

MEDIATION ANALYSIS FOR CASE-CONTROL STUDIES WITH SECONDARY AND
TERTIARY OUTCOMES

By

Zichun Cao

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics – Doctor of Philosophy

2023

ABSTRACT

Mediation analyses have been implemented in observational studies such as cross-sectional, case-control and both prospective and retrospective cohort studies. However, little has been done for mediation analysis in a case-control study with subsequently collected (secondary and/or tertiary) outcomes. This may happen in a case-control study with additional interests of a secondary/tertiary outcome and direct or indirect effects of exposure on the secondary (tertiary) outcome through the primary (secondary) outcome. Performing such mediation analysis ignoring the complex study design could lead to seriously biased estimates since the primary case-control sample was a biased sample. To address this, we propose to account for the sampling process with the ratio-of-mediator-probability weighting method and established two new weighting estimators. The proposed estimators work well with marginal structural models in estimating controlled direct effects, natural direct effects and natural indirect effects under complex study designs and allow for the existence of exposure-mediator interaction in relation to the outcome. We compare the proposed methods with existing and extended versions of existing methods and find that the proposed methods significantly reduce the computational resources when estimating standard errors compared to existing methods. We also illustrate the implementation of the proposed weighting estimators using real-world data. Finally, we explore the minimum detectable effect sizes and the optimal case-to-control ratio for proposed weighting estimators for case-control studies with secondary (tertiary) outcomes.

Copyright by
ZICHUN CAO
2023

For my family and friends

ACKNOWLEDGMENTS

I would like to express gratitude to my dissertation committee chairs, Dr. Zhehui Luo and Dr. Honglei Chen, for their guidance and inspiring comments for both scientific knowledge and life experience.

I would like to thank my other committee members: Dr. Chenxi Li and Dr. Yuehua Cui for their kindness, encouragement and the valuable input they give.

I would like to thank the faculty and staff in my department.

I would like to thank all the friends and coworkers I've met.

Lastly, I would like to thank my family, for loving and supporting me all the way along my PhD journey.

TABLE OF CONTENTS

CHAPTER 1	Mediation Analysis	1
1.1	Introduction	1
1.2	Causal Directed Acyclic Graphs	2
1.3	Neyman-Rubin Causal Model	4
1.4	Fundamental Assumptions	5
1.5	Casual Mediation Estimands	6
1.6	Estimation Strategies for Mediation Analysis	8
CHAPTER 2	Mediation Analyses for Case-Control Studies With Secondary Outcomes	16
2.1	Background	16
2.2	Settings	17
2.3	Existing Methods for the Scenario	19
2.4	Proposed Weighting Method (Weighting Estimator I)	23
2.5	Simulation	28
2.6	Simulation Results	32
2.7	Application: the Olfaction Sub-study of the Sister Study Cohort	37
2.8	Conclusion	40
CHAPTER 3	Mediation Analyses for Case-Control Studies With Secondary and Tertiary Outcomes	41
3.1	Background	41
3.2	Settings	42
3.3	Existing Methods for the Scenario	44
3.4	Proposed Weighting Method	49
3.5	Simulation	52
3.6	Simulation Results	54
3.7	Application: the Olfaction Sub-study of the Sister Study Cohort	60
3.8	Conclusion	62
CHAPTER 4	Monte Carlo Based Statistical Power Analysis for Mediation Analysis Using Case-Control Studies With Secondary and Tertiary Outcomes	64
4.1	Background	64
4.2	Settings	66
4.3	Estimation Method	72
4.4	Simulation Studies	74
4.5	Results	76
4.6	Conclusion	80
CHAPTER 5	Conclusions	81
BIBLIOGRAPHY		84

APPENDIX A PROOFS OF THEOREMS	90
APPENDIX B DATA GENERATING PROCESS FOR CHAPTER 4 . . .	104

CHAPTER 1 Mediation Analysis

1.1 Introduction

The research on the causal mediation analysis, including both methodology and application, has exploded in the past several decades (Ten Have & Joffe 2012, Baron & Kenny 1986, VanderWeele 2009, Huber 2014, Hong 2010, Tchetgen Tchetgen 2013). However, limited work has been done for mediation analysis with complex study designs. Most existing statistical methods have been developed assuming the exposure, mediator, and outcome are fully observed by study design, e.g., in a prospective cohort study with no missing data. Such designs are usually expensive and time-consuming to conduct. Compared with cohort studies, a case-control study design is less costly and has been widely used for studying associations between exposures and rare diseases/outcomes. Furthermore, followed by a case-control sampling, researchers may be interested in secondary analyses with additionally collected outcomes within case-control sample only and study the association between exposure with the additional outcomes (Schifano 2019). In these studies, the outcome for case-control sampling is usually called the primary outcome and the additionally collected variables are called secondary outcomes. For example, Weedon et al. studied genetic risk factors for adult and childhood height (secondary outcome) with a case-control sample drawn based on type 2 diabetes status (primary outcome) (Weedon et al. 2007). Weuve et al. studied the relationship between lead exposure and cognitive function using data from two case-control samples in which the primary outcomes were hypertension and osteoporosis, respectively (Weuve et al. 2009). While those studies mainly aimed for association instead of causation, several attempts have also been made to perform mediation analyses using case-control studies (VanderWeele & Vansteelandt 2010, VanderWeele & Tchetgen Tchetgen 2016, Kim & Kaplan 2014, Wang et al. 2019, Satten et al. 2022).

Treating the primary outcome (case-control status) as the mediator and the secondary outcome as the outcome of interest, Huber and Solovyeva proposed a weighting method to estimate the mediating effect of an exposure on the secondary outcome through

the primary outcome (Huber & Solovyeva 2020, Huber 2014). However, one big drawback of Huber and Solovyeva’s method is that part of the weight involves modeling the treatment as a function of the mediator and covariates, which does not have immediate substantive interpretations (Hong 2015). In addition, their approach relies on bootstrap resampling method to obtain the standard error and 95% confidence interval (CI), which could be computationally expensive. To address these drawbacks, we extend the work by Hong (Hong 2010, Lange et al. 2012) and propose a new weighting method.

Moving one-step forward, a tertiary outcome can be subsequently collected in addition to the secondary outcome in case-control studies. For example, when the secondary outcome, i.e., the mediator, is relatively expensive and researchers use a cheaper surrogate (primary outcome) to do the case-control sampling to improve the efficiency of the study, a tertiary outcome can further be collected in extended follow-up visits within the case-control sample while the study interest is the mediating effect of exposure on the tertiary outcome through the secondary outcome. Currently there is no readily available statistical method for estimating mediating effects in such scenario. Thus, we first extend Huber and Solovyeva’s work (Huber & Solovyeva 2020) by proposing a new weighted estimator, which was an extension of Hong’s work (Hong 2010, Lange et al. 2012), for mediation analyses which account for the complex study design. Finally, we explore the power and the optimal case-control sampling ratio for mediation analyses under complex study designs specified above.

1.2 Causal Directed Acyclic Graphs

Causal directed acyclic graphs (DAGs) have been introduced to help communicate researchers’ understanding of the potential interplay among different variables in mediation analysis, (Pearl 2009, Hernan & Robins 2020, Greenland et al. 1999, Lipsky & Greenland 2022). Briefly, the causal DAG is consisted of a set of vertices (or nodes) and a set of directed edges (or links) and contains no directed cycles.

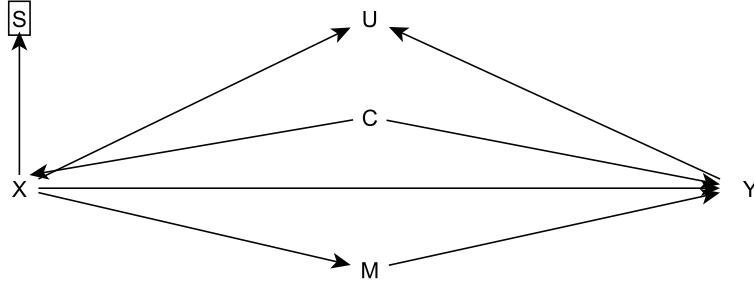


Figure 1.1: An Example of a Causal DAG

More specifically, the causal DAG visually lays out the assumptions of the data generating process (DGP), which helps researchers to minimize bias in causal inference analysis.

A causal DAG has the following characteristics (Lipsky & Greenland 2022):

a. Since causality implies the time ordering from cause to consequence, cycles (e.g., $X \rightarrow Y \rightarrow M \rightarrow X$) are not allowed.

b. The arrow reflects the causal relationship from cause to consequence. For example, $X \rightarrow Y$ in Figure 1.1 indicates that there is a direct causal effect of X on Y . However, the presence of an arrow does not guarantee that the relationship could be detected in the data since the effect can be small or negligible. And the lack of an arrow between any two variables indicates that there is no direct causal relationship between for any units.

c. A complete causal DAG should include all common causes for each pair of variables.

d. In order to reflect the study design, a causal DAG may also include a variable to indicate the (sample) selection process.

Taking Figure 1.1 as an example, if X is the exposure of interest and Y is the outcome of interest, the arrow from X to Y indicates the assumption of existence of direct causal effects of X on Y . There are several types of nodes (variables) in a causal DAG:

1) Confounder (common cause): variables that have causal effects on both the exposure and the outcome, i.e., C .

2) Mediator: intermediate variables on the causal path from the exposure to the outcome, i.e., M .

3) Collider (common consequence): the variable with two arrows pointing into it, i.e., U .

4) Selection indicator: researchers use a binary variable to indicate the sampling selection process, either selected (1) or unselected (0) (Didelez et al. 2010, Lipsky & Greenland 2022), i.e., S in Figure 1.1. Usually, a square outside the selection node suggests that the analysis is carried out among the selected sample. Practically, this can reflect the fact that in a case-control study we can only analyze data using the case-control sample, where $S = 1$.

1.3 Neyman-Rubin Causal Model

We discuss the causal effects under the potential (counterfactual) outcome framework first described by Neyman (Splawa-Neyman et al. 1923, 1990) and later developed by Rubin (Rubin 1974, 1978) and further extended to observational studies. Through out the entire dissertation, we adopt upper case to represent a random variable and lower case for a particular value of that random variable. For example, we use X to denote a binary exposure/treatment of interest, with the value of either 1 (exposed/treated) or 0 (unexposed/untreated). For an individual i , $Y_i^{x=1}$ denotes the potential outcome if he/she is exposed/treated, and $Y_i^{x=0}$ denotes the potential outcome if he/she is unexposed/untreated. For simplicity, we will use the term “potential outcome” for Y^x and “treatment” for X through the dissertation. Hence, the individual causal effect of treatment of subject i on the outcome can be defined by the difference between two potential outcomes (Rubin 1974, ?):

$$\triangle_i = Y_i^{x=1} - Y_i^{x=0} \quad (1.1)$$

Since we can only observe one of above two potential outcomes in the real world, individual causal effects cannot be identified, i.e., cannot be expressed as a function of the observed data (Hernan & Robins 2020). Now we will define the average causal/treatment

effects across a population. Taking a binary outcome Y (0/1) as an example, the marginal average treatment effects (ATE) are usually defined with three different scales (Hernan & Robins 2020):

Risk difference: $ATE_{RD} = P(Y^{x=1} = 1) - P(Y^{x=0} = 1)$;

Risk ratio: $ATE_{RR} = \frac{P(Y^{x=1}=1)}{P(Y^{x=0}=1)}$;

Odds ratio: $ATE_{OR} = \frac{P(Y^{x=1}=1)/P(Y^{x=1}=0)}{P(Y^{x=0}=1)/P(Y^{x=0}=0)}$.

A more general notation for marginal average treatment effects in terms of risk difference can be written in $ATE = E[Y^{x=1}] - E[Y^{x=0}]$ (Hong 2015, Hernan & Robins 2020). Some researchers may be interested in the conditional average treatment effects, e.g., $ATE = E[Y^{x=1}|\mathbf{C}] - E[Y^{x=0}|\mathbf{C}]$ where $\mathbf{C}$ is the conditioning set (VanderWeele 2015). The choice of marginal or conditional effects in reality depends on the study goal and purpose. In this dissertation work, we focus on the marginal effects for more general purposes. Even though the individual causal effects are not identifiable, the average causal/treatment effects in the population can sometimes be identified using observed data. To identify and estimate the average treatment effects, several basic assumptions are required and we will briefly introduce them in next section.

1.4 Fundamental Assumptions

The stable unit treatment value assumption (SUTVA) was first proposed and repeatedly emphasized by Rubin (Rubin 1980, 1986).

SUTVA (Stable Unit Treatment Value Assumption) (Rubin 1980, 1986, Imbens & Rubin 2015): The potential outcomes for any unit do not vary with the treatments assigned to other units (no interference), and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes (no hidden variations of treatments).

Consistency (Hernan & Robins 2020): For each subject, one of the potential outcomes is actually factual - the one corresponding to the treatment that the individual actually received. Formally, $Y_i = X_i Y_i^{x=1} + (1 - X_i) Y_i^{x=0}$ for a binary treatment X .

(Conditional) Exchangeability (Hernan & Robins 2020): The potential outcome is independent of the exposure (conditioning on covariates), i.e., $Y^x \perp\!\!\!\perp X|\mathbf{C}$.

Additional assumptions are required to identify target causal effects and will be discussed in later chapters.

For chapter 2 and 3, this dissertation work will mainly discuss the mediation analysis followed by the five steps below:

1. Define the population potential outcomes and the target causal/mediating effects of interest.
2. Identify the target causal effects using observable data under a set of assumptions. A population causal effect is identifiable only if the potential outcomes can be equated with some observable data even if such data were not actually available for all individuals in the population (Hong 2015).
3. Estimate the target causal effects from observed data.
4. Examine the performance of the estimation in simulation.
5. Implementation of the statistical methods into the real-world data.

The first three tasks are common for general causal inference analysis (Heckman & Vytlačil 2007) while the next two are for the evaluation of the estimation of the target causal effects (Hong 2015).

1.5 Casual Mediation Estimands

Unlike the average treatment effects, the mediation analysis studies the effect of an exposure on an outcome that operates through one (or multiple) intermediate process(es), referred as the indirect (or mediating effect). At the same time, we may also be interested in the effect of the treatment on an outcome that does not operate through specific intermediate process(es), referred as the direct (or unmediated) effect. Formally, let X be the exposure, M be the intermediate process, which can be a single mediator or a vector, and Y be the outcome of interest. Different types of causal contrasts have been proposed for mediation methodology such as controlled direct effects (CDE) (Pearl 2001), natural direct and indirect

effects (NDE/NIE) (Robins & Greenland 1992, Pearl 2001), and interventional (in)direct effects (VanderWeele et al. 2014), etc. In this dissertation, we mainly focus on CDE, NDE and NIE since they are the most commonly seen quantities in mediation research. Additionally, NDE/NIE decompose the total causal effect and provide insight about the underlying mechanisms that motivate researchers to conduct mediation analysis.

1.5.1 Controlled Direct Effects (CDE)

Holland was the first person to use Neyman-Rubin causal model to mediation problems (Holland 1988). He defined two causal effects of interest: $Y_i^{xm} - Y_i^{x^*m}$ and $Y_i^{xm} - Y_i^{xm^*}$. The first quantity measures the effect of a manipulable treatment while the second measures the effect of a manipulable mediator. Several years later, the first quantity was defined formally as the CDE (Pearl 2001, Robins & Greenland 1992). The individual CDE $Y_i^{xm} - Y_i^{x^*m}$ stands for the causal effect of X on Y for subject i when the mediator M is fixed at value m .

1.5.2 Natural Effects (NE)

Instead of fixing the value of the mediator at m , natural effects are the effect measures when one sets the value of the mediator by the random variable M^x . Pearl pointed out that the CDE is prescriptive since the mediator value is strictly under the experimenter's control and is not affected by the treatment assignment. However, the descriptive natural (in)direct effects (NDE/NIE) maintain the natural relationship between the treatment and the mediator (Pearl 2001, Hong 2015, Robins & Greenland 1992). To formally define the NDE/NIE for each subject, we use the following notations (Pearl 2001):

- (Pure) Natural direct effects: $\text{NDE}_i = Y_i^{xM_i^{x^*}} - Y_i^{x^*M_i^{x^*}}$, which represents the effect of X on Y when change the treatment value from x^* to x while setting the mediator to its natural value as if the treatment is set to the baseline value x^* .
- (Total) Natural indirect effects: $\text{NIE}_i = Y_i^{xM_i^x} - Y_i^{xM_i^{x^*}}$, the difference between the outcome Y of a treated subject i with his mediator set to the value as he would be treated ($Y_i^{xM_i^x}$), and the outcome Y of the treated subject i with his mediator set to the value as he would not be treated ($Y_i^{xM_i^{x^*}}$).

Similarly, we have another pair of natural effects:

- (Total) Natural direct effects: $\text{NDE}_i = Y_i^{xM_i^x} - Y_i^{x^*M_i^x}$.
- (Pure) Natural indirect effects: $\text{NIE}_i = Y_i^{x^*M_i^x} - Y_i^{x^*M_i^{x^*}}$.

The sum of (pure) NDE and (total) NIE, or the sum of (total) NDE and (pure) NIE, is the total effect of the treatment (X) on the outcome (Y): pure NDE + total NIE = total NDE + pure NIE = $Y^{xM^x} - Y^{x^*M^{x^*}} = Y^x - Y^{x^*}$. However, in general, the total effect cannot be divided in terms of CDEs and indirect effects. From now on, we refer to (pure) NDE simply as NDE and (total) NIE as NIE since they are more commonly used.

1.6 Estimation Strategies for Mediation Analysis

1.6.1 Path Analysis

The earliest mediation analysis is path analysis (Wright 1920, 1934), which used several paths to determine the piebald pattern of guinea-pigs. The path analysis approach was widely applied in social science research for mediation analysis (Alwin & Hauser 1975, Duncan 1966, Finney 1972).

1.6.2 Structural Equation Modeling (SEM)

In 1986, Baron and Kenny proposed the well-known structural equation modeling (SEM) approach to analyze direct and indirect effects (Baron & Kenny 1986).

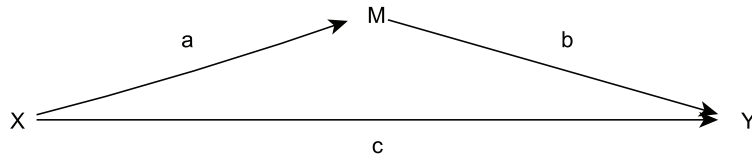


Figure 1.2: Causal DAG for SEM Approach

As shown in Figure 1.2, there are two causal paths from X to Y . One is $X \rightarrow Y$ (direct effect of X on Y) and the other one is $X \rightarrow M \rightarrow Y$ (indirect of X on Y through M). This approach used two linear regression models, one is for the mediator (M) and the

other for the outcome (Y):

$$M = \gamma_0 + aX + \varepsilon_M \quad (1.2)$$

$$Y = \beta_0 + bM + cX + \varepsilon_Y \quad (1.3)$$

Here, γ_0 and ε_M are the intercept and the error term in the mediator model while β_0 and ε_Y are the intercept and the error term for the outcome model. Researchers treat c from regression model 1.3 as the direct effect and the product ab from the regression models 1.2 and 1.3 as the indirect effect (Baron & Kenny 1986, Judd & Kenny 1981). Additionally, we can build a third regression model:

$$Y = \beta'_0 + c'X + \varepsilon'_Y \quad (1.4)$$

If treat $c' = ab + c$ from regression model 1.4 as the total effect of X on Y , we may then calculate the indirect effect as $c' - c$ instead of ab shown above. To determine if there exists an indirect effect, the Sobel test is widely used (Sobel 1982).

However, this approach has several disadvantages:

- The functional forms of the specified regression models have to be correct. For example, if M and Y are not linearly associated or the X and M do not additively influence Y , the direct and indirect effects calculated using above models could be incorrect.
- The total effect cannot be decomposed using controlled direct and indirect effects.
- Potential interaction between treatment (X) and mediator (M) may bias the direct and indirect effects of X on Y . Attempts have been made to use modified regression approaches to incorporate the treatment-mediator interaction in the outcome regression model (Preacher et al. 2007), but their proposed alternative conditional indirect effect does not correspond to the average natural indirect effect defined in section 1.5.2 without additional assumptions. Later on more researchers proposed different approaches to estimate the NDE with treatment-mediator interaction in the outcome regression

which requires much more complex and intensive calculations (Petersen et al. 2006, VanderWeele & Vansteelandt 2010, Valeri & Vanderweele 2013).

- When logistic regression models or probit models are used for binary outcomes, the product of coefficients approach (ab) and the difference of coefficients approach typically give different estimates of NIE defined in section 1.5.2, and neither is unbiased (MacKinnon 2007, Mackinnon & Dwyer 1993, MacKinnon et al. 2007).

Therefore, defining of NDE/NIE in terms of potential outcomes is more flexible and preferable than using the SEM.

1.6.3 Marginal Structural Models (MSM)

Marginal structural models can be used to estimate both controlled and natural effects, and allow for treatment-mediator interaction (VanderWeele 2009, Lange et al. 2012, Robins et al. 2000). To estimate CDE, we can use the following MSM (VanderWeele 2009):

$$E[Y^{xm}] = \alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3 xm \quad (1.5)$$

It can be easily shown that when change treatment value from x^* to x and fix the mediator value at m , and the MSM is correctly specified, the CDE in risk difference scale is $E[Y^{xm} - Y^{x^*m}] = \alpha_1(x - x^*) + \alpha_3(x - x^*)m$. To estimate such effect, researchers can use the inverse probability of treatment weighting procedure (VanderWeele 2009, Robins et al. 2000), with the first part of the weight as $\frac{P(X=x_i)}{P(X=x_i|\mathbf{C}=\mathbf{c}_i)}$ and the second part as $\frac{P(M=m_i|X=x_i)}{P(M=m_i|X=x_i, \mathbf{C}=\mathbf{c}_i, \mathbf{W}=\mathbf{w}_i)}$ where $\mathbf{C}$ is the treatment-outcome confounder set and $\mathbf{W}$ is the mediator-outcome confounder set, under the exchangeability assumptions. One advantage of this approach over the traditional regression approach is that it allows $\mathbf{W}$ to be a consequence of the treatment while the traditional approach doesn't (VanderWeele 2009).

To estimate NDE and NIE, we can use a similar MSM (Lange et al. 2012):

$$E[Y^{xM^{x^*}}] = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (1.6)$$

It can be easily shown that when changing treatment value from x^* to x , and the MSM is correctly specified, the NDE in risk difference scale is $\text{NDE}_{\text{RD}} = E[Y^{xM^{x^*}} - Y^{x^*M^{x^*}}] = (c_0 + c_1x + c_2x^* + c_3x \cdot x^*) - (c_0 + c_1x^* + c_2x^* + c_3x^* \cdot x^*) = c_1(x - x^*) + c_3x^*(x - x^*)$. and the NIE in risk difference scale is

$$\text{NIE}_{\text{RD}} = E[Y^{xM^x} - Y^{xM^{x^*}}] = (c_0 + c_1x + c_2x + c_3x \cdot x) - (c_0 + c_1x + c_2x^* + c_3x \cdot x^*) = c_2(x - x^*) + c_3x(x - x^*).$$

The major difference between model 1.5 and model 1.6 is how to handle the value of the mediator. The former assumes that the mediator M is manipulable so it specifies $E[Y^{xm}]$ as a function of x and m . However, for NDE/NIE, the mediator is usually assumed as not manipulable and $E[Y^{xM^{x^*}}]$ is specified with a function of the actually received treatment (x) and the counterfactual treatment (x^*).

1.6.4 Conditional Marginal Structural Models

Except the approach mentioned in section 1.6.3 and to address the potential treatment-mediator interaction issue in estimating NDE and NIE, VanderWeele proposed the conditional marginal structural models (VanderWeele 2009, 2015). He proposed to use two conditional marginal structural models to estimate conditional NDE and NIE:

$$E[Y^{xm} | \mathbf{C} = \mathbf{c}] = \theta_0 + \theta_1x + \theta_2m + \theta_3xm + \theta'_4\mathbf{c} \quad (1.7)$$

$$E[M^x | \mathbf{C} = \mathbf{c}] = \gamma_0 + \gamma_1x + \gamma'_2\mathbf{c} \quad (1.8)$$

where $\mathbf{C}$ is the conditioning set, e.g., treatment-outcome confounders.

With some calculation, the conditional NDE is $E[Y^{xM^{x^*}} - Y^{x^*M^{x^*}} | \mathbf{C} = \mathbf{c}] = (\theta_1 + \theta_3\gamma_0 + \theta_3\gamma_1x^* + \theta_3\gamma'_2\mathbf{c})(x - x^*)$ and the conditional NIE is $E[Y^{x^*M^x} - Y^{x^*M^{x^*}} | \mathbf{C} = \mathbf{c}] = \theta_2\gamma_1(x - x^*) + \theta_3\gamma_1x(x - x^*)$. As pointed out by VanderWeele, the conditional NIE in risk difference scale estimated using models 1.7 and 1.8 does not depend on covariate set $\mathbf{C}$ so the average marginal NIE would be same as the conditional NIE (VanderWeele 2009).

1.6.5 Resampling Approach

Researchers have also proposed Monte Carlo simulation algorithms for generating the distribution of potential outcomes $E[Y^{xM^{x*}}]$ and $E[Y^{x^*M^x}]$ to estimate NDE and NIE (Imai, Keele & Tingley 2010, Imai, Keele & Yamamoto 2010). However, these methods are computationally intensive. Thus, from the practical perspective, this method may not be easily implemented.

1.6.6 Imputation-based Approach

Like resampling approach, researchers also try to generate the unobservable potential outcomes using imputation-based methods (Vansteelandt et al. 2012). This approach is sometime attractive if unstable performance (due to extreme values in the weight) is concerned when using weighting-based estimators. However, since weighting-based methods are more well established and easy to implement in practical, we will focus on existing weighting-based methods and the extension of them.

1.6.7 Other Weighting Methods

Other weighting methods have been proposed based on the rationale of identifying population potential outcomes using observed data (Huber 2014, Huber & Solovyeva 2020, Tchetgen Tchetgen 2013, Hong 2010, Albert 2012, Nguyen et al. 2023). Interestingly, all weighting methods are mathematically equivalent or closely related (Hong 2010, Hong et al. 2015, Hong & Nomi 2012, Tchetgen & Shpitser 2012) except the Albert's method, which has advantages if the mediator is non-binary. For example, Hong's ratio-of-mediator-probability-weighting (RMPW) approach is valid under the following assumptions (Hong 2010):

- 1) $0 < P(X = x|C) < 1$
- 2) $0 < P(M^x = m|X, C) < 1$
- 3) $Y^{xm} \perp\!\!\!\perp M^x|X = x, C$
- 4) $M^x \perp\!\!\!\perp X|C$
- 5) $Y^{xm} \perp\!\!\!\perp M^{x*}|X = x, C$
- 6) $(Y^{xM^{x*}}, Y^{x^*M^{x*}}, Y^{x^*M^x}, Y^{xM^x}) \perp\!\!\!\perp X|C$

The population potential outcome $E[Y^{xM^{x^*}}]$ can be estimated using observed data in the absence of post-treatment covariates (Hong 2010):

$E[Y^{xM^{x^*}}] = E[wY|X = x]$ and $w = \frac{q^{(x^*)}(M^{x^*}=m|X=x^*,C)}{q^{(x)}(M^x=m|X=x,C)} \cdot \frac{P(X=x)}{P(X=x|C)}$, where $q^{(\cdot)}(\cdot)$ is the density function for the mediator. This method is quite flexible and can be easily implemented in practice. Lange et al. further extended Hong’s work with the combination of MSM to make it easier to obtain the robust standard error and 95% CI using standard statistical softwares (Lange et al. 2012).

1.6.8 Two-Phase and Multi-Phase Sampling

Unlike randomized control trials, the observational study does not have randomization process which makes it easily suffer from potential confounding issue. Many attempts have been made for performing causal mediation analyses in commonly seen observational study types such as cohort, case-control studies (Hong 2010, VanderWeele & Vansteelandt 2010, VanderWeele & Tchetgen Tchetgen 2016). However, limited amount of publications focus on mediation analyses with more complex study designs, e.g., two-phase sampling. In studies with two-phase sampling (a.k.a. double-sampling or two-stage sampling), some variables are measured on all individuals in a sample drawn from a target population, e.g., a cohort. Next, based on the values of one or multiple measured variables, a subsample is further drawn and the values of additional variables (secondary or tertiary outcomes) are further obtained only within the subsample (Neyman 1938, White 1982, Gustavo Amorim et al. 2017). Such designs are useful when the secondary and/or tertiary outcomes or covariates are expensive or difficult to measure, e.g., genetic data. With additional steps of subsampling, two-phase designs can be further extended to three-phase, or more general, a multi-phase sampling design (Whittemore & Halpern 1997).

In this dissertation, we are inspired by a real-world research question raised from the National Institute of Environmental Health Sciences (NIEHS) Sister Study cohort. While the details of the NIEHS Sister Study data will be introduced in detail later, the study design we focus on can be considered as a special case of the multi-phase sampling. Briefly, we consider

a case-control sample drawn from a pre-specified sampling pool derived from the Sister cohort. Additional variables (secondary and tertiary outcomes) are further collected only within the case-control sample. Therefore, we extend Hong’s and Lange’s work and proposed two sets of weighting-based estimators (weighting estimator I and II) for the mediation analysis under case-control study samples with secondary (and tertiary) outcome(s).

1.6.9 Aims and Organization

Aim 1: Mediation analyses of the secondary outcome with the primary case-control status as the mediator in case-control studies.

Specific Aim 1.1: Define and identify the causal estimand nonparametrically.

Specific Aim 1.2: Illustrate the necessity of considering the study design when performing the mediation analysis.

Specific Aim 1.3: Propose weighted estimators and compare the performance of proposed estimators with an existing estimator in simulation analysis.

Specific Aim 1.4: Implement the existing and proposed methods in real-world data.

Aim 2: Mediation analyses of the tertiary outcome with the secondary outcome as the mediator in case-control studies.

Specific Aim 2.1: Define and identify the causal estimand nonparametrically.

Specific Aim 2.2: Illustrate the necessity of considering the study design when performing the mediation analysis.

Specific Aim 2.3: Propose weighted estimators and compare the performance of proposed estimators with a modified version of an existing estimator in simulation analysis.

Specific Aim 2.4: Implement the existing and proposed methods in real-world data.

Aim 3: Monte Carlo simulation based power analysis and optimal case-control sampling ratio exploration for mediation analyses in case-control studies with secondary and/or tertiary outcomes.

Specific Aim 3.1: Perform power analysis for scenarios assumed in **Aim 1** under different sampling strategies with different effect sizes.

Specific Aim 3.2: Perform power analysis for scenarios assumed in **Aim 2** under different sampling strategies with different effect sizes.

CHAPTER 2 Mediation Analyses for Case-Control Studies With Secondary Outcomes

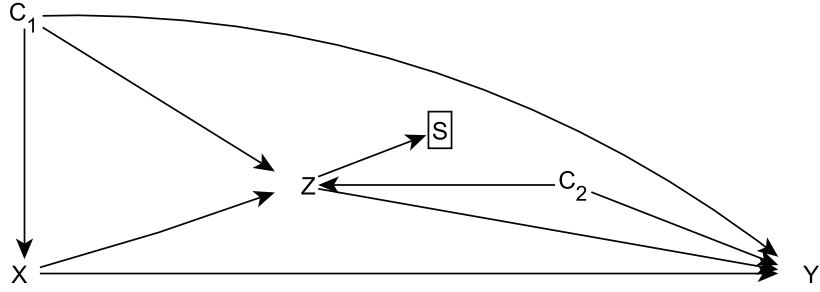
2.1 Background

In a recent scoping review (Rijnhart et al. 2021), researchers found that mediation analyses have been implemented in observational studies such as cross-sectional, case-control and both prospective and retrospective cohort studies. However, little has been done for mediation analysis in a case-control study with subsequently collected (secondary) outcomes. This may happen after a case-control study has been conducted but with additional interest of a secondary outcome and direct or indirect effects of exposure on the secondary outcome through the primary outcome. For example, Han et al. examined the different risk factors in relation to the skin cancer using a nested case-control sample from the Nurses' Health Study (Han et al. 2006). Later on, researchers might be further interested in how different risk factors are related to the death (a secondary outcome) through the skin cancer (the primary outcome) using the same case-control sample. Here, the the mortality (secondary outcome) might only be available within the case-control sample instead of the whole Nurses' Health Study cohort. However, performing such mediation analysis using existing methods could lead to seriously biased estimation since that the case-control sample was not designed for the mediation analysis purpose and potential bias could be an issue. Thus, appropriate methods which can account for the bias due to sampling should be considered. In this chapter, we will first describe the settings and notations, followed by applying a marginal structural model method to show the consequence of ignoring the complex study design. Then we review an existing method that is suitable for this scenario but with limitations. Next, we propose a new estimator (weighting estimator I) and compare its performance with the existing method in simulation analysis. Finally, we implement both the existing and the proposed method using the NIEHS Sister study data.

2.2 Settings

2.2.1 Notations and Causal DAG

Our study setting is the mediation analysis of a case-control sample with the case-control status as the mediator and a secondary outcome as the outcome of interest. Formally, let C_1 be the exposure-mediator, exposure-outcome, and mediator-outcome confounder, e.g., the age at study baseline. Let C_2 be the mediator-outcome confounder, e.g., comorbidity disease. Let Z , i.e., the variable used for selecting the case-control sample (0-control, 1-case), be the mediator of interest. We assume that the sample is selected from an existing cohort study, like in many epidemiological studies. Let S be an indicator of case-control sample selection status, where “ $S = 1$ ” means the unit is selected and “ $S = 0$ ” means the unit is unselected (Didelez et al. 2010). Let Y be the secondary outcome (the outcome of interest in the mediation analysis). Figure 2.1 shows the causal DAG for the data generating process of (C_1, C_2, X, Z, S, Y) , where the square outside S indicates that the analysis is performed using observed data, i.e., the selected case-control sample.



C_1 : confounder between exposure, mediator and outcome, e.g. participants' age at study baseline

C_2 : confounder between mediator and outcome, e.g. participant comorbidity

X : exposure/treatment

Z : case-control status (mediator)

S : case-control sample selection node

Y : (secondary) outcome

Figure 2.1: Causal DAG for Mediation Analysis

For simplicity, assume C_2 , X , Z and Y are all binary variables. In practice, our proposed approach can be extended to non-binary variables as well. If the case-control sample is selected from a cohort (source population), then (C_1, C_2, X, Z, S) are fully observed for all units whereas Y is only observed when $S = 1$, i.e., among units that are selected into the case-control sample (Cao et al. 2022). In our application of the Sister Study data, cases are 100% selected and controls are selected using simple random sampling (SRS) method with a pre-specified sampling rate (Cao et al. 2022). Our proposed approach may also be extended (with modifications) to other sampling method such as stratified sampling or proportion to size sampling as long as the sampling process is known by design. The aim here is to estimate the causal estimands (e.g., mediating effects) using the selected case-control sample. With such study design, X is fully observed from the cohort. We adopt Neyman-Rubin causal model (Rubin 1978, Splawa-Neyman et al. 1923) and invoke the SUTVA for simplicity (Rubin

1986). Let Z^x be the counterfactual mediator and Y^{xZ^x} be the counterfactual outcome when the exposure X is set to the value x . The goal is to study mediating effects from X to Y , through the mediator Z .

2.2.2 Controlled Direct Effects (CDE)

For binary Y and Z , the population (cohort) CDE can be defined in three different scales (Robins & Greenland 1992, Pearl 2001, VanderWeele 2015):

Risk difference: $\text{CDE}_{\text{RD}} = E[Y^{1z} - Y^{0z}]$, for $z \in \{0, 1\}$.

Relative risk: $\text{CDE}_{\text{RR}} = \frac{P(Y^{1z}=1)}{P(Y^{0z}=1)}$, for $z \in \{0, 1\}$.

Odds ratio: $\text{CDE}_{\text{OR}} = \frac{P(Y^{1z}=1)/P(Y^{1z}=0)}{P(Y^{0z}=1)/P(Y^{0z}=0)}$, for $z \in \{0, 1\}$.

As pointed out by Pearl (Pearl 2001), experimentally controlling the mediator at a fixed value reflects a prescriptive conceptualization that has limited practical meaning since it is hardly achievable in epidemiological studies. Thus, the use and interpretation of CDE needs to be careful.

2.2.3 Natural Effects (NEs)

Similar to CDE, for binary Y and Z , the population natural direct effects (NDE) and natural indirect effects (NIE) can also be defined in three different scales (Robins & Greenland 1992, Pearl 2001, VanderWeele 2015):

Risk difference: $\text{NDE}_{\text{RD}} = E[Y^{1Z^0} - Y^{0Z^0}]$ and $\text{NIE}_{\text{RD}} = E[Y^{1Z^1} - Y^{1Z^0}]$.

Relative risk: $\text{NDE}_{\text{RR}} = \frac{P(Y^{1Z^0}=1)}{P(Y^{0Z^0}=1)}$ and $\text{NIE}_{\text{RR}} = \frac{P(Y^{1Z^1}=1)}{P(Y^{1Z^0}=1)}$.

Odds ratio: $\text{NDE}_{\text{OR}} = \frac{P(Y^{1Z^0}=1)/P(Y^{1Z^0}=0)}{P(Y^{0Z^0}=1)/P(Y^{0Z^0}=0)}$ and $\text{NIE}_{\text{OR}} = \frac{P(Y^{1Z^1}=1)/P(Y^{1Z^1}=0)}{P(Y^{1Z^0}=1)/P(Y^{1Z^0}=0)}$.

2.3 Existing Methods for the Scenario

2.3.1 Weighting Method Ignoring the Study Design

There is a number of existing weighting methods which can estimate the mediating effects including both CDE and NEs (Coffman et al. 2023, Hong 2010, Huber 2014, Nguyen et al. 2023, Albert 2012, Tchetgen Tchetgen 2013, Lange et al. 2012, VanderWeele 2009). However, none of these takes the complex study design into consideration, except one recent paper by Huber and Solovyeva (Huber & Solovyeva 2020), that extended Huber's weighting method

(Huber 2014). Before looking into their method, I would like to point out the necessity for considering the study design while performing mediation analysis. Particularly in this section, I will use one of the most well-known weighting methods, proposed by VanderWeele (VanderWeele 2009), to illustrate the consequence of ignoring the secondary data collection process in terms of biased estimates in mediation analysis. Also, I will discuss the possible remedy for such scenario. The notation is consistent with the definitions in section 2.2.1.

VanderWeele proposed three marginal structural models (MSM) to estimate the causal mediating effects, for example in the continuous outcome and continuous mediator setting:

$$E[Y^{xz}] = g(x, z) = \alpha_0 + \alpha_1 x + \alpha_2 z + \alpha_3 xz \quad (2.1)$$

$$E[Z^x | \mathbf{C} = \mathbf{c}] = h(x, c) = \gamma_0 + \gamma_1 x + \gamma_2' \mathbf{c} \quad (2.2)$$

$$E[Y^{xz} | \mathbf{C} = \mathbf{c}] = g(x, z, c) = \theta_0 + \theta_1 x + \theta_2 z + \theta_3 xz + \theta_4' \mathbf{c} \quad (2.3)$$

The MSM 2.1 is used to estimate the CDE with assigning each individual with a weight of $w_x * w_z$, where $w_x = \frac{P(X=x_i)}{P(X=x_i | \mathbf{C}=\mathbf{c}_i)}$ and $w_z = \frac{P(Z=z_i | X=x_i)}{P(Z=z_i | X=x_i, \mathbf{C}=\mathbf{c}_i)}$. The MSM 2.2 and 2.3 are used to estimate the conditional NEs. For 2.2, the weight $w_2 = \frac{1}{P(X=x_i | \mathbf{C}=\mathbf{c}_i)}$ is assigned for each individual, and for 2.3, the same weight $w_x * w_z$ is used. One difference between 2.3 and 2.1 is that model 2.3 also conditions on the covariate set $\mathbf{C}$, i.e., (C_1, C_2) in our case. Thus, estimates from MSM 2.2 and 2.3 are called conditional NEs. To get marginal ones, we need to further take expectation regarding the covariate set $\mathbf{C}$.

VanderWeele further gave the explicit form of both CDE and marginal NEs:

$$\text{CDE}_{\text{RD}} = E[Y^{1z} - Y^{0z}] = \alpha_1 + \alpha_3 z \quad (2.4)$$

$$\text{pure NDE}_{\text{RD}} = E[Y^{1Z^0} - Y^{0Z^0}] = \theta_1 + \theta_3 \gamma_0 + \theta_3 \gamma_2' E[C] \quad (2.5)$$

$$\text{total NIE}_{\text{RD}} = E[Y^{1Z^1} - Y^{1Z^0}] = (\theta_2 + \theta_3) \gamma_1 \quad (2.6)$$

Table 2.1: Mediating Effects When Ignoring Study Design

Effect (RD)	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
CDE ₀	36491	0.027	0.027	0.0258	0.9941	0.0256	94.0
CDE ₁		0.208	0.208	0.0193	0.9840	0.0190	95.5
NDE		0.038	0.142	0.0153	0.9733	0.1054	0.0
NIE		0.038	0.081	0.0079	0.9760	0.0434	0.0

^a This is the ratio between empirical and bootstrap standard error.

From Table 2.1 we can see that even though the CDE can be unbiasedly estimated, there is serious bias in estimating NEs. One possible reason could be that when fixing the exposure X and the mediator Z (case-control status), $Y^{xz} \perp\!\!\!\perp S$ is achieved. Thus, ignoring the sample selection process would still give unbiased estimates of Y^{xz} from MSM 2.1 due to the independency. However, for NEs, even though we can fix X at x (or x^*), Z^x (or Z^{x^*}) is considered as a random variable. Thus, Y^{Z^x} (or $Y^{Z^{x^*}}$) is not independent of sample selection S in such scenario, which will lead to biased estimates of NEs if ignoring the study design. In such case, either the weight for MSM 2.3 needs to be adjusted to incorporate the sample selection process, or a different approach can be used to estimate the NEs. Below we give a potential remedy for this.

NDE can still be unbiasedly estimated if CDE estimates are unbiased since NDE is weighted averages of CDE in a cohort study (Rudolph et al. 2019, Petersen et al. 2006). In the case-control sample, if each mean of counterfactual outcome can be estimated unbiasedly, then so can all NEs.

$$\begin{aligned}
E[Y^{xZ^x}] &= E_{\mathbf{c}}[E_{Y^{xZ^x}|\mathbf{C}}[Y^{xZ^x}|\mathbf{C}]] \\
&= E_{\mathbf{c}}[\int y \cdot P(Y^{xZ^x} = y|\mathbf{C})dy] \\
&= E_{\mathbf{c}}[\int \int y \cdot P(Y^{xZ^x} = y, Z^x = z|\mathbf{C})dzdy] \\
&= E_{\mathbf{c}}[\int \int y \cdot P(Y^{xz} = y|Z^x = z, \mathbf{C}) \cdot P(Z^x = z|\mathbf{C})dzdy] \\
&= E_{\mathbf{c}}[\sum_z (E[Y^{xz}|\mathbf{C}]) \cdot P(Z^x = z|\mathbf{C})]
\end{aligned}$$

Similarly, we have $E[Y^{xZ^{x^*}}] = E_{\mathbf{c}}[\sum_z (E[Y^{xz}|\mathbf{C}]) \cdot P(Z^{x^*} = z|\mathbf{C})]$. Then NIE can be estimated using the difference between above two quantities. Although this is a separate topic and not of our primary focus, future research is needed to investigate further. Next, we will present an existing method that can give unbiased estimation of mediating effects in such complex study designs.

2.3.2 Existing Methods on Estimation of Causal Effects

Using notations from section 2.2.1 and data from mediator-dependent sampling, Huber and Solovyeva (2020) proposed to first estimate the means of the counterfactual outcomes Y^{xz}, Y^{xZ^x} and $Y^{xZ^{x^*}}$ (Huber & Solovyeva 2020). Their method and estimator will be called the HS method and estimator henceforth. Specifically, assuming the causal structure in Figure 2.1, they showed

$$\begin{aligned}\hat{E}[Y^{xz}] &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot I\{Z_i = z\} \cdot S_i}{P(X_i = x|C_{1i}, C_{2i}) \cdot P(Z_i = z|X_i, C_{1i}, C_{2i}) \cdot P(S_i = 1|X_i, Z_i, C_{1i}, C_{2i})} \\ &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot I\{Z_i = z\} \cdot S_i}{P(X_i = x|C_{1i}) \cdot P(Z_i = z|X_i, C_{1i}, C_{2i}) \cdot P(S_i = 1|Z_i)}\end{aligned}\quad (2.7)$$

$$\begin{aligned}\hat{E}[Y^{xZ^x}] &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|C_{1i}, C_{2i}) \cdot P(S_i = 1|X_i, Z_i, C_{1i}, C_{2i})} \\ &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|C_{1i}) \cdot P(S_i = 1|Z_i)}\end{aligned}\quad (2.8)$$

$$\begin{aligned}\hat{E}[Y^{xZ^{x^*}}] &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|Z_i, C_{1i}, C_{2i}) \cdot P(S_i = 1|X_i, Z_i, C_{1i}, C_{2i})} \cdot \frac{P(X_i = x^*|Z_i, C_{1i}, C_{2i})}{P(X_i = x^*|C_{1i}, C_{2i})} \\ &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|Z_i, C_{1i}, C_{2i}) \cdot P(S_i = 1|Z_i)} \cdot \frac{P(X_i = x^*|Z_i, C_{1i}, C_{2i})}{P(X_i = x^*|C_{1i})}\end{aligned}\quad (2.9)$$

The HS estimators of CDE and NEs naturally follow. Instead of directly plug in equation 2.7-2.9, they proposed to use normalized versions of the sample analogs (Huber & Solovyeva 2020).

However, to estimate NEs, the quantities $P(X_i = x^*|Z_i, C_{1i}, C_{2i})$ and $P(X_i = x|Z_i, C_{1i}, C_{2i})$ are required to be estimated using observed data. While such estimation is

usually based on parametric models, modeling the treatment as a function of the subsequent emerged mediator is counterintuitive and does not have immediate substantive interpretations since the treatment causally precedes rather than succeeds the mediator (Hong et al. 2015). The HS method estimates mean of counterfactual outcomes first and then estimates the target causal estimands as a function of those potential outcomes, which makes the SE estimation difficult. It generally relies on the bootstrap resampling method to obtain the standard error (SE) or confidence interval (CI) (Huber & Solovyeva 2020). To overcome these disadvantages, we extended Hong’s work on ratio-of-mediator-probability weighting (RMPW) (Hong 2010), VanderWeele’s MSM (VanderWeele 2009), and Lange’s unified approach (Lange et al. 2012), to propose a new weighting method (weighting estimator I) to estimate both CDE and NEs for a secondary outcome in a case-control sample drawn from a source population, e.g., a cohort. We will compare the performance between the HS estimator and weighting estimator I using the simulations, followed by an application using real-world data.

2.4 Proposed Weighting Method (Weighting Estimator I)

2.4.1 Nonparametric Identification

The central goal of this chapter is to identify and estimate the population causal mediating effects using the case-control sample. As mentioned in previous sections, counterfactual outcome Y^{xz} and nested counterfactual outcomes (Y^{xZ^x} , $Y^{xZ^{x^*}}$) play key roles in estimating CDE, NDE and NIE. Thus, to identify the causal mediating effects, we need to connect $E[Y^{xz}]$ with $E[Y|X = x, Z = z, S = 1]$, and $E[Y^{xZ^x}]$ and $E[Y^{xZ^{x^*}}]$ with $E[Y|X = x, S = 1]$. Before showing how to nonparametrically identify the population mediation effects using case-control sample, we lay out assumptions which are required for the identification, based on the causal structure shown in Figure 2.1.

Assumption 1a. $Y^{xz} \perp\!\!\!\perp (X, Z) | \mathbf{C}$ where $\mathbf{C} = (C_1, C_2)$, i.e., no unobserved confounders between the treatment and the outcome, and no unobserved confounders between the mediator and the outcome.

Assumption 1b. $Y^{xz} \perp\!\!\!\perp Z|(X, \mathbf{C})$.

Assumption 1c. $Y^{xz} \perp\!\!\!\perp Z^{x*}|(X, \mathbf{C})$.

Assumption 2. Conditional independence based on the study design and background knowledge.

$$Y \perp\!\!\!\perp S|(Z, X, \mathbf{C}) \quad (2.10)$$

$$S \perp\!\!\!\perp X|Z \quad (2.11)$$

$$S \perp\!\!\!\perp (X, \mathbf{C})|Z \quad (2.12)$$

$$X \perp\!\!\!\perp C_2|C_1 \quad (2.13)$$

$$0 < P(Z = z|X, \mathbf{C}), 0 < P(X = x|C_1) \text{ and } 0 < P(S = 1|Z) \quad (2.14)$$

The conditional independence assumptions 2.10 and 2.13 are primarily design-based. For example, in a randomized controlled trial, instead of $Y \perp\!\!\!\perp S|(Z, X, \mathbf{C})$, we have $Y \perp\!\!\!\perp S$. For observational studies such as general case-control and matched case-control designs, we may have $Y \perp\!\!\!\perp S|Z$ or $Y \perp\!\!\!\perp S|(Z, \mathbf{C})$ by design. It will change the weight slightly but will not affect the identification generally. For more details, please refer to the proof of Theorem 1 in Appendix A. Assumption 1a, 2.10-2.12 are required for the identification of CDE. To identify NDE and NIE, except for Assumption 1b and Assumption 1c, three sets additional assumptions are further required:

Assumption 3a. $Y^{xZ^x} \perp\!\!\!\perp X|(Z^x, \mathbf{C})$.

Assumption 3b. $Y^{xZ^{x*}} \perp\!\!\!\perp X|(Z^{x*}, \mathbf{C})$.

Assumption 4. $(Z^x, Z^{x*}) \perp\!\!\!\perp X|\mathbf{C}$.

Assumption 5. $Y^{xz} \perp\!\!\!\perp Z^{x*}|(X, \mathbf{C})$.

Next, we introduce 2.4.1, to show that the mean of (nested) counterfactual outcomes can be written using weighted conditional density of the observed outcome Y .

Identification of mean of (nested) counterfactual outcomes using case-control sample under assumptions 1 to 5.

$$\begin{aligned}
E[Y^{xz}] &= E[w_1 Y | X = x, Z = z, S = 1] \text{ where } w_1 = \frac{P(z|x)P(X=x)}{P(z|x, c_1, c_2)P(x|c_1)}. \\
E[Y^{xZ^x}] &= E[w_2 Y | X = x, S = 1] \text{ where } w_2 = \frac{P(S=1|X=x)P(X=x)}{P(S=1|Z=z)P(X=x|c_1)}. \\
E[Y^{xZ^{x^*}}] &= E[w_3 Y | X = x, S = 1] \text{ where } w_3 = \frac{P(Z=z|X=x^*, c_1, c_2)}{P(Z=z|X=x, c_1, c_2)} \cdot \frac{P(S=1|X=x)P(X=x)}{P(S=1|Z=z) \cdot P(X=x|C_1)}.
\end{aligned}$$

2.4.2 Estimation of the Causal Estimand

Compared to the Theorem 1 from (Huber & Solovyeva 2020), 2.4.1 rewrites $P(X|Z, \mathbf{C})$ into $P(Z|X, \mathbf{C})$ which makes it easier to be interpreted as the mediator causally succeeds the exposure (Hong 2015). So far, we can still estimate CDE and NEs using the same way as HS approach does but it is computationally expensive since it relies on the bootstrap resampling method to obtain SE and 95% CI (Huber & Solovyeva 2020). Therefore, we adopt a MSM modeling-based approach which assumes particular relationships between the mean of a (nested) counterfactual outcome and the observed exposure and mediator whenever appropriate (Robins et al. 2000, VanderWeele 2009, Lange et al. 2012). Following MSM will be used for estimating CDE (2.15) and NE (2.16):

$$g(E[Y^{xz}]) = c_0 + c_1 x + c_2 z + c_3 x \cdot z \quad (2.15)$$

$$g(E[Y^{xZ^{x^*}}]) = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (2.16)$$

where $g(\cdot)$ is the link function.

Assuming the functional form is correctly specified in 2.15 and 2.16, the causal mediating effects in terms of risk difference, relative risk and odds ratio can be expressed as a function of the parameters estimated from following MSM (Robins et al. 2000):

For CDE:

$$\text{linear model: } P(Y^{xz} = 1) = c_0 + c_1 x + c_2 z + c_3 x \cdot z \quad (2.17)$$

$$\text{log-linear model: } \log[P(Y^{xz} = 1)] = c_0 + c_1 x + c_2 z + c_3 x \cdot z \quad (2.18)$$

$$\text{logistic model: } \text{logit}[P(Y^{xz} = 1)] = c_0 + c_1 x + c_2 z + c_3 x \cdot z \quad (2.19)$$

For NE:

$$\text{linear model: } P(Y^{xZ^{x^*}} = 1) = c_0 + c_1x + c_2x^* + c_3x \cdot x^* \quad (2.20)$$

$$\text{log-linear model: } \log[P(Y^{xZ^{x^*}} = 1)] = c_0 + c_1x + c_2x^* + c_3x \cdot x^* \quad (2.21)$$

$$\text{logistic model: } \text{logit}[P(Y^{xZ^{x^*}} = 1)] = c_0 + c_1x + c_2x^* + c_3x \cdot x^* \quad (2.22)$$

where w_1 is used for estimating CDE in 2.17, 2.18 and 2.19 while w_3 is used for estimating NE in 2.20, 2.21 and 2.22. The interaction term $c_3x \cdot z$ or $c_3x \cdot x^*$ are added into above models to reflect the potential exposure-mediator interaction in relation to the outcome, and if the exposure-covariate interactions are of interest, we can further include $x \cdot \mathbf{C}$ and $x^* \cdot \mathbf{C}$ to capture the direct and indirect effects modified by covariates (Lange et al. 2012). In our scenario, to estimate CDE, MSM 2.17-2.19 can be directly applied. However, since MSM 2.20-2.22 involves not only the observed exposure x but also the counterfactual value of the exposure, i.e., x^* , we adopt (Lange et al. 2012) approach and construct a new dataset by repeating each observation in the original dataset twice, including an additional variable X^* , which is equal to the original exposure for the first replication and equal to the opposite of the actual exposure for the second replication. After computing weights for each individual (and the “duplicates”), we fit MSM 2.20-2.22 to estimate NEs.

Similar to (Lange et al. 2012), when assuming a MSM for the nested counterfactual outcomes, our method can be equated with the generalized estimating equations (GEE) method. For example, the estimating equation corresponding to the MSM in 2.16 is given by:

$$U(c, \alpha) = \sum_{i=1}^n \sum_{x^*=0}^1 d(X_i, x^*) (Y_i - c_0 - c_1X_i - c_2a^* - c_3X_ia^*) \frac{P(Z=z|X=x^*, c_1, c_2)}{P(Z=z|X=x, c_1, c_2)} \cdot \frac{P(S=1|X=x)P(X=x)}{P(S=1|Z=z) \cdot P(X=x|C_1)} \text{ with } d(X_i, x^*) = (1, X_i, x^*, X_ix^*)^T.$$

For the estimators to have good performance (i.e. to be regular), we further multiply the weight (w_3) by $P(X_i)$ like the stabilized weight in inverse probability weighting estimation (Lange et al. 2012). As Lange et al. argued, let model $\mathcal{M}_1$ be the parametric model for

$P(Z|X, \mathbf{C})$ (with the unknown parameter denoted α) and $\mathcal{M}_2$ be the model defined by the restrictions of model $\mathcal{M}_1$ and the additional assumption that the generalized linear MSM 2.16 holds. According to Lange et al., the condition under which conservative SE will be obtained is “that α is substituted by the maximum likelihood estimator under model $\mathcal{M}_1$ and that this is also an efficient estimator of α under model $\mathcal{M}_2$ ”. Although Lange et al. believe that the latter condition will almost always be (nearly) satisfied in practice, they recommend to obtain alternative SE (e.g., based on the bootstrap resampling method) if researchers are concerned about the validity of the additional condition, i.e., the generalized linear MSM is so overly restrictive that the restrictions it imposes on the outcome distribution carry information about the actual value of α .

Following the M-estimation theory, we know that:

$\sqrt{n}(\hat{c}_n - c_t) \xrightarrow{D} N(0, [E\{\frac{\partial U(c_t, \alpha_t)}{\partial c^T}\}]^{-1} \text{var}(U(c_t, \alpha_t)) [E\{\frac{\partial U(c_t, \alpha_t)}{\partial c^T}\}]^{-1T})$ where $\mathbf{c}_t$ and α_t stand for the true values of vectors $(c_0, c_1, c_2, c_3)^T$ and $(\alpha_0, \alpha_1, \alpha_2)^T$. And $(\alpha_0, \alpha_1, \alpha_2)^T$ is from the generalized linear model of Z on $(X, \mathbf{C})$ with link function $g(\cdot)$, i.e., $g(E[Z|X, \mathbf{C}]) = \alpha_0 + \alpha_1 X + \alpha_2 \mathbf{C}$. Since $\mathbf{c}_t$ and α_t are unknown, we instead substitute the GEE estimator $\hat{\mathbf{c}}_n$ for $\mathbf{c}_t$ and MLE estimator $\hat{\alpha}_n$ for α_t and to obtain the sandwich estimator for the asymptotic variance specified above, given by

$$[\hat{E}\{\frac{\partial U(\hat{c}_n, \hat{\alpha}_n)}{\partial c^T}\}]^{-1} \hat{\text{var}}(U(\hat{c}_n, \hat{\alpha}_n)) [\hat{E}\{\frac{\partial U(\hat{c}_n, \hat{\alpha}_n)}{\partial c^T}\}]^{-1T}.$$

So far, we have showed the validity of using MSM to estimate both CDE and NEs, including the sandwich estimators for SE. Following are estimates for mediating effects using MSMs 2.17-2.22:

$$\widehat{\text{CDE}}_{\text{RD},z} = \hat{c}_1(x - x^*) + \hat{c}_3 z(x - x^*).$$

$$\widehat{\text{CDE}}_{\text{RR},z} = \exp^{\hat{c}_1(x-x^*) + \hat{c}_3 z(x-x^*)}.$$

$$\widehat{\text{CDE}}_{\text{OR},z} = \exp^{\hat{c}_1(x-x^*) + \hat{c}_3 z(x-x^*)}.$$

$$\widehat{\text{NDE}}_{\text{RD}} = \hat{c}_1(x - x^*) + \hat{c}_3 x^*(x - x^*).$$

$$\widehat{\text{NIE}}_{\text{RD}} = \hat{c}_2(x - x^*) + \hat{c}_3 x(x - x^*).$$

$$\widehat{\text{NDE}}_{\text{RR}} = \exp^{\hat{c}_1(x-x^*) + \hat{c}_3 x^*(x-x^*)}.$$

$$\widehat{\text{NIE}}_{\text{RR}} = \exp^{\hat{c}_2(x-x^*) + \hat{c}_3x(x-x^*)}.$$

$$\widehat{\text{NDE}}_{\text{OR}} = \exp^{\hat{c}_1(x-x^*) + \hat{c}_3x^*(x-x^*)}.$$

$$\widehat{\text{NIE}}_{\text{OR}} = \exp^{\hat{c}_2(x-x^*) + \hat{c}_3x(x-x^*)}.$$

2.5 Simulation

2.5.1 Data Generating Process (DGP)

The parameters used in the data generating process are carefully chosen to achieve the similar marginal distributions as in the Sister Study data introduced in the section 2.7.

1) Generate the study confounders:

C_1 is a continuous confounder assessed at baseline (T_0), which is generated using the truncated normal distribution, $N(60.72, 54.05)$, with the interval $[44.1, 76.3]$. An example of C_1 could be the participants' age from the source population (cohort) at study baseline.

C_2 is a binary (0/1) confounder assessed at baseline and $C_2 \sim \text{Bernoulli}(P_{C_2})$ and $P_{C_2} \approx 12.71\%$. An example of C_2 could be the race/ethnicity where 0 is the White and 1 is the Black.

2) Generate the exposure:

X is the binary (0/1) exposure of interest and $X \sim \text{Bernoulli}(P_X)$, where $P_X = \{1 + \exp[-\log(0.991C_1) - \log(0.44)]\}^{-1}$ and $P(X = 1) \approx 20.32\%$.

3) Generate the primary outcome (case-control status):

Z is the binary (0/1) mediator, i.e., case-control status or the primary outcome, assessed at follow-up time T_1 , later than baseline, with $Z \sim \text{Bernoulli}(P_Z)$, where

$$P_Z = \{1 + \exp[-\log(1.033C_1) - \log(0.874C_2) - \log(2.7X) - \log(0.00882)]\}^{-1}$$

and $P(Z \approx 1) = 7.73\%$.

4) Generate the sampling selection indicator:

S is the indicator of being selected into the case-control sample (0-unselected and 1-selected), where for subjects with $Z = 1$, all units are selected and for units with $Z = 0$,

a simple random sampling is selected with sampling rate of 3.56%. This is very common in epidemiological studies where cases are completely selected while controls are selected with a small portion due to limited budget (Cao et al. 2022).

5) Generate the observed (secondary) outcome:

Y is the binary (0/1) outcome of interest and $Y \sim \text{Bernoulli}(P_Y)$, where

$$P_Y = \{1 + \exp[-k_Y]\}^{-1}$$

$$k_Y = \log(1.085C_1) + \log(1.57C_2) + \log(16.421X) + \log(4.832Z) + \log(0.961C_1X) \\ + \log(0.650C_2X) + \log(1.75XZ) + \log(0.000792)$$

and $P(Y = 1) \approx 14.91\%$.

6) Generate the potential outcomes:

For the mediator Z , $Z^x \sim \text{Bernoulli}(P_{Z^x})$, where P_{Z^x} is similar to P_Z with replacing X by the actual value x . For the outcome Y , $Y^{xZ^{x*}} \sim \text{Bernoulli}(P_{Y^{xZ^{x*}}})$, where $P_{Y^{xZ^{x*}}}$ is similar to P_Y with replacing X and Z by x and counterfactual outcome Z^{x*} .

2.5.2 Monte Carlo Simulation

The Monte Carlo (MC) simulation follows these steps:

1. Since exposure-covariate and exposure-mediator interactions are assumed, it would be cumbersome to directly calculate the true CDE and NEs, especially for log- and logit-link functions. Thus, the true mediating effects are calculated from a super population ($N_0 = 2000000$) based on the DGP specified in the section 2.5.1.
2. To better mimic the sampling process in the real world, a “cohort” with size N is separately simulated for each individual MC run using the DGP specified in the section 2.5.1.
3. A stratified (on Z) case-control sample is drawn from the “cohort” in step 2, with 100% sampling for cases and a simple random sampling with 3.56% sampling rate for controls. These sampling fractions are chosen to mimic the real-world data collection

in the section 2.7. Different sampling strategies can be applied as long as they are known to researchers.

4. CDE and NEs are estimated using the case-control sample in step 3, i.e., assuming the outcome Y is only observed within the case-control sample.
5. Repeat step 2 to step 4 for $R = 1000$ times.
6. The point estimates and bootstrap 95% CI of the CDE and NEs are calculated across R MC runs.

2.5.3 Bootstrap for Standard Errors

Both VanderWeele (VanderWeele 2009) and Huber (Huber & Solovyeva 2020) recommend to use bootstrap resampling method to calculate SE. For VanderWeele’s approach, two-step are used to estimate the NEs, which makes it difficult to have immediately available SE from the MSM. For HS method, $E[Y^{xz}]$, $E[Y^{xZ^x}]$, and $E[Y^{xZ^{x*}}]$ are separately estimated with weighting method and then CDE and NEs are further estimated by the function of those three quantities. In such case, a direct calculation of the SE can be cumbersome and not quickly achievable. Although our proposed method can obtain robust SE directly from the MSM, for comparison purposes, bootstrap SE will be used across three different methods. Later in section 2.6.2, we will explore comparisons between robust and bootstrap SE for the proposed method. Below are two-step bootstrap resampling method used to get the SE and the 95% CI:

1. Starting from the case-control sample at r -th MC simulation described in the step 3 in the section 2.5.2, we use stratified bootstrap sampling to ensure the same sample size for each category of case-control status (?). Specifically, assuming among the r -th case-control sample, there are n_r units with n_{r0} units for $Z = 0$ and n_{r1} units for $Z = 1$. We use stratified bootstrap method to resample separately among subjects with $Z = 0$ and $Z = 1$. Specifically, for each bootstrap sample b , we resample n_{r0b} from n_{r0} and n_{r1b} from n_{r1} with unrestricted random sampling and sampling ratio of 1, i.e., $n_{r0b} = n_{r0}$ and $n_{r1b} = n_{r1}$. The unrestricted random sampling assigns equal probability for each unit in the stratum and with replacement.

2. For the r -th MC run, we perform B bootstrap sampling runs ($B = 200$) and get the percentile-based bootstrap CI and the bootstrap SE (Zhang 2014). Specifically, the bootstrap SE is calculated using the following formula (Chernick & LaBudde 2011, Efron 1982):

$$SE_b = \sqrt{\frac{1}{B-1} \sum_b (\hat{\theta}_b^* - \hat{\theta}^*)^2}, \text{ where } \hat{\theta}_b^* \text{ is the estimated causal effect, e.g., CDE, estimated from the } b\text{-th bootstrap sample and } \hat{\theta}^* = \frac{\sum_b \hat{\theta}_b^*}{B}.$$

2.5.4 Simulation Performance Criteria

To evaluate the performance of estimators for CDE and NEs, we use the following criteria based on MC simulation (Sahiner et al. 2008, Hafdahl 2010, Morris et al. 2019):

Table 2.2: Simulation Evaluation Criteria

Criterion	Definition	Estimate	Reference
RMSE	$\sqrt{E[(\hat{\theta} - \theta)^2]}$	$\sqrt{\frac{1}{R} \sum_r (\hat{\theta}_r - \theta)^2}$	Sahiner et al. (2008)
Coverage Probability	the proportion of time the 95% CI covers the true parameter constructed from the estimates	$\frac{\sum_r I\{95\% \text{ CI covers } \theta\}}{R}$	Hafdahl (2010)
Empirical SE	Standard deviation of simulated point estimates	$\sqrt{\frac{1}{R-1} \sum_r (\hat{\theta}_r - \bar{\theta})^2}$	Morris et al. (2019)
Bootstrap SE	Average bootstrap SE	$\sqrt{\frac{1}{R} \sum_r \widehat{Var}_r(\theta)}$	Morris et al. (2019)
SE Ratio	Ratio between empirical and bootstrap SE	$\sqrt{\frac{\frac{1}{R-1} \sum_r (\hat{\theta}_r - \bar{\theta})^2}{\frac{1}{R} \sum_r \widehat{Var}_r(\theta)}}$	Morris et al. (2019)

RMSE: root mean squared error; SE: standard error; CI: confidence interval; θ : true causal effect.

$\hat{\theta}_r$: point estimate from r -th Monte Carlo run.

$\bar{\theta} = \frac{1}{R} \sum_r \hat{\theta}_r$: mean across R Monte Carlo simulation runs.

To explore how the sample size may affect the estimation, the simulation for comparing HS method and the proposed approach is also carried out under three different choices of “cohort” size N , i.e., 18000, 36491, and 72000. The second choice of sample size ($N = 36491$) is specifically to reflect what was happened in the empirical study sample described in the section 2.7. The sampling selection strategy is consistent across three scenarios, i.e., 100% sampling among subjects with $Z = 1$ and 3.56% simple random sampling among subjects with $Z = 0$. Therefor, the case-control sample sizes are approximately 2000, 4000, and 8000, respectively.

2.6 Simulation Results

2.6.1 Results Comparison Between Two Methods

Using MC simulation with 1000 runs, and under different presumed cohort size (18000, 36491, or 72000), we have the following result:

Table 2.3: Controlled Direct Effect Fixing $Z = 0$

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	2.722x10 ⁻²	2.876x10 ⁻²	0.0358	0.9724	0.0349	94.8
	36491		2.724x10 ⁻²	0.0249	0.9887	0.0246	94.0
	72000		2.608x10 ⁻²	0.0177	1.0123	0.0180	94.9
Proposed	18000		2.821x10 ⁻²	0.0371	0.9714	0.0360	94.3
	36491		2.688x10 ⁻²	0.0258	0.9941	0.0256	94.0
	72000		2.576x10 ⁻²	0.0184	1.0056	0.0186	95.0
Relative Risk							
Huber and Solovyeva	18000	1.232	1.284	0.3663	0.9368	0.3470	94.5
	36491		1.257	0.2434	0.9782	0.2393	94.3
	72000		1.242	0.1706	1.0028	0.1713	95.0
Proposed	18000		1.260	0.3523	0.9351	0.3305	94.1
	36491		1.238	0.2354	0.9840	0.2316	94.1
	72000		1.224	0.1652	0.9953	0.1645	95.5
Odds Ratio							
Huber and Solovyeva	18000	1.271	1.343	0.4548	0.9122	0.4208	94.5
	36491		1.305	0.2948	0.9658	0.2866	94.5
	72000		1.284	0.2043	0.9962	0.2039	95.0
Proposed	18000		1.318	0.4412	0.9099	0.4040	94.1
	36491		1.285	0.2876	0.9711	0.2795	94.0
	72000		1.265	0.1995	0.9900	0.1975	95.5

^a This is the ratio between empirical and bootstrap standard error.

From Table 2.3 we can see that our method is slightly better than the HS method for CDE when fixing the mediator Z at 0 for both RR and OR scale, in terms of smaller bias and bootstrap SE. For RD scale, the HS method is slightly better with smaller bootstrap SE and overall RMSE.

Table 2.4: Controlled Direct Effect Fixing $Z = 1$

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	0.208	0.199	0.0275	0.9844	0.0284	93.2
	36491		0.201	0.0193	0.9824	0.0203	91.8
	72000		0.200	0.0137	0.9844	0.0154	90.9
Proposed	18000		0.207	0.0275	0.9858	0.0272	94.2
	36491		0.208	0.0193	0.9840	0.0190	95.5
	72000		0.208	0.0137	0.9809	0.0134	95.5
Relative Risk							
Huber and Solovyeva	18000	1.558	1.528	0.0893	0.9856	0.0931	93.4
	36491		1.530	0.0623	0.9857	0.0677	91.1
	72000		1.528	0.0442	0.9896	0.0531	87.8
Proposed	18000		1.558	0.0923	0.9857	0.0909	94.1
	36491		1.560	0.0643	0.9854	0.0634	95.3
	72000		1.558	0.0457	0.9847	0.0450	95.3
Odds Ratio							
Huber and Solovyeva	18000	2.329	2.264	0.2651	0.9721	0.2657	93.2
	36491		2.267	0.1836	0.9767	0.1900	91.9
	72000		2.261	0.1295	0.9796	0.1442	90.7
Proposed	18000		2.337	0.2747	0.9730	0.2673	94.0
	36491		2.338	0.1902	0.9777	0.1860	95.6
	72000		2.333	0.1342	0.9761	0.1310	95.3

^a This is the ratio between empirical and bootstrap standard error.

From Table 2.4 we can see that our method is better than the HS method for CDE when fixing the mediator Z at 1. First, our method has smaller bias for RD, RR, and OR scales. Second, although our method has slightly larger bootstrap SE, the RMSE is still overall smaller compared to the HS estimates. Third, as the sample/cohort size increases, the coverage probability of our method gets closer to 95% whereas the HS estimated coverages actually decrease.

Table 2.5: Natural Direct Effects

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	3.770x10 ⁻²	3.864x10 ⁻²	0.0349	0.9686	0.0338	94.3
	36491		3.736x10 ⁻²	0.0243	0.9897	0.0240	93.8
	72000		3.623x10 ⁻²	0.0173	1.0045	0.0174	95.1
Proposed	18000		3.868x10 ⁻²	0.0348	0.9702	0.0338	94.8
	36491		3.738x10 ⁻²	0.0243	0.9926	0.0241	93.8
	72000		3.629x10 ⁻²	0.0173	1.0066	0.0175	95.1
Relative Risk							
Huber and Solovyeva	18000	1.283	1.304	0.2876	0.9488	0.2736	94.1
	36491		1.286	0.1947	0.9851	0.1918	94.0
	72000		1.274	0.1373	0.9964	0.1370	95.4
Proposed	18000		1.304	0.2867	0.9486	0.2727	94.0
	36491		1.286	0.1945	0.9879	0.1921	93.9
	72000		1.275	0.1373	0.9990	0.1373	95.0
Odds Ratio							
Huber and Solovyeva	18000	1.341	1.379	0.3740	0.9241	0.3475	94.1
	36491		1.351	0.2476	0.9725	0.2409	94.1
	72000		1.334	0.1726	0.9917	0.1713	95.2
Proposed	18000		1.378	0.3730	0.9242	0.3465	94.2
	36491		1.351	0.2473	0.9754	0.2413	94.0
	72000		1.334	0.1726	0.9942	0.1717	95.1

^a This is the ratio between empirical and bootstrap standard error.

For the NDE estimates in Table 2.5, two methods have almost exactly the same results, with our method has slightly larger RMSE (at 4 decimal places).

Table 2.6: Natural Indirect Effects

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	3.794x10 ⁻²	3.779x10 ⁻²	0.0047	0.9345	0.0044	95.6
	36491		3.790x10 ⁻²	0.0032	0.9818	0.0032	94.6
	72000		3.803x10 ⁻²	0.0023	0.9803	0.0023	94.2
Proposed	18000		3.794x10 ⁻²	0.0047	1.0740	0.0050	92.6
	36491		3.796x10 ⁻²	0.0032	1.1323	0.0037	91.7
	72000		3.805x10 ⁻²	0.0023	1.1384	0.0026	89.7
Relative Risk							
Huber and Solovyeva	18000	1.222	1.230	0.0752	0.8395	0.0636	95.1
	36491		1.227	0.0476	0.9513	0.0455	94.4
	72000		1.227	0.0325	0.9833	0.0324	93.3
Proposed	18000		1.230	0.0753	0.8607	0.0653	94.7
	36491		1.227	0.0477	0.9671	0.0464	94.7
	72000		1.227	0.0326	1.0052	0.0331	93.6
Odds Ratio							
Huber and Solovyeva	18000	1.280	1.289	0.0830	0.8462	0.0707	95.0
	36491		1.286	0.0528	0.9549	0.0507	94.6
	72000		1.286	0.0362	0.9846	0.0360	93.3
Proposed	18000		1.290	0.0832	0.8753	0.0734	94.5
	36491		1.286	0.0529	0.9811	0.0522	93.6
	72000		1.286	0.0362	1.0182	0.0373	93.1

^a This is the ratio between empirical and bootstrap standard error.

From the Table 2.6 we can see that overall, two methods are still very similar in simulation performance. For the RD scale, our method has lower coverage compared to the HS method, although both methods have very close point estimates and the bootstrap SE.

2.6.2 Uncertainty of the Estimators

Since researchers are usually interested in the NE parameters, we hereby discuss the SE for the NIE for our method in this section.

Although we have compared the bootstrap SE between the proposed and HS estimates in section 2.6.1, it is worth to take another look at the distribution of the estimator for NIE to examine the asymptotic normality. We would also like to compare the bootstrap SE with the robust SE using our method. This potentially helps researchers to understand how different it can be when using two methods to obtain SE in practice.

1) Distribution of the NIE estimator

Figure 2.2 shows the distribution of estimates for the NIE using the proposed method based on 1000 MC simulation. The top two histograms are for the RD scale for $N = 18,000$

and $N = 72,000$ while the bottom two are for the log-OR scale under two cohort sizes.

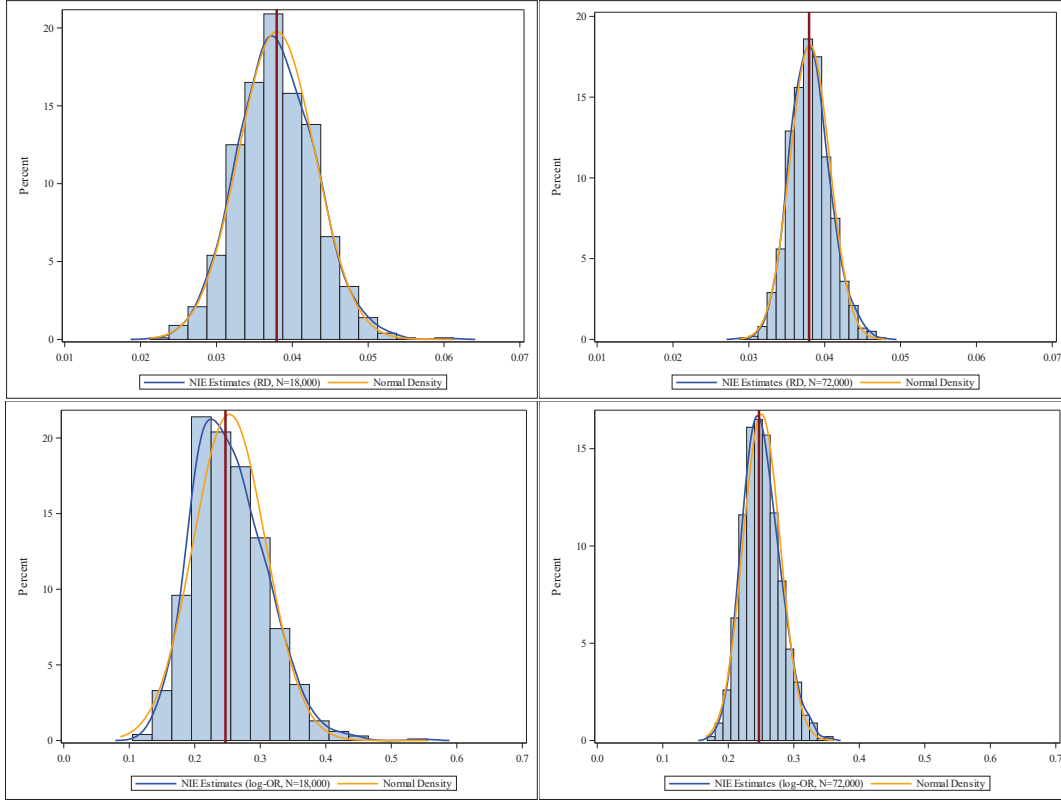


Figure 2.2: Distribution of NIE Estimates using Proposed Approach

From the visual check on the figure, we can see that the proposed approach gives estimates for NIE that is close to the normal density (orange curve). And when increasing the cohort size, the kernel density (blue curve) gets closer to the normal density, which supports the asymptotic normality assumption.

Figure 2.3 shows the distribution of estimates for the HS estimates of the NIE based on 1000 MC simulations.

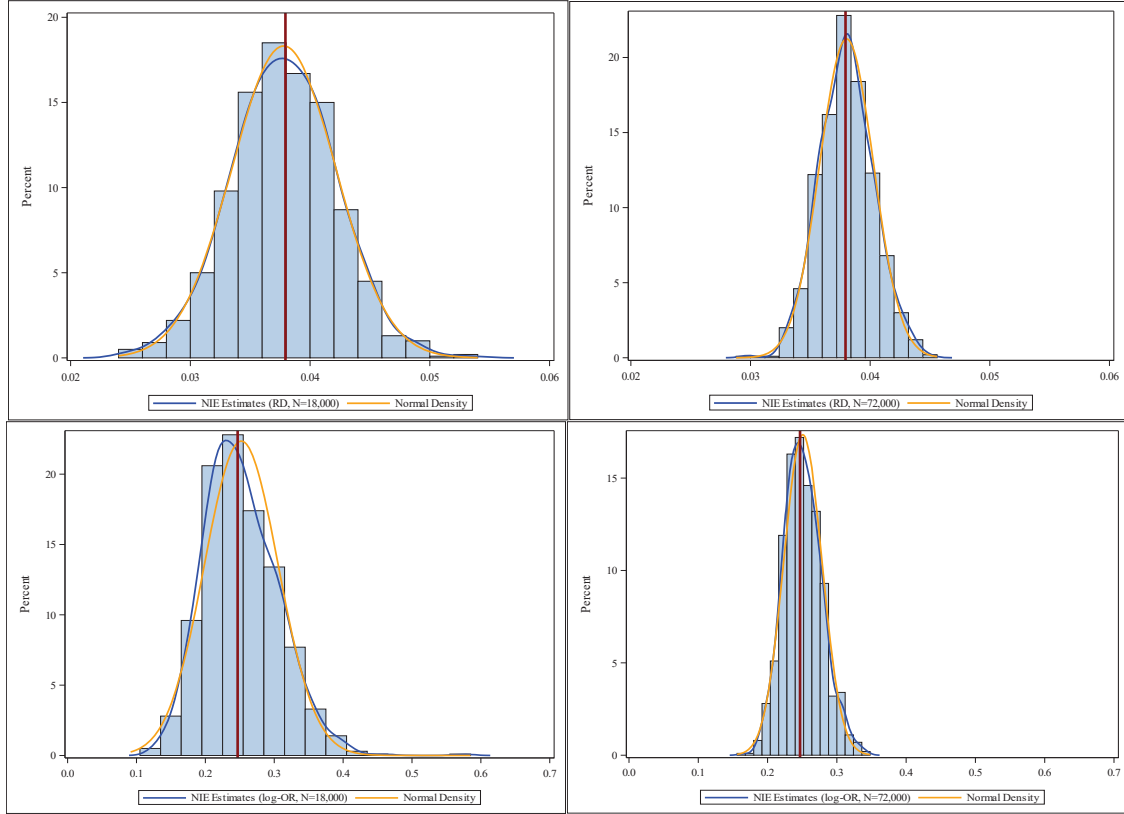


Figure 2.3: Distribution of NIE Estimates using HS Method

From the visual check on the figure, we can see that the HS method gives estimates that are also close to the normal. They point out that their approach is $\sqrt{n}$ -consistent and asymptotically normal under standard regularity condition (Huber & Solovyeva 2020). When increasing the cohort size from 18,000 to 72,000, we can see that the kernel density (blue curve) gets closer to the the normal density (orange curve).

With these supporting evidence, we consider that both the proposed and the HS method give asymptotically normally distributed estimators for the NIE.

2) Robust and bootstrap SE

Figure 2.4 shows the distribution of robust and bootstrap SE for our proposed method under simulation settings from section 2.5.2 and 2.5.3. The top two densities are for the RD scale for $N = 18,000$ and $N = 72,000$ while the bottom two are for the log-OR

scale under two cohort sizes.

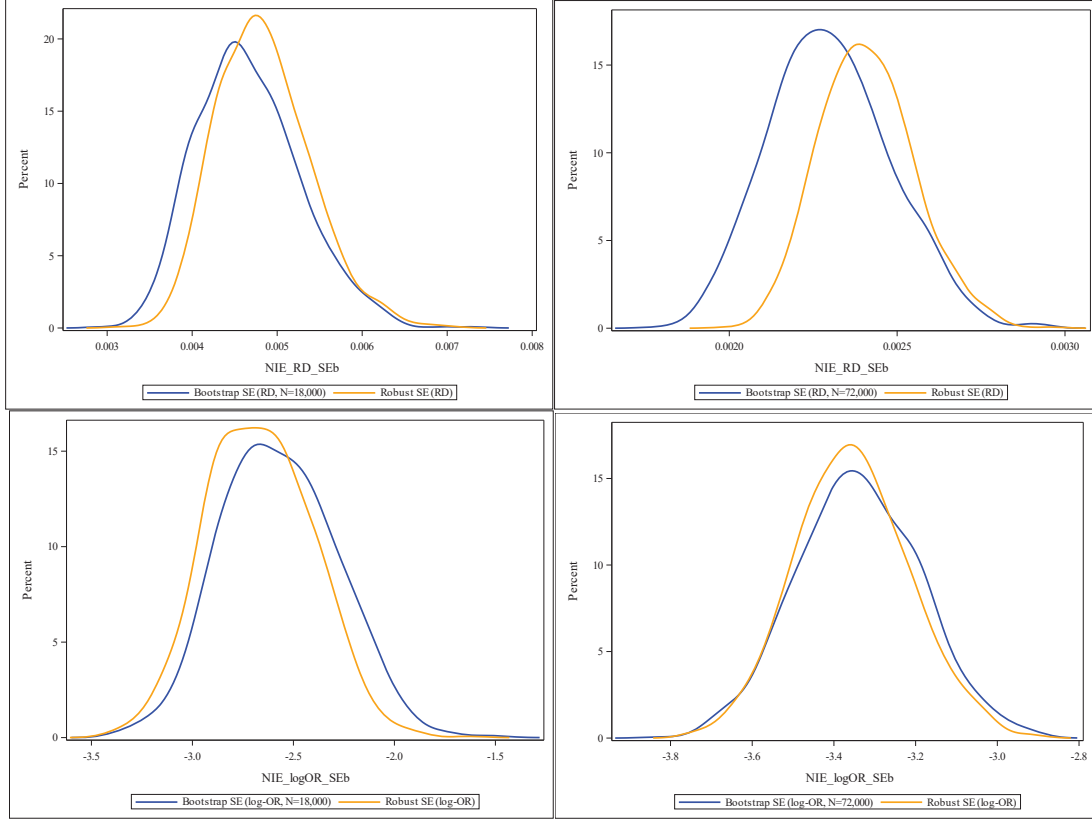


Figure 2.4: Distribution Comparison Between Robust and Bootstrap SE for Proposed Method

While the robust (orange curve) SE is slightly larger than and bootstrap (blue curve) SE for the NIE estimates in RD scale, their distribution are very close for the NIE estimates in log-OR scale. Since two methods give asymptotically normally distributed estimators, we believe both robust and bootstrap resampling based methods can be used to obtain the SE for the NIE estimates.

2.7 Application: the Olfaction Sub-study of the Sister Study Cohort

In this section, we implement both the HS method and our proposed method using the NIEHS Sister Study data. The Sister Study is an ongoing nationwide cohort to investigate environmental and genetic risk factors for breast cancer and other chronic diseases (Sandler

et al. 2017). Details of a case-control study design with a secondary outcome and data collection have been published elsewhere (Cao et al. 2022). In this example, we will use the same case-control study design as previously published (Cao et al. 2022), with a subsequently collected secondary outcome as the interest of the outcome, to implement two methods for estimating both CDE and NEs. Briefly, during 2012-2014 (second detailed-follow-up, i.e., DFU-2, considered as baseline in this example), the Sister Study collected data on whether the participants ever had a serious head injury that resulted in unconsciousness, coma, or hospitalization. At the third detailed-follow-up (DFU-3, 2014-2016), a stratified case-control sample was selected based on self-reported sense of smell (100% sampling for self-report “poor” and about 3.56% simple random sampling for self-report “normal”) (Cao et al. 2022). The self-report sense of smell is a surrogate of the olfactory function at DFU-3. Later in 2018-2019, study investigators mailed out objective olfaction test kits (12-item Brief Smell Identification Test, B-SIT) to those who were selected in the case-control sample only and collected the secondary outcome, B-SIT-tested olfaction. Previous publications have suggested the link between traumatic brain injury and olfactory impairment (Howell et al. 2018, Xydakis et al. 2015). We define the serious head injury for subjects who answered “YES” to the corresponding question and defined the mediator as self-report sense of smell at DFU-3. We further define the olfactory impairment in 2018-2019 using a cutoff of ≤ 9 for the B-SIT test (Cao et al. 2022).

In the illustrative example, we are interested in studying the effects of a serious head injury (X) on tested poor olfaction (secondary outcome Y) mediated through the self-report sense of smell (case-control status or the primary outcome Z) at DFU-3. Using the notations described in the section 2.2.1, we consider the participants’ age at DFU-2 as the common cause for the exposure, the mediator and the outcome, i.e., C_1 , the participants’ race/ethnicity as the common cause for the mediator and the outcome, i.e., C_2 . Table 2.7 below summarizes the definition of variables used in this analysis:

Table 2.7: Variable Definition for the Sister Study Example

Variables	Definition	Variable Type	Distribution (N=33,932)	Time of Assessment
Confounder (C_1)	Age	Continuous	60.7 (7.3)	DFU-2 (2012-2014)
Confounder (C_2)	Race/Ethnicity	White (0)/Other (1)	$P(C_2 = 1) = 12.4\%$	Enrollment (2003-2009)
Exposure (X)	History of serious head injury	Yes (1)/No (0)	$P(X = 1) = 7.4\%$	DFU-2 (2012-2014)
Mediator (Z)	Self-report olfaction	Poor (1)/Normal (0)	$P(Z = 1) = 6.6\%$	DFU-3 (2014-2016)
Outcome (Y)	B-SIT tested olfaction (≤ 9)	Poor (1)/Normal (0)	Poor: 970 out of 3,249	2018-2019

This analysis aims to illustrate the implementation of two methods for mediation analysis in the complex survey data and may not have epidemiological interpretation due to multiple reasons. First, the assessment of the mediator (olfactory status at DFU-3) is based on self-report, which suffers from potential measurement errors. Second, the mediator and the outcome are both surrogates for the true olfactory status at two timepoints so the clinical meaning of this mediating effects is questionable. Third, only a single question or test was given for the olfaction assessment for both the mediator and the outcome. This increase the possibility of having a measurement error for the true olfactory status.

This example includes 33,932 out of 36,491 participants from the cohort who had non-missing values for the covariate set $\mathbf{C}$, the exposure X , and the mediator (the primary case-control outcome) Z . Investigators used 100% sampling for cases and $\sim 3.2\%$ simple random sampling for controls. The complete case data has 3,249 participants in the case-control sample with no missingness in the outcome Y . Table 2.8 shows the estimates of the mediating effects in the RD, RR, and OR scales with 95% CIs.

Both methods give similar estimates on NEs and the CDE when fixing the mediator (self-report sense of smell) at value of 1. For the CDE when fixing the mediator at value of 0, two methods provide slightly different point estimates, while the bootstrap 95% CIs from two methods still give similar estimation. Comparing with the bootstrap 95% CI, the robust one gives similar estimation with slight efficiency gain (narrower). Again, this is just an example of showing how to implement two methods to the real-world data and no clinical conclusion should be drawn from this analysis.

Table 2.8: Method Comparison Using the Sister Study

Effect	Huber & Solovyeva		Proposed		
	Point Estimate	Bootstrap 95% CI	Point Estimate	Bootstrap 95% CI	Robust 95% CI
Risk Difference					
CDE ₀	-0.009	(-0.074, 0.071)	0.008	(-0.072, 0.102)	(-0.076, 0.093)
CDE ₁	0.034	(-0.034, 0.102)	0.039	(-0.029, 0.107)	(-0.029, 0.107)
NDE	0.009	(-0.065, 0.099)	0.011	(-0.062, 0.098)	(-0.066, 0.089)
NIE	0.007	(0.003, 0.011)	0.008	(0.004, 0.012)	(0.004, 0.012)
Relative Risk					
CDE ₀	0.92	(0.32, 1.73)	1.07	(0.36, 1.99)	(0.52, 2.20)
CDE ₁	1.09	(0.91, 1.29)	1.11	(0.92, 1.31)	(0.93, 1.32)
NDE	1.07	(0.50, 1.81)	1.09	(0.53, 1.82)	(0.62, 1.92)
NIE	1.05	(1.02, 1.15)	1.06	(1.02, 1.16)	(1.00, 1.11)
Odds Ratio					
CDE ₀	0.91	(0.29, 1.88)	1.08	(0.34, 2.24)	(0.48, 2.45)
CDE ₁	1.15	(0.86, 1.53)	1.18	(0.88, 1.56)	(0.89, 1.57)
NDE	1.08	(0.46, 2.04)	1.10	(0.49, 2.04)	(0.57, 2.13)
NIE	1.06	(1.02, 1.16)	1.07	(1.02, 1.17)	(1.01, 1.13)

2.8 Conclusion

In Chapter 2, we discussed the estimation of mediating effects (controlled direct effects and natural effects) using complex sampling data of a case-control sample drawn from the target population (e.g., a cohort) with the primary outcome as the mediator and a secondary outcome as the outcome of interest. We first showed that ignoring the study design would lead to serious bias in the natural (in)direct effect estimation. Second, we proposed a set of weighting estimators and compared it with Huber and Solovyeva's work using the Monte Carlo simulation study. For controlled direct effects, the proposed estimator has overall smaller root mean squared error and better coverage probability when fixing the mediator at value of 1. Two methods are very comparable for estimating natural (in)direct effects. The second advantage of the proposed method is that unlike the Huber and Solovyeva's approach, the proposed estimation of weights does not involve the probability model of the exposure given the mediator. Third, we showed that the validity of using robust standard error for the proposed estimator in addition to the bootstrap resampling based standard error. Fourth, we implemented and compare two approaches into a real-world data for illustration purpose. Overall, we recommend to use more than one method in practice, to provide researchers more information and comprehensive results for clinical interpretation.

CHAPTER 3 Mediation Analyses for Case-Control Studies With Secondary and Tertiary Outcomes

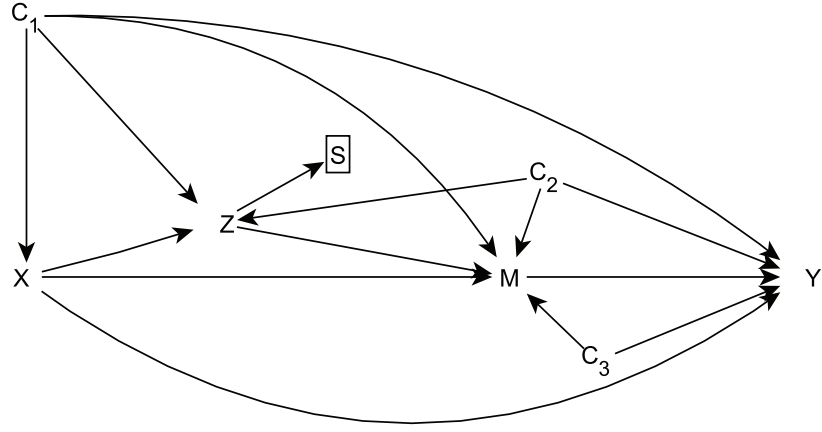
3.1 Background

In Chapter 2, we discussed the mediation analysis using case-control studies with the primary outcome as the mediator and a secondary outcome collected subsequently as the outcome of interest. In fact, with more follow-ups available to researchers, not only secondary but also tertiary outcomes can further be collected. This may happen after a case-control study has been conducted with additional interest of direct or indirect effects of exposure on the tertiary outcome through the secondary outcome. Unlike Chapter 2, here both the secondary outcome (the mediator) and the tertiary outcome (the outcome), and possibly some covariates, are observed only within the case-control sample. For example, Cao et al. (in press, *Environmental Health Perspectives*, DOI: 10.1289/EHP12066) studied exposure to ambient air pollutants in relation to a secondary outcome, i.e., olfactory impairment defined by a Brief Smell Identification Test (B-SIT), using a case-control sample from the NIEHS Sister Study. Using similar designs, researchers may be further interested in the mediating effects of exposure to air pollutants on a tertiary outcome, e.g., cognitive deficit, through the secondary outcome, i.e., olfactory impairment. However, performing such mediation analysis using existing methods could lead to seriously biased estimates since the case-control sample is an outcome-dependent sample. In this chapter, we will first describe the settings and notations, followed by a marginal structural model to show the consequence of ignoring the complex study design. Then we will discuss the extension of an existing estimator that is potentially suitable for this scenario. Next, we propose a new estimator (weighting estimator II) and compare its performance with the existing estimator. Finally, we implement two methods using a real-world example from the NIEHS Sister Study data.

3.2 Settings

3.2.1 Notations and Causal Directed Acyclic Graph (DAG)

In this chapter, our study setting is the mediation analysis of a case-control sample with a secondary outcome as the mediator and a subsequently collected tertiary outcome as the outcome of interest. We have (C_1, C_2, X, Z, S) with the same definition as described in Chapter 2. Additionally, let C_3 be the mediator-outcome confounder assessed post-baseline, e.g., comorbidity disease at the time of mediator assessment, prior to the assessment of the outcome. It is possible that C_3 is only assessed within the case-control sample. Let M be the mediator (secondary outcome) which is assessed after the collection of case-control sample and available only within the sample. We assume that the sample is selected from an existing cohort study, like in many epidemiological studies. Let Y be the tertiary outcome (the outcome of interest in the mediation analysis) assessed after the measurement of the mediator. Figure 3.1 shows the causal DAG for the data generating process of $(C_1, C_2, C_3, X, Z, M, S, Y)$.



- C_1 : confounder between the exposure, the mediator and the outcome
- C_2 : common cause for primary outcome, the mediator and the outcome
- C_3 : confounder between the mediator and the outcome
- X : exposure/treatment
- Z : case-control status
- M : secondary outcome (mediator)
- S : case-control sample selection node
- Y : (tertiary) outcome

Figure 3.1: Causal DAG for Mediation Analysis

For simplicity, assume C_2 , C_3 , X , Z , M and Y are all binary variables. In practice, our proposed approach can be extended to non-binary variables as well. If the case-control sample is selected from a cohort (source population) study, then (C_1, C_2, X, Z, S) are observed for all units whereas (C_3, M, Y) are only observed when $S = 1$, i.e., among units who are selected into the case-control sample. In our application of the Sister study data, cases are 100% selected and controls are selected using SRS with a pre-specified sampling rate (Cao et al. 2022). Similarly to the previous chapter, our proposed approach may also be extended to other sampling method such as stratified sampling or proportion to size sampling as long as the sampling process is known by design. The aim here is to estimate the causal parameter

(e.g., mediating effects) in the cohort using the selected case-control sample. With such study design, X is fully observed from the cohort. However, even if X is only observable within the case-control sample, using the approach discussed in section 3.3 we can still estimate the mediating effects in the cohort (source population). We adopt Neyman-Rubin causal model (Rubin 1978, Splawa-Neyman et al. 1923) and invoke the SUTVA for simplicity (Rubin 1986). Let M^x be the counterfactual mediator and Y^{xM^x} be the counterfactual outcome when the exposure X is set to the value x . The goal is to study mediating effects from X to Y , through the mediator M .

3.2.2 Controlled Direct Effect (CDE)

For binary Y and M , the population (cohort) CDE can be defined in three different scales (Robins & Greenland 1992, Pearl 2001, VanderWeele 2015):

$$\text{Risk difference: CDE}_{\text{RD}} = E[Y^{1m} - Y^{0m}], \text{ for } m \in \{0, 1\}.$$

$$\text{Relative risk: CDE}_{\text{RR}} = \frac{P(Y^{1m}=1)}{P(Y^{0m}=1)}, \text{ for } m \in \{0, 1\}.$$

$$\text{Odds ratio: CDE}_{\text{OR}} = \frac{P(Y^{1m}=1)/P(Y^{1m}=0)}{P(Y^{0m}=1)/P(Y^{0m}=0)}, \text{ for } m \in \{0, 1\}.$$

3.2.3 Natural Effects (NE)

Similar to CDE, for binary Y and M , the population natural (in)direct effects can also be defined in three different scales (Robins & Greenland 1992, Pearl 2001, VanderWeele 2015):

$$\text{Risk difference: NDE}_{\text{RD}} = E[Y^{1M^0} - Y^{0M^0}] \text{ and } \text{NIE}_{\text{RD}} = E[Y^{1M^1} - Y^{1M^0}].$$

$$\text{Relative risk: NDE}_{\text{RR}} = \frac{P(Y^{1M^0}=1)}{P(Y^{0M^0}=1)} \text{ and } \text{NIE}_{\text{RR}} = \frac{P(Y^{1M^1}=1)}{P(Y^{1M^0}=1)}.$$

$$\text{Odds ratio: NDE}_{\text{OR}} = \frac{P(Y^{1M^0}=1)/P(Y^{1M^0}=0)}{P(Y^{0M^0}=1)/P(Y^{0M^0}=0)} \text{ and } \text{NIE}_{\text{OR}} = \frac{P(Y^{1M^1}=1)/P(Y^{1M^1}=0)}{P(Y^{1M^0}=1)/P(Y^{1M^0}=0)}.$$

3.3 Existing Methods for the Scenario

3.3.1 Weighting Method Ignoring the Study Design

Similarly as in Chapter 2, I would like to point out the necessity for considering the study design while performing causal mediation analysis. The same model is used as in Chapter 2 (VanderWeele 2009), to illustrate the consequence of ignoring the secondary and tertiary data collection process in terms of biased estimates in mediation analysis. Also, I will discuss the possible remedy for such scenario. The notation is consistent with the definitions in section

3.2.1.

The adapt the notations and settings for this chapter, following MSMs are used:

$$E[Y^{xm}] = g(x, m) = \alpha_0 + \alpha_1 x + \alpha_2 m + \alpha_3 xm \quad (3.1)$$

$$E[M^x | \mathbf{C} = \mathbf{c}] = h(x, \mathbf{c}) = \gamma_0 + \gamma_1 x + \gamma'_2 \mathbf{c} \quad (3.2)$$

$$E[Y^{xm} | \mathbf{C} = \mathbf{c}] = g(x, m, \mathbf{c}) = \theta_0 + \theta_1 x + \theta_2 m + \theta_3 xm + \theta'_4 \mathbf{c} \quad (3.3)$$

The MSM 3.1 is used to estimate the CDE with assigning each individual with a weight of $w_x * w_m$, where $w_x = \frac{P(X=x)}{P(X=x|\mathbf{C}=\mathbf{c})}$ and $w_m = \frac{P(M=m|X=x)}{P(M=m|X=x, \mathbf{C}=\mathbf{c})}$. However, since M is only observed within the case-control sample, $P(M = m|X = x)$ and $P(M = m|X = x, \mathbf{C} = \mathbf{c})$ are not directly estimable using observed data. Here, we can take advantage of knowing the study design and estimate those two quantities indirectly. Specifically, we have

$$P(M = m|X = x) = \sum_z P(M = m|X = x, Z = z) \cdot P(Z = z|X = x) \quad (3.4)$$

$$P(M = m|X = x, \mathbf{C} = \mathbf{c}) = \sum_z P(M = m|X = x, \mathbf{C} = \mathbf{c}, Z = z) \cdot P(Z = z|X = x, \mathbf{C} = \mathbf{c}) \quad (3.5)$$

The MSM 3.2 and 3.3 are used to estimate the conditional NEs. For 3.2, the weight $w_2 = \frac{1}{P(X=x|\mathbf{C}=\mathbf{c})}$ is assigned for each individual, and for 3.3, the same weight $w_x * w_m$ is used. One difference between 3.3 and 3.1 is that model 3.3 also conditions on the covariate set $\mathbf{C}$, i.e., (C_1, C_2, C_3) in our case. Thus, estimates from MSM 3.2 and 3.3 are called conditional NEs. To get marginal ones, we need to further take expectation regarding the covariate set $\mathbf{C}$.

Since the case-control sampling is stratified by Z , $P(M|X, Z = z)$ and $P(M|X, \mathbf{C}, Z = z)$ can be correctly estimated. This is simply because for units with $Z = 1$, we have

100% sampling, which indicates that the probability mass function (pmf) $P(M|X, Z = 1)$ (or $P(M|X, \mathbf{C}, Z = 1)$) estimated from the case-control sample is the “true” pmf from the sub-population. Similarly, for subjects with $Z = 0$, the SRS also indicates that the pmf $P(M|X, Z = 0)$ (or $P(M|X, \mathbf{C}, Z = 0)$) can be correctly estimated using the case-control sample. And since Z is observed across the whole population (e.g., the cohort), $P(Z = z|X = x)$ and $P(Z = z|X = x, \mathbf{C} = \mathbf{c})$ can be correctly estimated using the whole population.

CDE and marginal NEs are estimated using the following:

$$\text{CDE}_{\text{RD}} = E[Y^{1m} - Y^{0m}] = \alpha_1 + \alpha_3 m \quad (3.6)$$

$$\text{pure NDE}_{\text{RD}} = E[Y^{1M^0} - Y^{0M^0}] = \theta_1 + \theta_3 \gamma_0 + \theta_3 \gamma_2' E[\mathbf{C}] \quad (3.7)$$

$$\text{total NIE}_{\text{RD}} = E[Y^{1M^1} - Y^{1M^0}] = (\theta_2 + \theta_3) \gamma_1 \quad (3.8)$$

Table 3.1: Mediating Effects When Ignoring Study Design

Effect (RD)	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
CDE ₀	36491	0.016	0.014	0.0148	0.9976	0.0148	94.6
CDE ₁		0.162	0.161	0.0269	0.9985	0.0269	94.6
NDE		0.034	0.050	0.0128	1.0032	0.0205	76.5
NIE		0.036	0.056	0.0067	1.0040	0.0209	11.6

^a This is the ratio between empirical and bootstrap standard error.

From Table 3.1 we can see that even though the CDE can be unbiasedly estimated, there is serious bias in estimating NEs. One possible reason could be that when fixing the exposure X and the mediator M with controlling for the covariates $\mathbf{C}$ (by weighting), $Y^{xm} \perp\!\!\!\perp S \mid \mathbf{C}$ is achieved. Thus, ignoring the sample selection process would still give unbiased estimates of Y^{xm} from MSM 3.1 due to the conditional independency. However, for NEs, even though we can fix X at x (or x^*), M^x (or M^{x^*}) is considered as a random variable. Thus, Y^{M^x} (or $Y^{M^{x^*}}$) is not independent of sample selection S in such scenario, which will lead to biased estimates of NEs if ignoring the study design. In such case, either

the weight for MSM 3.3 needs to be adjusted to incorporate the sample selection process, or a different approach can be used to estimate the NEs. Below we give a potential remedy for this.

NDE can still be unbiasedly estimated if CDE estimates are unbiased since NDE is weighted average of CDE (Rudolph et al. 2019, Petersen et al. 2006). In the case-control sample, if each counterfactual outcome mean can be estimated unbiasedly, then all NEs also can.

$$\begin{aligned}
E[Y^{xM^x}] &= E_{\mathbf{c}}[E_{Y^{xM^x}|\mathbf{C}}[Y^{xM^x}|\mathbf{C}]] \\
&= E_{\mathbf{c}}[\int y \cdot P(Y^{xM^x} = y|\mathbf{C})dy] \\
&= E_{\mathbf{c}}[\int \int y \cdot P(Y^{xM^x} = y, M^x = m|\mathbf{C})dmdy] \\
&= E_{\mathbf{c}}[\int \int y \cdot P(Y^{xm} = y|M^x = m, \mathbf{C}) \cdot P(M^x = m|\mathbf{C})dmdy] \\
&= E_{\mathbf{c}}[\sum_m (E[Y^{xm}|\mathbf{C}]) \cdot P(M^x = m|\mathbf{C})]
\end{aligned}$$

Similarly, we have $E[Y^{xM^{x*}}] = E_{\mathbf{c}}[\sum_m (E[Y^{xm}|\mathbf{C}]) \cdot P(M^{x*} = m|\mathbf{C})]$. Then NIE can be calculated using the difference between above two quantities. The key here is to unbiasedly estimate $P(M^x = m|\mathbf{C})$ and $P(M^{x*} = m|\mathbf{C})$. In the VanderWeele paper, MSM 3.2 is specifically used for estimating these two quantities while ignoring the study design would lead to biased estimation at this step, and further lead to the biased estimation in NEs. A thorough investigation of this will be a separate topic and is not our primary focus here. Next, we will present a modified version of an existing method that could potentially give unbiased estimation of mediating effects in such complex study design.

3.3.2 Modified Huber and Solovyeva's Method

With the notations and settings under this particular chapter, HS estimator can be written as following

$$\hat{E}[Y^{xm}] = \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot I\{M_i = m\} \cdot S_i}{P(X_i = x|C_{1i}) \cdot P(M_i = m|X_i, \mathbf{C}_i) \cdot P(S_i = 1|X_i, M_i, \mathbf{C}_i)} \quad (3.9)$$

$$\hat{E}[Y^{xM^x}] = \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|C_{1i}) \cdot P(S_i = 1|X_i, M_i, \mathbf{C}_i)} \quad (3.10)$$

$$\hat{E}[Y^{xM^{x*}}] = \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|M_i, \mathbf{C}_i) \cdot P(S_i = 1|X_i, M_i, \mathbf{C}_i)} \cdot \frac{P(X_i = 1 - x|M_i, \mathbf{C}_i)}{P(X_i = 1 - x|\mathbf{C}_i)} \quad (3.11)$$

where $\mathbf{C}_i = (C_{1i}, C_{2i}, C_{3i})$.

However, probabilities such as $P(S_i = 1|X_i, M_i, \mathbf{C}_i)$ cannot be directly estimated using observed data. And the stratified approach described in section 3.3.1 even fails since $P(Z_i = z|X_i, M_i, \mathbf{C}_i)$ is still unestimable using observed data only. Thus, we modified equations 3.9, 3.10 and 3.11 to the following with basic probability theory:

$$\hat{E}[Y^{xm}] = \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot I\{M_i = m\} \cdot S_i}{P(X_i = x|C_{1i}) \cdot P(M_i = m|X_i, \mathbf{C}_i, S_i = 1) \cdot P(S_i = 1|X_i, C_{1i}, C_{2i})} \quad (3.12)$$

$$\hat{E}[Y^{xM^x}] = \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i \cdot P(M_i = m|X_i, \mathbf{C}_i)}{P(X_i = x|C_{1i}) \cdot P(M_i = m|X_i, \mathbf{C}_i, S_i = 1) \cdot P(S_i = 1|X_i, C_{1i}, C_{2i})} \quad (3.13)$$

$$\begin{aligned} \hat{E}[Y^{xM^{x*}}] &= \frac{1}{n} \sum_i \frac{Y_i \cdot I\{X_i = x\} \cdot S_i}{P(X_i = x|M_i, \mathbf{C}_i)} \cdot \frac{P(X_i = 1 - x|M_i, \mathbf{C}_i)}{P(X_i = 1 - x|\mathbf{C}_i)} \\ &\quad \frac{P(M_i = m|X_i, \mathbf{C}_i)}{P(M_i = m|X_i, \mathbf{C}_i, S_i = 1) \cdot P(S_i = 1|X_i, C_{1i}, C_{2i})} \end{aligned} \quad (3.14)$$

And $P(X_i = x|M_i, \mathbf{C}_i) = \frac{P(M_i = m|X_i, \mathbf{C}_i) \cdot P(X|\mathbf{C}_i)}{P(M_i = m|\mathbf{C}_i)}$. For $P(M_i = m|X_i, \mathbf{C}_i)$ and $P(M_i = m|\mathbf{C}_i)$, we can further estimate them using the stratified approach described in section 3.3.1. This modified method and estimator will be called the modified HS method and estimator henceforth. The modified HS estimators of CDE and NEs naturally follow. Instead of directly plug in equation 3.12-3.14, they proposed to use normalized versions of the sample

analogs (Huber & Solovyeva 2020).

The modified HS method has a relatively complicated estimation process with many quantities to estimate, plus there is no easy way to obtain the standard error (SE) or confidence interval (CI) for both the CDE and NEs except using the bootstrap resampling method (Huber & Solovyeva 2020). To overcome these disadvantages, we extended Hong’s work on ratio-of-mediator-probability weighting (RMPW) (Hong 2010), VanderWeele’s MSM (VanderWeele 2009), and Lange’s unified approach for estimating NEs (Lange et al. 2012), to propose a new weighting method (weighting estimator II) to estimate both CDE and NEs for a tertiary outcome in the case-control sample drawn from a source population, e.g., a cohort. We will compare the performance between modified HS method and our proposed method using the simulation, followed by an application using real-world data.

3.4 Proposed Weighting Method

3.4.1 Nonparametric Identification

Similarly to Chapter 2, to identify the causal mediating effects for the population, we need to connect $E[Y^{xm}]$ with $E[Y|X = x, M = m, S = 1]$, and $E[Y^{xM^x}]$ and $E[Y^{xM^{x^*}}]$ with $E[Y|X = x, S = 1]$. Before showing how to nonparametrically identify the population mediating effects using case-control sample, we lay out assumptions which are required for the identification, based on the causal structure shown in Figure 3.1.

Assumption 1. $Y^{xm} \perp\!\!\!\perp (X, M) | \mathbf{C}$ where $\mathbf{C} = (C_1, C_2, C_3)$, i.e., no unobserved confounders between the treatment and the outcome, and no unobserved confounders between the mediator and the outcome.

Assumption 2. Conditional independency based on the study design and background knowledge.

$$Y \perp\!\!\!\perp S \mid (X, Z, M, \mathbf{C}) \quad (3.15)$$

$$S \perp\!\!\!\perp (X, \mathbf{C}) \mid Z \quad (3.16)$$

$$X \perp\!\!\!\perp (C_2, C_3) \mid C_1 \quad (3.17)$$

$$Y \perp\!\!\!\perp S \mid (X, M, \mathbf{C}) \quad (3.18)$$

$$S \perp\!\!\!\perp C_3 \mid (X, C_1, C_2) \quad (3.19)$$

$$0 < P(M = m \mid X, \mathbf{C}), 0 < P(M = m \mid X, \mathbf{C}, S = 1), 0 < P(S = 1 \mid Z) \\ 0 < P(S = 1 \mid X, C_1, C_2) \text{ and } 0 < P(X = x \mid C_1) \quad (3.20)$$

The conditional independence assumptions 3.15-3.19 are primarily design-based. For example, in a randomized controlled trial, instead of $Y \perp\!\!\!\perp S \mid (X, Z, M, \mathbf{C})$, we have $Y \perp\!\!\!\perp S$. For more details, please refer to the proof of Theorem 2 in Appendix A. Assumption 1 and 3.15-3.17 are required for identification of CDE. To identify NDE and NIE, 3.18, 3.19 and three additional assumptions are further required, as shown below:

Assumption 3. $(Y^{xM^x}, Y^{xM^{x^*}}) \perp\!\!\!\perp X \mid \mathbf{C}$.

Assumption 4. $Y^{xm} \perp\!\!\!\perp (M^x, M^{x^*}) \mid (X, \mathbf{C})$.

Assumption 5. $M^{x^*} \perp\!\!\!\perp X \mid (Z, \mathbf{C})$.

Next, we introduce 3.4.1, to show that the mean of population (nested) counterfactual outcomes can be written using weighted conditional density of the observed outcome Y .

Identification of population (nested) counterfactual outcomes using case-control sample.

$$E[Y^{xm}] = E[w_4 Y \mid X = x, M = m, S = 1] \text{ where } w_4 = \frac{P(M=m \mid x, S=1) \cdot P(S=1 \mid x) \cdot P(X=x)}{P(M=m \mid x, c_1, c_2, c_3) P(S=1 \mid z) P(X=x \mid c_1)} \cdot \\ E[Y^{xM^x}] = E[w_5 Y \mid X = x, S = 1] \text{ where } w_5 = \frac{P(M=m \mid c_1, c_2, c_3, x)}{P(M=m \mid x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1 \mid x) P(X=x)}{P(S=1 \mid x, c_1, c_2) P(X=x \mid c_1)} \cdot \\ E[Y^{xM^{x^*}}] = E[w_6 Y \mid X = x, S = 1] \text{ where } w_6 = \frac{P(M=m \mid c_1, c_2, c_3, x^*)}{P(M=m \mid x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1 \mid x) P(X=x)}{P(S=1 \mid x, c_1, c_2) P(X=x \mid c_1)} \cdot$$

3.4.2 Estimation of the Causal Estimand

We adopt the same MSM modeling-based approach as described in Chapter 2 (Robins et al. 2000, VanderWeele 2009, Lange et al. 2012). Following MSM will be used for estimating

CDE (3.21) and NEs (3.22):

$$g(E[Y^{xm}]) = c_0 + c_1x + c_2m + c_3x \cdot m \quad (3.21)$$

$$g(E[Y^{xM^{x^*}}]) = c_0 + c_1x + c_2x^* + c_3x \cdot x^* \quad (3.22)$$

where $g(\cdot)$ is the link function.

Assuming the functional form is correctly specified in 3.21 and 3.22, the causal mediating effects in terms of risk difference, relative risk and odds ratio can be expressed as a function of the parameters estimated from following MSM (Robins et al. 2000):

CDE can be estimated by replacing z by m in 2.17-2.19. NEs can be estimated using the similar model described in Chapter 2 (2.20-2.22). And w_4 is used for estimating CDE while w_5 and w_6 are used for estimating NEs. The interaction term $c_3x \cdot m$ or $c_3x \cdot x^*$ are added into above models to reflect the potential exposure-mediator interaction effect in relation to the final outcome, and if the exposure-covariate interactions are of interest, we can further include $x \cdot \mathbf{C}$ and $x^* \cdot \mathbf{C}$ to capture the (in)direct effects modified by covariates (Lange et al. 2012).

Similar to (Lange et al. 2012), when assuming a MSM for the nested counterfactual outcomes, our method can be equated with the generalized estimating equations (GEE) method. For example, the estimating equation corresponding to the MSM in 3.22 is given by:

$$U(c, \alpha) = \sum_{i=1}^n \sum_{x^*=0}^1 d(X_i, x^*) (Y_i - c_0 - c_1X_i - c_2a^* - c_3X_ia^*) \frac{P(M=m|c_1, c_2, c_3, x^*)}{P(M=m|x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1|x)P(X=x)}{P(S=1|x, c_1, c_2)P(X=x|c_1)} \text{ with } d(X_i, x^*) = (1, X_i, x^*, X_ix^*)^T.$$

As Lange et al. argued, let model $\mathcal{M}_1$ be the parametric model for $P(M|X, \mathbf{C})$ (with the unknown parameter denoted α) and $\mathcal{M}_2$ be the model defined by the restrictions of model $\mathcal{M}_1$ and the additional assumption that the generalized linear MSM 3.22 holds. According to Lange et al., the condition under which conservative SE will be obtained, is “that α is substituted by the maximum likelihood estimator under model $\mathcal{M}_1$ and that this

is also an efficient estimator of α under model $\mathcal{M}_2$ ". Although Lange et al. believe that the latter condition will almost always be (nearly) satisfied in practice, they recommend to obtain alternative SE (e.g., based on the bootstrap resampling method) if researchers are concerned about the validity of the additional condition, i.e., the generalized linear MSM is so overly restrictive that the restrictions it imposes on the outcome distribution carry information about the actual value of α .

Following from the M-estimation theory, we know that:

$$\sqrt{n}(\hat{c}_n - c_t) \xrightarrow{D} N(0, [E\{\frac{\partial U(c_t, \alpha_t)}{\partial c^T}\}]^{-1} \text{var}(U(c_t, \alpha_t)) [E\{\frac{\partial U(c_t, \alpha_t)}{\partial c^T}\}]^{-1^T})$$

where $\mathbf{c}_t$ and $\boldsymbol{\alpha}_t$ stand for the true values of vectors $(c_0, c_1, c_2, c_3)^T$ and $(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2)^T$. And $(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2)^T$ is from the generalized linear model of M on $(X, \mathbf{C})$ with link function $g(\cdot)$, i.e.,

$$E[M|X, \mathbf{C}] = g(\alpha_0 + \alpha_1 X + \boldsymbol{\alpha}_2 \mathbf{C}).$$

Since $\mathbf{c}_t$ and $\boldsymbol{\alpha}_t$ are unknown, we instead substitute the GEE estimator $\hat{\mathbf{c}}_n$ for $\mathbf{c}_t$ and MLE estimator $\hat{\boldsymbol{\alpha}}_n$ for $\boldsymbol{\alpha}_t$ and to obtain the sandwich estimator for the asymptotic variance specified above, given by

$$[\hat{E}\{\frac{\partial U(\hat{c}_n, \hat{\alpha}_n)}{\partial c^T}\}]^{-1} \hat{\text{var}}(U(\hat{c}_n, \hat{\alpha}_n)) [\hat{E}\{\frac{\partial U(\hat{c}_n, \hat{\alpha}_n)}{\partial c^T}\}]^{-1^T}.$$

So far, we have showed the validity of using MSM to estimate both CDE and NEs, including the sandwich estimators for SE. The estimates for mediating effects are similar to those described in Chapter 2.

3.5 Simulation

3.5.1 Data Generating Process (DGP)

Part of the parameters used in the DGP are same with those described in Chapter 2 section 2.5.1. The DGP for additional variables is described below:

1) Generate the study confounders:

Additional to C_1 and C_2, C_3 is a binary (0/1) confounder assessed at or after baseline, and $C_3 \sim \text{Bernoulli}(P_{C_3})$ and $P_{C_3} \approx 7.27\%$. An example of C_3 could be the comorbidity disease at the time of assessment of the secondary outcome (the mediator).

2) Generate the mediator (secondary outcome):

M is the mediator (secondary outcome), assessed at T_1 , with $M \sim \text{Bernoulli}(P_M)$, where

$$P_M = \{1 + \exp[-k_M]\}^{-1}$$

$$k_M = \log(1.079C_1) + \log(1.468C_2) + \log(1.439C_3) + \log(2.4X) + \log(4Z) + \log(0.001)$$

and $P(M = 1) \approx 14.90\%$.

3) Generate the observed (tertiary) outcome:

Y is the binary (0/1) outcome of interest assessed at T_2 , follow-up after T_1 and $Y \sim \text{Bernoulli}(P_Y)$, where

$$P_Y = \{1 + \exp[-k_Y]\}^{-1}$$

$$k_Y = \log(1.012C_1) + \log(1.187C_2) + \log(3.271C_3) + \log(1.1X) + \log(2.3M) \\ + \log(1.5C_3X) + \log(2XM) + \log(0.0503)$$

and $P(Y = 1) \approx 13.51\%$.

4) Generate the potential outcomes:

For the mediator M , $M^{x,Z^{x*}} \sim \text{Bernoulli}(P_{M^{x,Z^{x*}}})$, where $P_{M^{x,Z^{x*}}}$ is similar to P_M with replacing X and Z by x and counterfactual outcome Z^{x*} . For the outcome Y , $Y^{xM^{x*}} \sim \text{Bernoulli}(P_{Y^{xM^{x*}}})$, where $P_{Y^{xM^{x*}}}$ is similar to P_Y with replacing X and M by x and counterfactual outcome M^{x*} .

3.5.2 Monte Carlo, Bootstrapping and Performance Criteria

The Monte Carlo (MC) simulation, bootstrapping, and performance evaluation criteria methods are same as described in Chapter 2.

3.6 Simulation Results

3.6.1 Results Comparison Between Two Methods

Using 1,000 MC simulation with 200 bootstrap samples for each single run, and under different presumed cohort size (18,000, 36,491, or 72,000), we have the following result:

Table 3.2: Controlled Direct Effect Fixing $M = 0$

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	1.574x10 ⁻²	1.472x10 ⁻²	0.0212	0.9780	0.0207	94.8
	36491		1.568x10 ⁻²	0.0149	1.0016	0.0149	95.0
	72000		1.585x10 ⁻²	0.0106	1.0085	0.0107	94.2
Proposed	18000		1.566x10 ⁻²	0.0363	1.0025	0.0363	93.0
	36491		1.590x10 ⁻²	0.0255	1.0075	0.0257	93.4
	72000		1.677x10 ⁻²	0.0181	0.9958	0.0181	95.3
Relative Risk							
Huber and Solovyeva	18000	1.146	1.146	0.2100	0.9549	0.2004	94.6
	36491		1.150	0.1455	0.9949	0.1448	95.1
	72000		1.150	0.1024	1.0030	0.1027	94.1
Proposed	18000		1.167	0.3734	0.9648	0.3607	93.3
	36491		1.157	0.2509	0.9833	0.2468	93.7
	72000		1.162	0.1764	0.9864	0.1747	94.9
Odds Ratio							
Huber and Solovyeva	18000	1.167	1.170	0.2459	0.9462	0.2326	94.6
	36491		1.173	0.1695	0.9904	0.1680	95.0
	72000		1.172	0.1189	1.0003	0.1190	94.2
Proposed	18000		1.205	0.4541	0.9321	0.4247	93.4
	36491		1.186	0.2978	0.9684	0.2888	93.7
	72000		1.188	0.2069	0.9794	0.2037	95.0

^a This is the ratio between empirical and bootstrap standard error.

Table 3.3: Controlled Direct Effect Fixing $M = 1$

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	1.624x10 ⁻¹	1.628x10 ⁻¹	0.0394	1.0118	0.0399	94.7
	36491		1.627x10 ⁻¹	0.0275	0.9988	0.0275	94.6
	72000		1.626x10 ⁻¹	0.0196	1.0296	0.0202	94.1
Proposed	18000		1.599x10 ⁻¹	0.0906	1.0505	0.0951	93.0
	36491		1.650x10 ⁻¹	0.0642	1.0179	0.0654	94.2
	72000		1.626x10 ⁻¹	0.0461	0.9794	0.0451	95.1
Relative Risk							
Huber and Solovyeva	18000	1.758	1.783	0.2716	0.9641	0.2629	94.8
	36491		1.773	0.1823	0.9871	0.1805	94.2
	72000		1.767	0.1281	1.0081	0.1294	94.1
Proposed	18000		1.870	0.7979	0.8534	0.6899	93.5
	36491		1.836	0.4669	0.9475	0.4491	94.7
	72000		1.794	0.3052	0.9565	0.2941	95.2
Odds Ratio							
Huber and Solovyeva	18000	2.215	2.273	0.4867	0.9512	0.4663	94.7
	36491		2.246	0.3217	0.9819	0.3172	94.0
	72000		2.233	0.2248	1.0084	0.2273	93.9
Proposed	18000		2.500	1.6764	0.7804	1.3383	94.0
	36491		2.390	0.9009	0.9033	0.8320	94.4
	72000		2.292	0.5636	0.9263	0.5274	95.3

^a This is the ratio between empirical and bootstrap standard error.

From the Table 3.2 and 3.3, we can see that modified HS method generally has smaller bias, bootstrap SE and RMSE for the CDE. However, when increasing the cohort/sample size, the proposed method has a coverage probability that is closer to 95%, whereas the modified HS method doesn't.

Table 3.4: Natural Direct Effects

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	3.403x10 ⁻²	3.256x10 ⁻²	0.0189	0.9824	0.0186	94.2
	36491		3.325x10 ⁻²	0.0133	0.9979	0.0133	94.5
	72000		3.339x10 ⁻²	0.0094	1.0037	0.0095	94.3
Proposed	18000		3.425x10 ⁻²	0.0190	1.0192	0.0193	93.3
	36491		3.472x10 ⁻²	0.0133	1.0035	0.0134	93.4
	72000		3.512x10 ⁻²	0.0094	0.9822	0.0093	95.1
Relative Risk							
Huber and Solovyeva	18000	1.281	1.273	0.1691	0.9672	0.1637	94.6
	36491		1.276	0.1179	0.9929	0.1171	94.4
	72000		1.275	0.0833	0.9986	0.0833	93.8
Proposed	18000		1.290	0.1722	1.0086	0.1739	93.4
	36491		1.291	0.1197	1.0009	0.1202	93.3
	72000		1.292	0.0843	0.9808	0.0835	94.8
Odds Ratio							
Huber and Solovyeva	18000	1.332	1.326	0.2077	0.9600	0.1994	94.5
	36491		1.328	0.1441	0.9897	0.1426	94.5
	72000		1.327	0.1016	0.9967	0.1013	94.0
Proposed	18000		1.347	0.2119	0.9994	0.2122	93.2
	36491		1.346	0.1466	0.9965	0.1467	92.9
	72000		1.347	0.1029	0.9791	0.1018	94.8

^a This is the ratio between empirical and bootstrap standard error.

For NDE, Table 3.4 shows that the modified HS method and the proposed method

have overall similar performance while there are several notable observations. First, the modified HS method has overall slightly smaller bootstrap SE and RMSE (except for the risk difference scale at $N = 72,000$). Second, the true effect tends to be underestimated using the modified HS method while overestimated using the proposed method. Last, when increasing the cohort/sample size, the proposed method has a coverage probability that is closer to 95%, whereas the modified HS estimates does not.

Table 3.5: Natural Indirect Effects

Method	Cohort size	True	Point Estimate	SE _{bootstrap}	SE ratio ^a	RMSE	Coverage (%)
Risk Difference							
Huber and Solovyeva	18000	3.638x10 ⁻²	3.711x10 ⁻²	0.0055	1.2419	0.0068	88.4
	36491		3.709x10 ⁻²	0.0039	1.2719	0.0049	88.5
	72000		3.694x10 ⁻²	0.0027	1.2485	0.0035	89.0
Proposed	18000		3.689x10 ⁻²	0.0055	1.2413	0.0068	87.7
	36491		3.699x10 ⁻²	0.0038	1.3006	0.0050	86.1
	72000		3.679x10 ⁻²	0.0027	1.2971	0.0035	86.1
Relative Risk							
Huber and Solovyeva	18000	1.234	1.244	0.0521	1.0928	0.0577	91.3
	36491		1.241	0.0357	1.1574	0.0418	90.5
	72000		1.239	0.0250	1.1329	0.0286	90.2
Proposed	18000		1.242	0.0519	1.1279	0.0590	91.0
	36491		1.239	0.0354	1.1555	0.0412	90.5
	72000		1.237	0.0247	1.1686	0.0289	90.6
Odds Ratio							
Huber and Solovyeva	18000	1.290	1.301	0.0616	1.1194	0.0698	90.5
	36491		1.298	0.0423	1.1818	0.0506	90.3
	72000		1.295	0.0296	1.1564	0.0347	89.6
Proposed	18000		1.299	0.0613	1.1486	0.0710	90.3
	36491		1.297	0.0419	1.1843	0.0501	89.7
	72000		1.293	0.0293	1.1954	0.0351	89.8

^a This is the ratio between empirical and bootstrap standard error.

For NIE, Table 3.5 shows that two methods have overall similar results with a few interesting observations. First, the proposed method consistently has smaller bias compared to modified HS method. Second, the proposed method has slightly smaller bootstrap SE than modified HS method. Third, the proposed method has slightly larger RMSE compared to the modified HS method. Finally, both methods have poor but comparable coverages that are deviated from the 95%.

3.6.2 Uncertainty of the Estimator

Similarly to Chapter 2 2.6.2, we will discuss the SE for the NIE for our method in this section.

1) Distribution of the NIE estimates

Figure 3.2 shows the distribution of estimates for the NIE using the proposed method based on 1,000 MC simulation. The top two histograms are for the RD scale for $N = 18,000$ and $N = 72,000$ while the bottom two histograms are for the log-OR scale under two cohort sizes.

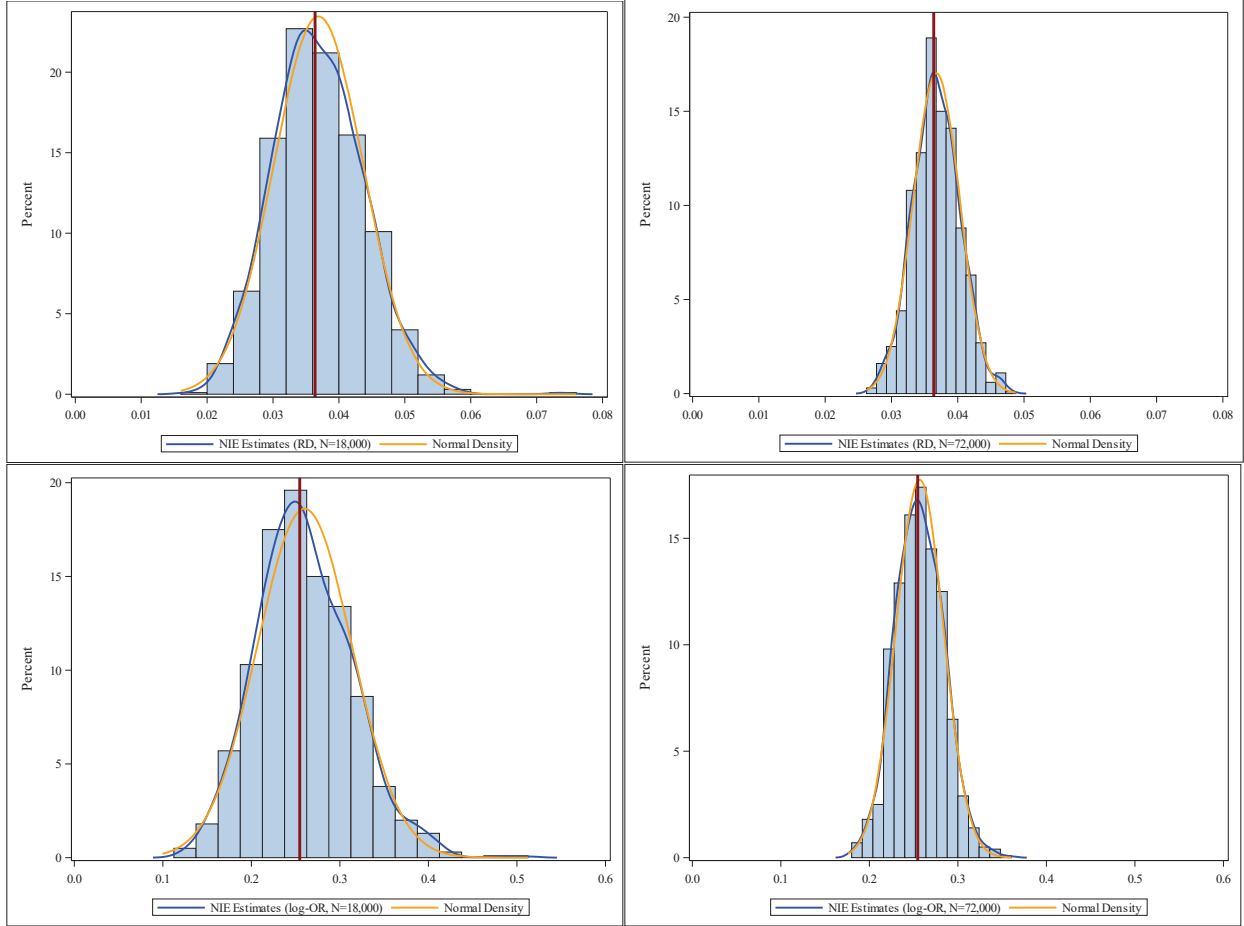


Figure 3.2: Distribution of NIE Estimates using Proposed Approach

From the visual check on the figure, we can see that the proposed approach gives estimates for NIE (blue curve) that is close to the normal density (orange curve). And when increasing the cohort size, the kernel density (blue curve) gets closer to the normal density (orange curve), which supports the asymptotic normality assumption.

Figure 3.3 shows the distribution of NIE estimates using modified HS method and

based on 1,000 MC simulation.

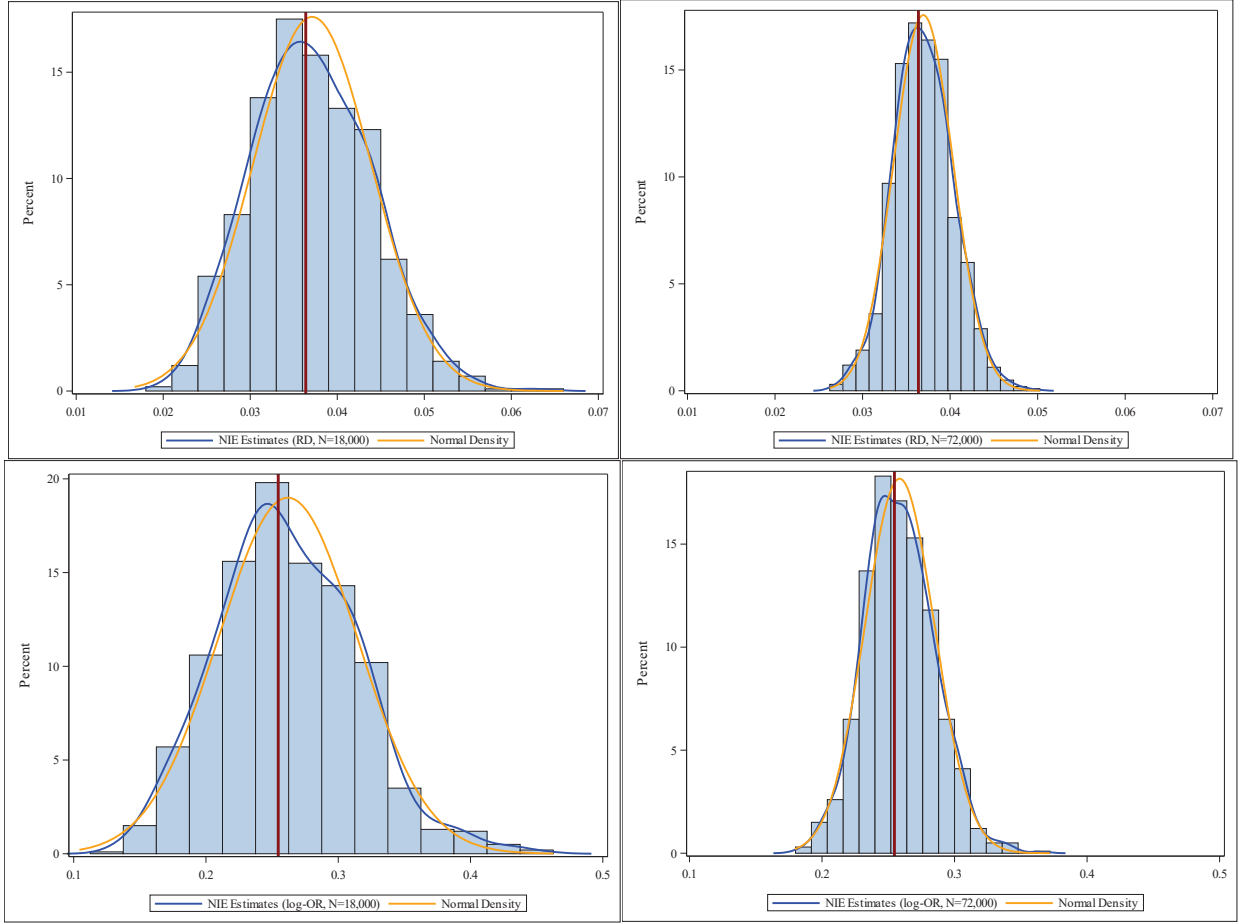


Figure 3.3: Distribution of NIE Estimates using Modified HS Method

From the visual check on the figure, we can see that the modified HS method gives NIE estimates (blue curve) that are also close to the normal density (orange curve). They point out that their original approach is $\sqrt{n}$ -consistent and asymptotically normal under standard regularity condition (Huber & Solov'yeva 2020). When increasing the cohort size from 18,000 to 72,000, we can see that the kernel density (blue curve) gets closer to with slight deviation from the the normal density (orange curve).

With these supporting evidence, we believe both our proposed method and the modified HS method give asymptotically normally distributed estimators for the NIE.

2) Robust standard errors and bootstrap standard errors

Figure 3.4 shows the distribution of robust and bootstrap SE for our proposed method under simulation settings from the section 3.5.2 and 2.5.3. The top two densities are for the RD scale for $N = 18,000$ and $N = 72,000$ while the bottom two are for the log-OR scale under two cohort sizes.

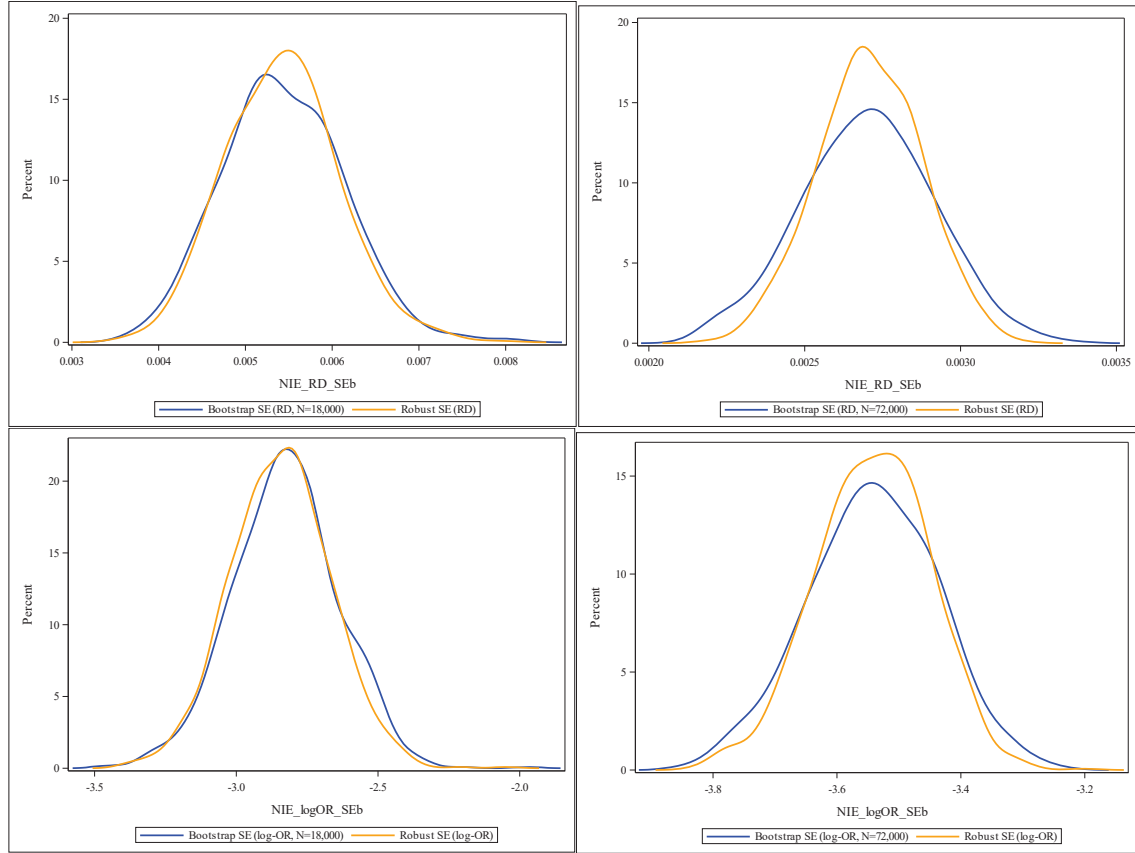


Figure 3.4: Distribution Comparison Between Robust and Bootstrap SE for Proposed Method

The distribution between robust and bootstrap SE for the NIE estimates are quite close to each other for both RD and log-OR scales. Since two methods give asymptotically normally distributed estimators, we believe both robust and bootstrap resampling based methods can be used to obtain the SE or the CI for the NIE.

3.7 Application: the Olfaction Sub-study of the Sister Study Cohort

In this section, we implement both the modified HS method and our proposed method using the NIEHS Sister study data. Details of the Sister study have been introduced in Chapter 2 2.7. In this example, we will use the same case-control study design as previously published (Cao et al. 2022), with the subsequently collected data as the tertiary outcome, to implement two methods for estimating both CDE and NEs. Briefly, during 2014-2016 (third detailed-follow-up, i.e., DFU-3, considered as baseline in this example), the Sister study collected data on whether the participants ever had a serious head injury that resulted in unconsciousness, coma, or hospitalization. And at the same visit, a stratified case-control sample was selected based on self-report sense of smell (100% sampling for self-report “poor” and ~3.56% simple random sampling for self-report “normal”) (Cao et al. 2022). The self-report sense of smell is a surrogate of the olfactory function at DFU-3. Later in 2018-2019, study investigators mailed out objective olfaction test kits (12-item Brief Smell Identification Test, B-SIT) to those who were selected in the case-control sample only and collected the secondary outcome, B-SIT-tested olfaction. At the Sister Study fourth detailed-follow-up (DFU-4, 2017-2019), a dementia screening interview was conducted using the AD-8 instrument (Galvin et al. 2006) and the sum score was calculated for each participant. Here, the self-report history of serious head injury is the exposure of interest. We treat the B-SIT tested olfactory status (secondary outcome) as the mediator of interest and we consider the cognitive function (tertiary outcome) collected at DFU-4 as the outcome of interest. Previous publications have suggested the link between the traumatic brain injury and olfactory impairment (Howell et al. 2018, Xydakis et al. 2015), head injury and dementia (Gu et al. 2022, Fann et al. 2018), and olfaction loss and dementia (Stanciu et al. 2014, Chen et al. 2021). We defined the serious head injury for subjects who answered “YES” to the corresponding question and defined the olfactory impairment using a cutoff of ≤ 9 for the B-SIT test score (Cao et al. 2022). The cognitive impairment/dementia was defined as AD-8 score ≥ 4 (Cai et al. 2021).

In this illustrative example, we are interested in studying the effects of a serious head

injury (X) on cognitive impairment/dementia (the tertiary outcome Y) mediated through the B-SIT tested olfaction (the secondary outcome M) while the primary outcome (case-control status) is self-report sense of smell (Z). Using the notation described in the section 3.2.1, we also consider the participants' age (continuous) at DFU-2 as the common cause for the exposure, the primary outcome, the mediator and the outcome, i.e. C_1 , the participants' race/ethnicity as the common cause for the primary outcome, the mediator and the outcome, i.e., C_2 , and the self-report overall health status in the past year assessed at DFU-3 as a surrogate for the the comorbidity disease (the confounder between the mediator and the outcome), i.e., C_3 . Table below summarizes the definition of variables used in this analysis.

Table 3.6: Variable Definition For The Sister Study Example

Variables	Definition	Variable Type	Distribution (N=33,742)	Time of Assessment
Confounder (C_1)	Age	Continuous	63.4 (7.3)	DFU-3 (2014-2016)
Confounder (C_2)	Race/Ethnicity	White (0)/Other (1)	$P(C_2 = 1) = 11.8\%$	Enrollment (2003-2009)
Confounder (C_3)	Self-report health at DFU-3	Good (0)/Fair-poor (1)	$P(C_3 = 1) = 6.0\%$	DFU-3 (2014-2016)
Exposure (X)	History of serious head injury	Yes (1)/No (0)	$P(X = 1) = 7.7\%$	DFU-3 (2014-2016)
Case-control status (Z)	Self-report olfaction	Poor (1)/Normal (0)	$P(Z = 1) = 5.1\%$	DFU-3 (2014-2016)
Mediator (M)	B-SIT tested olfaction (≤ 9)	Poor (1)/Normal (0)	Poor: 732 out of 2,502	2018-2019
Outcome (Y)	Cognitive impairment (AD-8 ≥ 4)	Yes (1)/No (0)	Yes: 371 out of 2,502	DFU-4 (2017-2019)

This analysis aims to illustrate the implementation of two methods for mediation analysis in the complex survey data and may not have epidemiological interpretation due to multiple reasons. First, the mediator is a surrogate for the true olfactory status but only measured at a single timepoint, plus the olfactory status is a multi-dimension assessment, e.g., discrimination, threshold and identification. Therefore, a single smell identification test may not reflect the true olfactory status at one particular time point. Third, the exposure is serious head injury history which is based on self-reports and may suffers from recall bias and it may also suffer from the measurement error compared to the definition of traumatic brain injury. Fourth, the choice of the cutoff for AD-8 in the definition of the cognitive impairment is higher than the common choice since the Sister cohort is made from relatively

healthy people. This may lead to the measurement error in the outcome. Fifth, there is a overlap in the data collection time period for the mediator (B-SIT tested olfaction) and outcome (cognitive function) which might not reflect the true causal relationship.

This example includes 33,742 out of 36,491 participants from the cohort who had non-missing values for the covariates $\mathbf{C}$, the exposure X , and the primary case-control outcome Z . Investigators used 100% sampling for cases and ~2.5% sampling for controls. The complete data has 2,502 participants in the case-control sample. Table 3.7 shows the estimates of the mediating effects in the RD, RR, and OR scales with 95% CIs:

Table 3.7: Method Comparison Using the Sister Study

Effect	Modified Huber & Solovyeva		Proposed		
	Point Estimate	Bootstrap 95% CI	Point Estimate	Bootstrap 95% CI	Robust 95% CI
Risk Difference					
CDE ₀	0.036	(-0.022, 0.099)	0.003	(-0.065, 0.084)	(-0.071, 0.077)
CDE ₁	0.065	(-0.030, 0.167)	-0.008	(-0.092, 0.119)	(-0.101, 0.086)
NDE	0.037	(-0.015, 0.092)	0.038	(-0.014, 0.093)	(-0.016, 0.092)
NIE	0.002	(-0.001, 0.005)	0.002	(-0.001, 0.005)	(-0.001, 0.005)
Relative Risk					
CDE ₀	1.28	(0.83, 1.80)	1.04	(0.26, 2.09)	(0.43, 2.49)
CDE ₁	1.40	(0.82, 2.12)	0.93	(0.37, 2.64)	(0.36, 2.39)
NDE	1.27	(0.89, 1.70)	1.28	(0.90, 1.71)	(0.93, 1.77)
NIE	1.01	(0.99, 1.03)	1.01	(0.99, 1.03)	(0.99, 1.03)
Odds Ratio					
CDE ₀	1.34	(0.81, 2.03)	1.04	(0.24, 2.30)	(0.40, 2.71)
CDE ₁	1.52	(0.79, 2.65)	0.92	(0.34, 3.03)	(0.32, 2.63)
NDE	1.33	(0.87, 1.90)	1.34	(0.88, 1.92)	(0.91, 1.98)
NIE	1.01	(0.99, 1.04)	1.01	(0.99, 1.04)	(0.99, 1.04)

Both methods give very close estimates on NEs. For CDE, two methods provide different point estimates and 95% CIs, regardless the choice of fixed value for the mediator M , i.e., the B-SIT tested olfaction status. Again, this is just an example of showing how to implement two methods to the real-world data and no clinical conclusion should be drawn from this analysis.

3.8 Conclusion

In Chapter 3, we discussed the estimation of mediating effects (controlled direct effect [CDE] and natural (in)direct effects [NEs]) using complex sampling data of a case-control sample drawn from the target population (e.g., a cohort) with a secondary as the mediator and a tertiary outcome as the outcome of interest. We first showed that ignoring the study design would lead to serious bias in NEs estimation. Second, we proposed a set of weighting

estimators and compared it with modified version of Huber and Solovyeva's work using the Monte Carlo simulation study. Overall, two methods have comparable performance while the proposed method is computationally less expensive. Third, we showed that the validity of using robust standard error for the proposed estimator in addition to the bootstrap resampling based standard error. Fourth, we implemented and compare two methods using real-world dataset for illustration purpose. Overall, we recommend to use more than one method in practice, to provide researchers more information and comprehensive results for clinical interpretation.

CHAPTER 4 Monte Carlo Based Statistical Power Analysis for Mediation Analysis Using Case-Control Studies With Secondary and Tertiary Outcomes

4.1 Background

Mediation analysis has exploded in past two decades, in both methodology and applications, especially for observational studies (Robins et al. 2000, Pearl 2001, VanderWeele 2009, Hong 2010, VanderWeele & Vansteelandt 2010, Lange et al. 2012, Richiardi et al. 2013, Tchetgen Tchetgen 2013, Valeri & Vanderweele 2013, Huber 2014, Hong 2015, VanderWeele 2015, Huber & Solovyeva 2020). However, less attention has been paid to the study design and sample size calculation for mediation analyses (VanderWeele 2020), primarily due to the fact that such analysis usually relies on crude approximation of the sample size (Rudolph et al. 2019). In such case, researchers may not be clear what is the minimum detectable mediating effects with particular statistical method and the study design they choose.

Two approaches have been used for calculating statistical power for the estimation of indirect effects, one based on generalized linear models (GLM) (Vittinghoff et al. 2009) and one based on simulation using the Baron and Kenny product estimator (Baron & Kenny 1986, Judd & Kenny 1981). Both approaches have significant limitations. The GLM method can only detect whether an indirect effect exists but cannot provide information on the power for a given effect size. The Baron and Kenny's method generally doesn't allow the exposure-mediator interaction and requires a linear relationship between the mediator and outcome. There are alternative methods for performing mediation analysis which allow for the potential exposure-mediator interactions and different functional forms between the outcome and the exposure, e.g., log- or logit-link (Lange et al. 2012, VanderWeele 2009, Huber & Solovyeva 2020). However, power analysis using such methods are rarely seen or discussed.

In addition to the statistical methods accounting for the exposure-mediator interaction and different functional forms, the complexity of the study design should also be taken into consideration. Mediation analysis has been implemented in different types of

observational studies, including cross-sectional, case-control and cohort studies (Rijnhart et al. 2021). However, little has been done for mediation analysis in a case-control study with subsequently collected (secondary and/or tertiary) outcomes. This may happen after a case-control study has been conducted but with additional interest of a secondary outcome and direct or indirect effects of exposure on the secondary outcome through the primary outcome. Similarly, subsequent (tertiary) outcomes can further be collected and the study interest is the mediating effects from the exposure to the tertiary outcome through a secondary outcome.

At the time of writing, there is no comprehensive publication discusses about the power analysis and sample size calculation using methods other than structural equation modelings. In fact, while most publications focus on the power analysis in mediation analysis using the Baron and Kenny’s approach (Fritz & MacKinnon 2007), there is no publication discussing the design of a mediation analysis involving complex sampling process with secondary (and tertiary) outcomes. Therefore, we would like to fill these two gaps in this chapter, by exploring the relationship between the effect size of the natural indirect effect and the statistical power from a recently proposed weighting method particularly for the mediation analysis in complex study designs with existence of secondary (and tertiary) outcomes, using simulation. With such study design, the exposure X is fully observed from the cohort. However, even if X is only observed within the case-control sample, using the approach discussed in Section 3.4-3.6 we can still estimate the mediating effects in the cohort (source population).

The first aim is to find the minimum detectable effect size for the natural indirect effect at a fixing target statistical power, i.e., 0.8, using a recently proposed weighting method. The second aim is to find the optimal case-to-control ratio with a fixed sample size to detect a given effect size.

In this chapter, we will first introduce study settings, notations, the data structure presented by causal directed acyclic graphs (DAG), a brief review of natural indirect effects,

the data generating process, Monte Carlo simulation, and the bootstrap resampling procedure. Next, we introduce the estimation method for the target causal parameter, i.e., the natural indirect effect, followed by simulation results. Lastly, we summarize our findings and draw the conclusion.

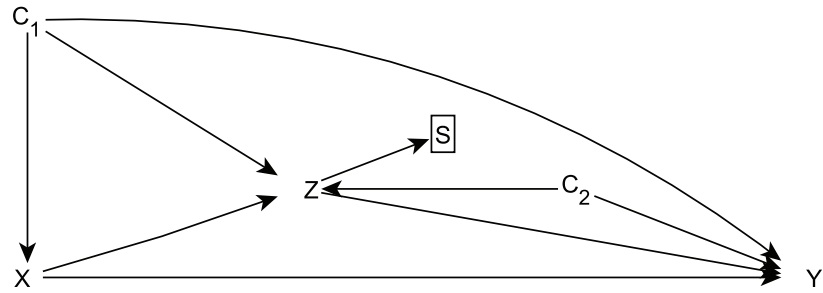
4.2 Settings

4.2.1 Notation and Causal DAG

In this chapter, we considered the mediation analysis using case-control samples drawn from a pre-defined target population, e.g., a cohort, under two different scenarios. The scenario 1 studies the mediating effect of the exposure to a secondary outcome through the variable for case-control sampling (primary outcome). The scenario 2 studies the mediating effect of the exposure to a tertiary outcome through the secondary outcome. Next, we introduce the variables involved in the two scenarios.

Formally, let C_1 be the exposure-mediator, exposure-outcome, and mediator-outcome confounder, e.g., the participants' age at the study baseline. Let C_2 be the common cause of the primary outcome (case-control status), the mediator and the outcome, e.g., race/ethnicity. Let C_3 be the mediator-outcome confounder assessed post-baseline, e.g., comorbidity disease at the time of mediator assessment, prior to the assessment of the outcome. C_1 and C_2 are assessed at the baseline (T_0). It is possible that C_3 is only assessed within the case-control sample (T_1). Let X be the exposure of interest assessed at baseline (T_0). Z is the variable used for selecting the case-control sample (0-control, 1-case). Let M be the secondary outcome which is assessed after collection of case-control sample (T_1) and available only within the case-control sample. We assume that the sample is selected from an existing cohort study. Let S be an indicator of case-control sample selection status, where " $S = 1$ " means the unit is selected and " $S = 0$ " means it is unselected (Didelez et al. 2010). Let Y be the tertiary outcome assessed after the measurement of the mediator (T_2). Figure 4.1 and 4.2 show the causal DAG for the data generating process for two scenarios, where the square outside S indicates the analysis is performed using observed data, i.e., the selected case-

control sample.



C_1 : confounder between exposure, mediator and outcome, e.g. participants' age at study baseline

C_2 : confounder between mediator and outcome, e.g. participant comorbidity

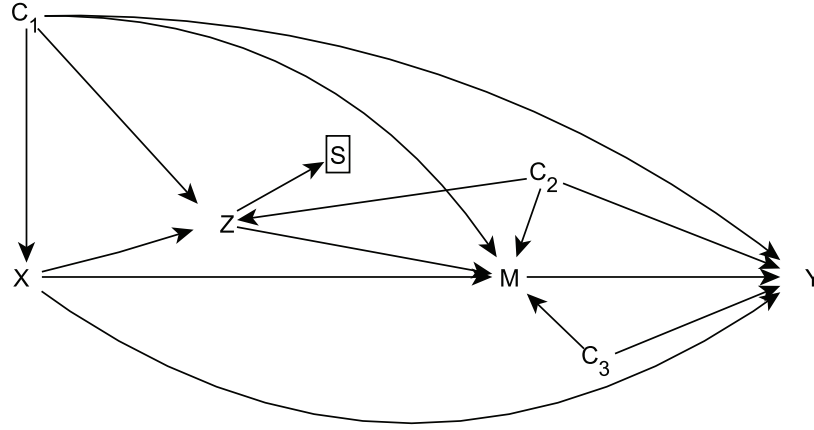
X : exposure/treatment

Z : case-control status (mediator)

S : case-control sample selection node

Y : (secondary) outcome

Figure 4.1: Causal DAG for Scenario 1



- C_1 : confounder between the exposure, the mediator and the outcome
 C_2 : common cause for primary outcome, the mediator and the outcome
 C_3 : confounder between the mediator and the outcome
 X : exposure/treatment assessed at T_0
 Z : case-control status assessed between T_0 and T_1
 M : secondary outcome (mediator) assessed at T_1
 S : case-control sample selection node
 Y : (tertiary) outcome assessed at T_2

Figure 4.2: Causal DAG for Scenario 2

For simplicity, assume C_2 , C_3 , X , Z , M and Y are all binary variables. In practice, they can be non-binary variables as well. If the case-control sample is selected from a cohort (source population) study as we assumed, then (C_1, C_2, X, Z, S) are observed for all participants whereas (C_3, M, Y) are only observed when $S = 1$, i.e., among subjects who are selected into the case-control sample.

4.2.2 Natural Indirect Effect (NIE)

Since NIE is generally more attractive than other causal effects (controlled direct effects and natural direct effects) in mediation analysis, we hereby focus on the NIE for this chapter. For binary mediator and outcome, the population NIE can be defined with two different

scales, separately for scenario 1 and 2 (Robins & Greenland 1992, Pearl 2001, VanderWeele 2015):

Scenario 1:

$$\text{Risk difference: } \text{NIE}_{\text{RD}} = E[Y^{1Z^1} - Y^{1Z^0}].$$

$$\text{Odds ratio: and } \text{NIE}_{\text{OR}} = \frac{P(Y^{1Z^1}=1)/P(Y^{1Z^1}=0)}{P(Y^{1Z^0}=1)/P(Y^{1Z^0}=0)}.$$

Scenario 2:

$$\text{Risk difference: } \text{NIE}_{\text{RD}} = E[Y^{1M^1} - Y^{1M^0}].$$

$$\text{Odds ratio: } \text{NIE}_{\text{OR}} = \frac{P(Y^{1M^1}=1)/P(Y^{1M^1}=0)}{P(Y^{1M^0}=1)/P(Y^{1M^0}=0)}.$$

4.2.3 Data Generating Process (DGP)

Since different case-control sampling ratios may affect the statistical power when fixing the sample size, we explore the power-sample size relationship for 10 different effect sizes of NIE with 4 different case-to-control sampling ratios (2:1, 1:1, 1:2 and 1:3). To achieve different effect sizes, we adopt different parameters to generate the data, as detailed below and in the Appendix B.

Scenario 1:

1) Generate the observed outcomes:

C_1 is a continuous confounder assessed at baseline (T_0), which is generated using the truncated normal distribution, $N(60.72, 54.05)$, with the interval $[44.1, 76.3]$. An example of C_1 could be the participants' age from the source population (cohort) at study baseline.

C_2 is a binary confounder assessed at baseline and $C_2 \sim \text{Bernoulli}(P_{C_2})$ and $P_{C_2} \approx 12.71\%$. An example of C_2 could be the race/ethnicity where 0 is the White and 1 is the Black.

X is a binary exposure of interest and $X \sim \text{Bernoulli}(P_X)$, where $P_X = \frac{1}{1 + \exp(-\beta_{C_1 X} C_1 - \beta_{0X})}$ and $P(X = 1) \approx 20\%$.

Z is a binary mediator (primary outcome) assessed at T_1 , later than baseline, with

$Z \sim \text{Bernoulli}(P_Z)$, where

$$P_Z = \{1 + \exp(-\beta_{c_1z}C_1 - \beta_{c_2z}C_2 - \beta_{xz}X - \beta_{0z})\}^{-1}$$

and $P(Z \approx 1) = 7.73\%$.

S is the indicator of being selected into the case-control sample (0-unselected and 1-selected), where for subjects with $Z = 1$, 100% is selected and for subjects with $Z = 0$, a simple random sampling is selected with sampling rate of 3.56%. This is very common in epidemiological studies where cases are completely selected while controls are selected with a small portion due to limited budget (Cao et al. 2022).

Y is the binary (0/1) outcome of interest and $Y \sim \text{Bernoulli}(P_Y)$, where

$$P_Y = \{1 + \exp(-k_Y)\}^{-1}$$

$$k_Y = \beta_{c_1y}C_1 + \beta_{c_2y}C_2 + \beta_{xy}X + \beta_{zy}Z + \beta_{c_1xy}C_1X + \beta_{c_2xy}C_2X + \beta_{xzy}XZ + \beta_{0y}$$

and $P(Y = 1) \approx 14.90\%$.

2) Generate the potential outcomes:

For the mediator Z , $Z^x \sim \text{Bernoulli}(P_{Z^x})$, where P_{Z^x} is similar to P_Z with replacing X by the actual value x . For the outcome Y , $Y^{xZ^{x*}} \sim \text{Bernoulli}(P_{Y^{xZ^{x*}}})$, where $P_{Y^{xZ^{x*}}}$ is similar to P_Y with replacing X and Z by x and counterfactual outcome Z^{x*} .

Scenario 2:

1) Generate the observed outcomes:

The DGP of C_1 , C_2 , X , Z , and S are exactly same as in scenario 1. Below are the DGP for three additional observed variables.

C_3 is a binary confounder assessed either at T_0 , between T_0 and T_1 , or at T_1 , and $C_3 \sim \text{Bernoulli}(P_{C_3})$ and $P_{C_3} \approx 7.27\%$. An example of C_3 could be the comorbidity disease at the time of assessment of the secondary outcome (the mediator).

M is the mediator (secondary outcome), assessed post-baseline (T_1), with $M \sim$

$Bernoulli(P_M)$, where

$$P_M = \{1 + \exp(-\beta_{c_1m}C_1 - \beta_{c_2m}C_2 - \beta_{c_3m}C_3 - \beta_{xm}X - \beta_{zm}Z - \beta_{0m})\}^{-1}$$

and $P(M = 1) \approx 14.91\%$.

Y is the binary (0/1) outcome of interest assessed at T_2 and $Y \sim Bernoulli(P_Y)$, where

$$P_Y = \{1 + \exp(-k_Y)\}^{-1}$$

$$k_Y = \beta_{c_1y}C_1 + \beta_{c_2y}C_2 + \beta_{c_3y}C_3 + \beta_{xy}X + \beta_{my}M + \beta_{c_3xy}C_3X + \beta_{xmy}XM - \beta_{0y}$$

and $P(Y = 1) \approx 13.51\%$.

2) Generate the potential outcomes:

For the primary outcome (case-control status) Z , $Z^x \sim Bernoulli(P_{Z^x})$, where P_{Z^x} is similar to P_Z with replacing X by the actual value x . For the mediator M , $M^{x,Z^{x*}} \sim Bernoulli(P_{M^{x,Z^{x*}}})$, where $P_{M^{x,Z^{x*}}}$ is similar to P_M with replacing X and Z by x and counterfactual outcome Z^{x*} . For the outcome Y , $Y^{xM^{x*}} \sim Bernoulli(P_{Y^{xM^{x*}}})$, where $P_{Y^{xM^{x*}}}$ is similar to P_Y with replacing X and M by x and counterfactual outcome M^{x*} .

To explore the relationship between different effect sizes for the NIE and statistical power across different case:control selection ratios, we manipulate primarily the coefficients for the exposure and mediator in the data generation process for the outcome Y . Meanwhile, we also modify the intercepts to ensure the desired population prevalence for Z , M and Y when applicable. Table 7.1 and 7.2 in Appendix show the choice of coefficients to generate data for scenario 1 and 2. Briefly, we start with no NIE, i.e., $NIE_{RD} = 0$ and $NIE_{OR} = 1$. And we keep increasing the NIE until the power reaches at least 90%.

4.3 Estimation Method

In Chapter 2 and 3, we proposed two weighting estimators to estimate the NIE for scenario 1 and 2 and we showed the validity to use both bootstrap resampling and the sandwich methods to estimate the standard error (SE) of proposed estimators. Additionally, an existing method (Huber & Solovyeva 2020) can be used to estimate the NEs in scenario 1 (HS method) and with some modification, it can be further used for scenario 2 as well (modified HS method). However, to estimate SE, (modified) HS method mainly relies on the bootstrap resampling method, which is computationally expensive. Therefore, we simulate the data with robust SE primarily, along with (modified) HS method with bootstrap SE for only 1 case-to-control ratio in each scenario, for comparison purpose.

We will lay out the brief introduction on estimation methods for two scenarios separately. For proof and more details, please refer to the Chapter 2, Chapter 3, and the Appendix A.

1) Scenario 1: primary case-control outcome as the mediator

Proposed method (weighting estimator I):

$$E[Y^{xZ^x}] = E[w_1 Y | X = x, S = 1] \text{ where } w_1 = \frac{P(S=1|X=x)P(X=x)}{P(S=1|Z=z)P(X=x|c_1)}.$$

$$E[Y^{xZ^{x^*}}] = E[w_2 Y | X = x, S = 1] \text{ where } w_2 = \frac{P(S=1|X=x)P(Z=z|X=x^*, c_1, c_2)P(X=x)}{P(S=1|Z=z)P(Z=z|X=x, c_1, c_2)P(X=x|c_1)}.$$

To estimate the NIE, we adopted MSM approach. The RD and OR can be expressed in terms of the parameters from MSMs with linear- or logit-link functions:

$$\text{linear model: } P(Y^{xZ^{x^*}} = 1) = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (4.1)$$

$$\text{logistic model: } \text{logit}[P(Y^{xZ^{x^*}} = 1)] = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (4.2)$$

And it is not difficult to show:

$$\hat{\text{NIE}}_{\text{RD}} = \hat{c}_2(x - x^*) + \hat{c}_3 x(x - x^*) \text{ where } (\hat{c}_2, \hat{c}_3) \text{ is estimated from the model 4.1.}$$

$$\hat{\text{NIE}}_{\text{OR}} = e^{\hat{c}_2(x-x^*) + \hat{c}_3 x(x-x^*)} \text{ where } (\hat{c}_2, \hat{c}_3) \text{ is estimated from the model 4.2.}$$

Huber & Solovyeva's (HS) method:

Taking risk difference as an illustrative example as shown in the original publication (Huber & Solovyeva 2020), we have

$$\hat{NIE}_{RD} = \hat{E}[Y^{1Z^1}] - \hat{E}[Y^{1Z^0}],$$

$$\text{where } \hat{E}[Y^{1Z^1}] = \frac{1}{n} \sum_i \frac{Y_i \cdot X_i \cdot S_i}{P(X_i=1|C_{1i}) \cdot P(S_i=1|Z_i)} / \frac{1}{n} \sum_i \frac{X_i \cdot S_i}{P(X_i=1|C_{1i}) \cdot P(S_i=1|Z_i)},$$

$$\text{and } \hat{E}[Y^{1Z^0}] = \frac{1}{n} \sum_i \frac{Y_i \cdot (1-X_i) \cdot S_i \cdot P(X_i=0|Z_i, C_{1i}, C_{2i})}{P(X_i=0|C_{1i}) \cdot P(X_i=1|Z_i, C_{1i}, C_{2i}) \cdot P(S_i=1|Z_i)} / \frac{1}{n} \sum_i \frac{(1-X_i) \cdot S_i \cdot P(X_i=0|Z_i, C_{1i}, C_{2i})}{P(X_i=0|C_{1i}) \cdot P(X_i=1|Z_i, C_{1i}, C_{2i}) \cdot P(S_i=1|Z_i)}.$$

1) Scenario 2: secondary outcome as the mediator

Proposed method (weighting estimator II):

$$E[Y^{xM^x}] = E[w_3 Y | S = 1] \text{ where } w_3 = \frac{P(M=m|c_1, c_2, c_3, x)}{P(M=m|x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1|x)P(X=x)}{P(S=1|x, c_1, c_2)P(X=x|c_1)}.$$

$$E[Y^{xM^{x^*}}] = E[w_4 Y | S = 1] \text{ where } w_4 = \frac{P(M=m|c_1, c_2, c_3, x^*)}{P(M=m|x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1|x)P(X=x)}{P(S=1|x, c_1, c_2)P(X=x|c_1)}.$$

Similarly, we use the following MSMs to estimate the NIE:

$$\text{linear model: } P(Y^{xM^{x^*}} = 1) = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (4.3)$$

$$\text{logistic model: } \text{logit}[P(Y^{xM^{x^*}} = 1)] = c_0 + c_1 x + c_2 x^* + c_3 x \cdot x^* \quad (4.4)$$

And we have:

$$\hat{NIE}_{RD} = \hat{c}_2(x - x^*) + \hat{c}_3 x(x - x^*) \text{ where } (\hat{c}_2, \hat{c}_3) \text{ is estimated from the model 4.3.}$$

$$\hat{NIE}_{OR} = e^{\hat{c}_2(x-x^*) + \hat{c}_3 x(x-x^*)} \text{ where } (\hat{c}_2, \hat{c}_3) \text{ is estimated from the model 4.4.}$$

Modified Huber & Solovyeva's (HS) method:

Taking risk difference as example, we have

$$\hat{NIE}_{RD} = \hat{E}[Y^{1M^1}] - \hat{E}[Y^{1M^0}], \text{ where}$$

$$\hat{E}[Y^{1M^1}] = \frac{1}{n} \sum_i \frac{Y_i \cdot X_i \cdot S_i \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i})}{P(X_i=1|C_{1i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i}, S_i=1) \cdot P(S_i=1|X_i, C_{1i}, C_{2i})} / s_1 \text{ and}$$

$$s_1 = \frac{1}{n} \sum_i \frac{X_i \cdot S_i \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i})}{P(X_i=1|C_{1i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i}, S_i=1) \cdot P(S_i=1|X_i, C_{1i}, C_{2i})}.$$

$$\hat{E}[Y^{1M^0}] = \frac{1}{n} \sum_i \frac{Y_i \cdot (1-X_i) \cdot S_i \cdot P(X_i=0|M_i, C_{1i}, C_{2i}, C_{3i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i})}{P(X_i=0|C_{1i}, C_{2i}, C_{3i}) \cdot P(X_i=1|M_i, C_{1i}, C_{2i}, C_{3i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i}, S_i=1) \cdot P(S_i=1|X_i, C_{1i}, C_{2i})} / s_2 \text{ and}$$

$$s_2 = \frac{1}{n} \sum_i \frac{(1-X_i) \cdot S_i \cdot P(X_i=0|M_i, C_{1i}, C_{2i}, C_{3i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i})}{P(X_i=0|C_{1i}, C_{2i}, C_{3i}) \cdot P(X_i=1|M_i, C_{1i}, C_{2i}, C_{3i}) \cdot P(M_i=m|X_i, C_{1i}, C_{2i}, C_{3i}, S_i=1) \cdot P(S_i=1|X_i, C_{1i}, C_{2i})} / s_2.$$

4.4 Simulation Studies

4.4.1 Monte Carlo (MC) simulation

The MC simulation follows these steps:

1. Since exposure-covariate and exposure-mediator interaction are assumed, it would be cumbersome to derive the true NIE, especially for the logit-link function. Thus, the true NIE is calculated from a super population ($N_0 = 2,000,000$) based on the DGP specified in the section 4.2.3.
2. Generate a cohort ($N = 36,491$) and four case-control samples with fixed sample size ($n = 4,000$) and different case-to-control ratios, 2 : 1, 1 : 1, 1 : 2 and 1 : 3. Both cases and controls are drawn from the generated cohort ($N = 36,491$) using stratified simple random sampling with different sampling rates within the strata defined by the primary case-control outcome.
3. Estimate the NIE within the selected samples using either proposed or (modified) HS method.
4. Test the statistical significance based on either robust SE from one single MSM (proposed method) or the bootstrap (HS method) confidence intervals (CI) with $\alpha = 0.05$. The details of calculating bootstrap (CI) will be described below.
5. Repeat step 2 to step 4 for $R = 1,000$ (number of MC runs) times, with r out of R runs show statistical significance. The power of the analysis is calculated as $1 - \beta = \frac{r}{R}$.

4.4.2 Bootstrap Resampling

Previously, we have shown the validity of using robust SE with weighting estimators I and II. For (modified) HS method, they recommend to use bootstrap SE since the estimation of the NIE is by separately estimating two quantities instead of from one single model. Below are two-step bootstrap resampling method used to get the bootstrap SE and the quantile-based 95% CI:

1. Starting from the case-control sample at the r -th run of the MC simulation described in the step 2 in the section 4.4.1, we use stratified bootstrap sampling to ensure case-

control ratio. Specifically, assuming among the r -th case-control sample, there is n_r subjects with n_{r0} subjects for $Z = 0$ and n_{r1} subjects for $Z = 1$. For each bootstrap sample b , we resample n_{r0b} from n_{r0} and n_{r1b} from n_{r1} with unrestricted random sampling such that $n_{r0b} + n_{r1b} = n_r$, the predetermined sample size (4000) by various case-control ratios. The unrestricted random sampling assigns equal probability for each individual in the same stratum and with replacement.

2. For each Monte Carlo run r , we run B bootstrap sampling runs ($B = 500$) and get the percentile bootstrap confidence interval and the bootstrap SE (Zhang 2014). Specifically, the bootstrap SE is calculated using the following formula (Chernick & LaBudde 2011, Efron 1982): $SE_b = \sqrt{\frac{1}{B-1} \sum_b (\hat{\theta}_b^* - \hat{\theta}^*)^2}$, where $\hat{\theta}_b^*$ is the causal effect, i.e., NIE, estimated from the b -th bootstrap sample and $\hat{\theta}^* = \frac{\sum_b \hat{\theta}_b^*}{B}$.

4.5 Results

4.5.1 Power of the proposed methods

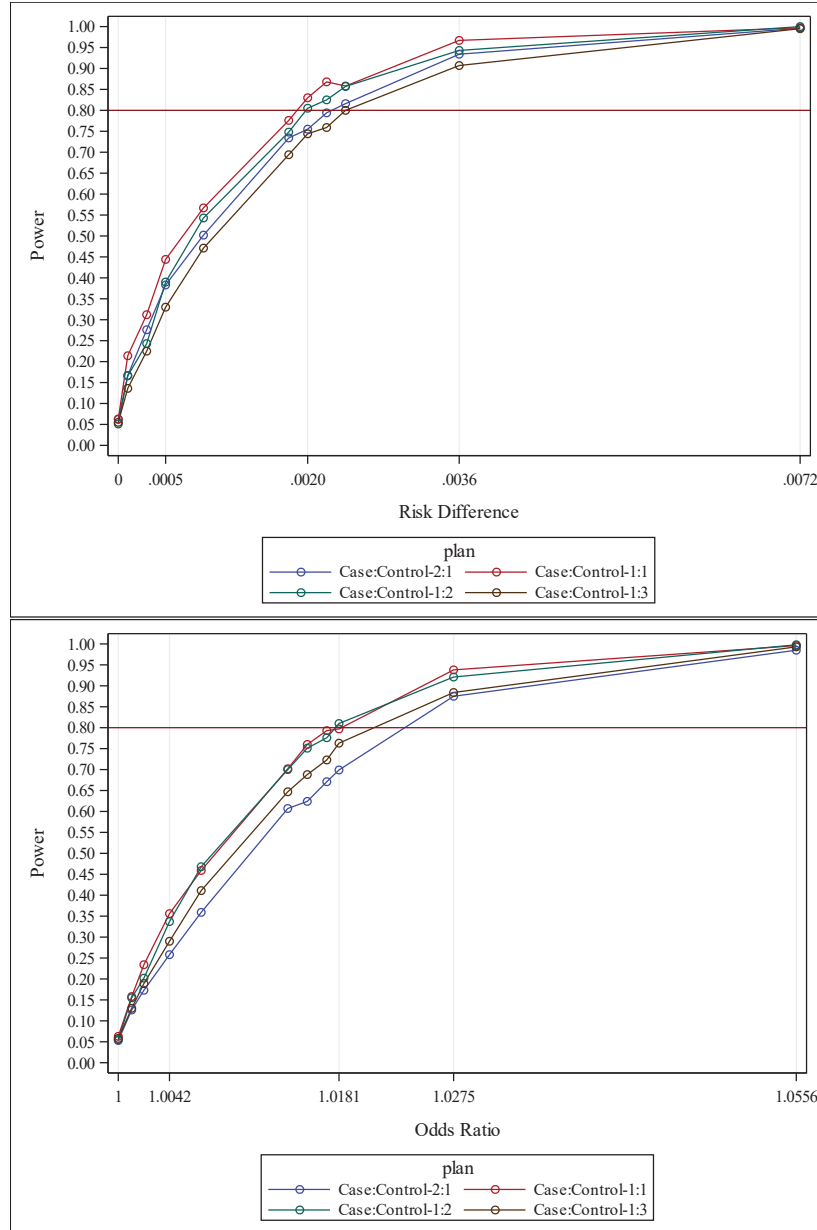


Table 4.1: Scenario 1 power analysis for natural indirect effects in risk difference scale (upper panel) and odds ratio scale (lower panel) using weighting estimator I with marginal structural models and robust standard errors

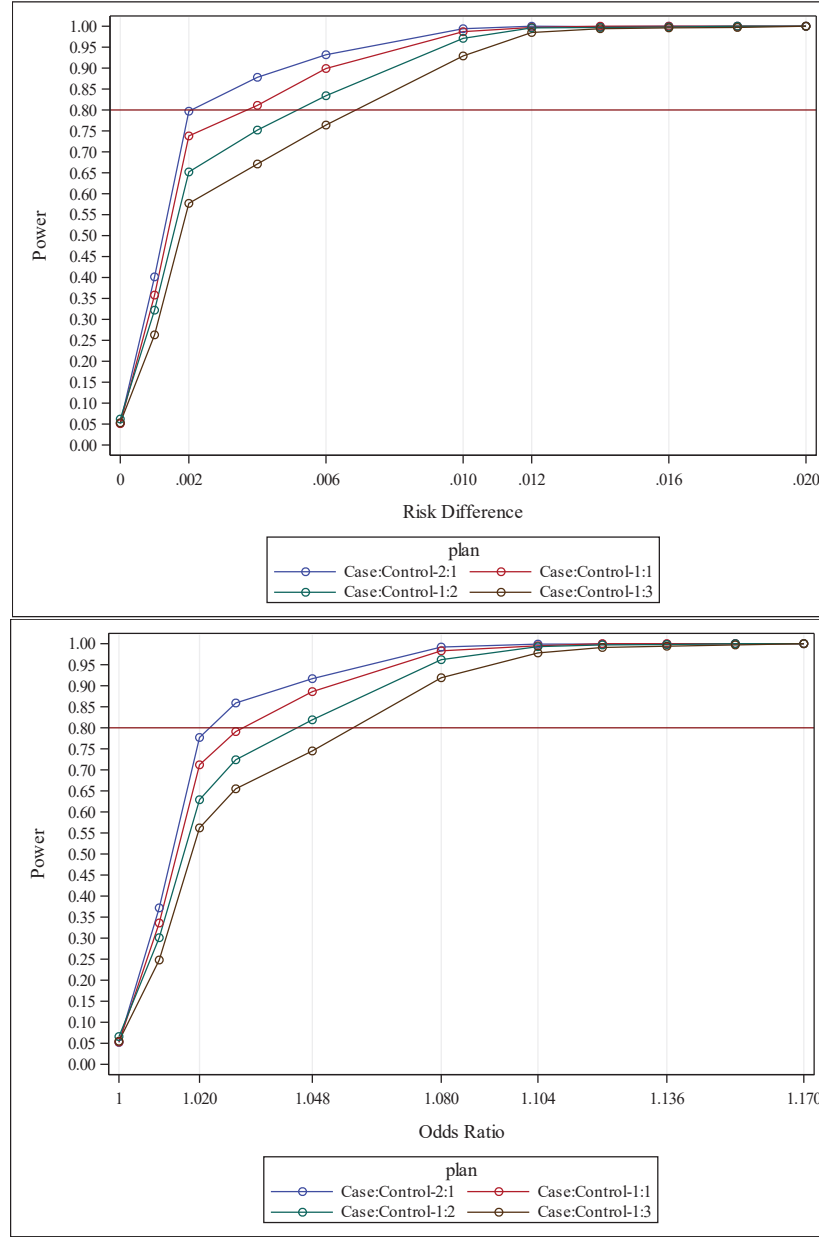


Table 4.2: Scenario 2 power analysis for natural indirect effects in risk difference scale (upper panel) and odds ratio scale (lower panel) using weighting estimator II with marginal structural models and robust standard errors

For scenario 1 (Table 4.1), there are several interesting observations. First, the statistical power of the analysis increases as the true effect (risk difference and odds ratio) gets larger, which is expected. Second, when fixing the sample size at 4000, both 2 : 1 and

1 : 1 have higher statistical power than the other two case-to-control ratios for odds ratio scale. For the risk difference scale, it seems that 1 : 1 is the optimal case-to-control ratio in terms of having a higher statistical power. Third, with a target statistical power of 0.8, we can see the minimum detectable NIE is about 0.0020 (risk difference) and 1.0181 (odds ratio), with the optimal case-to-control ratio of 2 : 1 or 1 : 1.

For scenario 2 (Table 4.2), the statistical power of the analysis also increases as the true effect (risk difference and odds ratio) gets larger. However, we found that the case-to-control ratio of 2 : 1 has the highest power regardless the effect measure scale. Under this case-to-control ratio and with a target statistical power of 0.8, the minimum detectable NIE is about 0.002 (risk difference) and 1.025 (odds ratio). The minimum detectable effect (OR) for scenario 2 is slightly larger than that in scenario 1 and this could be due to the estimation of mediator model using the case-control sample at scenario 2.

4.5.2 Comparison between the proposed methods and the HS method

We also compared (modified) HS method using bootstrap resampling based method with our proposed method (weighting estimator I and II) using robust SE under five different effect sizes for each scenario (Table 4.3).

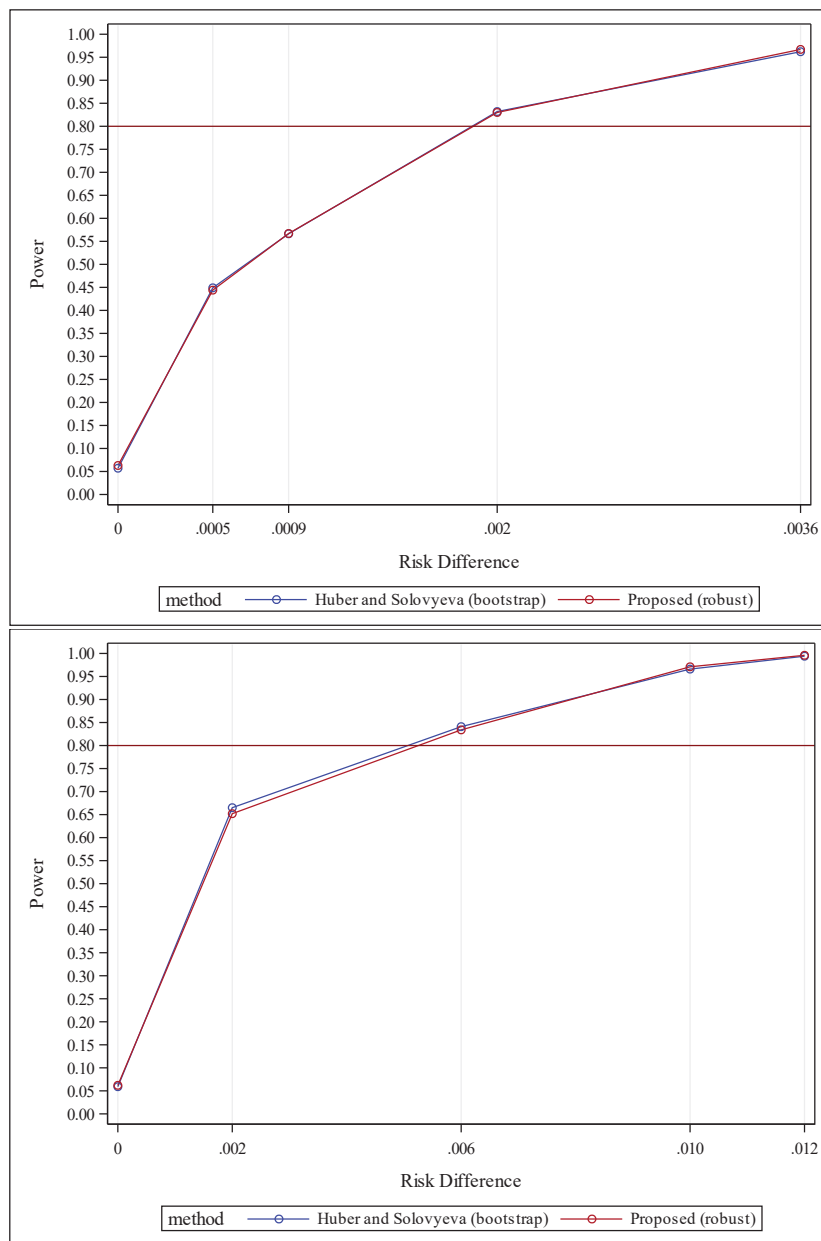


Table 4.3: Power analysis comparison for natural indirect effects for scenario 1 in which the primary case-control outcome is the mediator (upper panel) and scenario 2 in which the secondary outcome is the mediator (lower panel) in risk difference scale

We found that for scenario 1, both HS method and the proposed weighting estimator I have almost the same power across five different effect sizes (risk difference). Both methods have inflated type-I error rate (0.057 for HS method and 0.063 for weighting estimator I).

For scenario 2, we observed that modified HS method has slightly higher but overall close power than proposed weighting estimator II (risk difference). Also, both methods have inflated type-I error rate (0.059 for modified HS method and 0.062 for weighting estimator II). However, HS method relies on bootstrap resampling method which is computationally expensive. Our proposed MSM-based approach dramatically reduces the amount of computation regardless the small difference in minimum detectable effect. The price of large amount of computation may not be an issue in analyzing a real-world dataset but it would be costly when designing a study, especially for researchers to explore different scenarios such as the presence of the exposure-mediator interaction, sensitivity analysis, etc. Therefore, our proposed method will be more preferable for such situations.

4.6 Conclusion

In Chapter 4, we discussed the optimal case-to-control sampling ratio and the minimum detectable effect size for the natural indirect effects in a complex survey data of a case-control sample drawn from the target population (e.g., a cohort) with presence of a secondary (and tertiary) outcome(s). We first showed results using robust standard errors estimated from the weighted marginal structural model of our proposed methods (weighting estimator I and II), under different presumed effect sizes. Second, we compared results from bootstrap resampling method of a (modified) existing approach with our proposed method. Overall, using robust standard errors with the proposed methods could significantly reduce the time of simulation and performed very well when designing studies for mediation analysis.

CHAPTER 5 Conclusions

In this dissertation, we consider the mediating effects, i.e., the CDE and NEs, in case-control studies under two scenarios. In Chapter 2 (scenario 1), we are interested in the mediating effects from an exposure to a secondary outcome through the primary case-control outcome as the mediator where the exposure and mediator are observed through the whole target population (e.g., a cohort) but the secondary outcome is only observed within the case-control sample only. In Chapter 3 (scenario 2), we are interested in the mediating effects from an exposure to a tertiary outcome through a secondary outcome (mediator) where both the mediator and the outcome of interest are observed only within the case-control sample. We additionally explore the optimal case-to-control ratio and the minimum detectable effect size in RD and OR scales for both scenarios.

Using the Neyman-Rubin counterfactual outcome causal framework, we compare two existing methods, i.e., VanderWeele’s MSM (VanderWeele 2009) and Huber and Solovyeva’s weighting approaches (Huber & Solovyeva 2020), with our proposed weighting estimators which is an extension of Hong’s ratio-of-mediator-probability weighting (RMPW) method (Hong 2010, Lange et al. 2012, Hong 2015). When using the VanderWeele’s MSM approach and ignoring the complex study design, although the CDE can be unbiasedly estimated, directly applying his method lead to serious bias in estimation of the NEs. When only the secondary outcome is involved, HS method suffers from two main disadvantages. First, their estimation of the weight involves modeling the treatment as a function of the mediator and covariates, which does not have immediate substantive interpretations given that the treatment causally precedes rather than succeeds the mediator (Hong 2015). Second, their method relies on the bootstrap resampling method in estimating the SEs of the estimator, which is computationally expensive, especially at study design phase. When both the secondary and tertiary outcomes are involved, HS method cannot be directly applied since that the probability of being selected into the case-control sample given the exposure, mediator and covariates is not estimable under the particular study design. After some modification

to fit the scenario, their (modified) approach still suffers from above two disadvantages.

Our proposed method provides a solution in mediating effect estimation under complex study designs which can unbiasedly estimate both the CDE and NEs under two particular scenarios that are expected to be commonly seen in epidemiological studies. Extending Hong’s RMPW approach with the MSM estimation proposed by Lange et al. (Lange et al. 2012), our proposed method overcomes two disadvantages mentioned above. The estimation of the weight involves both sampling selection process and the mediator model and propensity score. For the estimation of SE, the robust/sandwich estimator from the MSM can be easily estimated using standard statistical software.

This dissertation work provides several meaningful insights. First, as far as we know, this is one of the first work that studies mediation analysis under complex study designs. Second, our proposed method combines the MSM and RMPW method, which is easily to implement in practical with standard statistical software and requires relatively fewer identification assumptions comparing with existing methods (Hong 2015). Third, comparing with the only available weighting-based method (Huber & Solovyeva 2020) that can fit with two scenarios, our proposed method is more interpretable in estimation of the weight and easier to estimate the SE with the help of MSM. Fourth, for the natural indirect effect, the proposed method has smaller bias and more stable when both the secondary and tertiary outcomes exist. Fifth, the proposed method can be easily used to design a mediation analysis study, even though with a relatively small target effect size (risk difference and odds ratio). Sixth, we also provide the optimal case-control ratio for performing and designing a mediation analysis using the proposed approach, which could further help researchers reduce the cost.

Although our proposed approach has many advantages, this work could still be extended in a number of directions. First, our discussion primarily focus on the estimation of the CDE and NEs with limited exploration on the SEs. Although a stratified bootstrap resampling method and the robust SE can be used, the weighting-based methods are usually criticized with less statistical efficiency. Therefore, different bootstrap resampling approaches

and other methods for estimating SEs could also be explored to increase the statistical efficiency. Second, other methods under complex study designs could be developed (e.g., semi-parametric estimators) to examine issues of efficiency and robustness to violations in identification assumptions. Third, we haven't considered the missing data issue which is very common in complex study designs. It might be worth to study how different missingness mechanisms would affect the estimation of mediating effects and the potential remedies for this. Additionally, there are several notable observations in our work that might require further exploration. For scenario 1, the coverage probability of the proposed method in estimating the natural indirect effect is relatively low (lower than 95%) with risk difference scale and it is not converging to 95% for HS method when increasing the sample size. For scenario 2, the coverage probability is generally lower than 95% in estimating the natural indirect effect for both our proposed and HS method.

BIBLIOGRAPHY

- Albert, J. M. (2012), ‘Distribution-free mediation analysis for nonlinear models with confounding’, *Epidemiology* **23**(6), 879–888.
- Alwin, D. F. & Hauser, R. M. (1975), ‘The decomposition of effects in path analysis’, *American Sociological Review* **40**(1), 37–47.
- 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.
- Cai, Y., Qiu, P., Wan, Y., Meng, S. S., Liu, T., Wang, Y., Rao, S. & Kuang, W. (2021), ‘Establishing cut-off scores for the self-rating AD8 based on education level’, *Geriatric Nursing* **42**(5), 1093–1098.
- Cao, Z., Yang, A., D’Aloisio, A. A., Suarez, L., Deming-Halverson, S., Li, C., Luo, Z., Pinto, J. M., Werder, E. J., Sandler, D. P. & Chen, H. (2022), ‘Assessment of self-reported sense of smell, objective testing, and associated factors in middle-aged and older women’, *JAMA Otolaryngology–Head & Neck Surgery* **148**(5), 408–417.
- Chen, Z., Xie, H., Yao, L. & Wei, Y. (2021), ‘Olfactory impairment and the risk of cognitive decline and dementia in older adults: A meta-analysis’, *Brazilian Journal of Otorhinolaryngology* **87**(1), 94–102.
- Chernick, M. R. & LaBudde, R. A. (2011), *An Introduction to Bootstrap Methods with Applications to R*, 1st edition edn, Wiley, Hoboken, N.J.
- Coffman, D. L., Schuler, M. S., Nguyen, T. Q. & McCaffrey, D. F. (2023), Weighting estimators for causal mediation, in ‘Handbook of Matching and Weighting Adjustments for Causal Inference’, CRC.
- Didelez, V., Kreiner, S. & Keiding, N. (2010), ‘Graphical models for inference under outcome-dependent sampling’, *Statistical Science* **25**(3), 368–387.
- Duncan, O. D. (1966), ‘Path analysis: Sociological examples’, *American Journal of Sociology* **72**(1), 1–16.
- Efron, B. (1982), The bootstrap, in ‘The Jackknife, the Bootstrap and Other Resampling Plans’, Society for Industrial and Applied Mathematics, pp. 27–36.
- Fann, J. R., Ribe, A. R., Pedersen, H. S., Fenger-Grøn, M., Christensen, J., Benros, M. E. & Vestergaard, M. (2018), ‘Long-term risk of dementia among people with traumatic brain injury in denmark: a population-based observational cohort study’, *The Lancet Psychiatry* **5**(5), 424–431.
- Finney, J. M. (1972), ‘Indirect effects in path analysis’, *Sociological Methods & Research* **1**(2), 175–186.

- Fritz, M. S. & MacKinnon, D. P. (2007), ‘Required sample size to detect the mediated effect’, *Psychological Science* **18**(3), 233–239.
- Galvin, J. E., Roe, C. M., Xiong, C. & Morris, J. C. (2006), ‘Validity and reliability of the AD8 informant interview in dementia’, *Neurology* **67**(11), 1942–1948.
- Greenland, S., Pearl, J. & Robins, J. M. (1999), ‘Causal diagrams for epidemiologic research’, *Epidemiology* **10**(1), 37–48.
- Gu, D., Ou, S. & Liu, G. (2022), ‘Traumatic brain injury and risk of dementia and alzheimer’s disease: A systematic review and meta-analysis’, *Neuroepidemiology* **56**(1), 4–16.
- Gustavo Amorim, Alastair J. Scott & Chris J. Wild (2017), Multi-Phase Sampling, in ‘Handbook of Statistical Methods for Case-Control Studies’, Chapman and Hall/CRC. Num Pages: 20.
- Hafdahl, A. R. (2010), ‘Random-effects meta-analysis of correlations: Monte carlo evaluation of mean estimators’, *British Journal of Mathematical and Statistical Psychology* **63**(1), 227–254.
- Han, J., Colditz, G. A. & Hunter, D. J. (2006), ‘Risk factors for skin cancers: A nested case-control study within the Nurses’ Health Study’, *International Journal of Epidemiology* **35**(6), 1514–1521.
- Heckman, J. J. & Vytlacil, E. J. (2007), Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation, in J. J. Heckman & E. E. Leamer, eds, ‘Handbook of Econometrics’, Vol. 6, Elsevier, pp. 4779–4874.
- Hernan, M. A. & Robins, J. M. (2020), *Causal Inference: What If*, CRC, Boca Raton, FL.
- Holland, P. W. (1988), ‘Causal inference, path analysis, and recursive structural equations models’, *Sociological Methodology* **18**, 449–484.
- Hong, G. (2010), ‘Ratio of mediator probability weighting for estimating natural direct and indirect effects’, p. 15.
- Hong, G. (2015), *Causality in a Social World: Moderation, Mediation and Spill-over*, Wiley, Chichester, England.
- Hong, G., Deutsch, J. & Hill, H. D. (2015), ‘Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction’, *Journal of Educational and Behavioral Statistics* **40**(3), 307–340.
- Hong, G. & Nomi, T. (2012), ‘Weighting methods for assessing policy effects mediated by peer change’, *Journal of Research on Educational Effectiveness* **5**(3), 261–289.
- Howell, J., Costanzo, R. M. & Reiter, E. R. (2018), ‘Head trauma and olfactory function’, *World Journal of Otorhinolaryngology - Head and Neck Surgery* **4**(1), 39–45.

- Huber, M. (2014), ‘Identifying causal mechanisms (primarily) based on inverse probability weighting’, *Journal of Applied Econometrics* **29**(6), 920–943.
- Huber, M. & Solovyeva, A. (2020), ‘Direct and indirect effects under sample selection and outcome attrition’, *Econometrics* **8**(4), 44.
- Imai, K., Keele, L. & Tingley, D. (2010), ‘A general approach to causal mediation analysis’, *Psychological Methods* **15**(4), 309–334.
- Imai, K., Keele, L. & Yamamoto, T. (2010), ‘Identification, inference and sensitivity analysis for causal mediation effects’, *Statistical Science* **25**(1).
- Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, Cambridge, UK.
- Judd, C. M. & Kenny, D. A. (1981), ‘Process analysis: Estimating mediation in treatment evaluations’, *Evaluation Review* **5**, 602–619.
- Kim, R. S. & Kaplan, R. C. (2014), ‘Analysis of secondary outcomes in nested case-control study designs’, *Statistics in Medicine* **33**(24), 4215–4226.
- Lange, T., Vansteelandt, S. & Bekaert, M. (2012), ‘A simple unified approach for estimating natural direct and indirect effects’, *American Journal of Epidemiology* **176**(3), 190–195.
- Lipsky, A. M. & Greenland, S. (2022), ‘Causal directed acyclic graphs’, *JAMA* **327**(11), 1083–1084.
- MacKinnon, D. (2007), *Introduction to Statistical Mediation Analysis*, Routledge, New York, NY.
- Mackinnon, D. P. & Dwyer, J. H. (1993), ‘Estimating mediated effects in prevention studies’, *Evaluation Review* **17**(2), 144–158.
- MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W. & Hoffman, J. M. (2007), ‘The intermediate endpoint effect in logistic and probit regression’, *Clinical Trials* **4**(5), 499–513.
- Morris, T. P., White, I. R. & Crowther, M. J. (2019), ‘Using simulation studies to evaluate statistical methods’, *Statistics in Medicine* **38**(11), 2074–2102.
- Neyman, J. (1938), ‘Contribution to the theory of sampling human populations’, *Journal of the American Statistical Association* **33**(201), 101–116.
- Nguyen, T. Q., Ogburn, E. L., Schmid, I., Sarker, E. B., Greifer, N., Koning, I. M. & Stuart, E. A. (2023), ‘Causal mediation analysis: from simple to more robust strategies for estimation of marginal natural (in)direct effects’, *Statistics Surveys* **17**.
- Pearl, J. (2001), Direct and indirect effects, in M. Kaufmann, ed., ‘Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence’, San Francisco, CA, p. 10.

- Pearl, J. (2009), *Causality*, Cambridge University Press, Cambridge, UK.
- Petersen, M. L., Sinisi, S. E. & van der Laan, M. J. (2006), ‘Estimation of direct causal effects’, *Epidemiology* **17**(3), 276–284.
- 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.
- Richiardi, L., Bellocco, R. & Zugna, D. (2013), ‘Mediation analysis in epidemiology: methods, interpretation and bias’, *International Journal of Epidemiology* **42**(5), 1511–1519.
- Rijnhart, J. J. M., Lamp, S. J., Valente, M. J., MacKinnon, D. P., Twisk, J. W. R. & Heymans, M. W. (2021), ‘Mediation analysis methods used in observational research: a scoping review and recommendations’, *BMC Medical Research Methodology* **21**, 226.
- Robins, J. M. & Greenland, S. (1992), ‘Identifiability and exchangeability for direct and indirect effects’, *Epidemiology* **3**(2), 143–155.
- Robins, J. M., Hernán, M. Á. & Brumback, B. (2000), ‘Marginal structural models and causal inference in epidemiology’, *Epidemiology* **11**(5), 550–560.
- Rubin, D. B. (1974), ‘Estimating causal effects of treatments in randomized and nonrandomized studies’, *Journal of Educational Psychology* **66**, 688–701.
- Rubin, D. B. (1978), ‘Bayesian inference for causal effects: The role of randomization’, *The Annals of Statistics* **6**(1), 34–58.
- Rubin, D. B. (1980), ‘Randomization analysis of experimental data: The fisher randomization test comment’, *Journal of the American Statistical Association* **75**(371), 591–593.
- Rubin, D. B. (1986), ‘Statistics and causal inference: comment: which ifs have causal answers’, *Journal of the American Statistical Association* **81**(396), 961–962.
- Rudolph, K. E., Goin, D. E., Paksarian, D., Crowder, R., Merikangas, K. R. & Stuart, E. A. (2019), ‘Causal mediation analysis with observational data: considerations and illustration examining mechanisms linking neighborhood poverty to adolescent substance use’, *American Journal of Epidemiology* **188**(3), 598–608.
- Sahiner, B., Chan, H.-P. & Hadjiiski, L. (2008), ‘Classifier performance prediction for computer-aided diagnosis using a limited dataset’, *Medical Physics* **35**(4), 1559–1570.
- Sandler, D. P., Hodgson, M. E., Deming-Halverson, S. L., Juras, P. S., D’Aloisio, A. A., Suarez, L. M., Kleeberger, C. A., Shore, D. L., DeRoo, L. A., Taylor, J. A., Weinberg, C. R. & Sister Study Research Team (2017), ‘The sister study cohort: Baseline methods and participant characteristics’, *Environmental Health Perspectives* **125**(12), 127003.
- Satten, G. A., Curtis, S. W., Solis-Lemus, C., Leslie, E. J. & Epstein, M. P. (2022), ‘Efficient estimation of indirect effects in case-control studies using a unified likelihood framework’, *Statistics in Medicine* **41**(15), 2879–2893.

- Schifano, E. D. (2019), ‘A review of analysis methods for secondary outcomes in case-control studies’, *Communications for Statistical Applications and Methods* **26**(2), 103–129.
- Sobel, M. E. (1982), ‘Asymptotic confidence intervals for indirect effects in structural equation models’, *Sociological Methodology* **13**, 290.
- Splawa-Neyman, J., Dabrowska, D. M. & Speed, T. P. (1923), ‘On the application of probability theory to agricultural experiments’, *Statistical Science* **5**(4), 465–472.
- Splawa-Neyman, J., Dabrowska, D. M. & Speed, T. P. (1990), ‘On the application of probability theory to agricultural experiments’, *Statistical Science* **5**(4), 465–472.
- Stanciu, I., Larsson, M., Nordin, S., Adolfsson, R., Nilsson, L.-G. & Olofsson, J. K. (2014), ‘Olfactory impairment and subjective olfactory complaints independently predict conversion to dementia: a longitudinal, population-based study’, *Journal of the International Neuropsychological Society* **20**(2), 209–217.
- Tchetgen, E. J. T. & Shpitser, I. (2012), ‘Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis’, *The Annals of Statistics* **40**(3), 1816–1845.
- Tchetgen Tchetgen, E. J. (2013), ‘Inverse odds ratio-weighted estimation for causal mediation analysis’, *Statistics in Medicine* **32**(26), 4567–4580.
- Ten Have, T. R. & Joffe, M. M. (2012), ‘A review of causal estimation of effects in mediation analyses’, *Statistical Methods in Medical Research* **21**(1), 77–107.
- Valeri, L. & Vanderweele, T. J. (2013), ‘Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros’, *Psychological Methods* **18**(2), 137–150.
- VanderWeele, T. J. (2009), ‘Marginal structural models for the estimation of direct and indirect effects’, *Epidemiology* **20**(1), 18–26.
- VanderWeele, T. J. (2015), *Explanation in Causal Inference: Methods for Mediation and Interaction*, Oxford University Press.
- VanderWeele, T. J. (2020), ‘Invited commentary: Frontiers of power assessment in mediation analysis’, *American Journal of Epidemiology* **189**(12), 1568–1570.
- VanderWeele, T. J. & Tchetgen Tchetgen, E. J. (2016), ‘Mediation analysis with matched case-control study designs’, *American Journal of Epidemiology* **183**(9), 869–870.
- VanderWeele, T. J. & Vansteelandt, S. (2010), ‘Odds ratios for mediation analysis for a dichotomous outcome’, *American Journal of Epidemiology* **172**(12), 1339–1348.
- VanderWeele, T. J., Vansteelandt, S. & Robins, J. M. (2014), ‘Effect decomposition in the presence of an exposure-induced mediator-outcome confounder’, *Epidemiology* **25**(2), 300–306.

- Vansteelandt, S., Bekaert, M. & Lange, T. (2012), ‘Imputation strategies for the estimation of natural direct and indirect effects’, *Epidemiologic Methods* **1**(1), 131–158.
- Vittinghoff, E., Sen, S. & McCulloch, C. E. (2009), ‘Sample size calculations for evaluating mediation’, *Statistics in Medicine* **28**(4), 541–557.
- Wang, J., Ning, J. & Shete, S. (2019), ‘Mediation analysis in a case-control study when the mediator is a censored variable’, *Statistics in Medicine* **38**(7), 1213–1229.
- Weedon, M. N., Lettre, G., Freathy, R. M., Lindgren, C. M., Voight, B. F., Perry, J. R. B., Elliott, K. S., Hackett, R., Guiducci, C., Shields, B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burt, N. P., Rayner, N. W., Saxena, R., Ardlie, K., Tobias, J. H., Ness, A. R., Ring, S. M., Palmer, C. N. A., Morris, A. D., Peltonen, L., Salomaa, V., Smith, G. D., Groop, L. C., Hattersley, A. T., McCarthy, M. I., Hirschhorn, J. N. & Frayling, T. M. (2007), ‘A common variant of HMGA2 is associated with adult and childhood height in the general population’, *Nature Genetics* **39**(10), 1245–1250.
- Weuve, J., Korrick, S. A., Weisskopf, M. A., Ryan, L. M., Schwartz, J., Nie, H., Grodstein, F. & Hu, H. (2009), ‘Cumulative exposure to lead in relation to cognitive function in older women’, *Environmental Health Perspectives* **117**(4), 574–580.
- White, J. E. (1982), ‘A two stage design for the study of the relationship between a rare exposure and a rare disease’, *American Journal of Epidemiology* **115**(1), 119–128.
- Whittemore, A. S. & Halpern, J. (1997), ‘Multi-stage sampling in genetic epidemiology’, *Statistics in medicine* **16**(2), 153–167.
- Wright, S. (1920), ‘The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs’, *Proceedings of the National Academy of Sciences* **6**(6), 320–332.
- Wright, S. (1934), ‘The method of path coefficients’, *The Annals of Mathematical Statistics* **5**(3), 161–215.
- Xydakis, M. S., Mulligan, L. P., Smith, A. B., Olsen, C. H., Lyon, D. M. & Belluscio, L. (2015), ‘Olfactory impairment and traumatic brain injury in blast-injured combat troops: a cohort study’, *Neurology* **84**(15), 1559–1567.
- Zhang, Z. (2014), ‘Monte carlo based statistical power analysis for mediation models: Methods and software’, *Behavior Research Methods* **46**(4), 1184–1198.

APPENDIX A PROOFS OF THEOREMS

A.1 Proof of Theorem 1

For the controlled direct effects, we need to recover the mean of the counterfactual outcome Y^{xz} . By recovering we mean using observed data to identify the target estimand.

$$E[Y^{xz}] = E_{C_1, C_2} [E(Y^{xz} \mid C_1, C_2)] \quad (\text{A.1})$$

$$= E_{C_1, C_2} [E(Y^{xz} \mid C_1, C_2, X = x, Z = z)] \quad (\text{A.2})$$

$$= E_{C_1, C_2} [E(Y \mid C_1, C_2, X = x, Z = z)] \quad (\text{A.3})$$

$$= E_{C_1, C_2} [E(Y \mid C_1, C_2, X = x, Z = z, S = 1)] \quad (\text{A.4})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2) dc_1 dc_2 dy \quad (\text{A.5})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) \frac{f(c_1, c_2 \mid x, z, S = 1)}{f(c_1, c_2 \mid x, z, S = 1)} f(c_1, c_2) dc_1 dc_2 dy \quad (\text{A.6})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2 \mid x, z, S = 1) \frac{P(x, z, S = 1)}{P(x, z, S = 1 \mid c_1, c_2) f(c_1, c_2)} f(c_1, c_2) dc_1 dc_2 dy \quad (\text{A.7})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2 \mid x, z, S = 1) \frac{P(S = 1 \mid x, z) P(x, z)}{P(S = 1 \mid x, z, c_1, c_2) P(x, z \mid c_1, c_2)} dc_1 dc_2 dy \quad (\text{A.8})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2 \mid x, z, S = 1) \frac{P(S = 1 \mid z) P(x, z)}{P(S = 1 \mid z) P(x, z \mid c_1, c_2)} dc_1 dc_2 dy \quad (\text{A.9})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2 \mid x, z, S = 1) \frac{P(z \mid x) P(x)}{P(z \mid x, c_1, c_2) P(x \mid c_1, c_2)} dc_1 dc_2 dy \quad (\text{A.10})$$

$$= \int_y \int_{c_1, c_2} y P(Y = y \mid c_1, c_2, x, z, S = 1) f(c_1, c_2 \mid x, z, S = 1) \frac{P(z \mid x) P(x)}{P(z \mid x, c_1, c_2) P(x \mid c_1)} dc_1 dc_2 dy \quad (\text{A.11})$$

$$= E[w_1 Y \mid X = x, Z = z, S = 1] \quad (\text{A.12})$$

where $w_1 = \frac{P(z|x) \cdot P(x)}{P(z|c_1, c_2, x) \cdot P(x|c_1)}$. For simplicity, the integral symbol is used while it represents the summation for taking expectation of non-continuous variables (same for the rest of proofs in Appendix). Equation (6.1) follows from the law of iterated expectations. Equation (6.2) follows from the assumption $Y^{xz} \perp\!\!\!\perp (X, Z) \mid (C_1, C_2)$. Equation (6.3) follows from the consistency. Equation (6.4) follows from the assumption $Y \perp\!\!\!\perp S \mid (X, Z, C_1, C_2)$. Equation (6.5) follows from the definition. Equation (6.6) is (6.5) multiplied by unity. Equations (6.7)-(6.8) follow from the laws of probabilities. Equation (6.9) follows under the assumptions $S \perp\!\!\!\perp X \mid Z$ and $S \perp\!\!\!\perp (X, C_1, C_2) \mid Z$. Equation (6.10)-(6.12) follow from the laws of probabilities and the definition of conditional expectation.

Next, we show the nonparametric identification of $E[Y^{xZ^x}]$ and $E[Y^{xZ^{x^*}}]$ so that the natural in(direct) effects are identified.

$$E[Y^{xZ^x}]$$

$$=E_{C_1, C_2} [E(Y^{xZ^x} \mid C_1, C_2)] \quad (\text{A.13})$$

$$=E_{C_1, C_2} [E_{Z^x \mid C_1, C_2} (E \{Y^{xZ^x} \mid C_1, C_2, Z^x\} \mid C_1, C_2)] \quad (\text{A.14})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^x} = y \mid c_1, c_2, Z^x = z) P(Z^x = z \mid c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \quad (\text{A.15})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^x} = y \mid c_1, c_2, Z^x = z, X = x) P(Z^x = z \mid c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \quad (\text{A.16})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^x} = y \mid c_1, c_2, Z^x = z, X = x) P(Z^x = z \mid X = x, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \quad (\text{A.17})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x) f(z \mid x, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \quad (\text{A.18})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z \mid x, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \quad (\text{A.19})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z, c_1, c_2 \mid x, S = 1) \frac{f(z \mid x, c_1, c_2) f(c_1, c_2)}{f(z, c_1, c_2 \mid x, S = 1)} dz dc_1 dc_2 dy \quad (\text{A.20})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z, c_1, c_2 \mid x, S = 1) \frac{f(z \mid x, c_1, c_2) f(c_1, c_2) P(S = 1 \mid x) P(X = x)}{P(S = 1 \mid z, c_1, c_2, x) P(X = x \mid z, c_1, c_2) p(z \mid c_1, c_2) f(c_1, c_2)} dz dc_1 dc_2 dy \quad (\text{A.21})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z, c_1, c_2 \mid x, S = 1) \frac{P(X = x \mid z, c_1, c_2) f(z, c_1, c_2) P(S = 1 \mid x) P(X = x)}{f(x, c_1, c_2) P(S = 1 \mid z) P(X = x \mid z, c_1, c_2) p(z \mid c_1, c_2)} dz dc_1 dc_2 dy \quad (\text{A.22})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z, c_1, c_2 \mid x, S = 1) \frac{P(S = 1 \mid x) P(X = x)}{P(S = 1 \mid z) P(X = x \mid c_1, c_2)} dz dc_1 dc_2 dy \quad (\text{A.23})$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, Z = z, X = x, S = 1) f(z, c_1, c_2 \mid x, S = 1)$$

$$\frac{P(S = 1 | x)P(X = x)}{P(S = 1 | z)P(X = x | c_1)} dz dc_1 dc_2 dy \quad (\text{A.24})$$

$$= E[w_2 Y | X = x, S = 1] \quad (\text{A.25})$$

where $w_2 = \frac{P(S=1|X=x)P(X=x)}{P(S=1|Z=z)P(X=x|c_1)}$. Equations (6.13)-(6.14) follow from the law of iterated expectations. Equation (6.15) follows by the definition. Equation (6.16) follows from the assumption $Y^{xZ^x} \perp\!\!\!\perp X | (Z^x, C_1, C_2)$. Equation (6.17) follows from the assumption $Z^x \perp\!\!\!\perp X | (C_1, C_2)$. Equation (6.18) follows from the consistency. Equation (6.19) follows from the the assumption $Y \perp\!\!\!\perp S | (X, Z, C_1, C_2)$. Equation (6.20) is (6.19) multiplied by unity. Equation (6.21) follows from the laws of probabilities. Equation (6.22) follows from the the assumption $S \perp\!\!\!\perp (X, C_1, C_2) | Z$. Equation (6.23) follows from the the assumption and $X \perp\!\!\!\perp C_2 | C_1$. Equation (6.24) follows from the definition.

The steps for identifying $E[Y^{xZ^{x^*}}]$ are similar.

$$\begin{aligned}
& E[Y^{xZ^{x^*}}] \\
&= E_{C_1, C_2} \left[E(Y^{xZ^{x^*}} \mid C_1, C_2) \right] \tag{A.26}
\end{aligned}$$

$$= E_{C_1, C_2} \left[E_{Z^{x^*} \mid C_1, C_2} \left(E \left\{ Y^{xZ^{x^*}} \mid C_1, C_2, Z^{x^*} \right\} \mid C_1, C_2 \right) \right] \tag{A.27}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^{x^*}} = y \mid c_1, c_2, Z^{x^*} = z) P(Z^{x^*} = z \mid c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.28}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^{x^*}} = y \mid c_1, c_2, Z^{x^*} = z, X = x) P(Z^{x^*} = z \mid c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.29}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xZ^{x^*}} = y \mid c_1, c_2, Z^{x^*} = z, X = x) P(Z^{x^*} = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.30}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xz} = y \mid c_1, c_2, Z^{x^*} = z, X = x) P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.31}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xz} = y \mid c_1, c_2, X = x) P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.32}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y^{xz} = y \mid c_1, c_2, X = x, Z = z) P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.33}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z) P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.34}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) dz dc_1 dc_2 dy \tag{A.35}$$

$$= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) f(z, c_1, c_2 \mid X = x, S = 1) \frac{P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2)}{f(z, c_1, c_2 \mid X = x, S = 1)} dz dc_1 dc_2 dy \tag{A.36}$$

$$\begin{aligned}
&= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) P(Z = z \mid X = x, S = 1) \\
&\quad \frac{P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) P(X = x, S = 1)}{P(Z = z, X = x, S = 1, c_1, c_2)} dz dc_1 dc_2 dy
\end{aligned} \tag{A.37}$$

$$\begin{aligned}
&= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) P(Z = z \mid X = x, S = 1) \\
&\quad \frac{P(Z = z \mid X = x^*, c_1, c_2) f(c_1, c_2) P(S = 1 \mid X = x) P(X = x)}{P(S = 1 \mid Z = z, c_1, c_2, x) P(Z = z \mid X = x, c_1, c_2) P(X = x \mid c_1, c_2) f(c_1, c_2)} dz dc_1 dc_2 dy
\end{aligned} \tag{A.38}$$

$$\begin{aligned}
&= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) P(Z = z \mid X = x, S = 1) \\
&\quad \frac{P(Z = z \mid X = x^*, c_1, c_2) P(S = 1 \mid X = x) P(X = x)}{P(S = 1 \mid Z = z) P(Z = z \mid X = x, c_1, c_2) P(X = x \mid c_1, c_2)} dz dc_1 dc_2 dy
\end{aligned} \tag{A.39}$$

$$\begin{aligned}
&= \int_y \int_{c_1, c_2} \int_z y P(Y = y \mid c_1, c_2, X = x, Z = z, S = 1) P(Z = z \mid X = x, S = 1) \\
&\quad \frac{P(Z = z \mid X = x^*, c_1, c_2) P(S = 1 \mid X = x) P(X = x)}{P(S = 1 \mid Z = z) P(Z = z \mid X = x, c_1, c_2) P(X = x \mid c_1)} dz dc_1 dc_2 dy
\end{aligned} \tag{A.40}$$

$$= E[w_3 Y \mid X = x, S = 1] \tag{A.41}$$

where $w_3 = \frac{P(S=1|X=x)P(Z=z|X=x^*,c_1,c_2)P(X=x)}{P(S=1|Z=z)P(Z=z|X=x,c_1,c_2)P(X=x|c_1)}$. Equations (6.25)-(6.27) follow from the law of iterated expectations. Equation (6.28) follows from the assumption $Y^{xZ^{x^*}} \perp\!\!\!\perp X \mid (Z^{x^*}, C_1, C_2)$. Equation (6.29) follows from the assumption $Z^{x^*} \perp\!\!\!\perp X \mid (C_1, C_2)$. Equation (6.30) follows from the consistency. Equation (6.31) follows from the assumption $Y^{xz} \perp\!\!\!\perp Z^{x^*} \mid (X, C_1, C_2)$. Equation (6.32) follows from the assumption $Y^{xz} \perp\!\!\!\perp Z \mid (X, C_1, C_2)$. Equation (6.33) follows from the consistency. Equation (6.34) follows from the assumption $Y \perp\!\!\!\perp S \mid (X, Z, C_1, C_2)$. Equation (6.35) if (6.34) multiplies by unity. Equations (6.36)-(6.37) follow from the laws of probabilities. Equation (6.38) follows from the assumption $S \perp\!\!\!\perp (X, C_1, C_2) \mid Z$. Equation (6.39) follows from the assumption $X \perp\!\!\!\perp C_2 \mid C_1$. Equation (6.40) follows from the definition.

□

A.2 Proof of Theorem 2

$$E[Y^{xm}] = \int_y \int_{\mathbf{c}} y P(Y^{xm} = y \mid \mathbf{c}) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.42})$$

$$= \int_y \int_{\mathbf{c}} y P(Y^{xm} = y \mid \mathbf{c}, X = x, M = m) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.43})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z) f(z \mid \mathbf{c}, x, m) f(\mathbf{c}) dz d\mathbf{c} dy \quad (\text{A.44})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m) f(\mathbf{c}) dz d\mathbf{c} dy \quad (\text{A.45})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) \frac{f(z \mid \mathbf{c}, x, m) f(\mathbf{c})}{f(z \mid \mathbf{c}, x, m, S = 1)} dz d\mathbf{c} dy \quad (\text{A.46})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) \frac{f(\mathbf{c}) f(\mathbf{c}, x, m, S = 1)}{P(S = 1 \mid z, \mathbf{c}, x, m) f(x, m \mid \mathbf{c}) f(\mathbf{c})} dz d\mathbf{c} dy \quad (\text{A.47})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) f(\mathbf{c} \mid x, m, S = 1) \frac{f(x, m, S = 1)}{P(S = 1 \mid z, \mathbf{c}, x, m) f(x, m \mid \mathbf{c})} dz d\mathbf{c} dy \quad (\text{A.48})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) f(\mathbf{c} \mid x, m, S = 1) \frac{P(M = m \mid x, S = 1) P(S = 1 \mid x) P(X = x)}{P(S = 1 \mid z) P(M = m \mid x, \mathbf{c}) P(X = x \mid \mathbf{c})} dz d\mathbf{c} dy \quad (\text{A.49})$$

$$= \int_y \int_{\mathbf{c}} \int_z y P(Y = y \mid \mathbf{c}, X = x, M = m, Z = z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) f(\mathbf{c} \mid x, m, S = 1) \frac{P(M = m \mid x, S = 1) P(S = 1 \mid x) P(X = x)}{P(M = m \mid x, c_1, c_2, c_3) P(S = 1 \mid z) P(X = x \mid c_1)} dz d\mathbf{c} dy \quad (\text{A.50})$$

$$= E[w_4 Y \mid X = x, M = m, S = 1] \quad (\text{A.51})$$

where $w_4 = \frac{P(M=m|x, S=1) \cdot P(S=1|x) \cdot P(X=x)}{P(M=m|x, c_1, c_2, c_3) P(S=1|z) P(X=x|c_1)}$. Equation (6.41) follows from the law of iterated expectations. Equation (6.42) follows from the assumption of $Y^{xm} \perp\!\!\!\perp (X, M) \mid (C_1, C_2, C_3)$.

Equation (6.43) follows from the consistency and basic probability theory. Equation (6.44) follows from the assumption of $Y \perp\!\!\!\perp S|(C_1, C_2, C_3, X, Z, M)$. Equation (6.45)-(6.47) follow from the basic probability theory. Equation (6.48) follows from the assumption of $S \perp\!\!\!\perp (X, C_1, C_2, C_3)|Z$ and basic probability theory. Equation (6.49) follows from the assumption of $X \perp\!\!\!\perp (C_2, C_3)|C_1$ and basic probability theory. Equation (6.50) follows from the basic probability theory.

Since M is only observed in selected sample, $P(M = m|C_1, C_2, C_3, X)$ is usually not directly estimable using case-control sample. However, we can further take advantage of the sampling scheme to estimate it by (6.51) below:

$$\begin{aligned}
P(M = m|X, C_1, C_2, C_3) &= \sum_z P(M = m|X, C_1, C_2, C_3, Z = z) \cdot P(Z = z|X, C_1, C_2, C_3) \\
&= P(M = m|X, C_1, C_2, C_3, Z = 0) \cdot P(Z = 0|X, C_1, C_2) + \\
&\quad P(M = m|X, C_1, C_2, C_3, Z = 1) \cdot P(Z = 1|X, C_1, C_2) \\
&= P(M = m|X, C_1, C_2, C_3, Z = 0, S = 1) \cdot P(Z = 0|X, C_1, C_2) + \\
&\quad P(M = m|X, C_1, C_2, C_3, Z = 1, S = 1) \cdot P(Z = 1|X, C_1, C_2) \tag{A.52}
\end{aligned}$$

Specifically, to estimate $P(M = m|C_1, C_2, C_3, X)$ using case-control sample, we can consider to do it stratified by the value of Z . Equation (6.51) shows the derivation when Z is a binary variable but it can be extended to other types of Z as well. In stratified sampling (by Z), researchers usually know the sampling rate by design. Here, for example, we have 100% sampling for subjects with $Z = 1$ and ~3.6% simple random sampling for units with $Z = 0$. In such case, $P(M = m|X, C_1, C_2, C_3, Z = 0, S = 1)$ and $P(M = m|X, C_1, C_2, C_3, Z = 1, S = 1)$ can be directly estimated using a parametric model among controls and cases, respectively. This approach is valid since both cases and controls are “representative” for the target population with $Z = 1$ and $Z = 0$.

Next, we show the nonparametric identification of $E[Y^{xM^x}]$ and $E[Y^{xM^{x^*}}]$ so that the natural in(direct) effects are identified.

$$E[Y^{xM^x}] = \int_y \int_{\mathbf{c}} y P(Y^{xM^x} = y \mid \mathbf{c}) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.53})$$

$$= \int_y \int_{\mathbf{c}} y P(Y^{xM^x} = y \mid \mathbf{c}, X = x) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.54})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xM^x} = y \mid \mathbf{c}, X = x, M^x = m) P(M^x = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.55})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xm} = y \mid \mathbf{c}, X = x, M = m) P(M = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.56})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y = y \mid \mathbf{c}, X = x, M = m) P(M = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.57})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y = y \mid \mathbf{c}, X = x, M = m, S = 1) P(M = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.58})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y, z \mid \mathbf{c}, X = x, M = m, S = 1) dz P(M = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.59})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid z, \mathbf{c}, X = x, M = m, S = 1) f(z \mid \mathbf{c}, X = x, M = m, S = 1) dz P(M = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.60})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x) f(\mathbf{c})}{P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1)} dz dm d\mathbf{c} dy \quad (\text{A.61})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x) f(\mathbf{c}) f(S = 1 \mid \mathbf{c}, x) P(S = 1 \mid x) f(x)}{P(S = 1 \mid m, \mathbf{c}, x) P(M = m \mid \mathbf{c}, x) P(S = 1 \mid \mathbf{c}, x) f(x \mid \mathbf{c}) f(\mathbf{c})} dz dm d\mathbf{c} dy \quad (\text{A.62})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x) P(S = 1 \mid x) f(x)}{P(S = 1 \mid m, \mathbf{c}, x) P(M = m \mid \mathbf{c}, x) f(x \mid \mathbf{c})} dz dm d\mathbf{c} dy \quad (\text{A.63})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1)$$

$$f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid c_1, c_2, c_3, x)}{P(M = m \mid x, c_1, c_2, c_3, S = 1)} \cdot \frac{P(S = 1 \mid x)P(X = x)}{P(S = 1 \mid x, c_1, c_2)P(X = x \mid c_1)} dz dm d\mathbf{c} dy \quad (\text{A.64})$$

$$= E[w_5 Y \mid X = x, S = 1] \quad (\text{A.65})$$

where $w_5 = \frac{P(M=m \mid c_1, c_2, c_3, x)}{P(M=m \mid x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1 \mid x)P(X=x)}{P(S=1 \mid x, c_1, c_2)P(X=x \mid c_1)}$. Equation (6.52) follows from the law of iterated expectations. Equation (6.53) follows from the assumption of $Y^{xM^x} \perp\!\!\!\perp X \mid (C_1, C_2, C_3)$. Equation (6.54) follows from the law of iterated expectations. Equation (6.55) and (6.56) follows from the consistency. Equation (6.57) follows from the assumption of $Y \perp\!\!\!\perp S \mid (C_1, C_2, C_3, X, M)$. Equation (6.58)-(6.62) follow from the basic probability theory. Equation (6.63) follows from the assumptions of $X \perp\!\!\!\perp (C_2, C_3) \mid C_1$ and $S \perp\!\!\!\perp C_3 \mid (X, C_1, C_2)$.

$$E[Y^{xM^{x^*}}] = \int_y \int_{\mathbf{c}} y P(Y^{xM^{x^*}} = y \mid \mathbf{c}) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.66})$$

$$= \int_y \int_{\mathbf{c}} y P(Y^{xM^{x^*}} = y \mid \mathbf{c}, X = x) f(\mathbf{c}) d\mathbf{c} dy \quad (\text{A.67})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xM^{x^*}} = y \mid \mathbf{c}, X = x, M^{x^*} = m) P(M^{x^*} = m \mid \mathbf{c}, X = x) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.68})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xM^{x^*}} = y \mid \mathbf{c}, X = x, M^{x^*} = m) P(M^{x^*} = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.69})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xm} = y \mid \mathbf{c}, X = x, M^{x^*} = m) P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.70})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y^{xm} = y \mid \mathbf{c}, X = x, M^x = m) P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.71})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y = y \mid \mathbf{c}, X = x, M = m) P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.72})$$

$$= \int_y \int_{\mathbf{c}} \int_m y P(Y = y \mid \mathbf{c}, X = x, M = m, S = 1) P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.73})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y, z \mid \mathbf{c}, X = x, M = m, S = 1) dz P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.74})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid z, \mathbf{c}, X = x, M = m, S = 1) f(z \mid \mathbf{c}, X = x, M = m, S = 1) dz P(M = m \mid \mathbf{c}, X = x^*) f(\mathbf{c}) dm d\mathbf{c} dy \quad (\text{A.75})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x^*) f(\mathbf{c})}{P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1)} dz dm d\mathbf{c} dy \quad (\text{A.76})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1)$$

$$f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x^*) f(\mathbf{c}) f(S = 1 \mid \mathbf{c}, x) P(S = 1 \mid x) f(x)}{P(S = 1 \mid m, \mathbf{c}, x) P(M = m \mid \mathbf{c}, x) P(S = 1 \mid \mathbf{c}, x) f(x \mid \mathbf{c}) f(\mathbf{c})} dz dm d\mathbf{c} dy \quad (\text{A.77})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid \mathbf{c}, x^*) P(S = 1 \mid x) f(x)}{P(S = 1 \mid m, \mathbf{c}, x) P(M = m \mid \mathbf{c}, x) f(x \mid \mathbf{c})} dz dm d\mathbf{c} dy \quad (\text{A.78})$$

$$= \int_y \int_{\mathbf{c}} \int_m \int_z y P(Y = y \mid \mathbf{c}, x, m, z, S = 1) f(z \mid \mathbf{c}, x, m, S = 1) P(M = m \mid \mathbf{c}, x, S = 1) f(\mathbf{c} \mid x, S = 1) \frac{P(M = m \mid c_1, c_2, c_3, x^*)}{P(M = m \mid x, c_1, c_2, c_3, S = 1)} \cdot \frac{P(S = 1 \mid x) P(X = x)}{P(S = 1 \mid x, c_1, c_2) P(X = x \mid c_1)} dz dm d\mathbf{c} dy \quad (\text{A.79})$$

$$= E[w_6 Y \mid X = x, S = 1] \quad (\text{A.80})$$

where $w_6 = \frac{P(M=m|c_1, c_2, c_3, x^*)}{P(M=m|x, c_1, c_2, c_3, S=1)} \cdot \frac{P(S=1|x)P(X=x)}{P(S=1|x, c_1, c_2)P(X=x|c_1)}$. Equation (6.65) follows from the law of iterated expectations. Equation (6.66) follows from the assumption of $Y^{xM^{x^*}} \perp\!\!\!\perp X \mid (C_1, C_2, C_3)$. Equation (6.67) follows from the law of iterated expectations. Equation (6.68) follows from the assumption of $M^{x^*} \perp\!\!\!\perp X \mid (Z, C_1, C_2, C_3)$. Equation (6.69) follows from the consistency. Equation (6.70) follows from the assumption of $Y^{xm} \perp\!\!\!\perp (M^x, M^{x^*}) \mid (X, C_1, C_2, C_3)$. Equation (6.71) again follows from the consistency. Equation (6.72) follows from the assumption of $Y \perp\!\!\!\perp S \mid (C_1, C_2, C_3, X, M)$. Equation (6.73)-(6.77) follow from the basic probability theory. Equation (6.78) follows from the assumptions of $X \perp\!\!\!\perp (C_2, C_3) \mid C_1$ and $S \perp\!\!\!\perp C_3 \mid (X, C_1, C_2)$. $\square$

A.3 Comparison of Assumption Used Between (Modified) Huber and Solovyeva and Proposed Methods

Table A.1: Assumption Comparison for Chapter 2

Proposed	Huber and Solovyeva
Y^{xz}	
$Y^{xz} \perp\!\!\!\perp X \mathcal{C}$	$\checkmark$
$/$	$Y^{xz} \perp\!\!\!\perp Z (X, \mathcal{C})$
$Y \perp\!\!\!\perp S (Z, X, \mathcal{C})$	$\checkmark$
Y^{xz^x}	
$Y^{xz^x} \perp\!\!\!\perp X (Z^x, \mathcal{C})$	$Y^{xz^x} \perp\!\!\!\perp X \mathcal{C}$
$Z^x \perp\!\!\!\perp X \mathcal{C}$	$/$
$Y \perp\!\!\!\perp S (Z, X, \mathcal{C})$	$\checkmark$
$Y^{xz^{x^*}}$	
$Y^{xz^{x^*}} \perp\!\!\!\perp X (Z^{x^*}, \mathcal{C})$	$/$
$Z^{x^*} \perp\!\!\!\perp X \mathcal{C}$	$/$
$Y^{xz} \perp\!\!\!\perp Z^{x^*} (X, \mathcal{C})$	$\checkmark$
$Y^{xz} \perp\!\!\!\perp Z (X, \mathcal{C})$	$\checkmark$
$/$	$Y^{xz} \perp\!\!\!\perp X \mathcal{C}$
$/$	$Y^{xz} \perp\!\!\!\perp X (Z^{x^*}, \mathcal{C})$
$Y \perp\!\!\!\perp S (Z, X, \mathcal{C})$	$\checkmark$

Table A.2: Assumption Comparison for Chapter 3

Proposed	Huber and Solovyeva
Y^{xm}	
$Y^{xm} \perp\!\!\!\perp (X, M) \mathcal{C}$	$\checkmark$
$/$	$Y^{xm} \perp\!\!\!\perp Z (X, M, \mathcal{C})$
$Y \perp\!\!\!\perp S (Z, M, X, \mathcal{C})$	$\checkmark$
Y^{xM^x}	
$Y^{xM^x} \perp\!\!\!\perp X \mathcal{C}$	$\checkmark$
$Y \perp\!\!\!\perp S (M, X, \mathcal{C})$	$\checkmark$
$Y^{xM^{x^*}}$	
$Y^{xM^{x^*}} \perp\!\!\!\perp X \mathcal{C}$	$/$
$M^{x^*} \perp\!\!\!\perp X (Z, \mathcal{C})$	$/$
$Y^{xm} \perp\!\!\!\perp (M^x, M^{x^*}) (X, \mathcal{C})$	$Y^{xm} \perp\!\!\!\perp X (M^{x^*}, \mathcal{C})$
	$Y^{xm} \perp\!\!\!\perp M^{x^*} (X, \mathcal{C})$
	$Y^{xm} \perp\!\!\!\perp X \mathcal{C}$
	$Y^{xm} \perp\!\!\!\perp M (X, \mathcal{C})$
$Y \perp\!\!\!\perp S (M, X, \mathcal{C})$	$\checkmark$

APPENDIX B DATA GENERATING PROCESS FOR CHAPTER 4

The tables below include the parameters used for simulations in Chapter 4.

Table B.1: Data generating process for scenario 1

Parameter	No effect	Effect 1	Effect 2	Effect 3	Effect 4	Effect 5	Effect 6	Effect 7	Effect 8	Effect 9	Effect 10
β_{c_1x}	.991										
β_{0x}	.44										
β_{c_1z}	1.033										
β_{c_2z}	.874										
β_{xz}	1	1.03	1.096	1.162	1.228	1.36	1.39	1.415	1.43	1.57	1.9
β_{0z}	.01164	.01157	.01141	.01127	.01113	.01086	.0108	.01076	.01073	.01047	.00993
β_{c_1y}	1.085										
β_{c_2y}	1.57										
β_{xy}	13.5										
β_{zy}	1	1.03	1.098	1.166	1.234	1.37	1.4	1.425	1.44	1.58	1.91
β_{c_1xy}	.961										
β_{c_2xy}	.65										
β_{xzy}	1	1.2									
β_{0y}	.001037	.0010305	.0010245	.0010185	.0010125	.0010015	.001	.0009982	.0009968	.0009868	.0009656

Effects are taken log scale.

Table B.2: Data generating process for scenario 2

Parameter	No effect	Effect 1	Effect 2	Effect 3	Effect 4	Effect 5	Effect 6	Effect 7	Effect 8	Effect 9	Effect 10
β_{c_1x}	0.991										
β_{0x}	0.4325										
β_{c_1z}	1.033										
β_{c_2z}	0.874										
β_{xz}	2.7										
β_{0z}	.00888										
β_{c_1m}	1.079										
β_{c_2m}	1.468										
β_{c_3m}	1.439										
β_{xm}	0	1.07	1.14	1.25	1.5	1.74	1.9	2	2.1	2.2	2.3
β_{zm}	4										
β_{0m}	.001245	.001227	.00121	.001186	.001136	.001095	.001069	.001054	.00104	.001026	.001013
β_{c_1y}	1.012										
β_{c_2y}	1.187										
β_{c_3y}	3.271										
β_{xy}	1.0002										
β_{zy}	0										
β_{my}	0	1.06	1.15	1.32	1.5	1.74	1.9	2	2.1	2.2	2.3
β_{c_3xy}	1.5										
β_{xmy}	0	1.3	1.42	1.3	1.2						
β_{0y}	.06455	.06325	.0621	.06085	.0596	.05775	.05663	.05597	.05529	.0547	.05411

Effects are taken log scale.