0^{'}, 0, 0, \beta_1, \beta_1 a, 0^{'}, 0)$ for the log of total natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*), $\Gamma^{MoM} = (\theta_3^{MoM}, \theta_3^{MoM}(a + a^*) + \theta_2^{MoM}, \theta_3^{MoM}c^{'}, 0, 1, \theta_3^{MoM}\sigma^{MoM2} + \beta_1, \beta_0 + \beta_1(a + a^*) + \beta_2^{'}c + \theta_2^{MoM}\sigma^{MoM2} + \theta_3^{MoM}\sigma^{MoM2}(a^2 - a^{*2}), 0^{'}, 0.5\theta_3^{MoM2}(a^2 - a^{*2}))$ for the logarithm of the total effect.

Standard errors of the method of moments and regression calibration estimators of the causal effects of interest can be obtained in absence of exposure-mediator interaction in a similar way by setting $\theta_3 = 0$.

In absence of exposure mediator interaction if we employ the approximate method of moments estimator for binary logistic or probit link (and is exact for log link), that coincides with the regression calibration estimator, the standard errors of the controlled and natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma^{MoM}\Sigma^{MoM}\Gamma^{MoM'}}|a - a^*|$$

with $\Gamma^{MoM} = (0, 0, 0^{'}, 0, 1, 0, 0^{'}, 0)$ for the log of controlled direct effect odds ratio and for the log pure natural direct effect odds ratio, $\Gamma^{MoM} = (0, \theta_2^{MoM}, 0^{'}, 0, 0, \beta_1, 0^{'}, 0)$ for the log of total natural indirect effect (the same expression holds for the pure natural indirect effect upon substituting a and a^*), $\Gamma^{MoM} = (0, \theta_2^{MoM}, 0', 0, 1, \beta_1, 0^{'}, 0)$ for the logarithm of the total effect.

and

$$\Sigma^{MoM} = \begin{bmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta^{MoM} & 0 \\ 0 & 0 & \Sigma_{\sigma^{2MoM}} \end{bmatrix}$$

where $\Sigma_{\sigma^{MoM2}} = \Sigma_{\sigma^{*2}}$ and

$$\Sigma_\theta^{MoM} = D^{MoM} \begin{bmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_{\theta^*} \end{bmatrix} D^{MoM^T}$$

with D^{MoM} defined as in the section of continuous outcome.

We have also seen that, when the exposure-mediator interaction is absent, the method of moments estimator can be improved when the mediator is binary and modeled with either logit or probit link. The standard errors for the improved method of moments estimators of the causal effects in absence of exposure-mediator interaction have the same form of the ones given above with D^{MoM} defined as follows. Recall that $\theta_2^{MoM} = \frac{\theta_2^*}{\lambda^2 - \theta_2^{*2}\lambda\sigma_u^2/S^2}$ and let

$$\frac{\partial\theta_0^{MoM}}{\partial\beta_0} = \frac{\partial\theta_1^{MoM}}{\partial\beta_1} = \frac{\partial\theta_4^{MoM'}}{\partial\beta_2'} = -\theta_2^{MoM}(1-\lambda)$$

$$\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} = \frac{(\lambda^2 - \theta_2^{*2}\lambda\sigma_u^2/S^2) + 2\lambda\sigma_u^2\theta_2^4/S^2}{(\lambda^2 - \theta_2^{*2}\lambda\sigma_u^2/S^2)^2}$$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_0^*} = \frac{\partial\theta_1^{MoM}}{\partial\theta_1^*} = \frac{\partial\theta_4^{MoM'}}{\partial\theta_4^{*'}} = (1 + [\theta_2^{MoM2}\lambda\sigma_u^2]/S^2)^{\frac{1}{2}}$$

$$\frac{\partial\theta_0^{MoM}}{\partial\theta_2^*} = \frac{1}{2}\theta_0^*(1 + [\theta_2^{MoM2}\lambda\sigma_u^2]/S^2)\frac{2\lambda\sigma_u^2}{S^2}\theta_2^{MoM}\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} - (1-\lambda)\beta_0\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*}$$

$$\frac{\partial\theta_1^{MoM}}{\partial\theta_2^*} = \frac{1}{2}\theta_1^*(1 + [\theta_2^{MoM2}\lambda\sigma_u^2]/S^2)\frac{2\lambda\sigma_u^2}{S^2}\theta_2^{MoM}\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} - (1-\lambda)\beta_1\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*}$$

$$\frac{\partial\theta_4^{MoM'}}{\partial\theta_2^*} = \frac{1}{2}\theta_4^*(1 + [\theta_2^{MoM2}\lambda\sigma_u^2]/S^2)\frac{2\lambda\sigma_u^2}{S^2}\theta_2^{MoM}\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} - (1-\lambda)\beta_2'\frac{\partial\theta_2^{MoM}}{\partial\theta_2^*}$$

Then,

$$D^{MoM} = \begin{bmatrix} \frac{\partial\theta_0^{MoM}}{\partial\beta_0} & 0 & 0 & \frac{\partial\theta_0^{MoM}}{\partial\theta_0^*} & 0 & \frac{\partial\theta_0^{MoM}}{\partial\theta_2^*} & 0 \\ 0 & \frac{\partial\theta_0^{MoM}}{\partial\beta_1} & 0 & 0 & \frac{\partial\theta_1^{MoM}}{\partial\theta_1^*} & \frac{\partial\theta_1^{MoM}}{\partial\theta_2^*} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\partial\theta_2^{MoM}}{\partial\theta_2^*} & 0 \\ 0 & 0 & \frac{\partial\theta_0^{MoM}}{\partial\beta_2'} & 0 & 0 & \frac{\partial\theta_4^{MoM'}}{\partial\theta_2^*} & \frac{\partial\theta_4^{MoM'}}{\partial\theta_4^{*'}} \end{bmatrix}.$$

Table 2.4: Simulations results for continuous outcome

($\sigma_u^2 = 0.1, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.093	0.002	0.002	0.005	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.000
NIE	-0.092	-0.002	-0.002	-0.005	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000
TE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
($\sigma_u^2 = 0.5, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.335	0.002	0.002	0.1	0.000	0.000	0.000	0.000	0.11	0.000	0.000	0.01
NIE	-0.333	0.000	0.000	-0.01	0.000	0.000	0.000	0.000	0.1	0.000	0.000	0.01
TE	0.002	0.002	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
($\sigma_u^2 = 0.1, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect ($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.153	-0.013	-0.002	0.002	0.000	0.000	0.000	0.000	0.024	0.000	0.000	0.000
NDE	0.156	0.010	0.001	0.005	0.001	0.001	0.001	0.001	0.025	0.001	0.000	0.001
NIE	0.015	0.184	-0.003	-0.003	0.001	0.002	0.001	0.001	0.002	0.036	0.001	0.001
TE	0.174	0.174	-0.003	0.001	0.002	0.002	0.002	0.002	0.032	0.032	0.002	0.002
($\sigma_u^2 = 0.5, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.588	-0.23	-0.004	0.161	0.001	0.001	0.001	0.001	0.347	0.057	0.001	0.027
NDE	0.586	-0.21	-0.002	0.159	0.001	0.002	0.002	0.002	0.34	0.056	0.002	0.02
NIE	0.084	0.89	-0.001	0.015	0.002	0.005	0.002	0.002	0.008	0.81	0.002	0.003
TE	0.67	0.66	-0.003	0.175	0.002	0.002	0.002	0.002	0.453	0.446	0.002	0.03

Table 2.5: Simulation results for binary (logistic link) outcome

($\sigma_u^2 = 0.1, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect ($\theta_3 = 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.117	-0.002	-0.002	-0.004	0.009	0.006	0.007	0.008	0.02	0.006	0.006	0.008
NIE	-0.275	0.006	-0.047	-0.02	0.01	0.01	0.01	0.01	0.08	0.01	0.01	0.01
TE	-0.069	-0.003	-0.069	-0.043	0.06	0.06	0.04	0.06	0.069	0.04	0.04	0.065
($\sigma_u^2 = 0.5, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
NDE	0.463	0.001	-0.012	0.142	0.01	0.007	0.01	0.01	0.23	0.007	0.01	0.03
NIE	-0.827	-0.007	-0.117	-0.316	0.003	0.015	0.01	0.01	0.68	0.015	0.029	0.11
TE	-0.189	-0.01	-0.189	0.069	0.05	0.04	0.05	0.06	0.09	0.04	0.09	0.06
($\sigma_u^2 = 0.1, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect ($\theta_3 \neq 0$)	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.163	0.008	0.016	-0.017	0.014	0.011	0.012	0.014	0.041	0.011	0.013	0.013
NDE	-0.122	-0.097	0.012	0.024	0.002	0.002	0.004	0.004	0.017	0.012	0.004	0.005
NIE	0.004	-0.026	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TE	-0.031	-0.04	0.000	-0.005	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000
($\sigma_u^2 = 0.5, n = 10,000$)	**Bias**				**Variance**				**MSE**			
Effect	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX	Naive	MoM	RC	SIMEX
CDE	0.452	0.169	-0.06	0.155	0.02	0.009	0.01	0.02	0.226	0.038	0.015	0.044
NDE	-0.323	-0.342	0.023	0.192	0.000	0.000	0.006	0.002	0.105	0.118	0.006	0.039
NIE	0.013	-0.035	0.018	0.046	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.002
TE	-0.08	-0.106	0.003	-0.036	0.000	0.000	0.000	0.000	0.008	0.011	0.000	0.001

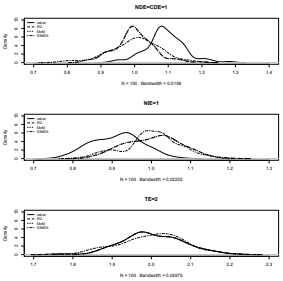

Figure 2.3: Density of causal effect estimators: small error, linear Y model and $\theta_3 = 0$.

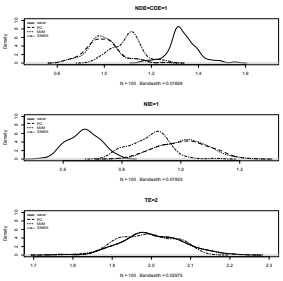

Figure 2.4: Density of causal effect estimators: moderate error, linear Y model and $\theta_3 = 0$.

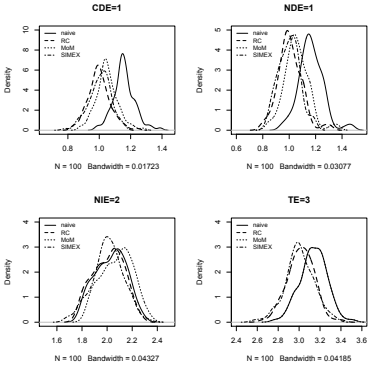

Figure 2.5: Density of causal effect estimators: small error, linear Y model and $\theta_3 \neq 0$.

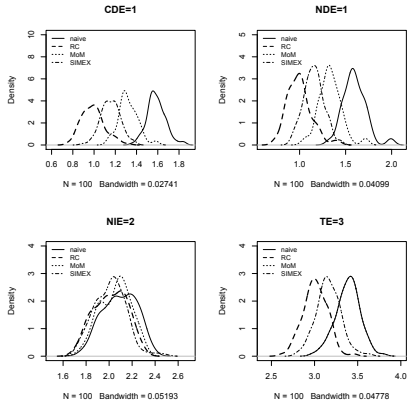

Figure 2.6: Density of causal effect estimators: moderate error, linear Y model and $\theta_3 \neq 0$.

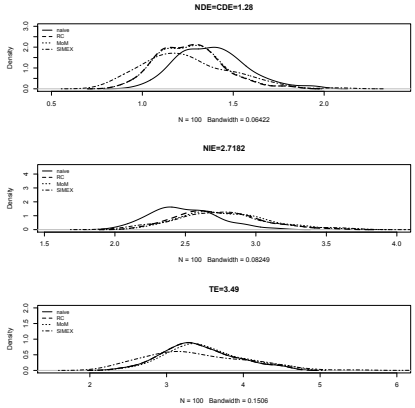

Figure 2.7: Density of causal effect estimators: small error, logistic Y model and $\theta_3 = 0$.

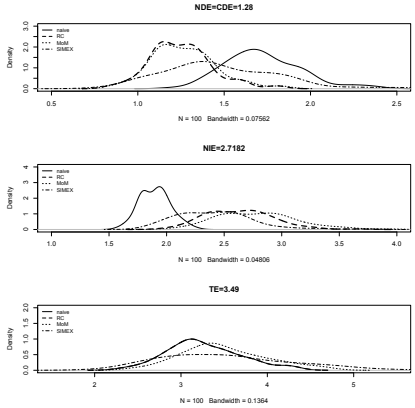

Figure 2.8: Density of causal effect estimators: moderate error, logistic Y model and $\theta_3 = 0$.

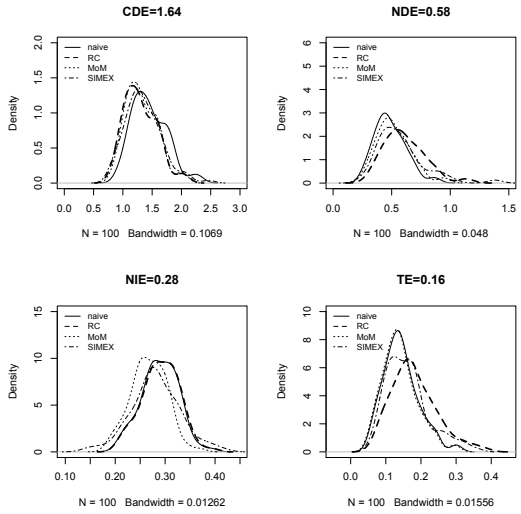

Figure 2.9: Density of causal effect estimators: small error, logistic Y model and $\theta_3 \neq 0$.

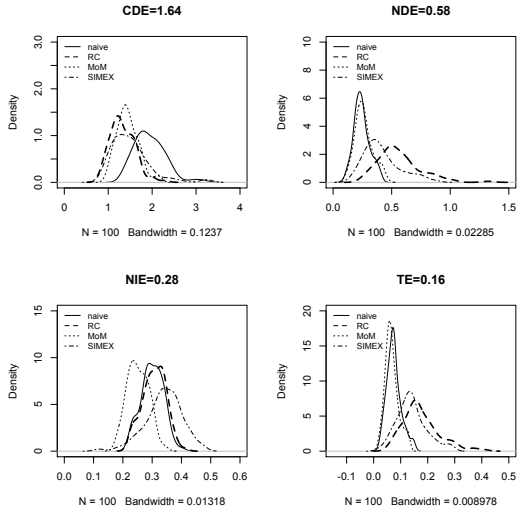

Figure 2.10: Density of causal effect estimators: moderate error, logistic Y model and $\theta_3 \neq 0$.

The estimation of direct and indirect causal effects in the presence of a misclassified binary mediator

[1]Linda Valeri and [1,2] Tyler J. VanderWeele

[1]Department of Biostatistics, Harvard School of Public Health and
[2]Department of Epidemiology, Harvard School of Public Health

Abstract

Mediation analysis serves to quantify the effect of an exposure on an outcome mediated by a certain intermediate, and to quantify the extent to which the effect is direct. When the mediator is misclassified the validity of mediation analysis can be severely undermined. The contribution of the present work is to study the effects of non-differential misclassification of a binary mediator in the estimation of direct and indirect causal effects when the outcome is either continuous or binary and exposure-mediator interaction can be present, and to allow for the correction of misclassification. A full maximum likelihood approach and an hybrid of likelihood-based and predictive value weighting method for misclassification correction coupled with sensitivity analyses are proposed and compared for which no validation samples or gold standard for the misclassified mediator are needed. The approaches are applied to a perinatal epidemiological study on the determinants of pre-term birth.

Keywords: Iteratively re-weighted least squares; Maximum likelihood; Mediation analysis; Misclassification; Predictive value weighting; Sensitivity analyses.

3.1 Introduction

Causal mediation analysis investigates the role of intermediate variables (mediators) in explaining the mechanisms through which an exposure variable exerts a causal effect on an outcome variable. A mediational model hypothesizes that the exposure variable causes the mediator variable, which in turn causes the outcome variable (MacKinnon, 2008). The use of mediation analysis in biomedical and social sciences is widespread and has been strongly influenced by the seminal paper of Baron and Kenny (1986). More recently, new advances in mediation analysis have been made by applying the counterfactual framework in this field (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010). The use of the counterfactual framework has allowed for definitions of direct and indirect effects and for decomposition of a total effect into direct and indirect effects even in models with interactions and non-linearities. The property of effect decomposition is particularly appealing because in many contexts investigators are interested in assessing whether most of the effect is mediated through a particular intermediate or the extent to which it is through other pathways. Decomposition of a total effect into direct and indirect effects accomplishes this goal. VanderWeele and Vansteelandt (2009, 2010) showed how the notion of direct and indirect causal effects from causal inference in the counterfactual framework (Greenland and Robins, 1992; Pearl, 2001) can extend the regression approach to mediation analysis of Baron and Kenny to settings in which there is an interaction term between exposure and mediator in the outcome regression.

Mediation analysis is often performed as a secondary study, after the causal effect of an exposure on an outcome has been investigated, to deepen the understanding of the mechanisms. A shortcoming of mediation analysis in such studies is that while effort may be made to measure the exposure and outcome with high precision, less attention may be given to correctly measuring the mediator variable. Therefore, it is of interest to consider the consequence of measurement error or misclassification of the mediator in the estimation of the direct and indirect causal effects.

The problem of measurement error in mediation analysis has been explored for the simple and linear mediation model by Hoyle and Kenny (1999). VanderWeele et al. (2012) and Valeri et al. (2012) studied the impact of measurement error on a continuous mediator variable when direct and indirect causal effects are estimated using generalized linear models in the presence of exposure-mediator interaction. Misclassification of a binary mediator has been considered by Ogburn et al. (2012) in a non-parametric setting. To our knowledge no rigorous study of the misclassification problem has been proposed when mediation analysis is carried out using a parametric approach, allowing for the presence of non-linearities such as exposure-mediator interaction.

In the context of a regression-based approach to mediation analysis, the investigator needs to estimate the parameters from the *outcome* and *mediator regressions*. Then, direct and indirect causal effects are recovered as functions of those regression parameters, provided the models have been correctly specified and the no confounding assumptions described below hold (VanderWeele and Vansteelandt, 2009, 2010). The reader can refer to the next section for an explanation of the results that have been derived for parametric inference for direct and indirect causal effects under a counterfactual framework. When a binary mediator is misclassified, in order to understand if the estimators of the causal effects of interest are still valid, it is crucial to investigate how misclassification affects the estimation of outcome and mediator regressions' parameters. In this study, we use results that have been derived about the consequences of misclassification on parameter estimators in parametric regression models when an outcome or a covariate is misclassified (Neuhaus, 1999; Gustafson, 2004; Carroll et al., 2006).

The present work makes two contributions. First, we study the implications of non-differential misclassification of the mediator variable on the validity of mediation analysis. Assuming a continuous outcome modeled using linear regression, we derive the asymptotic bias of direct and indirect causal effects estimators in closed form. The asymptotic bias formulae are given assuming that exposure and mediator may interact in their effect on the outcome. We demonstrate that even if the error is assumed to be non-differential, regression coefficient estimators obtained in mediation analysis ignoring misclassification can sometimes be severely biased and therefore induce bias in the estima-

tion of causal direct and indirect effects. The second contribution of the present work is to propose strategies for misclassification correction that yield consistent or approximately consistent estimators of the direct and indirect causal effects under non-differential misclassification model. We propose a correction strategy for misclassification of a binary mediator when the outcome is either continuous or binary, allowing for exposure-mediator interaction. The correction approach is coupled with sensitivity analyses when no gold standard or validation samples for the mis-measured mediator are available. In particular, we evaluate and compare the performance of misclassification-corrected estimators for direct and indirect causal effects using iteratively reweighed least squares (IRLS) approach for outcome misclassification (Neuhaus, 1999; Carroll et al., 2006) paired with a predictive value weighting (PVW) approach for covariate misclassification or a fully likelihood-based (ML) method (Lyles and Lin, 2010; Carroll et al., 2006).

The paper is organized as follows. Section 3.2 defines direct and indirect causal effects and discusses some results from mediation analysis. Section 3.3 describes the mediator misclassification model, and studies the asymptotic bias in direct and indirect causal effects when the binary mediator is misclassified. In Section 3.4 we describe the approaches for misclassification correction and we evaluate their performance in estimating direct and indirect causal effects via a simulation study. In Section 3.5 we apply the proposed methods to a perinatal epidemiological study, followed by discussion in Section 3.6.

3.2 Mediation analysis within the counterfactual framework in the absence of misclassification

Let A be an exposure or treatment, Y an outcome, M a mediator and $\boldsymbol{C}$ a k-dimensional vector of covariates. The causal diagram in Figure 3.1 captures how the role of a mediator variable can be conceptualized. In the figure, the exposure can have an effect on the outcome by either exerting a causal effect on the mediator which in turn is causally related to the outcome, or by affecting the level of the outcome independently of its impact on the intermediate variable.

Let Y_a and M_a denote respectively the values of the outcome and mediator that would

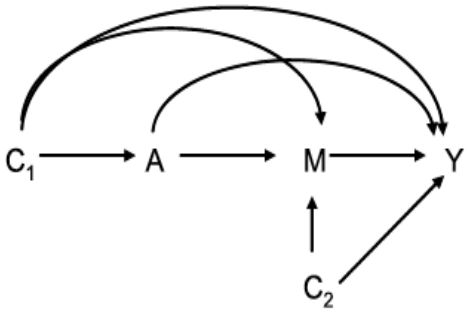

Figure 3.1: Mediation Directed Acyclic Graph (DAG)

have been observed had the exposure *A* been set to level *a*. Let Y_{am} denote the value of the outcome that would have been observed had the exposure, A, and mediator, M, been set to levels a and m, respectively. Given these counterfactual variables, the following causal effects can be non-parametrically defined. The controlled direct effect (CDE), defined by $E[Y_{am} - Y_{\tilde{a}m}|\boldsymbol{C}]$, expresses how much the outcome would change on average if the mediator were controlled at level m uniformly in the population but the treatment were changed from level $\tilde{a}$ to level a e.g. for a binary exposure from $\tilde{a} = 0$ to $a = 1$. The natural direct effect (NDE), defined by $E[Y_{aM_{\tilde{a}}} - Y_{\tilde{a}M_{\tilde{a}}}|\boldsymbol{C}]$, measures how much the mean of the outcome would change if the exposure were set at level *a* versus level $\tilde{a}$ but the mediator were kept at the level it would have taken under $\tilde{a}$. The natural indirect effect (NIE), defined by $E[Y_{aM_a} - Y_{aM_{\tilde{a}}}|\boldsymbol{C}]$, measures how much the mean of the outcome would change if the exposure were controlled at level a, but the mediator were changed from the level it would take under $\tilde{a}$ to the level it would take under *a*.

While controlled direct effects are often of greater interest in policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects are particularly of interest in eval-

uating the action of various mechanisms (Robins, 2003; Joffe et al., 2007). An important property of the natural indirect effect and the natural direct effect is that the total effect decomposes into the sum of these two effects, $TE = E[Y_a - Y_{\tilde{a}}|\boldsymbol{C}] = E[Y_{aM_{\tilde{a}}} - Y_{\tilde{a}M_{\tilde{a}}}|\boldsymbol{C}] + E[Y_{aM_a} - Y_{aM_{\tilde{a}}}|\boldsymbol{C}] = NDE + NIE$; this holds even in models with interactions or non-linearities (Pearl, 2001)

Let A and $\boldsymbol{C}$ be either continuous or categorical. In the context of a parametric approach to mediation analysis, for the case of a binary mediator and continuous outcome, the following regression models can be defined:

$$logit\{P[M = 1|A = a, \boldsymbol{C} = \boldsymbol{c}]\} = \beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c} \tag{3.1}$$

$$E[Y|A = a, M = m, \boldsymbol{C} = \boldsymbol{c}] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4'} \boldsymbol{c}. \tag{3.2}$$

Let $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ denote the vector of outcome and mediator regression parameters. Under models (3.1) and (3.2) controlled direct effect ($CDE(\boldsymbol{\theta})$), natural direct effect ($NDE(\boldsymbol{\theta}, \boldsymbol{\beta})$) and natural indirect effect ($NIE(\boldsymbol{\theta}, \boldsymbol{\beta})$) for a change in exposure from level $\tilde{a}$ to level a can be estimated by (Valeri and VanderWeele, 2012):

$$\begin{aligned}
CDE(\boldsymbol{\theta}) &= E[Y_{am} - Y_{\tilde{a}m}|C = c] = \{\theta_1 + \theta_3 m\}(a - \tilde{a}) \\
NDE(\boldsymbol{\theta}, \boldsymbol{\beta}) &= E[Y_{aM_{\tilde{a}}} - Y_{\tilde{a}M_{\tilde{a}}}|C = c] = \{\theta_1(a - \tilde{a})\} + \{\theta_3(a - \tilde{a})\} \frac{exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]}{1 + exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]} \\
NIE(\boldsymbol{\theta}, \boldsymbol{\beta}) &= E[Y_{aM_a} - Y_{aM_{\tilde{a}}}|C = c] = (\theta_2 + \theta_3 a)\{\frac{exp[\beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c}]}{1 + exp[\beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c}]} - \\
&\quad \frac{exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]}{1 + exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]}\}.
\end{aligned}$$

When the exposure A and mediator M do not interact in their effect on the outcome $\theta_3 = 0$ and direct and indirect causal effects can be estimated as

$$\begin{aligned}
NDE(\boldsymbol{\theta}) &= CDE(\boldsymbol{\theta}) = \theta_1(a - \tilde{a}) \\
NIE(\boldsymbol{\theta}, \boldsymbol{\beta}) &= \theta_2\{\frac{exp[\beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c}]}{1 + exp[\beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c}]} - \frac{exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]}{1 + exp[\beta_0 + \beta_1 \tilde{a} + \boldsymbol{\beta_2'} \boldsymbol{c}]}\}.
\end{aligned}$$

When the outcome is binary modeled with a logit link, equation (2) can be replaced by

$$logit\{P(Y=1|A=a,M=m,\boldsymbol{C}=\boldsymbol{c})\}=\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\boldsymbol{\theta_4'c}. \qquad (3.3)$$

If the outcome is case and rare, then from models (3.1) and (3.3) natural direct effect and natural indirect effect for a change in exposure from level $\tilde{a}$ to level a are given in terms of odds ratios by (Valeri and VanderWeele, 2012):

$$\begin{aligned}
OR^{CDE}(\boldsymbol{\theta}) &= exp[(\theta_1+\theta_3 m)(a-\tilde{a})] \\
OR^{NDE}(\boldsymbol{\theta},\boldsymbol{\beta}) &= \{\frac{exp[\theta_1 a](1+exp[\theta_2+\theta_3 a+\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}])}{exp[\theta_1\tilde{a}](1+exp[\theta_2+\theta_3\tilde{a}+\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}])}\} \\
OR^{NIE}(\boldsymbol{\theta},\boldsymbol{\beta}) &= \frac{[1+exp(\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c})][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\boldsymbol{\beta_2'c})]}{[1+exp(\beta_0+\beta_1 a+\boldsymbol{\beta_2'c})][1+exp(\theta_2+\theta_3 a+\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c})]}.
\end{aligned}$$

The expressions above in terms of regression coefficients will be equal to the counterfactual direct and indirect effects, and therefore have a causal interpretation, provided that the parametric models are correctly specified and that conditional on covariates $\boldsymbol{C}$ there is no unmeasured confounding of (i) the exposure-outcome relationship, (ii) the mediator outcome relationship, (iii) the exposure-mediator relationship, and (iv) that there is no variable affected by the exposure that confounds the mediator outcome relationship. In the counterfactual notation this is: (i) Y_{am} II $A|\boldsymbol{C}$, (ii) Y_{am} II $M|\boldsymbol{C}$, (iii) M_a II $A|\boldsymbol{C}$, (iv) Y_{am} II $M_{\tilde{a}}|\boldsymbol{C}$ (See Pearl (2001) and Robins and Richardson (2010) for further discussion of these assumptions).

3.3 Results on direct and indirect effects naive estimators when the mediator is misclassified

3.3.1 Mediator and outcome regressions when mediator is misclassified

Using the notation in section 3.2, assume that both A and $\boldsymbol{C}$, as well as the outcome Y, are correctly measured. Let M be the binary mediator at its true level and M^* the misclassified version of M. In the following we assume that the misclassification mechanism is independent of the outcome, the exposure, and the covariates ($P(M^*|M, Y, A, \boldsymbol{C}) = P(M^*|M)$, i.e. non-differential). Under this assumption the misclassification mechanism can be completely characterized by sensitivity ($SN = P(M^* = 1|M = 1)$) and specificity ($SP = P(M^* = 0|M = 0)$) and the prevalence of the latent mediator, $p = P(M = 1)$. When the true intermediate M is replaced by the observed intermediate M^* in models (3.1) and (3.2) an investigator operates with *observed* outcome and mediator regressions:

$$logit\{P[M^* = 1|A = a, \boldsymbol{C} = \boldsymbol{c}]\} = \beta_0^* + \beta_1^* a + \boldsymbol{c}, \tag{3.4}$$

$$E[Y|A = a, M^* = m^*, \boldsymbol{C} = \boldsymbol{c}] = \theta_0^* + \theta_1^* a + \theta_2^* m^* + \theta_3^* a m^* + \boldsymbol{\theta_4^{*\prime}} \boldsymbol{c}. \tag{3.5}$$

Misclassification typically causes parameter estimates of the mediator and outcome regression to be asymptotically biased (Carroll et al., 2006; Gustafson, 2004). We start by deriving the asymptotic limit for the coefficients' estimators of the mediator equation assuming a logistic model and outcome equation assuming a linear regression model allowing for mediator-exposure interaction. We then proceed giving the asymptotic bias of the naive direct and indirect causal effects estimators. All the derivations can be found in the online appendix.

3.3.2 Asymptotic limit of parameters of the mediator regression

Let $\hat{\boldsymbol{\beta}}^*$ be the vector of maximum likelihood (MLE) estimators of the parameters from (3.4) and let $\boldsymbol{\beta}^*$ be the asymptotic limit of those MLE estimators. Let $\boldsymbol{\beta}$ be the asymptotic limit of MLE estimators of the parameters from (3.1). Let SN and SP denote sensitivity and specificity parameters. Misclassification of the mediator causes a modification in the link function (Neuhaus, 1999). Therefore, to study the limit of the observed mediator regression parameters estimators we can borrow results from previous studies on the effect of model mis-specification on the validity of parameters estimation (Huber, 1967; Akaike, 1973; White, 1982). Extending the reasoning of Neuhaus (1999) of simple logistic regression to allow for the presence of multiple covariates, we find that $\hat{\boldsymbol{\beta}}^*$ will approximately converge in probability to

$$
\begin{aligned}
\beta_0^* &\approx logit\{(SN+SP-1)\frac{exp(\beta_0)}{1+exp(\beta_0)}+(1-SP)\} \\
\beta_1^* &\approx \beta_1 \frac{(SN+SP-1)exp(\beta_0+\boldsymbol{\beta_2'c})}{\{SNexp(\beta_0+\boldsymbol{\beta_2'c})+(1-SP)\}\{(1-SN)exp(\beta_0+\boldsymbol{\beta_2'c})+SP\}} \\
\boldsymbol{\beta_{2i}^{*\prime}} &\approx \boldsymbol{\beta_{2}'}_i \frac{(SN+SP-1)exp(\beta_0+\beta_1 a+\boldsymbol{\beta_2'}^{(-i)}\boldsymbol{c})}{\{SNexp(\beta_0+\beta_1 a+\boldsymbol{\beta_2'}^{(-i)}\boldsymbol{c})+(1-SP)\}\{(1-SN)exp(\beta_0+\beta_1 a+\boldsymbol{\beta_2'}^{(-i)}\boldsymbol{c})+SP\}},
\end{aligned}
$$

where the terms $\boldsymbol{\beta_{2i}^{*\prime}}$ and $\boldsymbol{\beta_{2}'}_i$ denote the i-th component of and $\boldsymbol{\beta_2'}$ respectively; and $\boldsymbol{\beta_2'}^{(-i)}$ refers to the vector $\boldsymbol{\beta_2'}$ with the i-th component of $\boldsymbol{\beta_2'}$ for which we want to give the asymptotic limit set to zero.

The approximation of the limit is valid provided that the binary mediator is modeled using a generalized linear model (GLM) and $SN+SP>1$, indicating that the procedure producing the observed classification M^* performs better than chance. Moreover, note that the asymptotic limits defined above depend on the assumption of independence of the measurement error mechanism with the covariates in the mediator regression.

Neuhaus (1999) shows that for link functions g for which $1/g'$ is concave, errors in the response lead to attenuated estimates of the covariates effects. Such links include logistic, probit, complementary log-log and any link function based on an inverse cumulative distribution function. The amount of attenuation induced by misclassification is dependent

upon the magnitude of sensitivity and specificity parameters, the true prevalence of the mediator in the population and the magnitude of the true effects of the covariates on the mediator.

3.3.3 Asymptotic limit of parameters of the outcome regression

Suppose that M is subject to non-differential misclassification and measured as M^* and that we fit the observed continuous outcome regression model (3.5) where M is replaced by M^*. We study the asymptotic limit of the naive estimators of the exposure, mediator and the exposure-mediator interaction coefficients and we denote them by θ_1^*, θ_2^* and θ_3^* respectively.

Let ($\hat{\theta}_1^*$, $\hat{\theta}_2^*$, $\hat{\theta}_3^*$) be the naive maximum likelihood estimators of the outcome regressors if M is replaced by M^*. Let (θ_1, θ_2, θ_3) be the true parameters of the regressors. Let U denote the misclassification error taking values $(-1, 0, 1)$ and $AU = A \times U$; let $X = (1, A, M, AM, \boldsymbol{C})$ and $X^* = (1, A, M^*, AM^*, \boldsymbol{C})$ denote the matrix of the true and observed covariates respectively. For arbitrary variables W and Z we define $\delta_{W,Z}$ as the inverse of the covariance of W and Z. Let $\gamma_1 = Cov(M^*, U)$, $\gamma_2 = Cov(AM^*, U)$, and $\gamma_3 = Cov(AM^*, AU)$, which are functions of sensitivity and specificity parameters and of the covariance between exposure and the true mediator, as shown in the online appendix.

For a continuous outcome modeled using the linear regression the asymptotic limit of the outcome regression parameters in the presence of exposure-mediator interaction is given by,

$$
\begin{aligned}
\theta_1^* &= \theta_1 - \theta_2\delta_{A,M^*}\gamma_1 - \theta_3\delta_{A,AM^*}\gamma_3 - \gamma_2\{\theta_2\delta_{A,AM^*} + \theta_3\delta_{A,M^*}\} \\
\theta_2^* &= \theta_2\{1 - \delta_{M^*,M^*}\gamma_1\} - \theta_3\delta_{M^*,AM^*}\gamma_3 - \gamma_2\{\theta_2\delta_{M^*,AM^*} + \theta_3\delta_{M^*,M^*}\} \\
\theta_3^* &= \theta_2\delta_{AM^*,M^*}\gamma_1 - \theta_3\{1 - \delta_{AM^*,AM^*}\gamma_3\} - \gamma_2\{\theta_3\delta_{AM^*,M^*} + \theta_2\delta_{AM^*,AM^*}\}.
\end{aligned}
$$

Note that the asymptotic limits of the naive outcome regression parameters estimators are complex functions of the true outcome regression parameters, the covariance between

the observed covariates and the misclassification error, and the correlation between the covariates. In the presence of exposure-mediator interaction, it is not clear the direction that the asymptotic bias of the naive outcome regression parameters could take.

3.3.4 Asymptotic bias of the direct and indirect causal effects

Given the limit of the naive outcome and mediator regression parameters estimators the asymptotic bias of the direct and indirect effects naive estimators can be derived.
Let the vector $\boldsymbol{\beta^*}$ and $\boldsymbol{\theta^*}$ denote the limit of the vector of the naive mediator and outcome regression parameters estimators $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\theta}}^*$. Let $\widehat{CDE}^* = CDE(\hat{\boldsymbol{\theta}}^*)$, $\widehat{NDE}^* = NDE(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\beta}}^*)$ and $\widehat{NIE}^* = NIE(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\beta}}^*)$ denote the naive estimators for the controlled direct effect, natural direct effect, and the indirect effect, respectively. Recall $\delta_{W,Z}$ is the inverse of the covariance of W and Z. Let $\gamma_1 = Cov(M^*, U)$, $\gamma_2 = Cov(AM^*, U)$, and $\gamma_3 = Cov(AM^*, AU)$ We then have that:

$$\begin{aligned} ABIAS(\widehat{CDE}^*) &= -[\theta_2\gamma_1(\delta_{A,M^*} + \delta_{AM^*,M^*}m) + \theta_3\gamma_3(\delta_{A,AM^*} + \delta_{AM^*,AM^*}m) + \\ &\quad +\gamma_2\{\theta_2(\delta_{A,AM^*} + \delta_{AM^*,AM^*}m) + \theta_3(\delta_{A,M^*} + \delta_{AM^*,M^*}m)\}] \end{aligned}$$

$$\begin{aligned} ABIAS(\widehat{NDE}^*) &\approx [\theta_3(expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}] - expit[\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}]) - \gamma_1\theta_2(\delta_{AM^*,AM^*} \times \\ &\quad \times expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}] + \delta_{A,M^*}) - \gamma_3\theta_3(\delta_{AM^*,AM^*}expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}] + \\ &\quad \delta_{A,AM^*}) - \gamma_2\{\theta_3(\delta_{AM^*,M^*} \times expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}] + \delta_{AM^*}) \\ &\quad +\theta_2(\delta_{AM^*,AM^*}expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}] + \delta_{A,AM^*})\}](a - \tilde{a}) \end{aligned}$$

$$\begin{aligned} ABIAS(\widehat{NIE}^*) &\approx \{expit[\beta_0^* + \beta_1^*a + \boldsymbol{c}] - expit[\beta_0^* + \beta_1^*\tilde{a} + \boldsymbol{c}]\}[\theta_2 + \theta_3 - \gamma_1\theta_2(\delta_{M^*,M^*} + \\ &\quad +a\delta_{AM^*,M^*}) - \gamma_3\theta_3(\delta_{M^*,AM^*} + a\delta_{AM^*,AM^*}) - \gamma_2\{\theta_2(\delta_{M^*,AM^*} + \\ &\quad +a\delta_{AM^*,AM^*}) + \theta_3(\delta_{M^*,M^*} + a\delta_{AM^*,M^*})\}] - \{expit[\beta_0 + \beta_1a + \boldsymbol{\beta_2'c}] + \\ &\quad -expit[\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}]\}(\theta_2 + \tilde{a}\theta_3). \end{aligned}$$

In the absence of exposure-mediator interaction the asymptotic bias formulae simplify to:

$$
\begin{aligned}
ABIAS(\widehat{NDE}^{*}) &= ABIAS(\widehat{CDE}^{*}) = -\theta_2\delta_{A,M^*}\gamma_1(a-\tilde{a}) \\
ABIAS(\widehat{NIE}^{*}) &\approx \theta_2[\{1-\delta_{M^*,M^*}\gamma_1\}\{\frac{exp[\beta_0^*+\beta_1^*a+\boldsymbol{c}]}{1+exp[\beta_0^*+\beta_1^*a+\boldsymbol{c}]}-\frac{exp[\beta_0^*+\beta_1^*\tilde{a}+\boldsymbol{c}]}{1+exp[\beta_0^*+\beta_1^*\tilde{a}+\boldsymbol{c}]}\}+ \\
&\quad -\{\frac{exp[\beta_0+\beta_1a+\boldsymbol{\beta_2'c}]}{1+exp[\beta_0+\beta_1a+\boldsymbol{\beta_2'c}]}-\frac{exp[\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}]}{1+exp[\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}]}\}],
\end{aligned}
$$

Misclassification of a binary mediator exerts its impact on the estimation of direct and indirect effects by inducing bias in both the mediator and the outcome regression parameters estimation. This contrasts with the effect of measurement error on a continuous mediator which typically induces bias on the outcome regression parameters naive estimators while leaving unbiased the naive estimators of the mediator linear regression parameters (Valeri et al., 2012).

In the absence of exposure-mediator interaction the asymptotic bias of direct and indirect effects have a particular direction. The direct effect naive estimator will be biased away from the null under the assumption that the effect of the exposure on the mediator, β_1, and the effect of the mediator on the outcome, θ_2, have the same sign. The bias will be larger as sensitivity (SN) and specificity (SP) decrease, as the effect of the mediator on the outcome increases (θ_2) as well as the more the variables in the model are correlated among each other and with the error (δ_{A,M^*}). The indirect effect will be diluted (biased towards the null) under the same assumption. The attenuation factor again depends on the magnitude of sensitivity and specificity parameters, on the effect of the mediator on the outcome, θ_2, and the correlation among the variables. Proofs are given in the online appendix.

In a non-parametric setting, VanderWeele et al. (2012) show that although measurement error in the mediator induces biased direct and indirect effects, the combination of these biased effects is in fact unbiased for the total effect. However, this result does not necessarily hold in a regression-approach to mediation analysis. This is because, misclassification of a binary mediator induces mis-specification in the mediator and outcome regressions and this creates a source of bias beyond simply measurement error in a parametric setting.

In the presence of exposure-mediator interaction the formulas for the asymptotic bias are more complex and with no intuitive interpretation. Note that the direction of the bias remains the same at least non-parametrically (Ogburn et al. 2012). In section 3.3.6 we illustrate how departures from this result can occur using parametric models for continuous or binary outcome via a numerical study.

3.3.5 Additional results on the behavior of direct and indirect causal effects naive estimators in the absence of exposure-mediator interaction

To explore further the behavior of naive causal effect estimators in the presence of misclassification (assuming the absence of exposure-mediator interaction) we evaluated the asymptotic bias formulae as functions of sensitivity and specificity parameters. We find that the asymptotic biases of both direct and indirect effects are maximized when sensitivity and specificity parameters take value $1/2$. The maximization results hold for each level of the prevalence of the true mediator. Let $NDE^* = NDE(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$ and $NIE^* = NIE(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$. We can show that

$$\begin{aligned} \frac{\partial(NDE^* - NDE)}{\partial SP} &= 0 \\ \frac{\partial(NDE^* - NDE)}{\partial SN} &= 0, \end{aligned}$$

gives $SP = SN = \frac{1}{2}$.

Also $SP = SN = \frac{1}{2}$ gives

$$\begin{aligned} \frac{\partial(NIE^* - NIE)}{\partial SP} &= 0 \\ \frac{\partial(NIE^* - NIE)}{\partial SN} &= 0. \end{aligned}$$

Via a partial derivative analysis we study the behavior of the asymptotic bias as sensitivity or specificity for the mediator measurement depart from their optimal level. As sensitivity or specificity for the mediator measurement depart from 1, the magnitude of the bias for the direct effect depends on the magnitude the effect of the mediator on the outcome and the prevalence of the mediator for sensitivity and on the magnitude of the effect of the mediator on the outcome and 1 minus the prevalence of the mediator for specificity. This is because:

$$\begin{aligned}
\frac{\partial(NDE^* - NDE)}{\partial SP}|_{SP=SN=1} &= -\theta_2\delta_{A,M^*}(1-p)^2 \\
\frac{\partial(NDE^* - NDE)}{\partial SN}|_{SP=SN=1} &= -\theta_2\delta_{A,M^*}p^2.
\end{aligned}$$

Likewise for the indirect effect it can be shown that:

$$\begin{aligned}
\frac{\partial(NIE^* - NIE))}{\partial SP}|_{SP=SN=1} &= \theta_2\{expit\{\beta_0 + \beta_1 a + \boldsymbol{\beta_2'c}\}[\frac{1-expit(\beta_0)}{expit(\beta_0)} - \frac{a}{e^{\beta_0+\boldsymbol{\beta_2'c}}} - \frac{c}{e^{\beta_0+\beta_1 a+\boldsymbol{\beta_2'c}}}] + \\
&\quad -expit\{\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}\}[\frac{\{1-expit(\beta_0)\}^2}{expit(\beta_0)} - \frac{\tilde{a}}{e^{\beta_0+\boldsymbol{\beta_2'c}}} - \frac{c}{e^{\beta_0+\beta_1+\tilde{a}\boldsymbol{\beta_2'c}}}] + \\
&\quad -\delta_{A,M^*}[expit\{\beta_0 + \beta_1 a + \boldsymbol{\beta_2'c}\} - expit\{\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}\}](1-p)^2\} \\
\frac{\partial(NIE^* - NIE)}{\partial SN}|_{SP=SN=1} &= \theta_2\{expit\{\beta_0 + \beta_1 a + \boldsymbol{\beta_2'c}\}[expit(\beta_0) - 1 - ae^{\beta_0+\boldsymbol{\beta_2'c}} - ce^{\beta_0+\beta_1 a+\boldsymbol{\beta_2'c}}] + \\
&\quad -expit\{\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}\}[expit(\beta_0) - 1 - \tilde{a}e^{\beta_0+\boldsymbol{\beta_2'c}} - ce^{\beta_0+\beta_1\tilde{a}+\boldsymbol{\beta_2'c}}] + \\
&\quad -\delta_{A,M^*}[expit\{\beta_0 + \beta_1 a + \boldsymbol{\beta_2'c}\} - expit\{\beta_0 + \beta_1\tilde{a} + \boldsymbol{\beta_2'c}\}]p^2\}.
\end{aligned}$$

The partial derivatives are not easily interpretable for indirect effect, however we can observe a similar finding. As sensitivity or specificity for the mediator measurement depart from 1, the magnitude of the bias for the indirect effect depends on the magnitude of the effect of the mediator on the outcome and the prevalence of the mediator for sensitivity and on the magnitude of the effect of the mediator on the outcome and 1 minus the prevalence of the mediator for specificity. Additionally, the change in the asymptotic bias for a departure of either sensitivity of specificity from optimality depends on the covariance between the exposure and the mediator.

Note also that if A and $\boldsymbol{C}$ were binary and the mediator regression was saturated, using

the property of effect decomposition that the direct and indirect effects sum to the total effect (and knowing that the total effect will be unbiased by measurement error), the partial derivatives for the NIE would simplify into

$$\begin{aligned}\frac{\partial(NIE^* - NIE)}{\partial SP}|_{SP=SN=1} &= \theta_2\delta_{A,M^*}(1-p)^2 \\ \frac{\partial(NIE^* - NIE)}{\partial SN}|_{SP=SN=1} &= \theta_2\delta_{A,M^*}p^2.\end{aligned}$$

Importantly, from the formulae above we can evince that the bias of direct and indirect effects will depend more on specificity for a low prevalence mediator and more on sensitivity for a high prevalence mediator.

3.3.6 Numerical bias analysis

Misclassification of a binary mediator induces bias in the estimators of direct and indirect effects if misclassification is ignored. In the previous sections we gave the asymptotic bias in closed form when the outcome is continuous allowing for the presence of exposure-mediator interaction. The theoretical results are rather complex and hard to interpret. Therefore, we carry out a numerical bias analysis to confirm our theoretical findings and to have clearer picture of how the magnitude and direction of the asymptotic bias of the causal effects can be influenced by the presence of non-linearities, such as interactions.
We consider $r = 100$ samples of size $n = 10,000$ and generate a binary exposure $A_i \sim Be(p_a)$ with $p_a = 0.4$ and a continuous covariate $C \sim N(1,1)$. The true binary mediator conditional on A and C is defined as $M|A,C \sim Be(p_M)$ with $p_M = exp(\beta_0 + \beta_1 A + \beta_2 C)/\{1 + exp(\beta_0 + \beta_1 A + \beta_2 C)\}$ and $\beta_0 = -0.25$, $\beta_1 = 0.2$, $\beta_2 = 0.2$. Under this scenario the mediator is common with a marginal probability of 50%. The observed mediator is defined so that $P(M^* = 1|M = 0)$ and $P(M^* = 0|M = 1)$ take values in the rectangular space $(0,1) \times (0,1)$. For continuous outcome we generate $Y|A,M,C \sim N(\mu_Y, \sigma^2_Y)$, where $\mu_Y = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C$ and $\sigma^2_Y = 1$,

with $\theta_0 = 0$, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = (0, 1)$, $\theta_4 = 1$. We carry out a numerical bias analysis assuming a binary outcome as well and we generate $Y|A, M, C \sim Ber(p_Y)$, where $p_Y = exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C)/(1 + exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C))$, with $\theta_0 = -0.8$, $\theta_1 = 1$, $\theta_2 = 1.7$, $\theta_3 = (0, 0.1)$, $\theta_4 = 1$

The observed outcome regression models are run assuming that the true model is known by the investigator, who just replaces the true M with the observed mediator M^*. In particular, if in the true model the exposure-mediator interaction term is present, then the interaction term is present in the naive model too; likewise, if in the true model the exposure-mediator interaction term is absent, the naive analysis is run without including the exposure-mediator interaction term.

Figures 3.2 and 3.3 give a three dimensional view of the relative bias of direct and indirect effects naive estimators when sensitivity and specificity take values in the interval $(0, 1)$. We define relative bias as the ratio of the asymptotic bias of the naive estimator of the causal effects over their true value.

Figure 3.2 displays the relative bias of naive estimators of direct and indirect causal effects for continuous outcome with exposure-mediator interaction either present (pink graphs) or absent (blue graphs) in the true model. In the absence of exposure-mediator interaction (blue graphs), the numerical results confirm the theoretical findings. We can observe first that with the true NDE equal to 1, the asymptotic relative bias for the naive estimator of NDE is positive and thus away from the null. We also note that with the true NIE equal to 0.05, the asymptotic relative bias of the naive indirect effect estimator is negative and thus towards the null. Therefore, naive estimators when the outcome is continuous estimate the direct effect away from the null and estimate the indirect effect towards the null. The asymptotic relative bias is maximized at $SN = SP = 1/2$. If both sensitivity and specificity were equal to zero we note that no bias would be found. This would coincide with the case in which the investigator recoded the mediator. We note that in the presence of exposure-mediator interaction (pink graphs), direct and indirect effects are biased in the same direction, but the magnitude of the bias for the direct effect is larger.

Figure 3.3 displays the relative bias of the naive estimators of direct and indirect causal

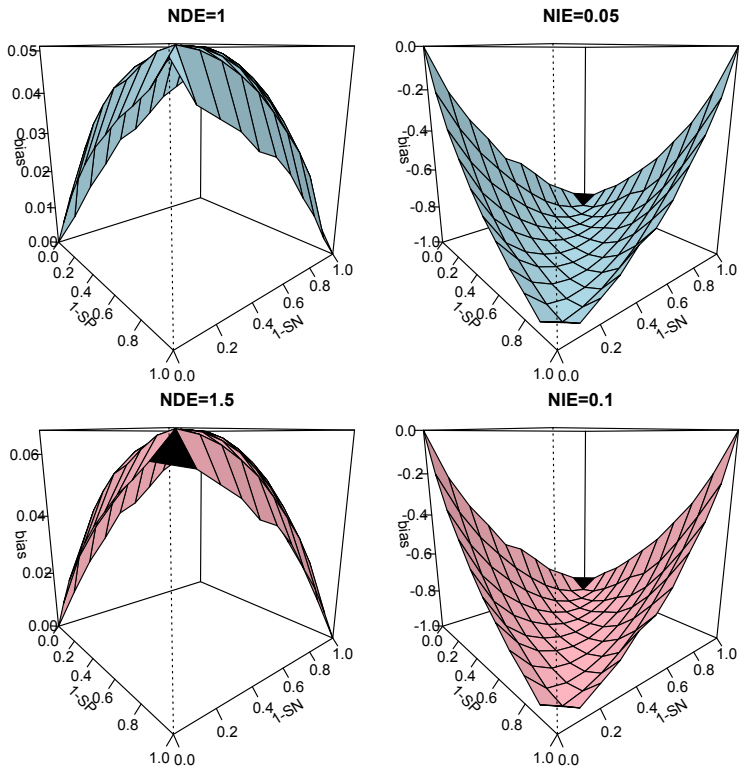

Figure 3.2: Relative bias of direct (NDE) and indirect (NIE) naive causal effects for the grid (SN,SP)=$(0.1, 0.9) \times (0.1, 0.9)$ for continuous outcome modeled using linear regression (BLUE-No exposure-mediator interaction/PINK-Exposure-mediator interaction)

effects for binary outcome when the interaction is present (yellow graph) or absent (green graph) in the true model. We observe that the naive direct effect is biased towards the null as well as the indirect effect. The result is found both in the presence and in the absence of exposure-mediator interaction in the true model. The magnitude of the bias of direct effect estimator increases in the presence of interaction (yellow graph).

We observe that in the presence of exposure-mediator interaction and/or binary outcome the asymptotic relative bias can take unintuitive directions. The magnitude and direction of the bias is now influenced even more by model mis-specification and by the magnitude and sign of the interaction term.

Moreover, the reader should be aware that when the direct and indirect effect naive estimators are biased in the same direction, the total effect will be asymptotically biased as well. This contrasts with the non-parametric result of VanderWeele et al. (2012). The

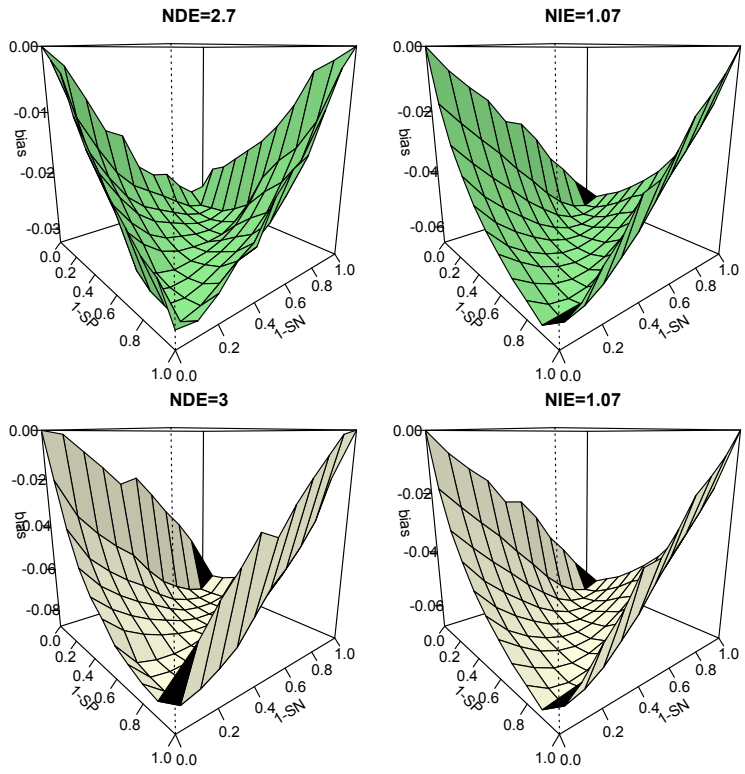

Figure 3.3: Relative bias of direct (NDE) and indirect (NIE) naive causal effects for the grid (SN,SP)=$(0.1, 0.9) \times (0.1, 0.9)$ for binary outcome modeled using logistic regression (GREEN-No exposure-mediator interaction/YELLOW-Exposure-mediator interaction)

reason for this result is that misclassification induces mis-specification of both the outcome and mediator regression and this mis-specification induces an additional source of bias. In the presence of interaction and when the outcome is binary the issue of mis-specification might be accentuated in certain settings, as the numerical analysis shows. We also observe that the numerical bias results for binary outcome are dramatically different from the ones that the asymptotic bias formulae for continuous outcome would have predicted both in the presence and in the absence of exposure-mediator interaction. Therefore, the asymptotic bias formulae that we give in the previous section cannot be used as approximations of the asymptotic bias of naive direct and indirect effects estimators when the outcome is binary.

3.4 Correction strategy for direct and indirect effects estimators

3.4.1 Iteratively Re-weighted Least Squares estimators for mediator regression

Asymptotically unbiased estimators for the vector of true mediator regression parameters $\boldsymbol{\beta}$ can be found by adjusting the link function and programming Iteratively Re-weighted Least Squares (IRLS) as proposed by Neuhaus (1999). This is a widely popular approach for misclassified outcome regression correction and is known to perform well in most settings (Lyles and Lin, 2010; Carroll et al., 2006).

Provided that the misclassification probabilities are known or set as sensitivity analysis parameters, the IRLS method entails constructing the likelihood for the probability that the true mediator is equal to one in terms of misclassification probabilities and the conditional probability of the mediator given the observed covariates using a logistic model and maximizing it with respect to the covariates' parameters.

We can define the misclassification probabilities as:

$$\begin{aligned} \gamma_0 &= P[M^* = 1|M = 0] \\ \gamma_1 &= P[M^* = 0|M = 1] \end{aligned}$$

It can be shown that the likelihood for the true mediator parameters given the observed misclassified mediator, misclassification probabilities, and the observed covariates A and $\boldsymbol{C}$ is given by:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \prod_{i=1}^{n}[\{\gamma_0 \frac{1}{1+exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)} + (1-\gamma_1)\frac{exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)}{1+exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)}\}^{M_i^*} \\ &\quad \times\{(1-\gamma_0)\frac{1}{1+exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)} + \gamma_1\frac{exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)}{1+exp(\beta_0+\beta_1 A_i+\boldsymbol{\beta_2'}\boldsymbol{C}_i)}\}^{(1-M_i^*)}] \end{aligned}$$

By setting $\gamma_0 = 1-SP$ and $\gamma_1 = 1-SN$ as sensitivity analysis parameters and maximizing this likelihood consistent estimators for the true parameters $\boldsymbol{\beta}$ can be easily recovered.

Standard errors for all parameters in the model can also be readily obtained via close

numerical approximations to the observed information matrix or using bootstrap method (Carroll et al., 2006).

3.4.2 Predictive Value Weighting estimators for outcome regression

Predictive value weighting (Lyles and Lin, 2010) is a method of correction for misclassification that shares similarities with imputation methods for missing data. We consider this methodology because it is intuitive, easy to program, and not too computationally intensive.

The approach consists of reconstructing data that might have been observed under no misclassification by using the observed data and assumptions about sensitivity and specificity parameters. Sensitivity analysis for different sensitivity and specificity values can be easily carried out. For each observed value of the binary mediator, the true mediator can take value zero or one with a certain probability. The probability depends on the estimated positive and negative predicted values.

The approach is implemented in two steps. The first step entails estimating positive and negative predicted values from the observed data and setting assumptions about sensitivity (SN) and specificity (SP).

$$\begin{aligned} PPV &= P(M=1|M^*=1,Y=y,A=a,\boldsymbol{C}=\boldsymbol{c}) \\ &= \frac{SN*P(M=1|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})}{SN*P(M=1|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})+(1-SP)*P(M=0|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})} \\ NPV &= P(M=0|M^*=0,Y=y,A=a,\boldsymbol{C}=\boldsymbol{c}) \\ &= \frac{SP*P(M=0|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})}{(1-SN)*P(M=1|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})+SP*P(M=0|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})} \end{aligned}$$

$P(M=m|Y,A,\boldsymbol{C})$ can be easily estimated conditioning on M^*, SN, and SP (Lyles and Lin, 2010). Therefore, estimators $\widehat{PPV}$ and $\widehat{NPV}$ can be obtained by setting SN and SP as sensitivity analysis parameters and estimating $\widehat{P}(M^*=1|Y=y,A=a,\boldsymbol{C}=\boldsymbol{c})$ from the data e.g. by running a logistic regression of the observed mediator on the outcome, the exposure and the additional covariates. The model associating the outcome Y with the true mediator M, the exposure A and the covariates $\boldsymbol{C}$ can be fitted to an expanded

dataset, where the observed mediator is replaced by the latent mediator, allowed to take two values with a certain probability dependent on the positive and negative predictive values estimated. The second step consists in running a weighted outcome regression, available in most software packages, from which the corrected estimators $\widehat{\boldsymbol{\theta}}^{PVW}$ are finally obtained.

Formally, a weighted maximum likelihood is maximized

$$l(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n}\sum_{m=0}^{1} w_{im} l_{im}(\boldsymbol{\theta}).$$

where w_{im} is a weight computed for each individual at the 2 possible levels of the true binary mediator.

Standard errors for all parameters in the model can be recovered using jackknife or bootstrap procedures (Lyles and Lin, 2010) or can be analytically derived as shown in the online appendix.

The approach can be extended to allow for the presence of exposure-mediator interaction and can be applied both to normal and binary outcomes.

Note that PVW will yield consistent estimators provided the SN and SP are correctly specified and provided that the predicted probability of the observed mediator given the outcome Y and the covariates, A and $\boldsymbol{C}$ ($P(M^* = 1|Y = y, A = a, \boldsymbol{C} = \boldsymbol{c})$) is consistently estimated. These are very crucial assumptions to which will return.

3.4.3 Likelihood-based approach for outcome and covariate misclassification

As an alternative procedure, we consider a direct maximum likelihood approach (Lyles and Lin, 2010; Carroll et al., 2006).

When fitting the outcome regression we wish to recover the vector of parameters $\boldsymbol{\theta}$ which characterizes the distribution of $Y|A, M, \boldsymbol{C}$. If the outcome is continuous, the conditional distribution of the outcome given the exposure, the mediator and the covariates can be defined as

$$f_{Y|A,M,\boldsymbol{C};\boldsymbol{\theta}} \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y-\mu_{Y|A,M,\boldsymbol{C}})},$$

where σ^2 is the conditional variance of Y given $A, M, \boldsymbol{C}$ and $\mu_{Y|A,M,\boldsymbol{C}} = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \boldsymbol{\theta_4'} \boldsymbol{C}$. When the outcome is binary we can define

$$f_{Y|A,M,\boldsymbol{C};\boldsymbol{\theta}} \;\; = \;\; p_y^Y (1-p_y)^{(1-Y)},$$

where $p_y = p_{Y|A,M,\boldsymbol{C}} = exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \boldsymbol{\theta_4'} \boldsymbol{C})/(1 + exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \boldsymbol{\theta_4'} \boldsymbol{C}))$. When the observed mediator M^* is a misclassified version of M the likelihood arising from the parametric models just given cannot be fit.

The maximum likelihood approach consists of specifying the likelihood for a measurement error model in which not only the outcome, but also the observed mediator is considered random rather than fixed. Assuming non-differential misclassification and sensitivity and specificity parameters known or set as sensitivity analysis parameters, the observed-data likelihood contribution can be defined in terms of the measurement error model (Carroll et al., 2006):

$$\begin{aligned}
f_{Y,M^*|A,\boldsymbol{C}} &= \sum_{m=0}^{1} f_{Y,M^*,M|A,\boldsymbol{C}}(y, m^*, m|a, \boldsymbol{c}) \\
&= \sum_{m=0}^{1} f_{Y|M^*,M,A,\boldsymbol{C}}(y|m^*, m, a, \boldsymbol{c}) P(M^* = m^*|M = m, A = a, \boldsymbol{C} = \boldsymbol{c}) \times \\
&\quad \times P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) \\
&= \sum_{m=0}^{1} f_{Y|M,A,\boldsymbol{C}}(y|m, a, \boldsymbol{c}) P(M^* = m^*|M = m) P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}).
\end{aligned}$$

Note that $P(M^* = m^*|M = m)$ for each combination of M and M^* is assumed to be known or specified in a sensitivity analysis, and in the mediation setting we are considering, the model for $P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c})$ is postulated as $P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}; \boldsymbol{\beta}) = expit(\beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c})$.

Numerically maximizing the log-likelihood determined by the contribution given above with respect to the true vector of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ provides the outcome and mediator regression parameter estimators and their standard errors are estimable by approximating the observed information matrix.

3.4.4 Corrected estimators for direct and indirect effects

Corrected estimators for direct and indirect causal effects ($\widehat{NDE} = NDE(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$, $\widehat{NIE} = NIE(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$) can be recovered by plugging in the formulas for direct and indirect effects the corrected mediator and outcome regression parameter estimators. If IRLS method is used to estimate the mediator regression parameters and the PVW approach is used to estimate the outcome regression parameters in the presence of misclassification, corrected direct and indirect causal effects estimators are given by:

$$
\begin{aligned}
\widehat{NDE} &= \{\hat{\theta}_1^{PVW}(a-\tilde{a})\} + \{\hat{\theta}_3^{PVW}(a-\tilde{a})\}\frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}\tilde{a} + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]}{1 + exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}\tilde{a} + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]} \\
\widehat{NIE} &= (\hat{\theta}_2^{PVW} + \hat{\theta}_3^{PVW}a)\{\frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]}{1 + exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]} + \\
&\quad - \frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}\tilde{a} + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]}{1 + exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}\tilde{a} + \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{'\,IRLS}\boldsymbol{c}]}\}.
\end{aligned}
$$

Standard errors for the corrected estimators can be obtained using jackknife or bootstrap procedure. Analytical standard errors for the corrected causal effects estimators can be obtained employing the multivariate delta method.

3.4.5 Simulations

We conducted simulation studies to evaluate and compare the estimates produced by the iteratively re-weighted least squares/predictive value weighting method (IRLS/PVW) and the maximum likelihood method (ML). Here we report the results of one such study for which the simulation setting is the same as the one used for the numerical bias analysis in section 3.3.6. The simulation implies a scenario under which the indirect effect of A on Y through M is small relative to the direct effect of A on Y.

The main concern about the applicability and performance of IRLS/PVW estimators is the validity of the estimation of the predictive probability of the observed mediator given

the exposure, the outcome, and the covariates $(M^*|Y, A, \boldsymbol{C})$ via a logistic regression. Lyles and Lin (2010) propose the following model

$$logit\{P[M^* = 1|Y = y, A = a, \boldsymbol{C} = \boldsymbol{c}\} = \xi_0 + \xi_1 y + \xi_2 a + \boldsymbol{c}. \tag{3.6}$$

Given the data generating process described in section 3.3.6, it can be shown that equation (3.6) is misspecified for the probability of $M^*|Y, A, \boldsymbol{C}$. We refer to this approach as IRLS/PVW. In order to overcome the issue of model mis-specification we considered a modified predictive value weighting estimator in which we used splines to model more flexibly the relationship between M^* and $(Y, A, \boldsymbol{C})$. We implemented this alternative approach by modeling each term in equation (3.6) and all possible interactions with splines as

$$\begin{aligned} logit\{P[M^* = 1|Y = y, A = a, \boldsymbol{C} = \boldsymbol{c}\} \quad &= \quad s(y, k) + s(a, k) + s(\boldsymbol{c}, k) + s(y * a, k) + s(y * \boldsymbol{c}, k) + \\ &\quad s(a * \boldsymbol{c}, k) \end{aligned}$$

where $s(\cdot)$ denotes a smoothing function and k indicates the number of knots (the more knots, the more flexible is the smoothing function). We chose $k = 4$. The analyses were performed using the package *gam* built in the R software. We refer to this approach as IRLS/**s**PVW.

Table 3.1 displays the simulation results. When the outcome is continuous, in the absence of exposure-mediator interaction ($\theta_3 = 0$), we observe that the proposed methods (IRLS/PVW, IRLS/**s**PVW and ML) improve over the naive estimators in terms of asymptotic bias, which reduces to be close to zero.

When the outcome is continuous in the presence of exposure-mediator interaction ML is found to outperform IRLS/PVW and IRLS/**s**PVW in terms of asymptotic bias. The persistence of a bias is due to the fact that, although the **s**PVW approach improves the estimation of mediator and exposure parameters in the outcome regression, the interaction parameter remains biased. Consequently, the corrected estimators of direct and indirect effects in the presence of exposure-mediator interaction do not appear consistent.

		naive	IRLS/PVW	IRLS/ **s**PVW	IRLS/**t**PVW	ML
No Int & $Y \sim N$						
SP=SN=0.9	NDE	0.02	0.01	0.01	0.00	0.00
	NIE	-0.02	-0.01	-0.01	-0.00	-0.00
	TE	0.00	0.01	0.01	0.00	0.00
SP=SN=0.85	NDE	0.03	0.02	0.02	0.00	0.00
	NIE	-0.03	-0.01	-0.01	-0.00	-0.00
	TE	0.00	0.01	0.01	0.00	0.00
SP=SN=0.8	NDE	0.03	0.03	0.03	0.00	0.00
	NIE	-0.03	-0.02	-0.02	-0.00	-0.00
	TE	0.00	0.01	0.01	0.00	0.00
No Int & $Y \sim Ber$						
SP=SN=0.9	NDE	-0.03	-0.02	-0.01	0.04	0.01
	NIE	-0.03	-0.01	-0.01	-0.00	-0.00
	TE	-0.10	-0.04	-0.03	0.04	0.01
SP=SN=0.85	NDE	-0.04	-0.04	-0.02	0.07	0.01
	NIE	-0.04	-0.01	-0.01	-0.00	-0.00
	TE	-0.14	-0.07	-0.06	0.06	0.01
SP=SN=0.8	NDE	-0.06	-0.05	-0.04	0.07	0.01
	NIE	-0.04	-0.02	-0.02	-0.00	-0.00
	TE	-0.18	-0.10	-0.09	0.07	0.01
Int & $Y \sim N$						
SP=SN=0.9	NDE	0.04	0.01	0.00	0.00	0.00
	NIE	-0.04	-0.01	-0.00	-0.00	-0.00
	TE	0.00	0.01	0.00	0.00	0.00
SP=SN=0.85	NDE	0.05	0.02	0.01	0.00	0.00
	NIE	-0.05	-0.01	-0.00	-0.00	-0.00
	TE	0.00	0.01	0.00	0.00	0.00
SP=SN=0.8	NDE	0.07	0.02	0.01	0.00	0.00
	NIE	-0.07	-0.01	-0.01	-0.00	-0.00
	TE	-0.00	0.01	0.00	0.00	0.00
Int & $Y \sim Ber$						
SP=SN=0.9	NDE	-0.11	-0.08	0.04	0.08	0.04
	NIE	-0.03	-0.00	-0.00	-0.00	-0.00
	TE	-0.19	-0.09	0.04	0.08	0.04
SP=SN=0.8	NDE	-0.14	-0.10	0.07	0.14	0.07
	NIE	-0.04	-0.00	-0.00	-0.00	-0.00
	TE	-0.26	-0.12	0.07	0.14	0.07
SP=SN=0.8	NDE	-0.18	-0.13	0.09	0.16	0.09
	NIE	-0.05	0.00	-0.00	-0.00	-0.00
	TE	-0.32	-0.15	0.08	0.16	0.08

Table 3.1: Asymptotic bias of naive, predictive value weighting (IRLS/PVW, IRLS/**s**PVW, IRLS/**t**PVW) and direct maximum likelihood (ML) of controlled direct effect (CDE) natural direct (NDE), natural indirect effect (NIE) and total effect (TE) when sample size is $n = 10,000$, marginal probability of the true mediator is 50%, and the outcome is continuous or binary.

We compared the correction strategy that employes PVW and **s**PVW for the outcome regression parameters estimation with one in which predictive value weighting approach is implemented by recovering and estimating the true $P(M^*|Y, A, \boldsymbol{C})$. For a data generating process typical in mediation analysis context this conditional probability is given by:

$$
\begin{aligned}
P(M^* = 1|Y, A, C) \quad = \quad & [SN \times P(Y|A, M = 1, C; \boldsymbol{\theta})P(M = 1|A, C; \boldsymbol{\beta}) + (1 - SP) \times \\
& \times P(Y|A, M = 0, C; \boldsymbol{\theta})P(M = 0|A, C; \boldsymbol{\beta})]/[\sum_{m=0}^{1} P(Y|A, M = m, C; \boldsymbol{\theta}) \times \\
& \times P(M = m|A, C; \boldsymbol{\beta})
\end{aligned}
$$

Given this probability, the likelihood function for the binary observed mediator conditional on the outcome, the exposure and the covariates, $M^*|Y, A, \boldsymbol{C}$, can be constructed and upon numerical maximization, the predicted probabilities are easily obtained. The PVW estimator in which we model the true distribution of $M^*|Y, A, \boldsymbol{C}$ (IRLS/**t**PVW) performs very similarly to the ML and they both improve over IRLS/PVW in the estimation of the indirect causal effect .

When the outcome is binary ML is found again to outperform IRLS/PVW, IRLS/**s**PVW, and IRLS/**t**PVW in the absence of exposure-mediator interaction. However, in the presence of exposure-mediator interaction the ML estimator does not completely eliminate the bias. This is because ML improves in the estimation of the exposure regression parameter (θ_1) but the exposure-mediator interaction term estimation remains problematic.

3.5 Example

3.5.1 Mother's age above 35, pre-eclampsia and preterm birth: background and data description

We apply the proposed correction methodologies for misclassification of a binary mediator to a perinatal epidemiological study on the causal mechanisms leading to preterm birth using NCHS birth certificate data.

Preterm birth is strongly associated with perinatal mortality and long-term morbidity in developed countries (McCormick, 1985). Preterm birth is clinically defined as birth at less than 37 week's gestational age after either spontaneous labour with intact membranes, preterm premature rupture of the membranes, or labour induction or cesarean delivery for maternal or fetal indications (Goldenberg et al. 2008). Pregnancy outcomes among

women in the age group of 35 years and more are considered to be less favorable than those of younger women and risk of preterm delivery has been found higher for these mothers (Jacobsson et al., 2004). Over time, more and more women decide to post-pone conception and this phenomenon might contribute to a cycle of reproductive disadvantage with far-reaching social and medical consequences. Unveiling the causal mechanisms that explain the effect of maternal age on preterm birth is critical for initiation of risk-specific treatments and to study more targeted interventions that may decrease the burden of the disease in women of this age group.

The mechanisms by which maternal age is related to preterm birth are still largely unknown. A potential intermediate of the age-preterm birth causal relationship is pre-eclampsia. Several studies confirmed pre-eclampsia as a risk factor for medically induced preterm birth (Goldenberg et al., 2008; Hnat et al., 2002). Maternal age, in turn, has been found to be a risk factor for pre-eclampsia (Lamminpaa et al., 2012). Pre-eclampsia is a multi system hypertensive disorder of pregnancy that affects approximately 3% to 5% of all pregnancies worldwide (World Health Organization Report, 2005). Due to the non-specificity of signs and symptoms, diagnosis of pre-eclampsia is typically subject to misclassification (Meads et al., 2008; Turner, 2010).

We therefore carry out mediation analysis to quantify the indirect causal effect of mother's age on preterm birth mediated by pre-eclampsia status as well as the direct effect of ethnicity on preterm birth through other pathways, independent of pre-eclampsia. Files from the National Center for Health Statistics (NCHS) for 2003 ($N = 3,918,542$) are employed to investigate this hypothesis. These publicly available de-identified files are derived from all birth certificates in the 50 states and District of Columbia in the US.

Preterm birth was categorized according to the gestational age variable in the NCHS data derived from the last menstrual period. The NCHS data have gestational age estimates both derived from last menstrual period and from clinical/obstetric information. Pre-eclampsia is diagnosed according to blood pressure and protein in the urine and is potentially subject to misclassification. We combined mother's age to form two categories: age above 35 and age below or equal to 35.

3.5.2 Mother's age above 35, pre-eclampsia and preterm birth: naive analyses

Let $Y = preterm$ be the binary outcome, and $A = ageabove35$ be the binary exposure variable. Let $M = preeclampsia$, be the latent mediator, and $M^* = preeclampsia^*$ be the observed pre-eclampsia status, potentially misclassified.

Before conducting mediation analysis as described in section 3.2, we need to verify that the findings reported from the literature are confirmed in our sample. In particular, we investigate whether (i) pre-eclampsia status is associated with preterm birth, and (ii) maternal age is associated with pre-eclampsia. It is also of interest to investigate whether age and pre-eclampsia interact in their effect on pre-term birth. In order to investigate these hypotheses in our sample, we adjust for factors that may confound age-preterm birth relationship, age and pre-eclampsia relationship, and pre-eclampsia and preterm birth relationship. We also assume the absence of pre-eclampsia and preterm birth relationship confounders that are affected by age. As potential confounders we consider mother's ethnicity (categorized as White Caucasian, Black non-Hispanic, Hispanic, Asian and Native American), marital status, as well as smoking status, drinking status, and whether the mother went to college.

In section 3.2 we illustrated how direct and indirect causal effects can be estimated when both the outcome and the mediator are binary, modeled using logistic regression, and exposure-mediator interaction may be present. The naive analyses consist of running models (3.1) and (3.3) replacing M with M^* (i.e. replacing the correct, unobserved pre-eclampsia disorder status, with the *observed*, possibly misclassified, pre-eclampsia status). We fit the following logistic models:

$$logit\{P[M^* = 1|A = a, \boldsymbol{C} = \boldsymbol{c}]\} = \beta_0 + \beta_1 a + \boldsymbol{\beta_2'}\boldsymbol{c} \tag{3.7}$$

$$logit\{P(Y = 1|A = a, M^* = m^*, \boldsymbol{C} = \boldsymbol{c})\} = \theta_0 + \theta_1 a + \theta_2 m^* + \theta_3 a m^* + \boldsymbol{\theta_4'}\boldsymbol{c}. \tag{3.8}$$

The naive mediator regression analysis (3.7) reveals a positive, but rather weak, effect of age on pre-eclampsia ($\beta_1 = 0.083$, $p - val = 0.01$). The naive outcome regression analysis

(3.8) confirms a strong, positive effect of both the exposure, age status, and the mediator, pre-eclampsia, on the risk of preterm birth ($\theta_1 = 0.257$, $p - value < 0.0001$; $\theta_2 = 0.978$, $p - value < 0.0001$) and indicates the presence of a positive exposure-mediator interaction ($\theta_3 = 0.170$, $p - val = 0.01$). We also run the naive analysis dropping the exposure-mediator interaction term.

$$logit\{P(Y = 1|A = a, M^* = m^*, \boldsymbol{C} = \boldsymbol{c})\} = \theta_0 + \theta_1 a + \theta_2 m^* + \boldsymbol{\theta_4'} \boldsymbol{c}. \tag{3.9}$$

The naive outcome regression analysis (3.9) when we ignore the presence of exposure-mediator interaction yields similar results on the effect of age and preeclampsia on the risk of preterm birth ($\theta_1 = 0.26$, $p - value < 0.0001$; $\theta_2 = 1.007$, $p - value < 0.0001$).
Computing naive direct and indirect causal effects allowing for exposure-mediator interaction we find that maternal age above 35 exerts a positive, significant direct effect on preterm birth through pathways independent of pre-eclampsia had the individual been a white, married woman, who smokes but does not drink, and has not attended college ($\widehat{OR}^{NDE} = 1.314$). We find that the indirect causal effect of age on preterm birth through pre-eclampsia is close to null ($\widehat{OR}^{NIE} = 1.005$). A useful measure to quantify the proportion of the total causal effect of an exposure on an outcome that is explained by the hypothesized mechanism is the proportion mediated (PM). When direct and indirect effects are expressed in terms of odds ratios $PM = OR^{NDE} \times (OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1)$. The naive analysis yields $\widehat{PM} = 2.3\%$, indicating that the proportion of the effect of maternal age on preterm birth explained by the pathway through preeclampsia is close to null. The naive analyses without including the exposure-mediator interaction term lead to a similar finding (Table 3.3).

3.5.3 Mother's age above 35, pre-eclampsia and preterm birth: mediation analysis corrected for misclassification

Aware that the naive analyses might be biased due to misclassification of the binary mediator, we carry out sensitivity analyses for misclassification employing the iteratively

re-weighted least squares/predictive value weighting (IRLS/PVW) method. We implement this approach modeling the probability $P(M^*|Y, A, \boldsymbol{C})$ using a saturated logistic model (i.e. including all the possible interaction terms between Y, A, and $\boldsymbol{C}$ in the logistic model (3.6) described in section 3.4.5).

Throughout the paper we have assumed non-differential misclassification. We make this assumption in this analysis as well; however, this assumption could be easily relaxed if predictive value weighting approach is adopted, as illustrated in Lyles and Lin (2010).

Values of sensitivity and specificity parameters are determined considering the constraints for which $max(P(M^* = 1|Y, A, \boldsymbol{C})) < SN$ and $min(P(M^* = 1|Y, A, \boldsymbol{C}) > 1 - SP$. Since $max(P(M^* = 1|Y, A, \boldsymbol{C})) = 0.058$ and $min(P(M^* = 1|Y, A, \boldsymbol{C}) = 0.0103$ in our sample, the range of plausible sensitivity and specificity values is $SN \in (0.05, 1)$ and $SP \in (0.99, 1)$. Therefore sensitivity analyses are run assuming $SP = 0.99$ and $SN = (0.8, 0.9, 0.95, 0.99)$. We obtain the misclassification-corrected direct and indirect effects, and proportion mediated presented in Table 3.2 and Table 3.3.

Estimates ($95\%CI$)	$\widehat{OR}^{NDE}$	$\widehat{OR}^{NIE}$	PM	$\widehat{\beta}_1$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$
Naive	1.31	1.005	2.3%	0.083	0.25	0.97	0.170
	(1.26,1.36)	(1.00,1.01)	-	(0.01, 0.14)	(0.21,0.29)	(0.92,1.03)	(0.02,0.3)
IRLS/PVW							
$SP = SN = 0.99$	1.302	1.003	1.6%	0.126	0.25	1.23	0.174
	(1.24, 1.35)	(0.99,1.01)	-	(0.04,0.21)	(0.21,0.29)	(1.19,1.30)	(0.03,0.3)
$SN = 0.95$	1.308	1.01	5.8%	0.247	0.25	1.23	0.174
	(1.25, 1.36)	(0.99,1.02)	-	(0.16,0.32)	(0.21,0.29)	(1.19,1.30)	(0.03,0.3)
$SN = 0.90$	1.307	1.02	9%	0.386	0.25	1.24	0.174
	(1.25, 1.36)	(1.00,1.04)	-	(0.30,0.46)	(0.21,0.29)	(1.19,1.30)	(0.03,0.3)
$SN = 0.80$	1.304	1.03	12%	0.484	0.25	1.25	0.174
	(1.25, 1.36)	(1.02,1.04)	-	(0.41,0.55)	(0.21,0.29)	(1.19,1.30)	(0.04,0.3)

Table 3.2: Naive and misclassification-corrected mediation analysis allowing for exposure-mediator interaction ($SP = 0.99$, $SN = (0.8, 0.9, 0.95, 0.99)$, CI obtained from delta method standard errors)

In Table 3.2 we display the results of misclassification-corrected mediation analysis, accounting for the presence of exposure-mediator interaction. We observe that the effect of the exposure, mother's age, on the intermediate, pre-eclampsia, increases in magnitude as the sensitivity parameter (SN) decreases. In the estimation of the outcome regression parameters we observe that nor the effect of the exposure nor the interaction term appear

Estimates ($95\%CI$)	$\widehat{OR}^{NDE}$	$\widehat{OR}^{NIE}$	PM
Naive	1.309	1.0002	0.09%
	(1.25,1.35)	(0.99,1.00)	-
IRLS/PVW			
$SP = SN = 0.99$	1.308	1.0002	0.05%
	(1.26,1.35)	(0.99,1.00)	-
$SP = 0.99$ & $SN = 0.95$	1.308	1.0003	0.13%
	(1.26,1.35)	(0.99,1.00)	-
$SP = 0.99$ & $SN = 0.90$	1.308	1.0007	0.3%
	(1.26,1.35)	(0.99,1.00)	-
$SP = 0.99$ & $SN = 0.80$	1.307	1.003	1.2%
	(1.26,1.35)	(1.001,1.004)	-

Table 3.3: Naive and misclassification-corrected mediation analysis assuming no exposure-mediator interaction ($SP = 0.99$, $SN = (0.8, 0.9, 0.95, 0.99)$, CI obtained from delta method standard errors)

severely biased for any of the sensitivity analysis parameter values. Finally, the effect of pre-eclampsia on pre term birth is found to be under estimated. The misclassification-corrected analyses accounting for interaction reveal that the indirect effect of maternal age on preterm birth mediated by pre-eclampsia status was underestimated in the naive analyses. The direct effect remains pretty constant.

When exposure-mediator interaction is not included in the analyses (Table 3.3) the sensitivity analyses indicate a slight over estimation of the direct effect and under estimation of the indirect effect. However, even in the case of severe misclassification, changes the in causal effects estimates are minimal.

Therefore, we can still conclude that for all the values considered in the sensitivity analysis the association of maternal age on preterm birth is primarily through pathways other than pre-eclampsia status (with proportion of the effect of age on preterm birth mediated through pre-eclampsia taking up to the values of 12% if exposure-mediator interaction is included and with proportion mediated up to 1.2% in the absence of exposure-mediator interaction).

3.6 Discussion

We studied the problem of misclassification of a binary mediator in the context of causal mediation analysis, when the outcome is either binary or continuous, allowing for the presence of exposure-mediator interaction. We demonstrated that when non-differential misclassification of a binary mediator is ignored in the analyses, the estimators of direct and indirect causal effects that have been employed can be severely biased. The theoretical results and a numerical study illustrate that the misclassification bias can take unintuitive directions in the presence of non-linearities.

VanderWeele et al. (2012) show that although measurement error in the mediator induces biased direct and indirect effects, the combination of these biased effects is in fact unbiased for the total effect. However, this is true only if the mediator and outcome models with M^* replacing M are correctly specified. In both the simulations and the data example above the total effect of the exposure on the outcome (computed as either the sum or the product of direct and indirect effects) was also biased. This is because misclassification of a binary mediator induces mis-specification in the mediator and outcome regression.

We considered a full maximum likelihood approach (ML) and an hybrid of likelihood-based and predictive value weighting method as possible strategies of correction for misclassification. We implemented the latter approach using three different estimators for $P(M^*|Y, A, \boldsymbol{C})$, namely logistic regression (IRLS/PVW), logistic regression with splines (IRLS/**s**PVW) and using the true model for the conditional distribution that we seek to estimate (IRLS/**t**PVW). We compared the performance of corrected estimators for direct and indirect effects in a simulation study. Although appealing for its ease of implementation and less computational burden, the hybrid of likelihood-based and predictive value weighting methods did not always eliminate misclassification bias. The approach is expected to perform better when all the variables in the analysis are dichotomous. In this setting, we would recommend to apply the predictive value weighting approach specifying a saturated model for $P(M^*|Y, A, \boldsymbol{C})$. The ML approach instead was found to be consistent is all simulation settings except for the case of binary outcome with exposure-

mediator interaction. Although in this case ML was able to substantially improve over the naive estimator, the bias was not eliminated. The reason for this is that the estimation of the exposure-mediator interaction term remains problematic when the outcome is binary. However, in general, ML is the approach that we recommend to adopt. Code for implementing ML and IRLS/(PVW,**s**PVW,**t**PVW) is available in the supplementary material.

In many instances auxiliary information on the mis-measured intermediate is not available in mediation studies. We illustrated in a real data example the correction strategy coupled with sensitivity analysis for the unknown sensitivity and specificity for which no validation data or replicates for the mis-measured mediator is needed. Although the correction strategy using sensitivity analysis does not require validation data or replicates for the mis-measured mediator, estimators for correction could make use of this information, if available.

Some possible extensions of our study should be mentioned. Derivation of closed form asymptotic bias formulae of direct and indirect effects when the outcome is binary in the presence of exposure-mediator interaction and continuous covariates is of interest. We make the strong assumption of independence between the misclassification mechanism and all the other variables measured without error. It would be of interest to study the bias of naive direct and indirect causal effects when misclassification of a binary mediator is differential and relaxing the assumption that the other variables are measured without error. Finally, more work is needed to improve the misclassification-corrected estimators for direct and indirect effects in the presence of interactions.

3.A Description of measurement error mechanism

We can write the measurement error mechanism in an additive form:

$$M^* = M + U.$$

When the latent variable is binary the measurement error, U, is not normally distributed and can take values $(-1, 0, 1)$ under certain probabilities and restrictions. Moreover, $Cov(U, M) \neq 0$ and $Cov(U, M^*) \neq 0$ that is the error must be correlated with both the true and the observed level of the mediator (Carroll et al. 2006).

The moments of the error can be completely characterized by the knowledge of the prevalence of the true mediator, the sensitivity and specificity parameters.

Let $p^* = P(M^* = 1)$, $p = P(M = 1)$, $q^* = 1 - p^*$, $q = 1 - p$. Moreover define the reclassification probabilities $\eta = P(M = 1|M^* = 0)$ and $\nu = P(M = 0|M^* = 1)$. Then the moments of the misclassification error are given by (Aigner, 1973)

$E(U) = \nu p^* - \eta q^*$, $Var(U) = \nu p^* + \eta q^* - (\nu p^* - \eta q^*)^2$, and $Cov(M^*, U) = (\nu + \eta)p^* q^*$.

Note that reclassification probabilities can be re-expressed in terms of misclassification probabilities. Define the misclassification probabilities as $\gamma_0 = P(M^* = 1|M = 0)$ and $\gamma_1 = P(M^* = 0|M = 1)$. Then,

$$\begin{aligned} \nu &= \gamma_0 \frac{p}{q^*} \\ \eta &= \gamma_1 \frac{q}{p^*} \end{aligned}$$

Note that misclassification probabilities can be expressed in terms of sensitivity ($SN = P(M^*1|M = 1)$) and specificity ($SP = P(M^*0|M = 0)$). In particular,

$$\begin{aligned} \gamma_0 &= 1 - SP \\ \gamma_1 &= 1 - SN. \end{aligned}$$

Finally note that the prevalence of the observed mediator can be expressed in terms of misclassification probabilities and true prevalence of the mediator.

$$p^* = (1 - \gamma_1)p + \gamma_0 q.$$

These facts will be used throughout.

We further assume that the outcome Y and the exposure A as well as the additional covariates $\boldsymbol{C}$ are correctly measured. We assume that the error is non differential (i.e. $Cov(U, Y) = 0$) and $Cov(U, A) = Cov(U, \boldsymbol{C}) = 0$. Moreover, in the context of mediation analysis covariates A and $\boldsymbol{C}$, which can be either continuous, count or categorical variables, can be correlated with the misclassified mediator.

3.B Probability Limit of MLE of Continuous Outcome Regression

Let $\hat{\theta}^*$ be the vector of MLE estimators of the parameters from the outcome regression. Rewrite the outcome regression (2) in terms of M^* exploiting the assumption of additive measurement error

$$\begin{aligned} Y &= \theta_0 + \theta_1 a + \theta_2(m^* - u) + \theta_3 a(m^* - u) + \theta_4^{'} c + \epsilon \\ &= \theta_0 + \theta_1 a + \theta_2 m^* + \theta_3 a m^* + \theta_4^{'} c + \epsilon - \theta_2 u - \theta_3 \xi \\ &= \theta_0 + \theta_1 a + \theta_2 m^* + \theta_3 a m^* + \theta_4^{'} c + \epsilon - \theta_2 u - \theta_3 \xi \\ &= \theta_0^* + \theta_1^* a + \theta_2^* m^* + \theta_3^* a m^* + \theta_4^{*'} c + \epsilon^*, \end{aligned}$$

with $\xi = a \times u$.

Let $X^* = (1, A, M^*, AM^*, C)^T$. Then the vector of MLE estimators of the outcome linear regression parameters is given by,

$$\begin{aligned} \hat{\theta^*} &= (X^{*'} X^*)^{-1} X^{'*} Y \\ &= (X^{*'} X^*)^{-1} X^{*'} (X^* \theta + \epsilon^*) \\ &= (X^{*T} X^*)^{-1} X^{*T} (X^* \theta + \epsilon - \theta_2 u - \theta_3 \xi). \end{aligned}$$

By rearranging the equation and taking the limit we obtain obtain a formula for the asymptotic bias of the outcome regression parameters estimators when M is replaced by M^*

$$ABIAS(\hat{\boldsymbol{\theta}}^*) = -\Sigma^{-1}_{x^*x^*}\{\theta_2\Sigma_{x^*u} + \theta_3\Sigma_{x^*\xi}\}$$

where,

$$\begin{aligned}
\Sigma^{-1}_{x^*x^*} &= plim(\frac{(X^{*T}X^*)^{-1}}{n^{-1}}) \\
\Sigma_{x^*u} &= plim(\frac{X^{*T}u}{n}) = (0, 0, \sigma_{M^*u}, \sigma_{AM^*u}, 0, ..., 0)^T \\
\Sigma_{x^*\xi} &= plim(\frac{X^{*T}\xi}{n}) = (0, 0, \sigma_{M^*\xi}, \sigma_{AM^*\xi}, 0, ..., 0)^T
\end{aligned}$$

with

$$\begin{aligned}
\sigma_{M^*u} &= Cov(M^*, U) = ((1-SN)\frac{q}{p^*} + (1-SP)\frac{p}{q^*})p^*q^* \\
\sigma_{AM^*u} &= Cov(AM^*, u) = E(AM^*U) - E(AM^*)E(U) \\
\sigma_{AM^*\xi} &= Cov(AM^*, \xi) = E(A^2M^*U) - E(AM^*)E(A)E(U) \\
\sigma_{M^*\xi} &= Cov(M^*, \xi) = Cov(AM^*, U).
\end{aligned}$$

The parameters σ_{AM^*u}, $\sigma_{M^*\xi}$, and $\sigma_{AM^*\xi}$ depend upon the specification of sensitivity and specificity parameters, the marginal probability of the latent mediator and the joint probability of the mediator and the exposure.

We can rewrite the asymptotic bias as

$$\begin{aligned}
ABIAS(\hat{\boldsymbol{\theta}}^*) = &-\theta_2 \times \begin{pmatrix}
\delta_{1,M^*}Cov(M^*, U) + \delta_{1,AM^*}Cov(AM^*, U) \\
\delta_{A,M^*}Cov(M^*, U) + \delta_{A,AM^*}Cov(AM^*, U) \\
\delta_{M^*,M^*}Cov(M^*, U) + \delta_{M^*,AM^*}Cov(AM^*, U) \\
\delta_{AM^*,M^*}Cov(M^*, U) + \delta_{AM^*,AM^*}Cov(AM^*, U) \\
\boldsymbol{\delta_{C,M^*}}Cov(M^*, U) + \boldsymbol{\delta_{C,AM^*}}Cov(AM^*, U)
\end{pmatrix} + \\
&-\theta_3 \times \begin{pmatrix}
\delta_{1,M^*}Cov(M^*, AU) + \delta_{1,AM^*}Cov(AM^*, AU) \\
\delta_{A,M^*}Cov(M^*, AU) + \delta_{A,AM^*}Cov(AM^*, AU) \\
\delta_{M^*,M^*}Cov(M^*, AU) + \delta_{M^*,AM^*}Cov(AM^*, AU) \\
\delta_{AM^*,M^*}Cov(M^*, AU) + \delta_{AM^*,AM^*}Cov(AM^*, AU) \\
\boldsymbol{\delta_{C,M^*}}Cov(M^*, AU) + \boldsymbol{\delta_{C,AM^*}}Cov(AM^*, AU),
\end{pmatrix}
\end{aligned}$$

where $\delta_{.M^*}$ and $\delta_{.AM^*}$ are columns of $\Sigma^{-1}_{x_1^*,x_1^*}$.

From the asymptotic bias formulae given above the probability limit of $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, and $\hat{\theta}_3^*$ can

be easily derived as

$$\begin{aligned}
\theta_1^* &= \theta_1 - \theta_2(\delta_{A,M^*}Cov(M^*,U) + \delta_{A,AM^*}Cov(AM^*,U)) - \theta_3(\delta_{A,M^*}Cov(M^*,AU) + \\
&\quad +\delta_{A,AM^*}Cov(AM^*,AU)) \\
\theta_2^* &= \theta_2 - \theta_2(\delta_{M^*,M^*}Cov(M^*,U) + \delta_{M^*,AM^*}Cov(AM^*,U)) - \theta_3(\delta_{M^*,M^*}Cov(M^*,AU) + \\
&\quad +\delta_{M^*,AM^*}Cov(AM^*,AU)) \\
\theta_3^* &= \theta_3 - \theta_2(\delta_{AM^*,M^*}Cov(M^*,U) + \delta_{AM^*,AM^*}Cov(AM^*,U)) - \theta_3(\delta_{AM^*,M^*}Cov(M^*,AU) + \\
&\quad +\delta_{AM^*,AM^*}Cov(AM^*,AU)).
\end{aligned}$$

The probability limit for the outcome regression coefficients in absence of exposure-mediator interaction is easily obtained setting $\theta_3 = 0$ and setting to zero all the covariance terms that involve the exposure-mediator interaction.

3.C Asymptotic limit of naive outcome and mediator regression parameters in the absence of exposure-mediator interaction (θ^*,β^*) in terms of misclassification probabilities and true prevalence of the mediator

$$\begin{aligned}
\theta_1^* &= \theta_1 - \theta_2\delta_{A,M^*}Cov(M^*,U) \\
&= \theta_1 - \theta_2\delta_{A,M^*}[\gamma_0 q\{1-(1-\gamma_1)p-\gamma_0 q\} + \gamma_1 p\{(1-\gamma_1)p-\gamma_0 q\}] \\
\theta_2^* &= \theta_2(1-\delta_{M^*,M^*}Cov(M^*,U)) \\
&= \theta_2(1-\delta_{33}[\gamma_0 q\{1+(1-\gamma_1)p-\gamma_0 q\} + \gamma_1 p\{(1-\gamma_1)p-\gamma_0 q\}])
\end{aligned}$$

$$\begin{aligned}
\beta_0^* &\approx logit\{(1-\gamma_0-\gamma_1)expit(\beta_0)+\gamma_0\} \\
\beta_1^* &\approx \beta_1 \frac{(1-\gamma_0-\gamma_1)e^{\beta_0+\beta_2' c}}{\{(1-\gamma_1)e^{\beta_0+\beta_2' c}+\gamma_0\}\{\gamma_1 e^{\beta_0+\beta_2' c}+(1-\gamma_0)\}} \\
\beta_2^* &\approx \beta_1 \frac{(1-\gamma_0-\gamma_1)e^{\beta_0+\beta_1 a+\beta_2^{(-)'} c}}{\{(1-\gamma_1)e^{\beta_0+\beta_1 a+\beta_2^{(-)'} c}+\gamma_0\}\{\gamma_1 e^{\beta_0+\beta_1 a+\beta_2^{(-)'} c}+(1-\gamma_0)\}}.
\end{aligned}$$

3.D Asymptotic bias of direct and indirect effects naive estimators

$$
\begin{aligned}
ABIAS(\widehat{NDE}^{*}) &= \{\theta_1^*(a-a^*)\}+\{\theta_3^*(a-a^*)\}\frac{exp[\beta_0^*+\beta_1 a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}-\{\theta_1(a-a^*)\}+\\
&\quad -\{\theta_3(a-a^*)\}\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\\
&= (a-a^*)[\theta_3\{(1-\delta_{AM^*,M^*}Cov(M^*,AU)-\delta_{AM^*,AM^*}Cov(AM^*,AU))\times\\
&\quad \times\frac{exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}-\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}+\\
&\quad -(\delta_{A,M^*}Cov(M^*,AU)+\delta_{A,AM^*}Cov(AM^*,AU))\}-\theta_2\{(\delta_{A,M^*}Cov(M^*,U)+\\
&\quad +\delta_{A,AM^*}Cov(AM^*,U))+(\delta_{AM^*,M^*}Cov(M^*,U)+\delta_{AM^*,AM^*}Cov(AM^*,U))\times\\
&\quad \times\frac{exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}\}]
\end{aligned}
$$

$$
\begin{aligned}
ABIAS(\widehat{NIE}^{*}) &= (\theta_2^*+\theta_3^* a)\{\frac{exp[\beta_0^*+\beta_1^* a+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a+\beta_2^{*'}c]}-\frac{exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}\}+\\
&\quad -(\theta_2+\theta_3 a)\{\frac{exp[\beta_0+\beta_1 a+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]}-\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}\\
&= \theta_2[\{(1-\delta_{M^*,M^*}Cov(M^*,U)-\delta_{M^*,AM^*}Cov(AM^*,U))-a(\delta_{AM^*,M^*}Cov(M^*,U)+\\
&\quad +\delta_{AM^*,AM^*}Cov(AM^*,U))\}\{\frac{exp[\beta_0^*+\beta_1^* a+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^*+\beta_2^{*'}c]}-\frac{exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}\}+\\
&\quad -\{\frac{exp[\beta_0+\beta_1 a+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]}-\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}]+\theta_3[\{(1+\\
&\quad -\delta_{AM^*,M^*}Cov(M^*,AU)-\delta_{AM^*,AM^*}Cov(AM^*,AU))a-\delta_{M^*,M^*}Cov(M^*,AU)+\\
&\quad -\delta_{M^*,AM^*}Cov(AM^*,AU)\}\{\frac{exp[\beta_0^*+\beta_1^* a+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a+\beta_2^{*'}c]}-\frac{exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}{1+exp[\beta_0^*+\beta_1^* a^*+\beta_2^{*'}c]}\}\\
&\quad -a^*\{\frac{exp[\beta_0+\beta_1 a+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a+\beta_2^{'}c]}-\frac{exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\}].
\end{aligned}
$$

3.E Other theoretical results for direct effect naive estimators

$$\frac{\partial(NDE(\theta^*)-NDE(\theta))}{\partial SP}=0$$

$$-\theta_2\delta_{A,M^*}[q-SNqp-2(1-SP)q^2-(1-SN)qp]=0$$

$$[q-SNqp-2(1-SP)q^2-(1-SN)qp]=0$$

$$q^2[1-2(1-SP)]=0$$

$$1-2SP=0$$

$$SP=1/2$$

$$\frac{\partial(NDE(\theta^*)-NDE(\theta))}{\partial SN}=0$$

$$-\theta_2\delta_{A,M^*}[(1-SP)qp-p^2-2(1-SN)p^2-(1-SN)qp]=0$$

$$[(1-SP)qp-p^2-2(1-SN)p^2-(1-SN)qp]=0$$

$$p^2-2(1-SN)p^2=0$$

$$SN=1/2.$$

Partial derivative results can be obtained by plugging in the values $SN=SP=1$ from the first derivation step.

3.F When the mediator is misclassified will the sum of the biased direct and indirect effects estimator still give an unbiased estimate of the total effect?

Non parametrically, YES. See VanderWeele et al. (2012)

Consider now to estimate direct and indirect effects modeling the outcome using linear regression (2) and the mediator using logistic regression (1).
In order to estimate the total effect we can run the following outcome regression model

$$E[Y|A=a,M=m,C=c] \;=\; \theta_0^\dagger+\theta_1^\dagger a+\theta_4^{\dagger\prime}c.$$

The total effect can be estimated by $\theta_1^\dagger$.
If the outcome and mediator regression models are correctly specified, the total effect can

be computed as the sum of direct and indirect effects.

$$TE \;\; = \;\; NDE + NIE = \theta_1^{\dagger}.$$

In absence of interaction and measurement error and assuming a binary exposure,

$$TE \;\; = \;\; \theta_1 + \theta_2[expit\{\beta_0 + \beta_1 + \beta_2' c\} - expit\{\beta_0 + \beta_2' c\}] = \theta_1^{\dagger}.$$

In the presence of measurement error on continuous mediator and in the absence of interaction the sum of the naive estimators of direct and indirect effects yields an unbiased estimator of the total effect. We investigate whether this property will continue to hold in the presence of misclassification.

Consider that the true mediator M is not observed, rather a misclassified version of it, M^* is known by the investigator.

Then, the total effect estimated as the sum of the naive direct and indirect effect estimators in the presence of misclassification will be asymptotically biased.

$$TE^* \;\; = \;\; \theta_1^* + \theta_2^*[expit\{\beta_0^* + \beta_1^* a + \beta_2^{'*} c\} - expit\{\beta_0^* + \beta_1^* a^* + \beta_2^{'*} c\}] \neq \theta_1^{\dagger}.$$

This is because misclassification of the mediator induces a mis-specification of the mediator model and we know that the property of effect decomposition relies on the assumption of correct specification of both the outcome and mediator regression models.

Simulation results show that, although small, a residual bias of the total effect remains (sample size of 10,000 and 1000 replications).

However, if all the variables were binary and the model for the mediator were completely saturated the total effect estimated as the sum of naive direct and indirect effect would be unbiased.

I demonstrate this claim using some results of Gustafson (2004).

Let $P(M^* = 1|M = 1) - P(M^* = 1|M = 0) = SN + SP - 1$. The large sample limiting coefficients from least-squares regression of Y on (A, M^*) (assuming no covariate C is present) can be expressed as

$$\begin{aligned}
\theta_1^* &= \theta_1 + \theta_2 \rho \frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A}[1 - (SN + SP - 1)(\frac{\theta_2^*}{\theta_2})] \\
\theta_2^* &= \theta_2(SN + SP - 1)[\frac{p(1-p)(1-\rho^2)}{p^*(1-p^*) - p(1-p)\rho^{*2}(SN + SP - 1)^2}]
\end{aligned}$$

where $p = P(M = 1)$, $p^* = P(M^* = 1)$, $\sigma_A = SD(A)$, and $\rho = Cor(M, A)$.

Then, we can rewrite the naive large sample limiting total effect as

$$\begin{aligned} TE^* &= NDE^* + NIE^* \\ &= \theta_1 + \theta_2\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A}[1-(SN+SP-1)(\frac{\theta_2^*}{\theta_2})] + \theta_2(SN+SP-1)\times \\ &\quad \times[\frac{p(1-p)(1-\rho^2)}{p^*(1-p^*)-p(1-p)\rho^{*2}(SN+SP-1)^2}]\{\frac{exp[\beta_0^*+\beta_1^*]}{1+exp[\beta_0^*+\beta_1^*]} - \frac{exp[\beta_0^*]}{1+exp[\beta_0^*]}\}. \end{aligned}$$

Since both M^* and A are binary the model for the mediator is completely saturated and we can re-write the equation as

$$\begin{aligned} &= \theta_1 + \theta_2\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A}[1-(SN+SP-1)(\frac{\theta_2^*}{\theta_2})] + \theta_2(SN+SP-1)\times \\ &\quad \times[\frac{p(1-p)(1-\rho^2)}{p^*(1-p^*)-p(1-p)\rho^{*2}(SN+SP-1)^2}]\{P(M^*=1|A=1)-P(M^*=1|A=0)\} \\ &= \theta_1 + \theta_2\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A} - \theta_2\{[(SN+SP-1)\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A} - (P(M^*=1|A=1)+ \\ &\quad -P(M^*=1|A=0)))](SN+SP-1)[\frac{p(1-p)(1-\rho^2)}{p^*(1-p^*)-p(1-p)\rho^{*2}(SN+SP-1)^2}]\}. \end{aligned}$$

We note that

$$\begin{aligned} (SN+SP-1)\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A} &= [P(M^*=1|M=1)-P(M^*=1|M=1)]\times\frac{Cov(M,A)}{\sigma_A} \\ &= [P(M^*=1|M=1)-P(M^*=1|M=0)]\times \\ &\quad \times[P(M=1|A=1)-P(M=1|A=0)] \\ &= [P(M^*=1|A=1)-P(M^*=1|A=0)]. \end{aligned}$$

Therefore a term in the equation drops and we obtain

$$TE^* = \theta_1 + \theta_2\rho\frac{\{p(1-p)\}^{\frac{1}{2}}}{\sigma_A} = \theta_1 + \theta_2[P(M=1|A=1)-P(M=1|A=0)] = TE.$$

Therefore, we conclude that, in the presence of misclassification of a binary mediator and when outcome and mediator are modeled using parametric models, the sum of naive direct and indirect effects will yield an unbiased total effect had the mediator regression been saturated.

3.G Standard errors of the PVW/IRLS estimators for direct and indirect causal effects

We now derive the standard errors of PVW/IRLS estimators for natural direct and indirect effects assuming that exposure-mediator interaction may be present.
Define the corrected PVW/IRLS estimators of the causal effects of interest when the outcome is continuous modeled using linear regression as

$$\begin{aligned}
\widehat{NDE}(\hat{\theta}^{PVW}, \hat{\beta}^{IRLS}) &= \{\hat{\theta}_1^{PVW}(a-a^*)\} + \{\hat{\theta}_3^{PVW}(a-a^*)\}\frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a^* + \hat{\beta}_2^{IRLS'}c]}{1+exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a^* + \hat{\beta}_2^{IRLS'}c]}. \\
\widehat{NIE}(\hat{\theta}^{PVW}, \hat{\beta}^{IRLS}) &= (\hat{\theta}_2^{PVW} + \hat{\theta}_3^{PVW}a)\{\frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a + \hat{\beta}_2^{IRLS'}c]}{1+exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a + \hat{\beta}_2^{IRLS'}c]} + \\
&\quad -\frac{exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a^* + \hat{\beta}_2^{IRLS'}c]}{1+exp[\hat{\beta}_0^{IRLS} + \hat{\beta}_1^{IRLS}a^* + \hat{\beta}_2^{IRLS'}c]}\}.
\end{aligned}$$

Define the corrected PVW/IRLS estimators of the causal effects of interest when the outcome is binary modeled using logistic regression as

$$\widehat{OR}^{NDE}(\hat{\theta}^{PVW}, \hat{\beta}^{IRLS}) = \{\frac{exp[\hat{\theta}_1^{PVW}a](1+exp[\hat{\theta}_2^{PVW}+\hat{\theta}_3^{PVW}a+\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a^*+\hat{\beta}_2^{IRLS'}c])}{exp[\hat{\theta}_1^{PVW}a^*](1+exp[\hat{\theta}_2^{PVW}+\hat{\theta}_3^{PVW}a^*+\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a^*+\hat{\beta}_2^{IRLS'}c])}\}.$$

$$\widehat{OR}^{NIE}(\hat{\theta}^{PVW}, \hat{\beta}^{IRLS}) = \frac{[1+exp(\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a^*+\hat{\beta}_2^{IRLS'}c)][1+exp(\hat{\theta}_2^{PVW}+\hat{\theta}_3^{PVW}a+\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a+\hat{\beta}_2^{IRLS'}c)]}{[1+exp(\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a+\hat{\beta}_2^{IRLS'}c)][1+exp(\hat{\theta}_2^{PVW}+\hat{\theta}_3^{PVW}a+\hat{\beta}_0^{IRLS}+\hat{\beta}_1^{IRLS}a^*+\hat{\beta}_2^{IRLS'}c)]}.$$

Suppose that the mediator regression has been fitted using Iteratively Reweighted Least Squares as described in section 3.4.1 and that the outcome regression has been fitted using the Predictive Value Weighting approach as described in section 3.4.2 and that the resulting estimates $\beta^{I\hat{R}LS}$ of $\beta^{IRLS} = (\beta_0^{IRLS}, \beta_1^{IRLS}, \beta_2^{IRLS'})'$ and $\theta^{P\hat{V}W}$ of $\theta^{PVW} = (\theta_0^{PVW}, \theta_1^{PVW}, \theta_2^{PVW}, \theta_3^{PVW}, \theta_4^{PVW'})'$ have covariance matrices $\Sigma_{\beta^{IRLS}}$ and $\Sigma_{\theta^{PVW}}$. Then the covariance matrix of $(\hat{\beta}^{IRLS'}, \hat{\theta}^{PVW'})$ is

$$\Sigma^{PVW/IRLS} = \begin{bmatrix} \Sigma_{\beta^{IRLS}} & 0 \\ 0 & \Sigma_{\theta^{PVW}} \end{bmatrix}$$

where $\Sigma_{\beta^{IRLS}} = E(-\frac{\partial^2 L(\beta)}{\partial\beta\beta^T})$, obtainable from the hessian of the optimization procedure; and $\Sigma_{\theta^{PVW}} = \Gamma_W^{-1}\Omega_W\Gamma_W^{-1}$. $\Sigma_{\theta^{PVW}}$ can be estimated by $\widehat{\Sigma}_{\theta^{PVW}} = \widehat{\Gamma}_W^{-1}\widehat{\Omega}_W\widehat{\Gamma}_W^{-1}$ where

$$\widehat{\Gamma}_W = \frac{1}{N}\sum_{i=1}^{N}(\frac{\partial E(Y_i)}{\partial\theta})^T|_{\hat{\theta}^{PVW}_{\widehat{W}}}\widehat{W}_i(\frac{\partial E(Y_i)}{\partial\theta})|_{\hat{\theta}^{PVW}_{\widehat{W}}}$$

and

$$\widehat{\Omega}_W = \frac{1}{N}\sum_{i=1}^{N}(\frac{\partial E(Y_i)}{\partial\theta})^T|_{\hat{\theta}^{PVW}_{\widehat{W}}}\widehat{W}_i\widehat{Var}(Y_i)\widehat{W}_i(\frac{\partial E(Y_i)}{\partial\theta})|_{\hat{\theta}^{PVW}_{\widehat{W}}}$$

For continuous outcome modeled with least squares regression

$$\widehat{\Gamma}_W = \frac{1}{N}\sum_{i=1}^{N}(X_i^T\widehat{W}_iX_i)$$

and

$$\widehat{\Omega}_W = \frac{1}{N}\sum_{i=1}^{N}(X_i^T\widehat{W}_i(Y_i - X_i\hat{\theta}^{PVW}_{\widehat{W}})(Y_i - X_i\hat{\theta}^{PVW}_{\widehat{W}})^T\widehat{W}_iX_i)$$

These standard errors are obtainable if the weighted regression procedure explained in section 3.4.2 is fitted using the generalized estimating equation framework (for example using the ”repeated” option PROC GENMOD in SAS outputs ”robust” standard errors).

Standard errors of the PVW/IRLS natural direct and indirect effects can be obtained (using the delta method) as

$$\sqrt{\Gamma^{PVW/IRLS}\Sigma^{PVW/IRLS}\Gamma^{PVW/IRLS\prime}}$$

When the outcome Y is continuous modeled using linear regression $\Gamma^{PVW/IRLS} = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the pure natural direct effect, where

$$d_1 = \frac{\theta_3^{PVW} exp[[\beta_0^{IRLS} + \beta_1^{IRLS} a^*_{\beta 2}{}^{IRLS'} c](1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]) - \theta_3^{PVW} \{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]\}^2}{(1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c])^2}$$

$$d_2 = \frac{\theta_3^{PVW} a^* exp[[\beta_0^{IRLS} + \beta_1^{IRLS} a^*_{\beta 2}{}^{IRLS'} c](1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]) - \{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]\}^2}{(1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c])^2}$$

$$d_3 = \frac{\theta_3^{PVW} c^{'} exp[[\beta_0^{IRLS} + \beta_1^{IRLS} a^*_{\beta 2}{}^{IRLS'} c](1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]) - \{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]\}^2}{(1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c])^2}$$

$$d_4 = 0$$

$$d_5 = 1$$

$$d_6 = 0$$

$$d_7 = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]}{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* \beta_2^{IRLS'} c]}$$

$$d_8 = 0^{'}$$

For the natural indirect effect let

$$A = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]\} - \{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]\}^2}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]\}^2}$$

$$B = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\} - \{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\}^2}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\}^2}$$

$$K = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a + \beta_2^{IRLS'} c]\}}$$

$$D = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\}}$$

and

$\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$, where

$$d_1 = \{\theta_2^{PVW} + \theta_3^{PVW} a\}[A - B]$$

$$d_2 = \{\theta_2^{PVW} + \theta_3^{PVW} a\}[aA - a^* B]$$

$$d_3 = \{\theta_2^{PVW} + \theta_3^{PVW} a\} c^{'} [A - B]$$

$$d_4 = 0$$

$$d_5 = 0$$

$$d_6 = K - D$$

$$d_7 = a[K - D]$$

$$d_8 = 0^{'}$$

When the outcome Y is binary modeled using logistic regression $\Gamma^{PVW/IRLS} = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the logarithm of pure natural direct effect, where let

$$A = \frac{exp[\theta_2^{PVW} + \theta_3^{PVW} a + \beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]}{\{1 + exp[\theta_2^{PVW} + \theta_3^{PVW} a + \beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\}}$$

$$B = \frac{exp[\theta_2^{PVW} + \theta_3^{PVW} a^* + \beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]}{\{1 + exp[\theta_2^{PVW} + \theta_3^{PVW} a^* + \beta_0^{IRLS} + \beta_1^{IRLS} a^* + \beta_2^{IRLS'} c]\}}$$

and

$$d_1 = A - B$$

$$d_2 = a^*(A - B)$$

$$d_3 = c^{'}(A - B)$$

$$d_4 = 0$$

$$d_5 = (a - a^*)$$

$$d_6 = A - B$$

$$d_7 = aA - a^* B$$

$$d_8 = 0^{'}$$

for the logarithm of the natural indirect effect let

$$A = \frac{exp[\theta_2^{PVW} + \theta_3^{PVW}a + \beta_0^{IRLS} + \beta_1^{IRLS}a + \beta_2^{IRLS'}c]}{\{1 + exp[\theta_2^{PVW} + \theta_3^{PVW}a + \beta_0^{IRLS} + \beta_1^{IRLS}a + \beta_2^{IRLS'}c]\}}$$

$$B = \frac{exp[\theta_2^{PVW} + \theta_3^{PVW}a + \beta_0^{IRLS} + \beta_1^{IRLS}a^* + \beta_2^{IRLS'}c]}{\{1 + exp[\theta_2^{PVW} + \theta_3^{PVW}a + \beta_0^{IRLS} + \beta_1^{IRLS}a^* + \beta_2^{IRLS'}c]\}}$$

$$K = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS}a + \beta_2^{IRLS'}c]}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS}a + \beta_2^{IRLS'}c]\}}$$

$$D = \frac{exp[\beta_0^{IRLS} + \beta_1^{IRLS}a^* + \beta_2^{IRLS'}c]}{\{1 + exp[\beta_0^{IRLS} + \beta_1^{IRLS}a^* + \beta_2^{IRLS'}c]\}}$$

and

$\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ where

$$d_1 = (D + A) - (K + B)$$

$$d_2 = a^*[D - B] + a[A - K]$$

$$d_3 = c^{'}[(D + A) - (K + B)]$$

$$d_4 = 0$$

$$d_5 = 0$$

$$d_6 = A - B$$

$$d_7 = a[A - B]$$

$$d_8 = 0^{'}.$$

Standard errors of the PVW/IRLS estimators of the causal effects of interest can be obtained in absence of exposure-mediator interaction in a similar way by setting $\theta_3 = 0$.

3.H Direction of asymptotic bias of naive direct and indirect effects estimators in the absence of exposure-mediator interaction

Proposition 1: For continuous outcome, in the absence of exposure mediator interaction, the naive direct effect estimator is biased away from the null if positive and towards the null if negative, under the assumption that the effect of the exposure on the mediator (β_1) and the effect of the mediator on the outcome (θ_1) have the same sign.

Proof:

In the absence of exposure-mediator interaction the asymptotic bias of the naive estimator of the direct effect can be expressed as

$$ABIAS(NDE(\hat{\theta}^*)) = -\theta_2 \delta_{A,M^*} Cov(M^*, U)(a - a^*).$$

Assume we are studying the direct effect of the exposure A on the outcome Y for an increase of the level of the exposure such that $(a - a^*) > 0$. Assume also that the procedure producing the observed classification of the binary mediator performs better than chance (i.e. $SN + SP > 1$) which implies that $0 < Cov(M^*, U) < 1$.
The term δ_{A,M^*} is an element of the inverse of the variance-covariance matrix of the observed variables (A, M^*, C) which, provided $Cov(A, M) \times Var(C) > Cov(A, C) \times Cov(M, C)$, will be negative if $Cov(A, M) > 0$ (i.e. $\beta_1 > 0$) and positive if $Cov(A, M) < 0$ (i.e. $\beta_1 < 0$).
We can show that $Cov(A, M) \times Var(C) \geq Cov(A, C) \times Cov(C, M)$ in the following way:

The LHS of the inequality can be re-written as

$Cov(A,M) \times Var(C) = [E(AM) - E(A)E(M)] \times Var(C)$
$= [E\{E(AM|C)\} - E\{E(A|C)\}E\{E(M|C)\}] \times Var(C)$
$= \{\sum_c \sum_a (a - \mu_{a|c}) \sum_m (m - \mu_{m|a,c}) p_{m|a,c} p_{a|c} p_c\} \times \{\sum_c (c - \mu_c)^2 p_c\}.$

The RHS of the inequality can be re-written as

$Cov(A,C) \times Cov(M,C) = [E(AC) - E(A)E(C)] \times [E(MC) - E(M)E(C)]$
$= [E\{E(AC|C)\} - E\{E(A|C)\}E\{E(C|C)\}] \times [E\{E(MC|C)\} - E\{E(M|C)\}E\{E(C|C)\}]$
$= \{\sum_c (c - \mu_c) \sum_a (a - \mu_{a|c}) p_{a|c} p_c\} \times \{\sum_c (c - \mu_c) \sum_a \sum_m (m - \mu_{m|a,c}) p_{m|a,c} p_{a|c} p_c\}.$

Note that given that $Var(C) > 0$ and assuming $Cov(A,M)$ has the same sign of θ_2, if $Cov(A,C)$ and $Cov(M,C)$ have different signs then the inequality immediately follows. Assume now that $Cov(A,C)$ and $Cov(M,C)$ have equal signs. We can note that by Holder's inequality

$Cov(A,M) \times Var(C) \geq \{\sum_c (c - \mu_c)^2 \sum_a (a - \mu_{a|c}) \sum_m (m - \mu_{m|a,c}) p_{m|a,c} p_{a|c} p_c\}$

$\geq \{\sum_c (c - \mu_c) \sum_a (a - \mu_{a|c}) p_{a|c} p_c\} \times \{\sum_c (c - \mu_c) \sum_a \sum_m (m - \mu_{m|a,c}) p_{m|a,c} p_{a|c} p_c\}$

$= Cov(A,C) \times Cov(M,C),$

and the inequality follows.

Thus, if $\theta_2 > 0$ and $\beta_1 > 0$ or if $\theta_2 < 0$ and $\beta_1 < 0$ then $ABIAS(NDE(\hat{\theta}^*)) > 0$.
Therefore, we conclude that the bias of natural direct effect will be biased away from the null if $NDE > 0$ and the natural direct effect will be biased towards the null if $NDE < 0$.

Proposition 2: For continuous outcome, in the absence of exposure mediator interaction, the naive indirect effect estimator is biased towards the null if positive and biased away from the null if negative under the assumption that the effect of the exposure on the mediator (β_1) and the effect of the mediator on the outcome (θ_1) have the same sign.

Proof:

In the absence of exposure-mediator interaction the asymptotic bias of the naive estimator of the indirect effect can be expressed as

$$ABIAS(NIE(\hat{\theta}^*, \hat{\beta}^*)) = \theta_2[\{\frac{exp[\beta_0^* + \beta_1^* a + \beta_2^{*'} c]}{1 + exp[\beta_0^* + \beta_1^* a + \beta_2^{*'} c]} - \frac{exp[\beta_0^* + \beta_1^* a^* + \beta_2^{*'} c]}{1 + exp[\beta_0^* + \beta_1^* a^* + \beta_2^{*'} c]}\}+$$

$$-\{\frac{exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2' c]}\}]$$

$$-\theta_2 \delta_{M^*,M^*} Cov(M^*, U)\{\frac{exp[\beta_0^* + \beta_1^* a + \beta_2^{*'} c]}{1 + exp[\beta_0^* + \beta_1^* a + \beta_2^{*'} c]} - \frac{exp[\beta_0^* + \beta_1^* a^* + \beta_2^{*'} c]}{1 + exp[\beta_0^* + \beta_1^* a^* + \beta_2^{*'} c]}\}.$$

Assume again that we are studying the indirect effect of the exposure A on the outcome Y through the mediator for an increase of the level of the exposure such that $a > a^*$. Assume also that the procedure producing the observed classification of the binary mediator performs better than chance (i.e. $SN + SP > 1$), which implies that $0 < Cov(M^*, U) < 1$. The term δ_{M^*,M^*} is element on the diagonal of inverse of the variance-covariance matrix of the observed variables (A, M^*, C) which will be always positive.
Consider the third line of the formula. This term is going to be negative if $\beta_1 > 0$ and $\theta_2 > 0$ or if $\beta_1 < 0$ and $\theta_2 < 0$. Now consider the first two lines of the equation. Since the misclassification induces a dilution of the covariate effects on the mediator, the change in probability of the mediator for a change in the exposure will be smaller when the naive estimator substitutes the true one. Therefore, the term in the first two lines of the equation

is negative if $\beta_1 > 0$ and $\theta_2 > 0$ or $\beta_1 < 0$ and $\theta_2 < 0$. Therefore, if θ_2 and β_1 have the same sign the asymptotic bias is going to be negative. We conclude that the asymptotic bias of the indirect effect is towards the null if the indirect effect is positive, and away from the null if the indirect effect is negative.

3.I Predictive Value Weighting approach with correctly specified model for $M^*|Y, A, C$

Recover the model for $M^*|Y, A, C$

$M^*|Y, A, C \sim Ber(p_{M^*|Y,A,C})$.

Where,

$P(M^* = 1|Y, A, C) = \sum_{m=0}^{1} P(M^* = 1|M = m, Y, A, C)P(M = m|Y, A, C)$.

Assuming non-differential misclassification

$P(M^* = 1|Y, A, C) = SN \times P(M = 1|Y, A, C) + (1 - SP) \times P(M = 0|Y, A, C)$

Bayes Theorem

$= SN\frac{P(Y,A,C|M=1)P(M=1)}{P(Y,A,C)} + (1 - SP)\frac{P(Y,A,C|M=0)P(M=0)}{P(Y,A,C)}$

Law of Total Probability and Bayes Theorem

$= \frac{[SNP(Y|A,M=1,C)P(M=1|A,C)+(1-SP)P(Y|A,M=0,C)P(M=0|A,C)]}{\sum_{m=0}^{1} P(Y|A,M=m,C)P(M=m|A,C)}$.

We observe that the probability just derived depends on both the mediator and outcome regression parameters we wish to estimate.

Therefore, maximizing the likelihood arising from the conditional distribution $M^*|Y, A, C$ with respect to (θ, β) will directly allow us to estimate the parameters of interest.

Given that consistent estimators for the parameters of interest must be recovered in order

to estimate consistently $P(M^*|Y, A, C)$, there is no need to continue estimating weights and implementing the weighted regression approach.

For both binary and continuous outcome we can obtain estimators for outcome and mediator regression parameters (θ and β) by maximizing the following likelihood

$$max_{\theta,\beta}\mathcal{L}(\theta, \beta) = max_{\theta,\beta}\prod_{i=1}^{n} P(M_i^* = 1|Y_i, A_i, C_i)^{M_i^*}(1 - P(M_i^* = 1|Y_i, A_i, C_i))^{(1-M_i^*)}.$$

3.J Maximum Likelihood (″Direct Method″) approach to misclassification correction in the outcome regression

When fitting the outcome regression we wish to recover the vector of parameters θ which characterizes the distribution of $Y|A, M, C$. When the outcome is continuous we define

$$f_{Y|A,M,C;\theta} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(Y-\mu_{Y|A,M,C})},$$

where σ^2 is the conditional variance of Y given A, M, C and $\mu_{Y|A,M,C} = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C$.

When the outcome is binary we define

$$f_{Y|A,M,C;\theta} = p_y^Y(1 - p_y)^{(1-Y)},$$

where $p_y = p_{Y|A,M,C} = exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C)/(1 + exp(\theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4 C))$.

When the observed mediator M^* is a misclassified version of M the likelihood arising from the parametric models just given cannot be fit.

The maximum likelihood approach consists of specifying the likelihood for a measurement error model in which not only the outcome, but also the observed mediator is consider random, rather than fixed.

Let's determine the joint distribution of $Y, M^*|A, C$ given that we assume sensitivity and

specificity known.

$f_{Y,M^*|A,C}(y, m^*|a, c) = \sum_{m=0}^{1} f_{Y,M^*,M|A,C}(y, m^*, m|a, c)$
$= \sum_{m=0}^{1} f_{Y|M^*,M,A,C}(y|m^*, m, a, c)P(M^* = m^*|M = m, A = a, C = c)P(M = m|A = a, C = c)$.

Assuming non differential misclassification

$= \sum_{m=0}^{1} f_{Y|M,A,C}(y|m, a, c)P(M^* = m^*|M = m)P(M = m|A = a, C = c)$.

Note that $P(M^* = m^*|M = m)$ for each combination of M and M^* is assumed to be known and in a mediation setting the model for $P(M = m|A = a, C = c)$ is easy to postulate as $P(M = m|A = a, C = c; \beta)$. Given the knowledge of sensitivity and specificity the vector of parameters β can be recovered.
The observed data distribution can be defined in terms of the measurement error model

$$f_{Y|M^*,A,C;\theta^*}(y|m^*, a, c) = \frac{f_{Y,M^*|A,C;\theta,\beta}(y, m^*|a, c)}{\int_y f_{Y,M^*|A,C;\theta,\beta}(y, m^*|a, c)dy}$$

and we can maximize the likelihood with respect to the true vector of parameters θ and β

$$max_{\theta,\beta}\mathcal{L}(\theta, beta) = max_{\theta,\beta}\frac{f_{Y,M^*|A,C;\theta,\beta}(y, m^*|a, c)}{\int_y f_{Y,M^*|A,C;\theta,\beta}(y, m^*|a, c)dy} = max_{\theta,\beta}f_{Y,M^*|A,C;\theta,\beta}(y, m^*|a, c).$$

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. N. Petrov and F. Czaki, pp. 267-81. Budapest: Akademiai Kiado.

[2] Alwin, D.F. & Hauser, R.M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, **40**: 37-47.

[3] Ananth, C.V. and VanderWeele, T.J. (2011). Placental abruption and perinatal mortality with preterm delivery as a mediator: disentangling direct and indirect effects. *American Journal of Epidemiology*, **174**:99-108.

[4] Armstrong, B. (1985). Measurement error in generalized linear models. *Communications in Statistics, Series B*, **14**:529-544.

[5] 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*, **51**:1173-1182.

[6] Carroll, R.J., Ruppert, D., Stefanski, and L.A., Crainiceanu, C.M. (2006). Measurement error in non-linear models. *Chapman & Hall/CRC*.

[7] Cochran, W.G. (1968). Errors of measurement in statistics. *Technometrics* **10**(4),637-666. cole Cole, D. A., Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. Journal of Abnormal Psychology, **112**: 558-577.

[8] Fuller, W.A., (2006). Measurement Error Models. *Wiley's Series in Probability and Statistics*.

[9] Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology*. Chapman & Hall/CRC.

[10] Goldenberg R.L., Culhane, J.F., Iams, J.D., Romero, R. (2008). Epidemiology and causes of preterm birth. *Lancet*, **371**(9606):75-84.

[11] Hafeman, D.M. and VanderWeele, T.J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, **22**:753-764.

[12] Hernán, M.A. (2004). A definition of causal effect for epidemiological studies. *Journal of Epidemiology and Community Health*, **58**:265-271.

[13] Hoyle, R.H., and Kenny, D.A. (1999). Sample size, reliability, and tests of statistical mediation. In: Hoyle RH, editor. *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage. pp. 195-222.

[14] Hnat,M.D., Baha, D., Sibai,M. et al. (2002). Perinatal outcome in women with recurrent preeclampsia compared with women who develop preeclampsia as nulliparas *American Journal of Obstetrics and Gynecology*, **186**(3): 422-426.

[15] Huang, L., Wang, H. , and Cox, C. (2005). Assessing Interaction Effects in Linear Measurement Error Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**(1):21-30.

[16] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Math. Statist.*, 1, Ed. L. M. Le Cam and J. Neyman, pp. 221-233. Berkeley: University of California Press.

[17] Hyman, H.H. (1955). *Survey design and analysis: Principles, cases and procedures.* Glencoe, IL: Free Press.

[18] Imai, K., Keele L., & Tingley, D. (2010a). A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4):309-334.

[19] Imai, K., Keele L.,Tingley, D., & Yamamoto, T. (2010b). Causal Mediation Analysis Using R, *Advances in Social Science Research Using R*, ed. H. D. Vinod, New York: Springer 129-154.

[20] James, L. R. & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, **69**:307-321.

[21] Jacobsson B., Ladfors L., Milson I. (2004). Advanced maternal age and adverse perinatal outcome. *Obstet Gynecol.*, **104**:727-33.

[22] Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, **13**:314-336.

[23] Joffe, M., Small, D., & Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*, **22**:74-97, DOI: 10.1214/088342306000000655.

[24] Judd, C.M. & Kenny, D.A. (1981). Process analysis: estimating mediation in treatment evaluations. *Evaluation Review*, **5**:602-619.

[25] Judd, C.M., Kenny, D.A., & McClelland, G.H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, **6**:115–34.

[26] Kraemer, HC, Kiernan M, Essex M,& Kupfer DJ. (2008). How and why the criteria defining moderators and mediators differ between the Baron Kenny and MacArthur Approaches. *Health Psychology*, **27**(2 Suppl.):S101-S108.

[27] Lamminpaa, R., Vehvilainen-Julkunen, K., Gissler M., Heinonen, S. (2012). Preeclampsia complicated by advanced maternal age: a registry-based study on primiparous women in Finland 1997-2008.*BMC Pregnancy Childbirth*, **12**(1):47.

[28] Lyles, R.H., Lin, J. (2010). Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in medicine*, **29**(22):2297-309.

[29] McCallum, B.T. (1972). Relative Asymptotic Bias from Errors of Omission and Measurement. *Econometrica*, **40**(4),757-758.

[30] McCormick, M.C. (1985).The contribution of low birth weight to infant mortality and childhood morbidity. *N. Engl. J. Med.*, **312**: 82-90.

[31] MacKinnon, D.P. & Dwyer, J.H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, **17**:144-158.

[32] MacKinnon, D.P., Warsi, G., & Dwyer, J.H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, **30**:41-62.

[33] MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York: Erlbaum.

[34] Meads, C.A., Cnossen, J.S., Meher, S., Juarez-Garcia, A., Riet, G., Duley, L., et al. (2008). Methods of prediction and prevention of pre-eclampsia: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess*, **12**(6).

[35] Merrill, R.M. (1994). *Treatment effect evaluation in non-additive mediation models.* Unpublished doctoral dissertation, Arizona State University, Tempe.

[36] Morgan-Lopez A.A. & MacKinnon, D.P. (2006). Demonstration and evaluation of a method to assess mediated moderation. *Behavioral Research Methods*, **38**:77–87.

[37] Muller, D., Judd & C.M., Yzerbyt, V.Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, **89**:852-63.

[38] Muthén, B. (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Submitted for publication.

[39] Murad, H., and Freedman, L.S. et al. (2007). Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Statistics in Medicine*, **26**:4293-4310. DOI: 10.1002/sim.2849.

[40] Neuhaus, J.M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**(4): 843-855.

[41] Neuhaus, J.M., and Jewell, N.P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**, 807-16.

[42] Ogburn, E.L. and VanderWeele, T.J. (2012). Analytic results on the bias due to non-differential misclassification of a binary mediator. *American Journal of Epidemiology*, **176**:555-561.

[43] Pearl, J. (2001). Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 411-420.

[44] Preacher, K. J. & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, **36**:717-731.

[45] Preacher, K.J., Rucker, D.D., & Hayes, A.F. 2007. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, **42**(1):185-227.

[46] Robins, J.M. (2003). *Semantics of causal DAG models and the identification of direct and indirect effects*. In Highly Structured Stochastic Systems, Eds. P. Green, N.L. Hjort, and S. Richardson, 70-81. Oxford University Press, New York.

[47] Robins, J.M., and Richardson, T.S. (2010). Alternative graphical causal models and the identification of direct effects. To appear in *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. P. Shrout, Editor. Oxford University Press.

[48] Robins, J.M. & Greenland S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**:143-155.

[49] Rosner, B., Spiegelman, D., and Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Americam Journal of Epidemiology*, **132**:734-745.

[50] Rosner, B., Spiegelman, D., and Willett, W.C. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within person measurement error. *Americam Journal of Epidemiology*, **132**:734-745.

[51] Rosner, B., Willett, W.C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates for systematic within-person measurement error. *Statistics in Medicine*, **8**:1051-1069.

[52] Shpitser, I. and VanderWeele, T.J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *International Journal of Biostatistics*, **7**, Article 16:1-24.

[53] Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology*, 290-312. San Francisco: Jossey-Bass.

[54] Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, **33**:230-251.

[55] Spiegelman, D., McDermott, A., and Rosner, B. (1997). Regression calibration method for correcting measurement error bias in nutritional epidemiology. *American Journal of Clinical Nutrition*, **65**(suppl):1179s-1186s.

[56] Stefanski, L., and Cook, J. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, **90**:1247-1256.

[57] Turner, J.A. (2010). Diagnosis and management of pre-eclampsia: an update. *International Journal of Women's Health*, **2**(1):327-337.

[58] Valeri, L., Lin, X., and VanderWeele, T.J. (2012). Mediation analysis in generalized linear models when the mediator is measured with error. *Technical Report*.

[59] Valeri, L., and VanderWeele, T. J. (2012).(In Press). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*.

[60] VanderWeele T. (2011). Causal Mediation Analysis with Survival Data. *Epidemiology*, **22**:575-581.

[61] VanderWeele, T. J. & Vansteelandt S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, **2**(4):457-468.

[62] VanderWeele, T. J. & Vansteelandt S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *Am. J. Epidemiol.*, **172**(12):1339-1348. DOI: 10.1093/aje/kwq332

[63] VanderWeele, T.J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, **21**:540-551.

[64] VanderWeele, T.J. et al. (2012a).(In Press). Genetic variants on 15q25.1, smoking and lung cancer: an assessment of mediation and interaction. *American Journal of Epidemiology.*.

[65] VanderWeele T.J., Valeri, L., Ogburn, E.L. (2012). The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology*, **23**: 561-564.

[66] Wang, N., Lin, X., Gutierrez, R.G., and Carroll, R.J.(1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of the American Statistical Association*, **93**(441):249-261.

[67] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**:1-25.

[68] World Health Organization. (2005). World health report 2005: make every mother and child count. Geneva: WHO. page 63.