Causal Inference with Selection and Confounding Variables

Linbo Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Xiao-Hua (Andrew) Zhou, Chair

Thomas S. Richardson, Chair

Peter B. Gilbert

Program Authorized to Offer Degree:
UW Biostatistics

University of Washington

Abstract

Causal Inference with Selection and Confounding Variables

Linbo Wang

Co-Chairs of the Supervisory Committee:
Professor Xiao-Hua (Andrew) Zhou
Biostatistics

Professor Thomas S. Richardson
Statistics

Most complex observational and randomized studies are motivated by the potential of drawing causal statements. However, a usual statistical analysis may yield estimates that do not have causal interpretations. In fact, unlike most other parameters, identification of causal parameters usually relies on untestable assumptions. Moreover, even under these identification assumptions, estimation of causal parameters often relies on nuisance models. The parameter estimation in the nuisance models is crucial to obtain robust causal effect estimates. My research attempts to address these methodological challenges.

In Chapter 2 we study robust estimation of propensity score weights. The propensity score plays a central role in inferring causal effects from observational studies. In particular, weighting and subclassification are two principal approaches to estimate the average causal effect based on estimated propensity scores. Unlike the conventional version of the propensity score subclassification estimator, if the propensity score model is correctly specified, the weighting methods offer consistent and possibly efficient estimation of the average causal effect. However, this theoretical appeal may be diminished in practice by sensitivity to misspecification of the propensity score model. In contrast, subclassification methods are usually more robust to model misspecification. We hence propose to use subclassification

for robust estimation of propensity score weights. Our approach is based on the intuition that the inverse probability weighting estimator can be seen as the limit of subclassification estimators as the number of subclasses goes to infinity. By formalizing this intuition, we propose novel propensity score weighting estimators that are both consistent and robust to model misspecification. Empirical studies show that the proposed estimators perform favorably compared to existing methods.

In Chapter 3 we study identification and estimation of causal effects with outcomes truncated by death. It is common that in medical studies, the outcome of interest is truncated by death, meaning that a subject had died before the outcome could be measured. In this case, restricted analysis among survivors may be subject to selection bias. It is hence of interest to estimate the survivor average causal effect (SACE), defined as the average causal effect among subjects who would survive under either exposure. In this chapter, we consider the identification and estimation problems of the SACE. We propose to identify a substitution variable for the latent membership of the always-survivor group. The identifiability conditions required for a substitution variable are similar in idea to conditions for an instrumental variable. We show that the SACE is identifiable with use of a substitution variable, and propose novel model parameterizations for estimation of the SACE under our identification assumptions. Our approaches are illustrated via simulation studies and two data analyses.

In Chapter 4, we study causal analysis of ordinal treatments and binary outcomes under truncation by death. It is common that in multi-arm randomized trials, the outcome of interest is truncated by death, meaning that it is only observed or well-defined conditioning on an intermediate outcome. In this case, in addition to pairwise contrasts, the joint inference for all treatment arms is also of interest. Under a monotonicity assumption we present methods for both pairwise and joint causal analyses of ordinal treatments and binary outcomes in presence of truncation by death. We illustrate via examples the appropriateness of our assumptions in different scientific contexts.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to my advisors, Xiao-Hua Zhou and Thomas Richardson, for introducing me to the field of causal inference and for their inspiration, guidance and support along the way. I would also like to extend my appreciation to my committee members, Peter Gilbert and Stephen Hawes, for their constructive comments and support during my dissertation research. I am also grateful for Li Hsu for her valuable suggestions at the early stage of my graduate career.

I thank the National Alzheimer's Coordinating Center (NACC) for providing the majority of the financial support during my graduate career, and the opportunity to work closely with epidemiologists and scientists in the wonderful environment of the Center.

I owe thanks to Marco Carone for extensive discussions on my work in Chapter 2, Peng Ding and Eric Tchetgen Tchetgen for insightful suggestions on my work in Chapter 3 and members of the HVTN Ancillary Study Committee for valuable comments on an earlier draft of Chapter 4. Research reported in Chapter 4 was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number UM1AI068635 and the Office of Naval Research (ONR) under Award Number N000141512672. The content is solely my responsibility and does not necessarily represent the official views of NIH or ONR. I thank the NIAID-funded HIV Vaccine Trials Network for providing the dataset from the HVTN 503 trial. Furthermore, I thank the participants, investigators, and sponsors of the HVTN 503 trial.

Finally, special appreciation go to my parents Feng Wang and Bizhen Lin, and my girlfriend Ailin Fan for their love and tolerance. My life becomes more colorful with their support.

I wish everyone lives a happy life.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Overview

Most complex observational and randomized studies are motivated by the potential of drawing causal statements. However, there are two sources of bias that commonly occur in estimating causal effects from such studies.

Confounding bias frequently occurs in observational studies and causes the unadjusted estimate of treatment effect to be biased. For example, suppose we are interested in estimating the effect of participation in meal programs on childhood obesity. However, participation in the program may depend on many baseline factors such as socio-economic status, and these factors can impact childhood obesity independently of the meal programs. It is well known that these baseline factors, known as confounders, can create association between the treatment and outcome variable even if the treatment has no causal effect on the outcome.

Truncation by death occurs when some study subjects die before the outcome of interest is measured. For example, suppose we are interested in estimating the effect of smoking on memory decline in an aged population. If a subject dies before the follow-up memory test is administered, then his/her memory score at the follow-up visit is undefined. Direct comparisons between smokers and non-smokers among observed survivors are subject to selection bias as nonsmokers are more likely to survive to the follow-up assessment (Rosenbaum, 1984; Robins and Greenland, 1992). More fundamentally, direct comparisons among observed survivors are not causally interpretable as they compare outcomes from different subpopulations at baseline (Rubin, 2006).

This dissertation deals with the aforementioned problems in various settings as summa-

rized in Table 1.1.

Table 1.1: Summary of chapters in my dissertation

Chapter	Setting			Problem
	Randomization	Truncation by death	Exposure type	
2	no	no	binary	estimation
3	no	yes	binary	identification and estimation
4	yes	yes	ordinal	partial identification and estimation

1.2 Robust estimation of propensity score weights

Propensity score weighting is a popular class of causal effect estimation methods under the assumption of no unmeasured confounders. Performance of these estimators relies crucially on the model specification for a nuisance parameter, i.e. the propensity score; here the propensity score is defined as the probability of receiving treatment conditioning on baseline covariates. The classical propensity score weighting estimators are unbiased and possibly efficient under correct specification of the propensity score model. However, they can also be very sensitive to misspecification of the propensity score model. To address this problem, we propose a rank-based approach that subclassifies model-based propensity score estimates. Here the propensity score model is only used for subclassification, but not for the subsequent estimation. To avoid the drawback of the conventional propensity score subclassification method, we increase the number of subclasses with the sample size at an appropriate rate such that the subclassified propensity scores approach the individual propensity scores in the limit. This method is simple to implement in practice; at the same time, it allows robust yet consistent estimation of causal effects. This work is described in detail in Chapter 2.

1.3 Causal inference with truncation by death

In this problem the "natural" statistical parameter has no causal interpretation, while the causal parameter of interest is not identifiable without untestable assumptions. Consider a randomized study for which the outcome of interest (such as memory decline) is well-defined conditioning on an intermediate outcome (such as survival). Naively researchers may restrict their analysis to the observed outcomes. However, this naive comparison is not causally interpretable as the comparison groups correspond to different populations at baseline. Instead one may restrict one's analysis to specific subgroups whose members would survive under more than one treatment assignment. The treatment effects in these subgroups, called the principal stratum direct effects (PSDEs) are causally interpretable as memberships of these subgroups are defined at baseline. However, PSDEs are in general not identifiable as memberships of the corresponding subgroups cannot be determined from observed data. Many previous researchers have considered the identification of the PSDE in two-arm randomized studies. In this dissertation, we develop identification methods for the PSDE that are applicable to more complex studies. In Chapter 3, we propose identification assumptions and methods that allow researchers to incorporate baseline covariate information in causal analyses for two-arm observational studies with truncation by death. In Chapter 4, we consider the identification problem in a multi-arm randomized trial. This is challenging as in addition to the pairwise treatment effects, the simultaneous contrasts among treatment arms are also of interest. We build a framework to systematically analyze PSDEs in a general multi-arm trial and derive the optimal large sample bounds / tests for PSDEs. This work provides the basis for future identification methods.

Chapter 2

ROBUST ESTIMATION OF PROPENSITY SCORE WEIGHTS VIA SUBCLASSIFICATION

2.1 Introduction

Observational studies are often used to infer the treatment effect in medical research. In these studies, the treatment assignment may be associated with observed variables, causing the unadjusted estimate of the treatment effect to be biased. This bias is widely known as the confounding bias. In their seminal work, Rosenbaum and Rubin (1983) showed that the propensity score, defined as the probability of assignment to a particular treatment conditioning on observed covariates, plays a central role in obtaining unbiased causal effect estimates from observational studies. Since then, many propensity score adjustment methods have been proposed for causal effect estimates. One popular approach is subclassification, which groups units into several subclasses based on their estimated propensity scores, so that the propensity scores are *approximately* balanced in treatment groups within each subclass (Rosenbaum and Rubin, 1984). The current convention is to subclassify at *quintiles* of estimated propensity scores (even for substantial sample sizes), in the hope that it will remove over 90% of the confounding bias due to observed covariates (Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004). Alternatively, weighting methods based on propensity scores, such as the inverse probability weighting (IPW) estimators (e.g., Rosenbaum, 1987) and the (classical) doubly robust (DR) estimator (Robins et al., 1994) can be used to correct for the confounding bias.

Compared to the conventional version of the propensity score subclassification estimator, weighting estimators are more appealing theoretically. For example, under correct specification of the propensity score model, one can show that the IPW estimators and the classical

DR estimator are all consistent for estimating the average causal effect (ACE). The latter attains the semiparametric efficiency bound if the analyst correctly specifies an additional outcome regression model. However, these weighting methods have often been criticized for their sensitivity to misspecification of the propensity score model (e.g. Drake, 1993; Kang and Schafer, 2007). Although in principle, non-parametric models can be used for the propensity score (e.g. Hirano et al., 2003; McCaffrey et al., 2004), the curse of dimensionality may be a problem. Consequently some researchers favor the subclassification methods as they are more robust to model misspecification (Drake, 1993; Kang and Schafer, 2007) and likely to have smaller variance in large samples (Williamson et al., 2012).

Over the past decade, there have been many endeavors to make the weighting estimators more stable and robust to model misspecification, especially for the classical DR estimator proposed by Robins et al. (1994). Most of these methods construct robust weights by deliberately incorporating the outcome data. See Rotnitzky and Vansteelandt (2014) for a review. However, as advocated by Rubin (2007), the design stage, including analysis of data on the treatment assignment, should be conducted prior to seeing any outcome data. The separation of the design stage from the analysis stage helps prevent selecting models that favor "publishable" results, thereby ensuring the objectivity of the design. Moreover, this separation widens the applicability of a robust weighting scheme as, in principle, it could be applied to any IPW-based estimator (Imai and Ratkovic, 2014). Hence in this article, we are primarily interested in robust estimation of propensity score weights without using the outcome data.

Specifically, we propose to use subclassification for estimating the propensity score weights. Our approach is based on the fact that the subclassification estimator can be seen as an IPW estimator with weights coarsened via subclassification (Imbens and Rubin, 2015, §17.8). On the other hand, the IPW estimators can be seen as the limit of subclassification estimators as the number of subclasses goes to infinity (Rubin, 2001). The main difficulty in formalizing this intuition, however, is that due to residual confounding, the conventional version of the propensity score subclassification estimator, based on a fixed number of subclasses, can lead

to inconsistent estimation of the ACE. Many authors have sought theoretical justifications for increasing the number of subclasses with sample size (e.g., Imbens, 2004; Lunceford and Davidian, 2004; Stuart, 2010; Williamson et al., 2012; Hattori and Henmi, 2014). In this paper, we study the rate at which the number of subclasses should increase with the sample size and show that the subclassification estimator is (root-N) consistent under certain rate conditions. The key insight here is that over-smoothing of the subclassification estimator leads to root-N consistency. By filling this important theoretical gap, we formalize the idea of subclassifying model-based propensity score weights. In particular, we propose a novel full subclassification method for which our rate conditions are satisfied. The full subclassification method can be used for robust estimation of propensity score weights, so that the resulting weighting estimators are both consistent and robust to model misspecification, thereby enjoying the advantages of both the classical weighting and subclassification methods.

In contrast to existing methods that construct robust propensity score weights by balancing empirical covariate moments (Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2015), the full subclassification method employs a rank-based approach to improve robustness of the causal effect estimate in the design stage. The parametric propensity score model is used only for subclassification, but not for the subsequent estimation. This leads to several attractive properties of the full subclassification weighting method. First, under correct propensity score model specification (and the positivity assumption), it leads to consistent estimation of the ACE regardless of the response pattern. Second, it dramatically improves upon model-based estimates in term of both weight stability and covariate balance, especially when the propensity score model is misspecified. Third, with the full subclassification weights, different weighting estimators tend to give similar answers; in particular, two popular IPW estimators coincide with each other (see Proposition 2.5). As we discuss later in Section 2.3.4, none of the existing methods have all these properties, which makes the full subclassification weighting method an appealing alternative in practice.

The rest of this article is organized as follows. In Section 2.2, we give a brief overview of relevant propensity score adjustment methods. In Section 2.3, we introduce the full

subclassification weighting method and discuss its theoretical properties. We also relate our approach to covariate balancing weighting schemes in the literature, and discuss further beneficial properties of our method. Sections 2.4 and 2.5 contain simulations and an illustrative data analysis. We end with a discussion in Section 2.6.

2.2 Background

2.2.1 The propensity score

Let Z be the treatment indicator (1=active treatment, 0=control) and $\boldsymbol{X}$ denote baseline covariates. We assume each subject has two potential outcomes $Y(1)$ and $Y(0)$, defined as the outcomes that would have been observed had the subject received the treatment and control, respectively. We make the consistency assumption such that the observed outcome Y satisfies

$$Y = ZY(1) + (1 - Z)Y(0).$$

Suppose that we independently sample N units from the joint distribution of $(Z, \boldsymbol{X}, Y)$, and denote them as triples $(Z_i, \boldsymbol{X}_i, Y_i), i = 1, \ldots, N$. We are interested in estimating the ACE, namely

$$\Delta = E\{Y(1)\} - E\{Y(0)\},$$

where $E\{\cdot\}$ denotes expectation in the population.

Remark 2.1. *An alternative estimand is the multiplicative causal effect defined by* $E\{Y(1)\}/E\{Y(0)\}$. *We note that all the estimators considered in this article estimate* $E\{Y(1)\}$ *and* $E\{Y(0)\}$ *separately. Although we mainly discuss the estimation problem of* Δ*, all the methodologies introduced here apply to estimating the multiplicative causal effect as well.*

The key assumption for identifying Δ from an observational study is the ignorable treatment assignment assumption (Rosenbaum and Rubin, 1983), which we maintain throughout this paper:

Assumption 2.1. *Strong Ignorability of Treatment Assignment: the treatment assignment is uninformative of the potential outcomes given observed covariates. Formally,* $Z \perp (Y(0), Y(1)) \mid \boldsymbol{X}$.

Remark 2.2. *Although results in this paper only rely on the weaker assumption that* $Z \perp Y(z) \mid \boldsymbol{X}, z = 0, 1$, *we keep Assumption 2.1 to follow convention.*

Rosenbaum and Rubin (1983) introduced the propensity score $e(\boldsymbol{X}) = pr(Z = 1 \mid \boldsymbol{X})$ as the probability of receiving the active treatment conditioning on observed covariates. They showed that adjusting for the propensity score is sufficient for removing confounding bias under Assumption 2.1. It is worth noting that while the covariates may be high dimensional, the propensity score is always one-dimensional and lies within the unit interval. This dimension reduction property of the propensity score is partly responsible for its popularity among applied researchers.

2.2.2 Weighting estimators

Weighting estimators provide ways to obtain unbiased estimates of the ACE using the propensity score. In its simplest form, the IPW estimator is known as the Horvitz-Thompson estimator and weights individual observations by the reciprocal of the estimated propensity score (Horvitz and Thompson, 1952; Rosenbaum, 1987):

$$\hat{\Delta}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i},$$

where $\hat{e}(\boldsymbol{X})$ is the estimated propensity score and $\hat{e}_i = \hat{e}(\boldsymbol{X}_i)$. There have been many estimators based on refinements of $\hat{\Delta}_{HT}$, including the Hájek estimator which normalizes the weights in the Horvitz-Thompson estimator within the treatment and control group

(Hájek, 1971):

$$\hat{\Delta}_{Hájek} = \frac{\sum_{i=1}^{N} Z_i Y_i/\hat{e}_i}{\sum_{i=1}^{N} Z_i/\hat{e}_i} - \frac{\sum_{i=1}^{N}(1-Z_i)Y_i/(1-\hat{e}_i)}{\sum_{i=1}^{N}(1-Z_i)/(1-\hat{e}_i)},$$

and the doubly robust estimator (Robins et al., 1994):

$$\hat{\Delta}_{DR} = \frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i - (Z_i - \hat{e}_i) m_1(\boldsymbol{X}_i, \hat{\alpha}_1)}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1-Z_i)Y_i - (Z_i - \hat{e}_i) m_0(\boldsymbol{X}_i, \hat{\alpha}_0)}{1-\hat{e}_i},$$

where $m_z(\boldsymbol{X}, \hat{\alpha}_z)$ is the model estimate for $E(Y|Z=z, \boldsymbol{X})$ obtained via outcome regression.

The propensity score weighting estimators have very attractive theoretical properties. For example, under correct model specifications, $\widehat{\Delta}_{DR}$ attains the semi-parametric efficiency bound; furthermore, it is doubly robust in the sense that it is consistent if either the propensity score model or the outcome regression model is correctly specified. However, this theoretical appeal may be diminished in practice by sensitivity to model misspecification (Kang and Schafer, 2007). Instead, the subclassification estimators are more robust to model misspecification (Drake, 1993; Kang and Schafer, 2007).

2.2.3 Subclassification estimators

The propensity score subclassification estimator involves stratifying units into subclasses based on estimated propensity scores, and then directly comparing treated and control units within the same subclass (Rosenbaum and Rubin, 1984). Formally, let $[\hat{e}_{min}, \hat{e}_{max}]$ be the range of estimated propensity scores; $\hat{C}_k = [\hat{q}_{k-1}, \hat{q}_k)$ $(k = 1, \ldots, K)$ be disjoint divisions of the interval $[\hat{e}_{min}, \hat{e}_{max})$; $n_k = \sum_{i=1}^{N} I(\hat{e}(\boldsymbol{X}_i) \in \hat{C}_k)$ and $n_{zk} = \sum_{i=1}^{N} I(\hat{e}(\boldsymbol{X}_i) \in \hat{C}_k) I(Z_i = z), z = 0, 1$. Then the subclassification estimator is

$$\hat{\Delta}_S = \sum_{k=1}^{K} \frac{n_k}{N} \left\{ \frac{1}{n_{1k}} \sum_{i=1}^{N} Z_i Y_i I(\hat{e}_i \in \hat{C}_k) - \frac{1}{n_{0k}} \sum_{i=1}^{N} (1-Z_i) Y_i I(\hat{e}_i \in \hat{C}_k) \right\}.$$

Note that due to strong ignorability of the propensity score, we have

$$\Delta = E_{e(\boldsymbol{X})}\left\{E[Y|e(\boldsymbol{X}), Z=1] - E[Y|e(\boldsymbol{X}), Z=0]\right\}. \tag{2.1}$$

The subclassification estimator can hence be viewed as a histogram approximation to (2.1).

Most applied publications choose $K = 5$ based on Rosenbaum and Rubin (1984)'s recommendation, in which case the cut-off points are often chosen as sample quintiles. It is well-known that when K is fixed, $\hat{\Delta}_S$ is biased and inconsistent for estimating Δ due to residual bias (see e.g., Lunceford and Davidian, 2004).

2.3 Methodology

2.3.1 A hybrid estimator

In this section, we study a hybrid of the subclassification estimator $\hat{\Delta}_S$ and the Horvitz-Thompson estimator $\hat{\Delta}_{HT}$. The hybrid estimator provides a way to consistently estimate the ACE using propensity score subclassification. Furthermore, as we describe in Section 2.3.3, the hybrid estimator motivates a novel robust weighting scheme that improves upon model-based propensity score weights. This improvement is achieved independently of the outcome data, and hence can be combined with any weighting method.

The key to constructing the hybrid estimator is the intrinsic connection between the two seemingly unrelated estimators: $\hat{\Delta}_S$ and $\hat{\Delta}_{HT}$. Specifically, as noted by Imbens and Rubin (2015, §17.8), $\hat{\Delta}_S$ can be seen as a coarsened version of $\hat{\Delta}_{HT}$. In fact, if we denote $p_k = pr(Z = 1 \mid \hat{e}_i \in \hat{C}_k)$, then the equality $p_k E(n_k) = E(n_{1k})$ suggests the *infeasible* estimator

$$\hat{\Delta}_{S-HT} = \sum_{k=1}^{K}\left\{\frac{1}{Np_k}\sum_{i=1}^{N} Z_i Y_i I(\hat{e}_i \in \hat{C}_k) - \frac{1}{N(1-p_k)}\sum_{i=1}^{N}(1-Z_i)Y_i I(\hat{e}_i \in \hat{C}_k)\right\} \tag{2.2}$$

may also provide a good approximation for Δ. Note p_k is well-defined as $\hat{e}_1, \ldots, \hat{e}_N$ are identically distributed. However, it is generally unknown in practice, and thus $\hat{\Delta}_{S-HT}$ is

infeasible. Note that $\hat{\Delta}_S$ can be seen as a result of substituting the empirical estimate of p_k (i.e. n_{1k}/n_k) into (2.2). On the other hand, (2.2) can be rewritten in a compact form:

$$\hat{\Delta}_{S-HT} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i I(Z_i=1)}{p_{\hat{k}_i}} - \frac{Y_i I(Z_i=0)}{1-p_{\hat{k}_i}}\right), \tag{2.3}$$

where $\hat{k}_i = k$ if $\hat{e}(\boldsymbol{X}_i) \in \hat{C}_k$. (2.3) has a similar form to $\hat{\Delta}_{HT}$ except that it uses the same weights for all units in the same subclass.

The properties of $\hat{\Delta}_S$ and $\hat{\Delta}_{HT}$ derive from their approach to estimate p_i (or $p_{\hat{k}_i}$). The subclassification estimator uses the same weights for all subjects in the same subclass, which reduces the variance but has bias from coarsening. Moreover, as $\hat{\Delta}_S$ only uses the rank information from the model-based PS estimates (see Figure 2.1), it is robust to misspecification of the propensity score model. In contrast, the Horvitz-Thompson estimator uses a separate model-based estimate for each individual weight. As a result, the corresponding estimator is consistent if the model is correctly specified, but can be highly variable especially when the propensity score is close to zero.

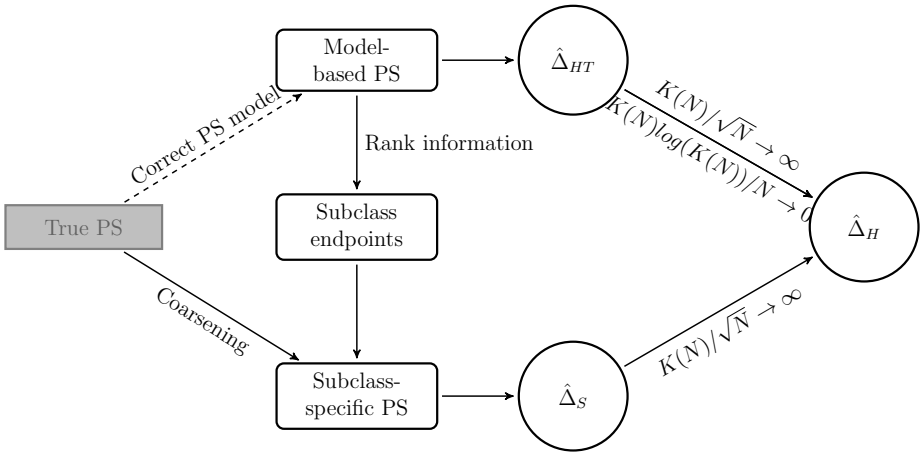

Figure 2.1: Conceptual comparison between $\hat{\Delta}_{HT}$, $\hat{\Delta}_S$ and $\hat{\Delta}_H$.

The motivation for the hybrid estimator is to find a balance for this bias-variance trade-off. Note that a larger number of subclasses would reduce bias, but potentially leads to higher

variance. We hence consider increasing the number of subclasses in $\hat{\Delta}_S$ with sample size such that with large enough sample size the coarsened weights can approximate the individual (model-based) weights to an arbitrary level, while with small sample size the coarsened weights are much more stable than the individual (model-based) weights. Formally, we define a hybrid estimator as follows:

$$\hat{\Delta}_H = \sum_{k=1}^{K(N)} \frac{n_k}{N} \left\{ \frac{1}{n_{1k}} \sum_{i=1}^{N} Z_i Y_i I(\hat{e}_i \in \hat{C}_k) - \frac{1}{n_{0k}} \sum_{i=1}^{N} (1 - Z_i) Y_i I(\hat{e}_i \in \hat{C}_k) \right\},$$

where we write $K = K(N)$ to emphasize that the number of subclasses K is a function of the sample size N. With slight abuse of notation, we define $\hat{C}_k, n_k$ and n_{zk} as in Section 2.2.3, with $K(N)$ replacing K in the original definitions. Following convention, we stratify at *quantiles* of estimated propensity scores such that $n_1 \approx \cdots \approx n_{K(N)} \approx N/K(N)$.

The key to the theoretical justification of $\hat{\Delta}_H$ is studying the rate at which the number of subclasses should increase with the sample size, to which we now turn.

2.3.2 Theoretical properties

We now discuss choice for the number of subclasses in $\hat{\Delta}_H$. Intuitively $K(N)$ should increase fast enough with N so that the residual bias is negligible asymptotically. This is formalized in Theorem 2.3, with proof in Appendix B.

Theorem 2.3. *Assume that Assumption 2.1 and the regularity conditions in Appendix A hold, $\hat{\Delta}_H$ is well-defined and additionally as $N \to \infty$,*

$$K(N) \to \infty. \tag{2.4}$$

Then $\hat{\Delta}_H$ is a consistent estimator for Δ, i.e. $\hat{\Delta}_H \to_p \Delta$. If we assume additionally that as $N \to \infty$,

$$K(N)/\sqrt{N} \to \infty, \tag{2.5}$$

then $\hat{\Delta}_H$ *is a root-N consistent estimator for* Δ.

Recall that the subclassification estimator essentially uses histograms to approximate Δ. The key insight given by Theorem 2.3 is that to achieve root-N consistency, smaller bandwidths are needed in the histogram approximation. This is similar in spirit to kernel density estimation methods that use under-smoothing to achieve root-N consistency (e.g. Newey, 1994; Newey et al., 1998; Paninski and Yajima, 2008).

On the other hand, for the hybrid estimator to be well-defined, the number of subclasses should grow slowly enough so that for all subclasses, there is at least one observation from each treatment group. This is formalized in Theorem 2.4, with proof in Appendix C.

Theorem 2.4. *Assume that the regularity conditions in Appendix A hold and additionally as* $N \to \infty$,

$$(K(N))log(K(N))/N \to 0, \tag{2.6}$$

then $\hat{\Delta}_H$ *is asymptotically well defined:* $pr(n_{zk} > 0, \text{ for all } z, k) \to 1$.

Theorem 2.3 and 2.4 provide theoretical guidelines for the choice of $K(N)$. In practice, we propose to choose the maximal number of subclasses such that the hybrid estimator is well-defined:

$$K_{max} = \max\{K : \Delta_H \text{ is well-defined}\}.$$

In other words, we choose the largest K such that for all subclasses $\hat{C}_1, \ldots, \hat{C}_K$ there is at least one observation from each treatment group. The resulting estimator is called the *full subclassification* estimator:

$$\hat{\Delta}_{FS} = \sum_{k=1}^{K_{max}} \frac{n_k}{N} \left\{ \frac{1}{n_{1k}} \sum_{i=1}^{N} Z_i Y_i I(\hat{e}_i \in \hat{C}_k) - \frac{1}{n_{0k}} \sum_{i=1}^{N} (1 - Z_i) Y_i I(\hat{e}_i \in \hat{C}_k) \right\}.$$

It is easy to see that $\hat{\Delta}_{FS}$ satisfies the rate conditions in Theorem 2.3. We emphasize that the definition of K_{max} does not use information from the outcome data, and is thus aligned with the original spirit of propensity score adjustment (Rubin, 2007).

The full subclassification estimator is closely related to the full matching estimator (Rosenbaum, 1991; Hansen, 2004; Stuart, 2010), which creates multiple matched sets such that each matched set contains either one treated subject and more than one control subjects, or one control subject and more than one treated subjects. The full matching estimator is essentially a subclassification estimator with the maximal number of subclasses. Our approach differs from full matching in that we subclassify by quantiles of the observed data, thereby achieving subclasses with (approximately) equal number of observations. In contrast, the full matching estimator can have different number of units in different subclasses. In addition, given a parametric propensity score model, the full subclassification estimator is unique, while the optimal full matching estimator depends on the distance measure used for matching.

2.3.3 The full subclassification weighting method

The hybrid estimator, and in particular the full subclassification estimator motivates a novel robust weighting scheme via subclassifying the model-based propensity score weights. As discussed in Section 2.3.1, $\hat{\Delta}_H$ can be written as a weighting estimator with weights defined by

$$w_H = \begin{cases} 1/\hat{p}_{\hat{k}_i} & \text{if } Z = 1 \\ 1/(1-\hat{p}_{\hat{k}_i}) & \text{if } Z = 0 \end{cases}, \tag{2.7}$$

where $\hat{p}_k = n_{1k}/n_k$ and $\hat{k}_i = k$ if $\hat{e}(\boldsymbol{X}_i) \in \hat{C}_k$. In particular, when we set $K(N) = K_{max}$, (2.7) is called the *full subclassification weight*. Compared with the standard inverse probability weight, the (full) subclassification weight can be viewed as replacing $\hat{e}_i$ with the coarsened estimate $\hat{p}_{\hat{k}_i}$.

Following Rubin (2007), the (full) subclassification weight is constructed independently of the outcome data. Therefore in principle, it can be applied to any IPW-based estimator. In fact, as the reciprocal of the (full) subclassification weight always lies within the unit interval, it can be regarded as an estimator for the propensity score itself. In what follows,

we use superscript H or FS to denote the corresponding weighting scheme.

As advocated by Rubin (2007), the propensity scores should be estimated in a way such that different model-based adjustments tend to give similar answers. In Proposition 2.5, we show that based on the (full) subclassification weight, the Horvitz-Thompson estimator coincides with the Hájek estimator. This is appealing as the latter has better statistical properties in terms of both efficiency and robustness (Lunceford and Davidian, 2004), while the former is easier to describe and arguably more widely used in practice. The proof is given in Appendix D.

Proposition 2.5. *If we estimate the propensity score with $\hat{p}_{\hat{k}_i}$, then the Horvitz-Thompson estimator coincides with the Hájek estimator: $\hat{\Delta}^H_{HT} = \hat{\Delta}^H_{Hájek}$. In particular, $\hat{\Delta}^{FS}_{HT} = \hat{\Delta}^{FS}_{Hájek}$.*

2.3.4 Relation to covariate balancing weighting schemes

The full subclassification weighting method is closely related to the covariate balancing weighting methods, which also aim to achieve robust estimates of the ACE without using the outcome data (Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2015). These methods are designed to reduce empirical covariate imbalance between the treatment groups, as it may result in severe bias in the final causal effect estimates. Prior to these methods, practitioners often try multiple propensity score models until a sufficiently balanced solution is found; this cyclic process is known as the propensity score tautology (Imai et al., 2008). To avoid this, the covariate balancing methods directly weight observations in a way that the empirical moments of pre-specified covariates are balanced in the weighted sample. These methods are appealing to many practitioners as they often achieve improved or even exact empirical balance for commonly-used moment conditions.

However, balancing certain moment conditions does not imply balancing the multivariate *covariate distributions* in treatment groups, which is required for unbiased estimation of the ACE with any response pattern. In contrast, within strata defined by specific values of the propensity score, the multivariate covariates distributions are balanced between treatment

and control groups. Therefore the covariate balancing conditions may be used as a supplement to propensity score adjustment methods, but caution should be exercised in applying these methods as they can create an illusion in the balance of covariate distributions. For example, constructions of just-identified covariate balancing propensity score (CBPS) (Imai and Ratkovic, 2014) and stable balancing weights (SBW) (Zubizarreta, 2015) rely solely on certain covariate balancing conditions. Consequently, the validity of these methods depends on shape of the response surface, something that cannot possibly be checked from data at the design stage. In contrast, the over-identified CBPS explicitly incorporates a propensity score model, and the empirical balancing (EB) weights (Hainmueller, 2012; Zhao and Percival, 2015) as well as the empirical balancing calibration (CAL) weights (Chan et al., 2015) implicitly fit a logistic model for the propensity score model. These methods yield consistent estimates for the ACE if the corresponding propensity score model is correct.

Furthermore, as pointed out by Zubizarreta (2015), tighter covariate balance generally comes at a cost in term of weight stability. Although the covariate balancing conditions can be used to *eliminate* biases due to imbalance in moment conditions (Hainmueller, 2012; Chan et al., 2015), as we show later in empirical studies, they can give rise to extreme weights even with a correct propensity score model. This instability of weight estimates not only increase the variance of the final causal effect estimates, but also make them highly sensitive to outliers in the outcome data. In contrast, the full subclassification weighting method often achieves a good compromise for this covariate balance-stability trade-off.

The full subclassification weighting method has several additional features compared to individual covariate balancing methods. First, based on the full subclassification weighting method, the Horvitz-Thompson estimator is consistent for estimating the ACE (with a correct propensity score model). In contrast, even under a linear response pattern, SBW only yields an approximately unbiased estimate of the ACE. Second, the reciprocal of full subclassification weights have the interpretation of coarsened propensity scores; in particular, they always lie within the unit interval. Consequently, the full subclassification weighting methods can be conceptualized as creating a pseudo population through inverse probability

weighting. In contrast, although the reciprocal of normalized EB and CAL weights imply propensity scores asymptotically, they can be greater than 1 or even negative in small sample settings. This is concerning for many practitioners given the "black box" involved in estimating these weights. Third, we allow any parametric form for the posited propensity score model, whereas the default version of the EB and CAL method both implicitly assume a logistic model. Fourth, calculating the full subclassification weight is a convex problem as long as the parameter estimation in the propensity score model is convex. In contrast, it was reported in the literature that even with a logistic regression model, the optimization problem of CBPS might be non-convex, so that it may be difficult to find the global solution in practice (Zhao and Percival, 2015).

2.4 Simulation Studies

In this section we evaluate the finite sample performance of the proposed full subclassification weighting method. We compare it to classical subclassification and weighting estimators, as well as various covariate balancing weighting schemes. Our simulation setting is similar to that of Kang and Schafer (2007), which has become a standard setting for evaluating the performance of propensity score weighting methods (e.g., Imai and Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2015). We also modify Kang and Schafer (2007)'s setting to evaluate the sensitivity of simulation results to the shapes of the propensity score and response surface.

Specifically, our simulation data consist of N independent samples from the joint distribution of $(Z, \boldsymbol{X}, Y, \boldsymbol{W})$. The covariates $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ follow a standard multivariate normal distribution $N(0, I_4)$, where I_4 is a 4×4 identity matrix. The treatment variable Z follows a Bernoulli distribution with mean $\text{expit}(\boldsymbol{X}\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (-1, 0.5, -0.25, -0.1)^T$. Conditional on $\boldsymbol{X}$, the potential outcome $Y(z)$ is defined by the linear model $Y(z) = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon$, where the independent error term ϵ follows a standard normal distribution. The observed outcome is generated following consistency: $Y = Y(z)$ if $Z = z$. Following Kang and Schafer (2007), we consider combinations of whether the propensity score and the outcome regression model is correctly specified. To correctly specify the

propensity score model, we (correctly) include $\boldsymbol{X}$ in the posited logistic regression model. Otherwise we include covariates $\boldsymbol{W}$, which are non-linear transformations of $\boldsymbol{X}$ given by $W_1 = \exp(X_1/2), W_2 = X_2/(1+\exp(X_1))+10, W_3 = (X_1X_3/25+0.6)^3, W_4 = (X_2+X_4+20)^2$. Similarly for specifications of the outcome regression model. We are interested in estimating the ACE, whose true value is 0. All the simulation results are based on average of 1000 random samples.

We first compare the full subclassification estimator with the classical subclassification estimator $\hat{\Delta}_S$ (with $K = 5$) and the Hájek estimator $\hat{\Delta}_{Hájek}$. $\hat{\Delta}_{HT}$ is not included as it performs uniformly worse than $\hat{\Delta}_{Hájek}$, and $\hat{\Delta}_{DR}$ is included later as its performance depends on an additional outcome regression model. For completeness, we include the full matching estimator, which is implemented with the default options in `R` package `MatchIt`. As pointed out by Stuart (2010), these four estimators represent a continuum in terms of the number of subclasses formed. Figure 2.2 presents the results. When the propensity score model is correctly specified, the classical subclassification estimator is not consistent; in fact, its bias stabilizes with increasing sample size. All the other three estimators are consistent for the ACE. Among them, $\hat{\Delta}_{FS}$ has the smallest RMSE, with comparable performance only when the sample size is very small. This shows that $\hat{\Delta}_{FS}$ achieves a good balance for the bias-variance trade-off discussed in Section 2.3.1. When the propensity score model is misspecified, the Hájek estimator is severely biased: the bias and RMSE grow with sample size! Consistent with previous findings in the literature, the other three estimators are more robust to model misspecification. Among them, $\hat{\Delta}_{FS}$ and $\hat{\Delta}_{FM}$ exhibit better performance than $\hat{\Delta}_S$ in term of both bias and RMSE.

We then compare various weighting schemes for the three classical weighting estimators introduced in Section 2.2.2: $\hat{\Delta}_{HT}, \hat{\Delta}_{Hájek}$ and $\hat{\Delta}_{DR}$. The weights we consider include true weights; logit weights, obtained by inverting propensity score estimates from a logistic regression model; trimmed weights, obtained by trimming logit weights at their 95% percentiles; (over-identified) covariate balancing propensity score (CBPS) weights of Imai and Ratkovic (2014); empirical balancing calibration weights of Chan et al. (2015) implied by exponential

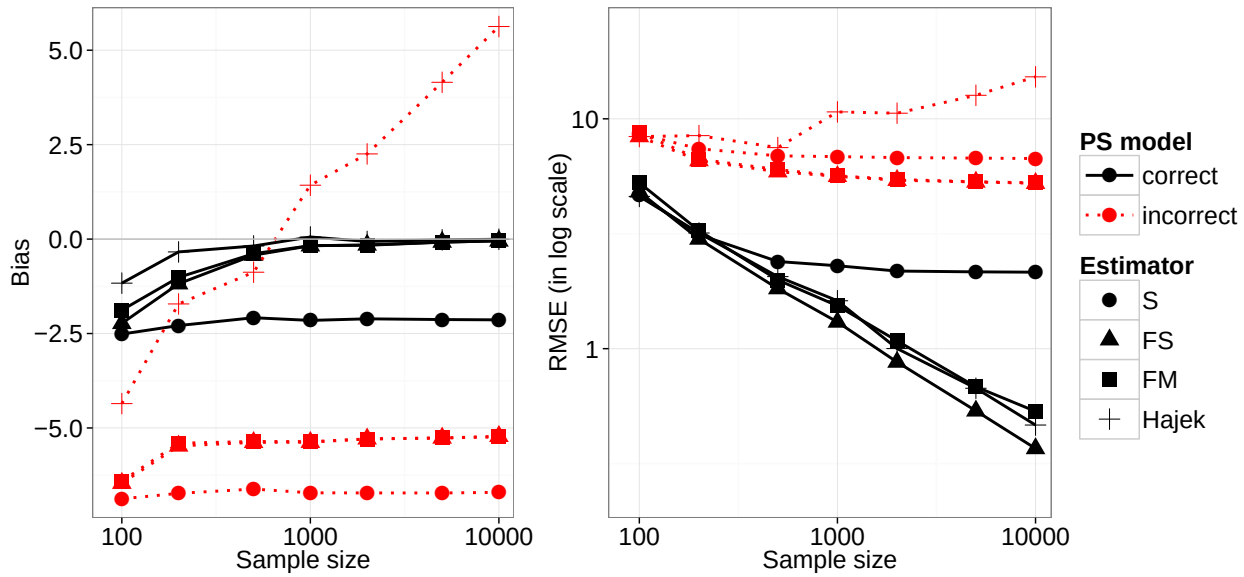

Figure 2.2: Bias and root mean squared error (RMSE) of the classical subclassification estimator (S), the full subclassification estimator (FS), the full matching estimator (FM) and the Hájek estimator (Hajek) under Kang and Schafer (2007)'s setting.

tilting (CAL-ET) or quadratic loss (CAL-Q) and the proposed full subclassification (FS) weights. We use the default options of `R` packages `CBPS` and `ATE` for calculating the CBPS and CAL weights, respectively.

As part of the design stage, we use Figure 2.3 to visualize the weight stability of various weighting schemes, and Table 2.1 to assess the covariate balance after weighting. The covariate balance is measured using the standardized imbalance measure (Rosenbaum and Rubin, 1985; Chan et al., 2015):

$$Imb = \left\{ \left(\frac{1}{N} \sum_{i=1}^{N} [(Z_i w_{1i} - (1 - Z_i) w_{0i}) \boldsymbol{X}_i]^T \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} (Z_i w_{1i} - (1 - Z_i) w_{0i}) \boldsymbol{X}_i \right) \right\}^{1/2},$$

where w_{1i} are weights for the treated, and w_{0i} are weights for the controls. One can see that logit weights perform reasonably well with a correct propensity score model. However, with the misspecified propensity score model, they become highly unstable and cause severely imbalanced covariate distributions between treatment groups. The CAL weights may look

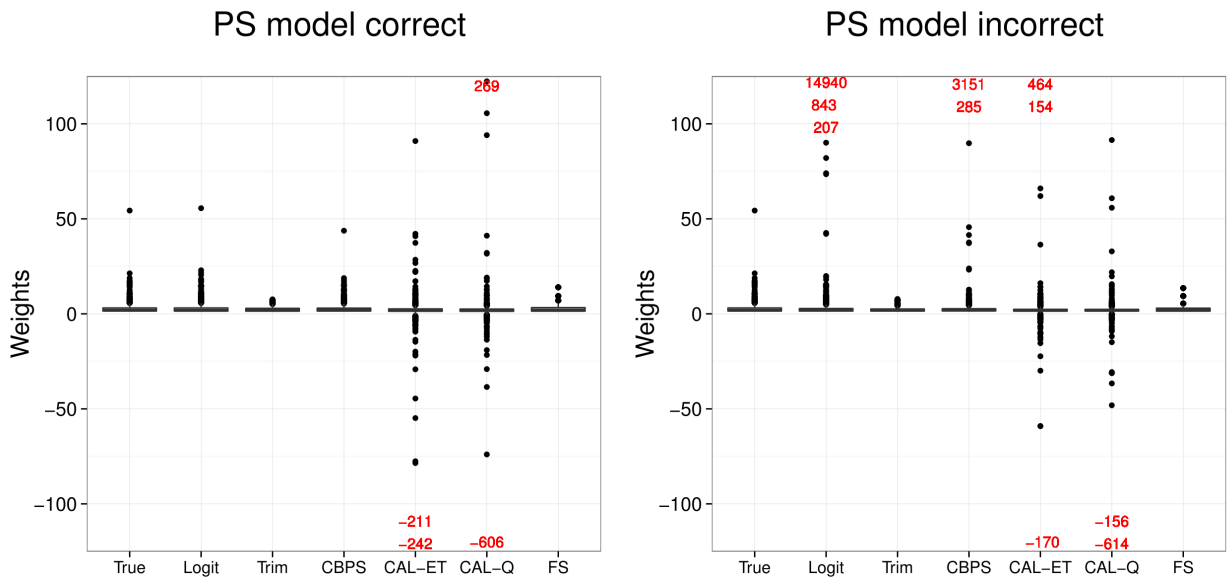

Figure 2.3: Distributions of weight estimates with various weighting scheme with a random simulated data set of sample size 1000. Weights outside of the plot range are annotated on the boarders.

very appealing as by design, they achieve exact balance in the first moment conditions (and hence the standardized imbalance measure) between treatment groups. However, as one can see from Figure 2.3, they are highly unstable even under correct propensity score model specification. Consequently the causal effect estimate may be driven by some highly influential observations. In contrast, CBPS and FS weights improve upon the logit weights in term of both stability and covariate balance, with FS exhibiting uniformly better performance than CBPS. The performance of FS on covariate balance is particularly impressive as it does not (directly) target at achieving covariate balance between treatment groups.

Table 2.2 summarizes the performance of various weighting schemes when applied to the three classical weighting estimators. For brevity, we only show results with sample size fixed at 1000. Consistent with the findings of Kang and Schafer (2007), logit weights are sensitive to misspecification of the propensity score model, regardless of whether the weighting estimator is doubly robust or not. Use of the full subclassification weights or the

Table 2.1: Standardized imbalance measures of various weighting schemes under Kang and Schafer (2007)'s setting. We consider both correct (✓) and incorrect (×) specifications of the propensity score (PS) model*.

Sample size	Model	Weighting scheme				
	PS	logit	CBPS	CAL-ET	CAL-Q	FS
200	✓	0.16	0.19	0.00	0.00	0.16
	×	0.52	0.19	0.00	0.00	0.17
1000	✓	0.07	0.09	0.00	0.00	0.07
	×	0.70	0.11	0.00	0.00	0.08
5000	✓	0.03	0.04	0.00	0.00	0.03
	×	6.09	0.12	0.00	0.00	0.06

*: For the covariate balancing weighting schemes, we say the propensity score model is "correctly specified" if we impose balancing conditions on $\boldsymbol{X}$, and say the propensity score model is "misspecified" if we impose balancing conditions on $\boldsymbol{W}$.

covariate balancing weights (CBPS, CAL-ET and CAL-Q) greatly improves upon the naive weights obtained from a logistic regression model. Among them, FS, CAL-ET and CAL-Q weights perform better than CBPS weights under all simulation settings. Moreover, the IPW estimator coincides with the Hájek estimator with the former three weights. Within these three, CAL-ET and CAL-Q tend to perform better when the propensity score model is correctly specified, while FS performs better otherwise. As argued by many previous researchers (Zubizarreta, 2015; Chan et al., 2015), it is very likely that the posited models are wrong in practice. Hence the robustness to model misspecification may be worth more attention than performance under correct model specification.

There has been a conjecture that a small subset of samples with extremely large weights are partly responsible for the bad performance of logit weights in this setting (e.g., Robins et al., 2007b). To gain insights into the improved performance of CBPS, CAL-ET, CAL-Q and FS weights, we compare these weights to the trimmed version of logit weights, which excludes the largest 5% weights from logit weights. The RMSE of CBPS weights is comparable or worse than that of trimmed weights, suggesting that the improvement of CBPS

Table 2.2: Bias and RMSE of classical weighting estimators with various weighting schemes under Kang and Schafer (2007)'s setting. We consider both correct (✓) and incorrect (×) specifications of the propensity score (PS) model* or the outcome regression (OR) model. The sample size is 1000.

Estimator	Model		Weighting scheme						
	PS	OR	True	logit	Trim	CBPS	CAL-ET	CAL-Q	FS
Bias									
$\hat{\Delta}_{HT}$	✓	–	-0.30	0.02	-0.08	-1.45	0.00	0.00	-0.21
$\hat{\Delta}_{HT}$	×	–	–	33.18	-5.25	3.28	-5.77	-5.82	-5.37
$\hat{\Delta}_{Hájek}$	✓	–	-0.23	-0.06	-0.09	-1.47	0.00	0.00	-0.21
$\hat{\Delta}_{Hájek}$	×	–	–	0.89	-5.96	-5.93	-5.77	-5.82	-5.37
$\hat{\Delta}_{DR}$	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\Delta}_{DR}$	✓	×	-0.12	-0.03	-0.05	-0.56	0.00	-0.25	-0.01
$\hat{\Delta}_{DR}$	×	✓	–	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\Delta}_{DR}$	×	×	–	-10.73	-4.83	-6.03	-5.77	-5.82	-4.39
RMSE									
$\hat{\Delta}_{HT}$	✓	–	16.44	6.48	4.97	5.85	0.07	0.07	1.25
$\hat{\Delta}_{HT}$	×	–	–	135.01	7.05	8.06	5.93	5.97	5.61
$\hat{\Delta}_{Hájek}$	✓	–	3.22	1.53	1.38	1.97	0.07	0.07	1.25
$\hat{\Delta}_{Hájek}$	×	–	–	9.96	6.15	6.24	5.93	5.97	5.61
$\hat{\Delta}_{DR}$	✓	✓	0.08	0.08	0.07	0.08	0.07	0.07	0.08
$\hat{\Delta}_{DR}$	✓	×	1.58	1.10	0.91	1.11	0.58	0.61	0.83
$\hat{\Delta}_{DR}$	×	✓	–	0.26	0.07	0.08	0.08	0.07	0.08
$\hat{\Delta}_{DR}$	×	×	–	43.43	4.94	6.28	5.93	5.97	4.55

*: For the covariate balancing weighting schemes, we say the propensity score model is "correctly specified" if we impose balancing conditions on $\boldsymbol{X}$, and say the propensity score model is "misspecified" if we impose balancing conditions on $\boldsymbol{W}$.

weights over logit weights is mainly due to stabilizing the extreme weights. In contrast, FS weights outperform trimmed weights in all scenarios, and CAL-ET and CAL-Q weights are better than trimmed weights except when both the propensity score and outcome regression models are misspecified.

We can also see that CAL-ET generally performs better than CAL-Q. In particular, when only the propensity score model is correctly specified, the bias of $\hat{\Delta}_{DR}^{CAL-Q}$ is much larger in magnitude than that of $\hat{\Delta}_{DR}^{CAL-ET}$ and $\hat{\Delta}_{DR}^{FS}$; here the superscripts denote the weighting schemes used to estimate the propensity score weights. Moreover, this bias stabilizes with increasing sample size (results not shown), showing that $\hat{\Delta}_{DR}^{CAL-Q}$ is not doubly robust (in the usual sense). This can be explained by the implicit correspondence between the objective function used in calculating the CAL weights and the posited propensity score model. Specifically, the objective function of CAL-ET corresponds to fitting a logistic regression model for the propensity score, while the objective function of CAL-Q does not. Since the propensity score here follows a logistic model, it is not surprising that CAL-Q is not "doubly robust."

To further illustrate this implicit correspondence and its implications, we consider an alternative simulation setting, where the treatment variable Z follows a Bernoulli distribution with mean $1-\exp(-\exp(\boldsymbol{X}\boldsymbol{\gamma}))$ and the potential outcome $Y(z)$ is defined by the linear model $Y(z) = \log(210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4) + \epsilon$. Here we ignore the fact that the linear term inside the logarithm may be non-positive as it is extremely unlikely to happen under our simulation setting. The true value for the ACE remains 0. The complementary log-log link used here is an asymmetric alternative to the logit link for modeling binary data. As (at least for practitioners) it is difficult to modify the objective function in Chan et al. (2015) to fit a complementary log-log model, we still compare the three estimators: $\hat{\Delta}_{DR}^{FS}$, $\hat{\Delta}_{DR}^{CAL-ET}$ and $\hat{\Delta}_{DR}^{CAL-Q}$. Figure 2.4 presents the comparison results with the outcome regression model misspecified. When the sample size is 100, $\hat{\Delta}_{DR}^{CAL-ET}$ occasionally fails to produce an estimate: 2.4% with a correct propensity score model, and 1.0% with an incorrect propensity score model. These scenarios are omitted from these plots. As one can see from

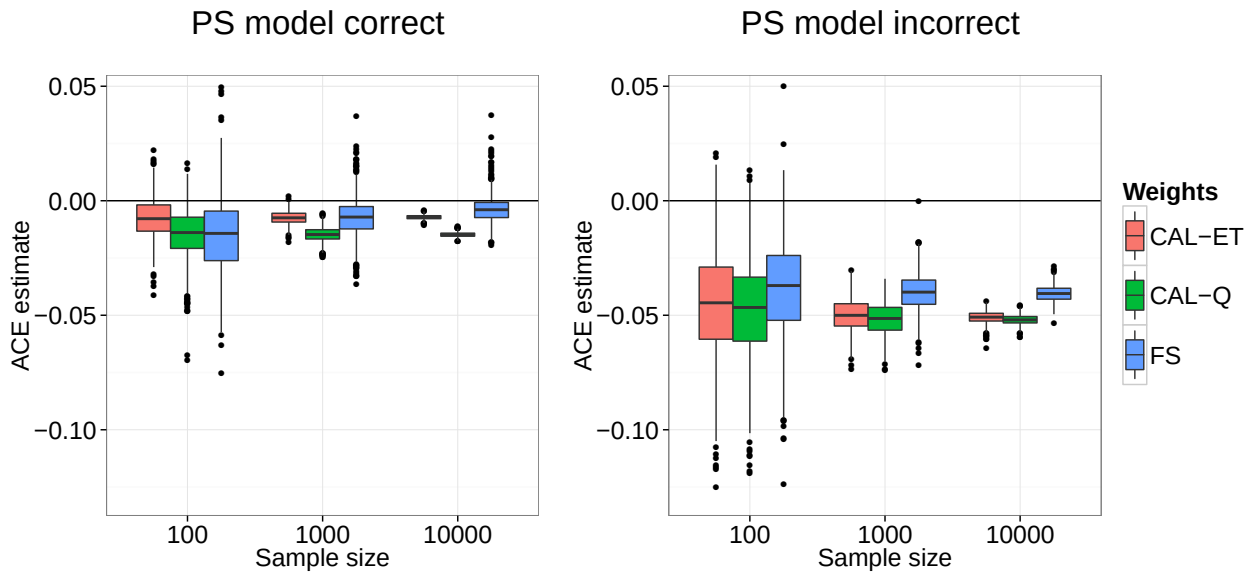

Figure 2.4: Boxplots of ACE estimates obtained with $\hat{\Delta}_{DR}^{CAL-ET}$, $\hat{\Delta}_{DR}^{CAL-Q}$ and $\hat{\Delta}_{DR}^{FS}$. The outcome regression model is misspecified for both plots; the propensity score model is correctly specified for the left panel, and misspecified for the right panel. The horizontal line at zero corresponds to the true value of the ACE.

the boxplots, CAL-ET and CAL-Q are not consistent even with correct propensity score model specification. When the propensity score model is misspecified, consistent with our findings before, $\hat{\Delta}_{DR}^{FS}$ performs better in term of both bias and RMSE.

2.5 Application to a childhood nutrition study

We illustrate the use of the proposed full subclassification weighting method using data from the 2007-2008 National Health and Nutrition Examination Survey (NHANES), which is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The data set we use were created by Chan et al. (2015), which contains observations on 2330 children aged from 4 to 17. Of these children, 55.1% participated in the National School Lunch or the School Breakfast programs. These are federally funded meal programs primarily designed to provide meals for children from poor

neighborhoods in the United States. However, there have been concerns that meals provided through the program may cause childhood obesity (Stallings et al., 2010; Woo Baidal and Taveras, 2014). Hence here we study how participation in these meal programs contributes to childhood obesity as measured by body mass index (BMI). Following Chan et al. (2015), we control for the following potential confounders in our analysis: child age, child gender, child race (black, Hispanic versus others), coming from a family above 200% of the federal poverty level, participation in Special Supplemental Nutrition (SSN) Program for Women Infants and Children, participation in the Food Stamp Program, childhood food security as measured by an indicator of two or more affirmative responses to eight child-specific questions in the NHANES Food Security Questionnaire Module, any health insurance coverage, and the age and gender of the survey respondent (usually an adult in the family).

Table 2.4 (in Appendix E) summarizes baseline characteristics and outcome measure by participation status in the school meal programs. Children participating in the school meal programs are more likely to be black or Hispanic, and come from a family with lower social economic status. Respondents for such children also tend to be younger and female. These differences in baseline characteristics suggest that the observed mean difference in BMI, that is 0.53 kg/m^2 (95% CI [0.09, 0.98]), may not be fully attributable to the school meal programs.

We then apply various weighting and subclassification methods to estimate the effect of participation in the meal programs. We consider two models for the propensity score: a logistic model and a complementary log-log model. We also consider a linear outcome regression model on the log-transformed BMI. All the covariates enter the propensity score model or the outcome regression model as linear terms. Figure 2.5 visualizes the distributions of propensity score weights and their reciprocals. Results with the complementary log-log propensity score model are similar to those with the logistic regression model and are omitted. We can see that the reciprocals of propensity score weights estimated using the full subclassification method or the CBPS method lie within the unit interval. In contrast, the reciprocals of CAL weights can be negative or greater than 1. Hence these weights cannot

be interpreted as propensity scores. Furthermore, consistent with our findings in Figure 2.3, the CAL weights are much more dispersed than the other weights. The 5 most extreme weights estimated by CAL-ET are $1350, 1259, -959, -943, 677$, and those for CAL-Q are $324, 153, -97, -97, 85$. As these weights are obtained independently of the outcome data, the final causal estimates are highly sensitive to these outliers.

Table 2.3 summarizes the standardized imbalance measure and causal effect estimates. As advocated by Rubin (2007), propensity score weights should be constructed such that the final causal effect estimate is insensitive to the weighting estimator used. However, the propensity score weights estimated with a parametric model or the CBPS method tend to give different answers with different weighting estimators. With these weighting methods, an Horvitz-Thompson estimator would suggest that participation in the school meal programs led to a significantly lower BMI. The Hájek and DR estimator instead yield estimates that are much closer to zero. In contrast, the subclassification weights (both the classical ones and the FS weights) and the CAL weights have a consistent implication with different weighting estimators that participation in school meal programs have negligible effects on the BMI. Moreover, we note that although different parametric propensity score models may give rise to very different causal effect estimates with an Horvitz-Thompson estimator, they yield much closer estimates with a (full) subclassification estimator. These results show that the subclassification methods are robust against propensity score model misspecification.

2.6 Discussion

Propensity score weighting and subclassification methods are among the most popular tools for drawing causal inference from observational studies. To choose among these methods, practitioners often face a bias-variance trade-off as the weighting methods can be consistent while the subclassification methods are more robust to model misspecification. In this article, we connect these two approaches by increasing the number of subclasses in the subclassification estimators. We show that the bias of the propensity score subclassification estimator can be eliminated asymptotically if the number of subclasses increases at a certain rate with

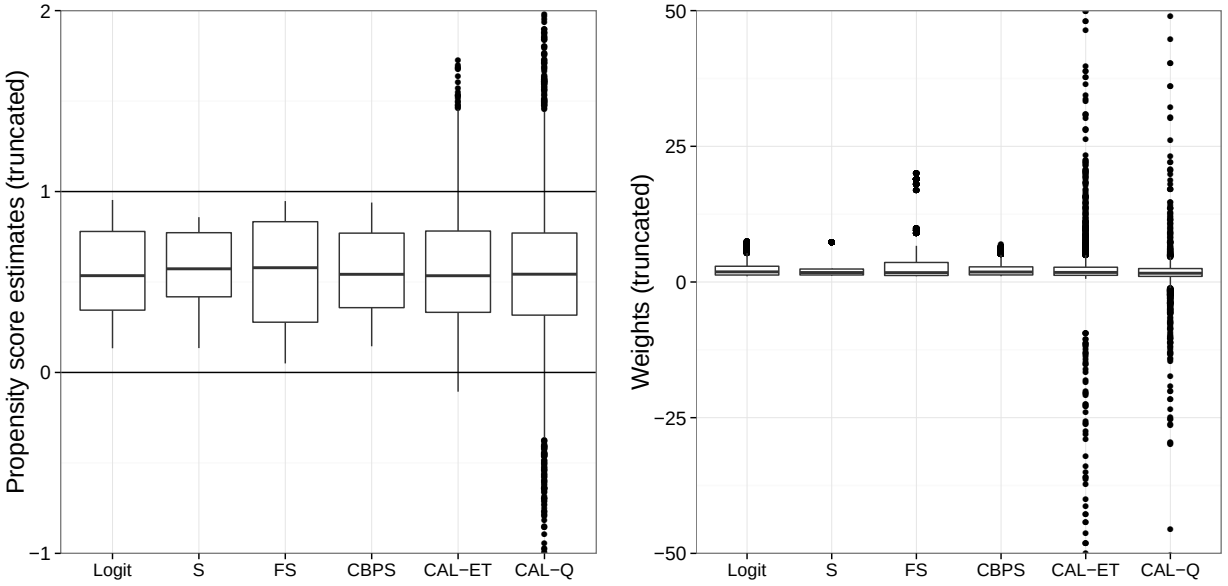

Figure 2.5: Distributions of propensity score estimates and weights with the NHANES data.

sample size. In particular, we propose a novel full subclassification estimator that inherits the advantages of both the classical IPW and subclassification method.

Moreover, we show that the full subclassification method can be used for robust estimation of propensity score weights. As discussed in detail by Zubizarreta (2015), a covariate balance-stability trade-off is key to constructing robust propensity score weights. Through extensive empirical studies, we show that the full subclassification weighting method achieves a good compromise in this trade-off, and dramatically improves upon model-based propensity score weights in both aspects, especially when the propensity score model is misspecified.

In this article, we have primarily focused on obtaining a good point estimate for the ACE. Although an explicit variance formula is available for the classical subclassification estimator with a fixed number of subclasses, it is likely to be complex in real-life situations and previous researchers have suggested using bootstrap estimates instead (Williamson et al., 2012). The explicit variance formula for the full subclassification estimator is challenging due to the uncertainty in the number of subclasses. Hence in practice, we recommend using bootstrap

Table 2.3: The average causal effect estimates associated with participation in the school meal programs. The 95% Wald-type confidence intervals in brackets are computed based on bootstrap estimates.

	Imbalance	HT	Hajek	DR
Naive	1.04	0.53 (0.09,0.98)	0.53 (0.09,0.98)	0.18 (-0.24,0.60)
Logit	0.10	-1.52 (-2.47,-0.56)	-0.16 (-0.63,0.32)	0.02 (-0.43,0.46)
S	0.08	-0.12 (-0.60,0.37)	-0.12 (-0.60,0.37)	0.00 (-0.45,0.45)
FS	0.12	-0.20 (-0.76,0.36)	-0.20 (-0.76,0.36)	-0.08 (-0.58,0.42)
Cloglog*	0.15	-2.26 (-3.70,-0.82)	-0.23 (-0.76,0.30)	0.04 (-0.44,0.51)
S + Cloglog	0.16	-0.05 (-0.54,0.43)	-0.05 (-0.54,0.43)	0.03 (-0.42,0.49)
FS + Cloglog	0.15	0.01 (-0.55,0.58)	0.01 (-0.55,0.58)	0.10 (-0.40,0.61)
CBPS	0.10	-1.25 (-2.12,-0.39)	-0.05 (-0.50,0.39)	0.03 (-0.41,0.47)
CAL-ET	0.00	-0.05 (-0.48,0.39)	-0.05 (-0.48,0.39)	0.08 (-0.35,0.51)
CAL-Q	0.00	-0.02 (-0.45,0.42)	-0.02 (-0.45,0.42)	0.10 (-0.33,0.53)

*: Cloglog indicates that the propensity scores are estimated with a complementary log-log model.

or subsampling methods to calculate the standard error and associated confidence intervals.

The full subclassification method in this article could be applied to address the missing data problem under the missing at random (MAR) assumption (see e.g. Rubin, 1978; Gelman and Meng, 2004; Kang and Schafer, 2007). It also extends directly to estimating the average treatment effect on the treated (ATT), and estimation of the generalized propensity score for multi-arm treatments (Imbens, 2000). Furthermore, since the full subclassification weights are constructed independently of the outcome data, it can potentially be applied to improve propensity score estimation in other contexts, such as causal inference with a marginal structural model (Robins et al., 2000) and in presence of interference (Tchetgen Tchetgen and VanderWeele, 2012).

Appendix

A Regularity conditions for Theorem 2.3

We now introduce the regularity assumptions needed for proving consistency of the hybrid estimator.

Assumption 2.2. *(Uniform positivity assumption) The support of* $e(\boldsymbol{X})$ *can be written as* $[e_{min}, e_{max}]$*, where* $e_{min} > 0$*,* $e_{max} < 1$*, and the quantile distribution of* $e(\boldsymbol{X})$ *is Lipschitz continuous.*

Assumption 2.2 implies that the cumulative distribution function of $e(\boldsymbol{X})$ has no flat portions between $[e_{min}, e_{max}]$, or the quantile distribution of $e(\boldsymbol{X})$ is continuous on $[0, 1]$. Violation of this assumption will cause some subclasses to be always empty, and the subclassification estimator to be ill-defined. This problem may be solved by considering only non-empty subclasses in constructing the subclassification estimator. For simplicity, we do not get into discussion of this issue here.

Assumption 2.3. *(Uniform consistency of estimated propensity scores) The propensity score model is correctly specified such that for all* $N, \hat{e}_i$ $(i = 1, \ldots, N)$ *is uniformly convergent in probability to* e_i $(i = 1, \ldots, N)$ *at* $\sqrt{N}$ *rate. Formally,* $\sqrt{N} \max_{1 \le i \le N} |\hat{e}_i - e_i| = O_p(1)$.

Under a smooth parametric model, the uniformity part in Assumption 2.2 and 2.3 can usually be inferred from uniform boundedness of the maximal norm of covariates, $|\boldsymbol{X}_i|_\infty$ (i=1,. . . ,N). The latter assumption holds if the support of the covariate $\boldsymbol{X}$ is a bounded set in $\mathbb{R}^p$, where p is the dimension of $\boldsymbol{X}$. This assumption has been widely used in the causal inference literature (for example, see Hirano et al., 2003).

As an example, suppose the true propensity model is the logit model:

$$e(\boldsymbol{X}) = \text{expit}(\boldsymbol{X}\beta) = \frac{\exp(\boldsymbol{X}\beta)}{1 + \exp(\boldsymbol{X}\beta)}.$$

In this case, $e(\boldsymbol{X}_i)$ is uniformly bounded away from 0 and 1 if $|\boldsymbol{X}_i|_\infty$ $(i = 1, \ldots, N)$ are uniformly bounded. At the same time, by the mean value theorem,

$$\sqrt{N}|\hat{e}(\boldsymbol{X}_i) - e(\boldsymbol{X}_i)| = \sqrt{N}|\text{expit}(\boldsymbol{X}_i\tilde{\beta})(1 - \text{expit}(\boldsymbol{X}_i\tilde{\beta}))\boldsymbol{X}_i(\hat{\beta} - \beta)| \leq \sqrt{N}\|\boldsymbol{X}_i\|_\infty\|\hat{\beta} - \beta\|_\infty,$$

where $\hat{\beta}$ is a consistent estimator of β, $\tilde{\beta}$ lies between $\hat{\beta}$ and β. Hence $\hat{e}(\boldsymbol{X}_i) - e(\boldsymbol{X}_i)$ $(i = 1, \ldots, N)$ is uniformly convergent in probability to zero at $\sqrt{N}$ rate if if $|\boldsymbol{X}_i|_\infty$ $(i = 1, \ldots, N)$ are uniformly bounded.

B Proof of Theorem 2.3

Under our assumptions, one can show via standard M-estimation theory that

$$\sqrt{N}(\hat{\Delta}_{HT} - \Delta) \to_d N(0, \Sigma_{HT}),$$

where Σ_{HT} is computed in Lunceford and Davidian (2004). To prove Theorem 2.3, we connect $\hat{\Delta}_H$ and $\hat{\Delta}_{HT}$ with an intermediate (infeasible) estimator $\hat{\Delta}_{S-HT}$. In the first step, Lemma 2.6 shows that the difference between $\hat{\Delta}_{S-HT}$ and $\hat{\Delta}_{HT}$ tends to zero. We defer the proof of Lemma 2.6 to the end of this section. In the second step, we show that the difference between $\hat{\Delta}_H$ and $\hat{\Delta}_{S-HT}$ tends to zero.

Lemma 2.6. *Under Assumption 2.1, the regularity conditions in Appendix A and condition* (2.4)*,*

(i) $\hat{\Delta}_{S-HT}$ *is consistent for estimating* Δ*:* $\hat{\Delta}_{S-HT} - \Delta = o_p(1)$;

(ii) if we assume additionally that (2.5) *holds, then* $\hat{\Delta}_{S-HT}$ *is* $\sqrt{N}$*-consistent for estimating* Δ*:* $\sqrt{N}(\hat{\Delta}_{S-HT} - \Delta) = O_p(1)$.

We now turn to the second step, in which we show

$$\sqrt{N}(\hat{\Delta}_H - \hat{\Delta}_{S-HT}) = O_p(1). \tag{2.8}$$

By symmetry, we only show (2.8) for the active treatment group, i.e.

$$\sqrt{N}\sum_{k=1}^{K}\frac{n_k}{N}\left\{\left(1-\frac{n_{1k}}{n_k p_k}\right)\frac{1}{n_{1k}}\sum_{i=1}^{N}Z_i Y_i I(\hat{e}_i\in\hat{C}_k)\right\}=O_p(1), \tag{2.9}$$

where K is used as a shorthand for $K(N)$.

Without loss of generalizability, we assume $n_1=\cdots=n_K=N/K\triangleq n$. As $\hat{\Delta}_H$ is well-defined, $n_{1k}\neq 0, k=1,\ldots,K$. Thus $n_{1k}\sim tBin(n,p_k)$, where $tBin$ denotes truncated binomial distribution with range $(0,n]$.

We will use Lyapunov central limit theorem to show (2.9). We denote

$$\begin{aligned}
h_k &= \frac{1}{1-p_k^n}-\frac{n_{1k}}{np_k}=c_{Nk}-\frac{n_{1k}}{np_k}\ (\text{note } E[h_k]=0);\\
m_{1k} &= \frac{1}{n_{1k}}\sum_{i=1}^{N}Z_i Y_i I(\hat{e}_i\in\hat{C}_k);\\
S_{NK} &= \frac{1}{K(N)}\sum_{k=1}^{K(N)}h_k m_{1k}.
\end{aligned}$$

Also let $e_{thres}=\min\{e_{min}/2,(1-e_{max})/2\}$. Note that equation (2.15) (see the proof of Lemma 2.6) implies $c_{Nk}-1=p_k^n/(1-p_k^n)=O((1-e_{thres})^n)$, which in turn implies

$$\sqrt{N}\sum_{k=1}^{K}\frac{n_k}{N}(1-c_{Nk})m_k=o_p(1).$$

Hence it suffices to show

$$\sqrt{N}S_{NK}=O_p(1).$$

By symmetry, we have

$$\begin{aligned}
E[m_{1k}|n_{1k}] &= E[Y_1|Z_1 I(\hat{e}_1)\in\hat{C}_k=1]\triangleq\mu_{1k},\\
E[m_{1k}^2|n_{1k}] &= \mu_{1k}^{12}+a_2\frac{1}{n_{1k}},
\end{aligned}$$

where a_2 is a linear combination of $\mu_{1k}^{11} \triangleq E[Y_1^2|Z_1 I(\hat{e}_1 \in \hat{C}_k) = 1]$ and $\mu_{1k}^{12} \triangleq E[Y_1 Y_2|Z_1 I(\hat{e}_1 \in \hat{C}_k) Z_2 I(\hat{e}_2 \in \hat{C}_k) = 1]$. Without loss of generalizability, we assume $\mu_{1k}^{11} \geq 1$ for all k (otherwise we could add 1 to all observed outcomes).

We hence have

$$E[h_k m_{1k}] = E[E(h_k m_{1k}|n_{1k})] = E[h_k]\,\mu_{1k} = 0,$$

$$\sigma_{Nk}^2 \triangleq E[(h_k m_{1k})^2] = \mu_{1k}^{11} \frac{q_k}{np_k} + O(1/n^2),$$

where $q_k = 1 - p_k$, and we use the fact that $E[1/n_{1k}] = \dfrac{np_k}{(np_k + q_k)^2} + O(1/n^2)$ (Znidaric, 2005). At the same time, we can show that

$$\gamma_{Nk} \triangleq E[(h_k m_{1k})^4] = O(1/n).$$

Let $\sigma_N^2 = \sum_{k=1}^{K} \sigma_{Nk}^2$ and $\gamma_N = \sum_{k=1}^{K} \gamma_{Nk}$. The Lyapunov CLT says as long as

$$\gamma_N / \sigma_N^4 \to 0,$$

we have $KS_{NK}/\sigma_N \to_d N(0,1)$. Due to what we have shown, $\sigma_N^2 = \sum_{k=1}^{K} \sigma_{Nk}^2 \geq e_{thres} \dfrac{K}{n} + O(K/n^2)$, and $\gamma_N = O(K/n)$. Hence the Lyapunov condition holds if $(K/n)/(K^2/n^2) \to 0$, or $K/\sqrt{N} \to \infty$.

In this case, $K/\sigma_N = O(\sqrt{N})$, and hence by Lyapunov central limit theorem, $\sqrt{N} S_{NK} = O_p(1)$.

Proof of Lemma 2.6

For simplicity we only prove claim (ii). Proof of claim (i) can be obtained following similar arguments. Due to the asymptotic normality of Δ_{HT}, it suffices to show that

$$\sqrt{N}(\hat{\Delta}_{S-HT} - \hat{\Delta}_{HT}) = O_p(1).$$

Let $e_1, \ldots, e_n$ be independent samples of $e(\boldsymbol{X})$, and $F^{-1}(\cdot)$ be the quantile distribution of $e(\boldsymbol{X})$. For $t \in (0,1)$, the empirical quantile distribution is defined as

$$\mathbb{F}_N^{-1}(t) = \inf\{x : \mathbb{F}_N(x) \geq t\},$$

where $\mathbb{F}_N(x) = \sum_{i=1}^{N} 1_{(-\infty,x]}(e_i)/N$ is the empirical distribution function.

Using standard empirical process theory (Shorack and Wellner, 1986), we can show that

$$\sqrt{N}\|\mathbb{F}_N^{-1} - F^{-1}\|_0^1 = \sqrt{N} \sup_{0\leq t\leq 1} |\mathbb{F}_N^{-1}(t) - F^{-1}(t)| = O_p(1). \tag{2.10}$$

As $F^{-1}(t)$ is Lipschitz continuous,

$$K \max_{1\leq k\leq K}\{q_k - q_{k-1}\} = O(1), \tag{2.11}$$

where $q_k = F^{-1}(k/K)$. Assumption (2.5), results (2.10) and (2.11) together imply

$$\sqrt{N} \max_{1\leq k\leq K} |\tilde{q}_k - q_k| = O_p(1) \quad \text{and} \quad \sqrt{N} \max_{1\leq k\leq K}\{\tilde{q}_k - \tilde{q}_{k-1}\} = O_p(1), \tag{2.12}$$

where $\tilde{q}_k = \mathbb{F}_n^{-1}(k/K)$ $(k = 1, \ldots, K)$, the sample quantiles of the (true) propensity scores. Now let $\hat{q}_k$ be the sample quantiles of the estimated propensity scores, Assumption 3 and result (2.12) imply

$$\sqrt{N} \max_{1\leq k\leq K} |\hat{q}_k - q_k| = O_p(1) \quad \text{and} \quad \sqrt{N} \max_{1\leq k\leq K}\{\hat{q}_k - \hat{q}_{k-1}\} = O_p(1). \tag{2.13}$$

Denote $\hat{e}_{min} = \min\{\hat{e}_i, i = 1, \ldots, N\}$, $\hat{e}_{max} = \max\{\hat{e}_i, i = 1, \ldots, N\}$ and $e_{thres} = \min\{e_{min}/2, (1 - e_{max})/2\}$, Assumption 2.2 and 2.3 imply that for large enough N,

$$e_{thres} < min(\hat{e}_{min}, 1 - \hat{e}_{max}), \tag{2.14}$$

Moreover, if we let $\delta = \max_{1\le k\le K}\{\hat{q}_k - \hat{q}_{k-1}\}$, then

$$\begin{aligned} p_{\hat{k}_i} &= P(Z=1|\hat{e}_i \in [\hat{q}_{\hat{k}_i-1}, \hat{q}_{\hat{k}_i}]) \\ &= E_{e_i|\hat{e}_i\in[\hat{q}_{\hat{k}_i-1},\hat{q}_{\hat{k}_i}]} E\left[P(Z=1|e_i,\hat{e}_i\in[\hat{q}_{\hat{k}_i-1},\hat{q}_{\hat{k}_i}])|\hat{e}_i\in[\hat{q}_{\hat{k}_i-1},\hat{q}_{\hat{k}_i}]\right] \\ &= E_{e_i|\hat{e}_i\in[\hat{q}_{\hat{k}_i-1},\hat{q}_{\hat{k}_i}]} E\left[e_i|\hat{e}_i\in[\hat{q}_{\hat{k}_i-1},\hat{q}_{\hat{k}_i}]\right] \\ &\in [\hat{q}_{\hat{k}_i-1}-\delta, \hat{q}_{\hat{k}_i}+\delta]. \end{aligned}$$

On the other hand, $\hat{e_i} \in [\hat{q}_{\hat{k}_i-1}, \hat{q}_{\hat{k}_i}] \subset [\hat{q}_{\hat{k}_i-1}-\delta, \hat{q}_{\hat{k}_i}+\delta]$, hence

$$|p_{\hat{k}_i} - \hat{e}_i| \le \max_{1\le k\le K}\{\hat{q}_k - \hat{q}_{k-1}\} + 2\max_{1\le i\le N}|\hat{e}_i - e_i|.$$

Combining (2.13), Assumption 2.3 and (2.14), we have for large enough N,

$$e_{thres} \le p_k \le 1 - e_{thres}. \tag{2.15}$$

On the other hand, For large enough N, we have

$$\begin{aligned} \sqrt{N}|\hat{\Delta}_{S-HT} - \hat{\Delta}_{HT}| &\le \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}|Y_i|I(Z_i=1)\left|\frac{1}{p_{\hat{k}_i}} - \frac{1}{\hat{e}_i}\right| + \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}|Y_i|I(Z_i=0)\left|\frac{1}{1-p_{\hat{k}_i}} - \frac{1}{1-\hat{e}_i}\right| \\ &\le \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}|Y_i|\frac{|p_{\hat{k}_i} - \hat{e}_i|}{e_{thres}^2} \\ &\le \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}|Y_i|\frac{\max_{1\le k\le K}\{\hat{q}_k - \hat{q}_{k-1}\} + 2\max_{1\le i\le N}|\hat{e}_i - e_i|}{e_{thres}^2} \\ &= O_p(1). \end{aligned}$$

Hence we complete the proof of Lemma 2.6.

C Proof of Theorem 2.4

When N is large enough, by uniform convergence of $\hat{e}_i$ $(i = 1, \ldots, N)$ and uniform convergence of sample quantitles $\hat{q}_k$ $(k = 1, \ldots, K)$ (see Section B for detailed proof), we have for large enough N,

$$e_{thres} \leq pr(Z = 1 \mid \hat{e}(\boldsymbol{X}) \in \hat{C}_k) \leq 1 - e_{thres}.$$

Then

$$\begin{aligned}
& pr(\text{exists } z, k, \text{ such that } n_{zk} = 0) \\
\leq & \sum_{z=0}^{1} \sum_{k=1}^{K} pr(n_{zk} = 0) \\
\leq & \sum_{k=1}^{K} \left(pr(Z = 1 \mid \hat{e}_i \in \hat{C}_k)^{n_k} + pr(Z = 0 \mid \hat{e}_i \in \hat{C}_k)^{n_k} \right) \\
\leq & \sum_{k=1}^{K} 2(1 - e_{thres})^{N/K - 1} \\
= & \, exp\left\{ log(2K) + \left(\frac{N}{K} + 1 \right) log(1 - e_{thres}) \right\} \\
\to & \, 0.
\end{aligned}$$

This completes the proof of Theorem 2.4.

D Proof of Proposition 2.5

The proof is straightforward by noting that

$$\sum_{i=1}^{N} Z_i / \hat{p}_{\hat{k}_i} = \sum_{i=1}^{N} n_{\hat{k}_i} Z_i / n_{1\hat{k}_i} = \sum_{k=1}^{K} \left((n_k / n_{1k}) \sum_{i:\hat{k}_i = k} Z_i \right) = \sum_{k=1}^{K} ((n_k / n_{1k}) n_{1k}) = N$$

and similarly

$$\sum_{i=1}^{N}(1-Z_i)/(1-\hat{p}_{\hat{k}_i}) = N.$$

E Descriptive statistics for the NHANES data

Please see Table 2.4.

Table 2.4: Baseline characteristics and outcome measure by participation status in the school meal programs

	Participated (N=1284)	Not participated (N=1046)
Child Age, mean (SD)	10.1 (3.5)	9.9 (4.4)
Child Male, N (%)	657 (51.2%)	549 (52.5%)
Black, N (%)	396 (30.8%)	208 (19.9%)
Hispanic, N (%)	421 (32.8%)	186 (17.8%)
Above 200% of poverty level, N (%)	317 (24.7%)	692 (66.2%)
Participation in SSN program, N (%)	328 (25.5%)	115 (11.0%)
Participation in food stamp program, N (%)	566 (44.1%)	122 (11.7%)
Childhood food security, N (%)	418 (32.6%)	155 (14.8%)
Insurance Coverage, N (%)	1076 (83.8%)	927 (88.6%)
Respondent Age, mean (SD)	38.6 (10.4)	40.3 (9.7)
Respondent Male, N (%)	506 (39.4%)	526 (50.3%)
BMI, mean (SD)	20.4 (5.5)	19.8 (5.4)

Chapter 3

IDENTIFICATION AND ESTIMATION OF CAUSAL EFFECTS WITH OUTCOMES TRUNCATED BY DEATH

3.1 Introduction

In medical studies, researchers are often interested in evaluating risk factors for a non-mortality outcome such as memory decline. However, the non-mortality outcome may be truncated by death if some subjects die before the follow-up assessment, leaving their non-mortality outcomes to be undefined. For example, suppose we are interested in estimating the effect of smoking on memory decline in an aged population. If a subject dies before the follow-up memory test is administered, then his/her memory score at the follow-up visit is undefined. Direct comparisons between smokers and non-smokers among observed survivors are subject to selection bias as nonsmokers are more likely to survive to the follow-up assessment (Rosenbaum, 1984; Robins and Greenland, 1992). More fundamentally, direct comparisons among observed survivors are not causally interpretable as they compare outcomes from different subpopulations at baseline (Rubin, 2006). Alternatively, Rubin (2000) proposed to estimate the average causal effect in the always-survivor group, the group of subjects who would survive if they choose to receive either exposure at baseline. The resulting estimand is termed the survivor average causal effect (SACE).

The SACE is only partially identifiable without further assumptions (Zhang and Rubin, 2003). Large sample bounds for the SACE have been derived under minimal assumptions (Zhang and Rubin, 2003; Imai, 2008). In order to identify the SACE, a common strategy is to perform a sensitivity analysis by assuming a class of identifiability conditions indexed by an interpretable sensitivity parameter (Gilbert et al., 2003; Shepherd et al., 2006; Hayden et al., 2005; Egleston et al., 2007; Chiba and VanderWeele, 2011). Alternatively, identification of

the SACE can be based on covariate information. For example, Tchetgen Tchetgen (2014) identified a *variant* of SACE to be applied when risk factors of survival are available in post-exposure follow-up. The resulting causal contrast is relative to the given post-exposure risk factors. Alternatively, to identify the SACE in a randomized study setting, Ding et al. (2011) proposed a semiparametric identification approach based on a baseline variable whose distribution is informative of the membership of the always-survivor group. With this baseline variable, they showed that the SACE was identifiable under their assumptions. However, as pointed out by Tchetgen Tchetgen (2014), their assumption essentially requires that there are no common causes of the potential survivals and potential outcomes, which is very unlikely even in randomized studies.

In this chapter, we relax Ding et al. (2011)'s identification assumptions by employing more detailed covariate information. In comparison to previous works, our causal estimand is defined independently of the incorporated covariates. Furthermore, the proposed approach is applicable to both randomized trials and observational studies, *and* allows for measured common causes of potential survivals and the potential outcomes. In case there is unmeasured dependence between the survival and outcome processes, we propose an alternative population level no-interaction assumption. This assumption is different from the individual level no-interaction assumption discussed in Nolen and Hudgens (2011, §7), which assumes that the causal effect is constant in the always-survivor stratum. We also develop novel parameterizations for our distributional assumptions. This is challenging because unlike standard observational study setting, in our identification framework the baseline covariates play the dual role of potential confounders and common causes of $A, (S(1), S(0))$ and $(Y(1), Y(0))$. As we explain later in Section 3.4, this dual role of covariates make the standard propensity score methods inappropriate in our setting. On the other hand, our distributional assumptions need to comply with constraints on the observed data law due to our identifiability assumptions.

The rest of the chapter is organized as follows. In Section 3.2, we define our parameter of interest, namely the SACE. We first introduce our identification assumptions in Section

3.3, and then propose model parameterizations for estimation of the SACE in Section 3.4. Finite sample performance of the proposed estimators is evaluated by simulation studies described in Section 3.5. In Section 3.6, we apply our method to compare two prostate cancer treatments with data from a Southwest Oncology Group (SWOG) Trial and estimate the effect of smoking on memory decline with data from the Health and Lifestyle Study (HALS). We end with a discussion in Section 3.7.

3.2 Data structure, notation and causal estimand

Consider a medical study with a single follow-up visit. Let Z be the exposure indicator and W denote observed covariates at baseline. We assume that each subject has two potential survivals $S(1)$ and $S(0)$, defined as the survival status at the follow-up visit that would have been observed if the subject would have been exposed and unexposed, respectively. Similarly we let $Y(1)$ and $Y(0)$ denote the potential outcomes under exposure and non-exposure, respectively. We assume that for $z = 0, 1$, $Y(z)$ takes real values only if $S(z) = 1$. We extend the definition of $Y(z)$ so that it takes a constant value $*$ if $S(z) = 0$. With slight abuse of terminology, we still say $Y(z)$ is well-defined only if $S(z) = 1$.

We use G to denote the survival type as defined in Table 3.1. One can see from Table 3.1 that there exists a one-to-one mapping between the survival type and the bivariate potential survivals $(S(1), S(0))$. As a result, G can be considered as an abbreviation for $(S(1), S(0))$.

We adopt the axiom of consistency such that the observed outcome Y satisfies $Y = ZY(1) + (1 - Z)Y(0)$ and the observed survival S satisfies $S = ZS(1) + (1 - Z)S(0)$. The observed samples $O_i = (Z_i, W_i, S_i, Y_i), i = 1, \ldots, N$ are independently drawn from an infinite super-population.

Throughout this chapter, we assume that there is no interference between study subjects regarding both the survival status S and the response Y, and there is only one version of exposure (SUTVA, Rubin (1980)).

Rubin (2000) noted that the observed survivors in the exposed group are from a mixture of always-survivor and protected strata, while the observed survivors in the non-exposed group

Table 3.1: Patient survival types

$S(1)$	$S(0)$	Survival type	G	Description
1	1	always-survivor	LL	The subject always survives, regardless of exposure status.
1	0	protected	LD	The subject survives if exposed, but dies if not exposed.
0	1	harmed	DL	The subject dies if exposed, but survives if not exposed.
0	0	doomed	DD	The subject always dies, regardless of exposure status.

are from a mixture of always-survivor and harmed strata. As a result, direct comparisons between different exposure groups among observed survivors are not causally meaningful as these people are from different subpopulations at baseline. Instead, as always-survivor is the only stratum such that both $Y(1)$ and $Y(0)$ are well-defined, we define our causal estimand to be the average causal effect in the always-survivor stratum:

$$SACE = E[Y(1) - Y(0)|G = LL].$$

3.3 Identification of the SACE

In general, the SACE is not identifiable without further untestable assumptions as it depends on the potential outcomes $Y(1), Y(0), S(1)$ and $S(0)$. In this section, we propose conditions to identify the SACE. We first introduce several standard assumptions in the causal inference literature and show that the SACE is not identifiable under these assumptions. We then propose additional identification assumptions on the structure of baseline variables and show that the SACE is identifiable with these additional assumptions.

A Non-identifiability of the SACE

We first introduce some commonly made assumptions in the causal inference literature.

A.1 (Monotonicity) $S_i(1) \geq S_i(0)$ for all $i = 1, \ldots, N$.

The monotonicity assumption may be plausible in some observational studies. For example, in studies evaluating the effect of smoking on memory decline, it is widely believed

that smoking is bad for the overall health, and hence the overall survival. This assumption tends to be shaky in randomized clinical trials with acute diseases because (typically) a clinical trial would be unethical if the researchers believe that one treatment benefits survival *a priori*. To address this issue, we relax this assumption later in Section D.

A.2 (S-Ignorability) The treatment assignment is independent of the potential survivals given the observed covariates W, denoted as $Z \perp (S(1), S(0))|W$.

In fact, in all the following derivations, A.2 can be relaxed to the weaker assumption that $Z \perp S(z)|W; z = 0, 1$. We keep A.2 nevertheless to facilitate our illustration in Section C.

A.3 (Y-Ignorability) The treatment assignment in the always-survivor stratum is marginally independent of the potential outcomes given observed covariates W, denoted as $Z \perp Y(z)|W, G = LL; z = 0, 1$.

A.2 and A.3 are similar to the (weakly) ignorable treatment assignment assumption. Under A.3, we have

$$E[Y(z)|G = LL] = \frac{E_W\left\{\mu^z_{LL,W}\pi_{LL|W}\right\}}{E_W\{\pi_{LL|W}\}}, \tag{3.1}$$

where $\mu^z_{g,W} = E[Y|Z = z, G = g, W], \pi_{g|W} = P(G = g|W)$. To see this, note

$$\begin{aligned} E[Y(z)|G = LL] &= \frac{E[Y(z)I(G = LL)]}{P(G = LL)} \\ &= \frac{E_W\left\{E[Y(z)|W, G = LL]P(G = LL|W)\right\}}{E_W\left\{P(G = LL|W)\right\}} \\ &= \frac{E_W\left\{E[Y|Z = z, W, G = LL]P(G = LL|W)\right\}}{E_W\left\{P(G = LL|W)\right\}}. \end{aligned}$$

Under A.1 and A.2, $\pi_{LL|W}$ and $\pi_{LD|W}$ can be identified by

$$\begin{aligned} \pi_{LL|W} &= P(S = 1|Z = 0, W), \\ \pi_{LD|W} &= P(S = 1|Z = 1, W) - P(S = 1|Z = 0, W); \end{aligned} \tag{3.2}$$

$\mu^0_{LL,W}$ can be identified by

$$\mu^0_{LL,W} = E[Y|Z=0, S=1, W]. \tag{3.3}$$

However, $\mu^1_{LL,W}$ is not identifiable from the observed data. In fact, even if the outcome Y has bounded support, one can only partially identify $\mu^1_{LL,W}$ from the equation

$$E[Y|Z=1, S=1, W] = \frac{\pi_{LL|W}}{P(S=1|Z=1,W)}\mu^1_{LL,W} + \frac{\pi_{LD|W}}{P(S=1|Z=1,W)}\mu^1_{LD,W}.$$

Therefore, A.1 - A.3 alone are not sufficient for identifying the SACE.

B Identifying the SACE using a substitution variable

To identify the SACE, without loss of generality, we assume that the baseline covariates W can be written as (X, A). In our identification framework, the role of X is similar to a "confounder," whereas the role of A is similar to an "instrument." Specifically, we make the following assumptions on A and X:

A.4 (Exclusion restriction) $A \perp Y(1)|Z=1, X, G$.

A.5 (Substitution relevance) $A \not\perp G|X, Z=1, S=1$.

We note that the survival type G is a baseline variable that satisfies A.4 and A.5. For this reason, any variable satisfying A.4 and A.5 is called a *substitution variable* for G. The conditions for a substitution variable are similar to those for an instrumental variable (Angrist et al., 1996). For example, A.4 is similar to the exclusion restriction assumption in an instrumental variable analysis in that they both capture the notion that A has no "direct effect" on Y, and A.5 is similar to the instrumental relevance assumption. Assumption A.5 also prevents A to be an irrelevant variable. As we illustrate later in Section C, even in a randomized study setting, the inclusion of covariate information X makes A.4 and A.5 more plausible.

The following Theorem 3.1 states that the SACE is identifiable with a substitution vari-

able for the survival type. The proof is left to Appendix A.

Theorem 3.1. *Under A.1 - A.5, the SACE is identifiable and is given by (3.1), where* $\pi_{LL|W}$ *and* $\mu^0_{LL,W}$ *can be identified from* (3.2) *and* (3.3) *respectively, and* $\mu^1_{LL,W}$ *can be identified from the following equations:*

$$\begin{aligned} E[Y|Z=1,S=1,X,A=a_1] &= p^1_{LL|X,a_1}\mu^1_{LL,X} + p^1_{LD|X,a_1}\mu^1_{LD,X}, \\ E[Y|Z=1,S=1,X,A=a_0] &= p^1_{LL|X,a_0}\mu^1_{LL,X} + p^1_{LD|X,a_0}\mu^1_{LD,X}, \end{aligned} \tag{3.4}$$

where a_1 *and* a_0 *are two distinct values in the range of* A, $p^z_{g|X,a} \equiv P(G=g|X,A=a)/P(S=1|Z=z,X,A=a)$ *is identifiable from the observed data, and* $\mu^z_{g,X} \equiv E[Y|Z=z,G=g,X]$.

C A graphical illustration

The key assumptions in our identification model are the conditional independence assumptions A.2 - A.4. In this part, we illustrate these assumptions with a directed acyclic graph (DAG) submodel (Lauritzen, 1996; Pearl, 2014). This submodel imposes more conditional independence assumptions than A.2 - A.4, but it is useful for understanding the causal structure underlying our identification assumptions, and facilitates comparison with previously proposed methods.

In the literature, the DAG is often used to describe a set of conditional independence relationships following the rule of d-separation (Pearl, 2014) (a.k.a. global Markov property). In particular, each missing edge in a DAG represents a conditional independence relationship. DAGs that describe exactly the same conditional independence assumptions form a Markov equivalence class, which can be uniquely represented by a completed partially directed acyclic graph (CPDAG) (Andersson et al., 1997). Figure 3.1(a) gives the CPDAG representation of A.2 - A.4. The CPDAG therein is denoted as $\mathcal{G}_1$. The missing edges between Z and $(S(1),S(0))$, Z and $Y(1)$,$Y(0)$ and A and $Y(1)$ correspond to the conditional independence assumptions A.2, A.3 and A.4, respectively. Figure 3.1(b) describes one possible DAG for which A and X are pre-exposure covariates.

In comparison to the identification assumptions in Ding et al. (2011), we have incorporated baseline covariates X into our identification framework. Hence our approach is applicable to observational studies due to the presence of edges $X \to Z$, $X \to Y(0)$ and $X \to Y(1)$. Moreover, we allow for edges $X \to (S(1), S(0))$, $X \to Y(1)$ and $X \to Y(0)$, hereby avoiding the assumption that there are no common causes of the potential survivals and potential outcomes, which is very restrictive even for randomized studies. Lastly, we allow for the edge $X \to A$, hereby avoiding the assumption that there is no common cause of the substitution variable and potential outcomes.

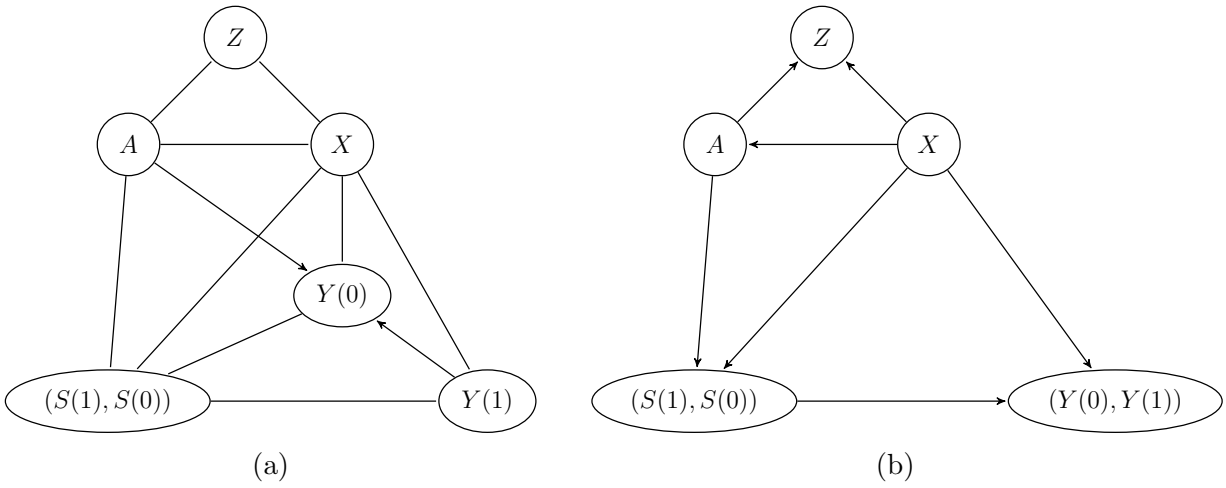

Figure 3.1: (a): the CPDAG representing the factorization of the joint distribution of nodes. (b): one possible DAG under (a) if A and X are defined pre-exposure.

D Relaxing the monotonicity assumption

In this part we consider an alternative stochastic monotonicity assumption to relax the monotonicity assumption. This assumption has been used previously by Roy et al. (2008) in the context of non-compliance and Lee et al. (2010) in the context of truncation by death.

A.6 (Stochastic monotonicity)

$$P(S(1)=1|S(0)=1,W)=P(S(1)=1|W)+\rho(U(W)-P(S(1)=1|W)), \tag{3.5}$$

where $U(W)=\min\left\{1,\dfrac{P(S(1)=1|W)}{P(S(0)=1|W)}\right\}$ is the upper bound for the left side of (3.5), and ρ is a sensitivity parameter ranging from 0 to 1.

A.6 implies the following restriction on potential survivals:

$$P(S(1)=1|S(0)=1,W)\geq P(S(1)=1|W).$$

In other words, the stochastic monotonicity requires that the potential survival under active treatment is not negatively correlated with the potential survival under control.

Also note that (3.5) is equivalent to the following equation:

$$\rho=\frac{P(G=LL|W)-P(S(1)=1|W)P(S(0)=1|W)}{\min\{P(S(1)=1|W),P(S(0)=1|W)\}-P(S(1)=1|W)P(S(0)=1|W)}.$$

Hence $\rho=0$ if and only if conditional on observed covariates, $S(1)$ is not correlated with $S(0)$. On the other hand, $\rho=1$ if and only if $P(G=LL|W)=\min\{P(S(1)=1|W),P(S(0)=1|W)\}$ Note A.6 does not reduce to A.1 when $\rho=1$; in particular, it does not specify which treatment is more beneficial for survival.

Theorem 3.2 states that under A.6, if we assume that A.4 and A.5 also hold in the control arm ($Z=0$), then the SACE is identifiable from the observed data.

Theorem 3.2. *If we assume A.2 - A.6, and the additional conditions that*

(i) $A\perp Y(0)|Z=0,X,G=LL$,

(ii) $A\not\perp G|X,Z=0,S=1$.

then the SACE is identifiable and is given by (3.1), where $\pi_{g|W}$ *can be identified from* (3.5)

and the following equations:

$$P(S(1)=1|W) = P(S=1|Z=1,W) = \pi_{LL|W} + \pi_{LD|W};$$
$$P(S(0)=1|W) = P(S=1|Z=0,W) = \pi_{LL|W} + \pi_{DL|W};$$
$$\pi_{LL|W} = P(S(1)=1|W)P(S(1)=1|S(0)=1,W),$$

and $\mu^z_{LL,X}$ *can be identified from the following equations:*

$$E[Y|Z=z,S=1,X,A=a_1] = p^z_{LL|X,a_1}\mu^z_{LL,X} + p^z_{LD|X,a_1}\mu^z_{LD,X};$$
$$E[Y|Z=z,S=1,X,A=a_0] = p^z_{LL|X,a_0}\mu^z_{LL,X} + p^z_{LD|X,a_0}\mu^z_{LD,X}.$$

E Relaxing the exclusion restriction assumption

In practice, the ignorability assumptions A.2 and A.3 may be questionable if there are unmeasured common causes of Z and the potential survivals or potential outcomes. Similarly, the exclusion restriction assumption A.4 may be questionable if the observed covariates X is not rich enough to contain all common causes of A, $(S(1),S(0))$ and $(Y(1),Y(0))$. There has been extensive literature on relaxing the ignorability assumptions for causal inference in observational studies. Hence in this chapter, we focus on alternative assumptions to relax the exclusion restriction assumption.

Specifically, we propose an alternative no-interaction assumption A.7, which assumes that conditioning on the baseline variables X, neither Z nor G modifies the effect of A on Y.

A.7 (no interaction) If $a_0 \neq a_1$, then

$$\begin{aligned}
&E[Y|Z=1,G=LL,X,A=a_1] - E[Y|Z=1,G=LL,X,A=a_0] \\
&= E[Y|Z=1,G=LD,X,A=a_1] - E[Y|Z=1,G=LD,X,A=a_0] \\
&= E[Y|Z=0,G=LL,X,A=a_1] - E[Y|Z=0,G=LL,X,A=a_0].
\end{aligned}$$

A.7 is similar in spirit to the no-interaction assumption in causal mediation analysis (Robins and Greenland, 1992). A.7 holds under a wide range of semiparametric models. For example, a sufficient condition for A.7 is the following additive model:

$$E[Y|Z, G, X, A] = \eta_1(Z, G, X) + \eta_2(A, X), \tag{3.6}$$

where $G = 1$ for the LL group and $G = 0$ for the LD group; η_1 and η_2 are two functions that are completely unspecified. However, A.7 may still hold even if (3.6) were not true. Another feature of A.7 is that it can be used to identify the SACE with binary substitution variables. In comparison, to identify the SACE in a randomized study under exclusion restrcition violation, Ding et al. (2011) assumed a parametric model on $E[Y|Z = 1, G, A]$ and their approach requires that the substitution variable A has at least three categories or is continuous.

Theorem 3.3 states that if we replace the exclusion restriction assumption A.4 with A.7, then the SACE is still identifiable. The proof is very similar to the proof for Theorem 3.1 and is hence omitted.

Theorem 3.3. *Under assumptions A.1 - A.3, A.5 and A.7, the SACE is identifiable and is given by (3.1), where*

$$\mu^0_{LL,W} = E[Y|Z = 0, S = 1, W]$$

and $\mu^1_{LL,X}$ *can be identified from the following equations:*

$$\begin{aligned}
E[Y|Z=1,S=1,X,A=a_1] &= p^1_{LL|X,a_1}\mu^1_{LL,X,a_1} + p^1_{LD|X,a_1}\mu^1_{LD,X,a_1};\\
E[Y|Z=1,S=1,X,A=a_0] &= p^1_{LL|X,a_0}\mu^1_{LL,X,a_0} + p^1_{LD|X,a_0}\mu^1_{LD,X,a_0};\\
\mu^1_{LL,X,a_1} - \mu^1_{LL,X,a_0} &= \mu^1_{LD,X,a_1} - \mu^1_{LD,X,a_0};\\
\mu^1_{LL,X,a_1} - \mu^1_{LL,X,a_0} &= \mu^0_{LL,X,a_1} - \mu^0_{LL,X,a_0};\\
E[Y|Z=0,S=1,X,A=a_1] &= \mu^0_{LL,X,a_1};\\
E[Y|Z=0,S=1,X,A=a_0] &= \mu^0_{LL,X,a_0};\\
\mu^1_{LL,X} &= p^1_{a_1|LL,X}\mu^1_{LL,X,a_1} + p^1_{a_0|LL,X}\mu^1_{LL,X,a_0},
\end{aligned}$$

where $\mu^z_{g,X,a} = E[Y|Z=z,G=g,X,A=a]$ *and* $p^z_{a|g,X} = P(A=a|Z=z,G=g,X)$ *is identified as in Theorem 3.1.*

3.4 Model parameterization

In the previous section we have shown that the SACE is identifiable under various identification assumptions. To estimate the SACE in practice, unless the covariates W only take a few discrete values, we need to impose additional distributional assumptions. We first note that the identification assumptions imply certain constraints on the observed data law. Proposition 3.4 summarizes these constraints, with the proof in Appendix C.

Proposition 3.4.

(I) The assumptions of Theorem 3.1 imply the following constraints on the law of (Z,W,S,Y)*:*

$$P(S=0|Z=1,X,A) \le P(S=1|Z=1,X,A); \tag{3.7}$$

$$\text{for all } x, E[Y|Z=1,S=1,A,X=x] \text{ is bounded}; \tag{3.8}$$

$$P(S=1|Z=0,X,a)/P(S=1|Z=1,X,a) \text{ is not a constant of } a. \tag{3.9}$$

(II) The assumptions of Theorem 3.2 imply the following constraints on the law of (Z, W, S, Y)*: either* (3.9) *holds, or* $\forall z, x, P(S = 1|Z = z, X = x, A = a)$ *is not a constant of* a*.*

(III) The assumptions of Theorem 3.3 imply (3.7) *and that*

$$\textit{for all } x, E[Y|Z = 1, S = 1, A, X = x] - E[Y|Z = 0, S = 1, A, X = x] \textit{ is bounded.} \tag{3.10}$$

To ensure that the modeling assumptions are compatible with the constraints in Proposition 3.4, we avoid imposing distributional assumptions on the observed data directly. Instead, we make the following distributional assumptions on the law of $(Z, W, S(1), S(0), Y(1), Y(0))$. For example, to estimate the SACE under the assumptions of Theorem 3.1, we need the following distributional assumptions:

M.1 $E[Y(0)|Z = 0, G = LL, X, A]$ is known up to a finite-dimensional parameter α_1; that is, $E[Y(0)|G = LL] = m_1(X, A; \alpha_1)$, where $m_1(\cdot, \cdot; \alpha_1)$ is a known function and α_1 is an unknown parameter. Specifically, for our simulations and data examples we consider

$$m_1(X, A; \alpha_1) = \alpha_{10} + X^T\alpha_{11} + A\alpha_{12}. \tag{3.11}$$

M.2 $E[Y(1)|Z = 1, G, X, A]$ is known up to a finite-dimensional parameter α_2; that is, $E[Y(1)|Z = 1, G = g, X, A] = m_2(X, G; \alpha_2)$, where G takes value in $\{LL, LD\}$, $m_2(\cdot, \cdot; \alpha_1)$ is a known function and α_2 is an unknown parameter. Note that due to A.4, $E[Y(1)|Z = 1, G, X, A]$ does not depend on the value of A. Specifically, for our simulations and data examples we code LL to be 1, LD to be 0, and consider

$$m_2(X, G; \alpha_2) = \alpha_{20} + X^T\alpha_{21} + G\alpha_{23}. \tag{3.12}$$

M.3 $P(S(1) = 1|X, A)$ is known up to a finite-dimensional parameter β_1; that is $P(S(1) = 1|X, A) = \theta_1(X, A; \beta_1)$, where $\theta_1(\cdot, \cdot; \beta_1)$ is a known function and β_1 is an unknown

parameter. Specifically, for our simulations and data examples we consider

$$\theta_1(X, A; \beta) = \text{expit}(\beta_{10} + X^T\beta_{11} + A\beta_{12}). \tag{3.13}$$

M.4 $P(S(0) = 1|X, A)/P(S(1) = 1|X, A)$ is known up to a finite-dimensional parameter γ; that is $P(S(0) = 1|X, A)/P(S(1) = 1|X, A) = \theta_{0|1}(X, A; \gamma)$, where $\theta_{0|1}(\cdot, \cdot; \gamma)$ is a known function and γ is an unknown parameter. Specifically, for our simulations and data examples we consider

$$\theta_{0|1}(X, A; \gamma) = \text{expit}(\gamma_0 + X^T\gamma_1 + A\gamma_2). \tag{3.14}$$

Rather than M.3 and M.4, some previous researchers impose distributional assumptions on $P(S(z)|X, A)$, $z = 0, 1$ directly (e.g. Lee et al., 2010). However, due to the monotonicity assumption, $P(S(0) = 1|X, A)$ resides in the range $[0, P(S(1) = 1|X, A)]$. Hence the model parameters for $P(S(1) = 1|X, A)$ live in a constrained space, making estimation and asymptotic analysis difficult. To avoid such constraints, we re-parameterize our parameters as in M.3 and M.4.

To derive the maximum likelihood estimator, we note that M.1 - M.4 correspond to the following modeling constraints on the observed data distribution:

$$\begin{aligned}
&P(S = 1|Z = 1, X, A) = \theta_1(X, A; \beta_1),\\
&P(S = 1|Z = 0, X, A) = \theta_1(X, A; \beta_1)\theta_{0|1}(X, A; \gamma),\\
&E[Y|Z = 1, S = 1, X, A] = \theta_{0|1}(X, A; \gamma)m_2(X, 1; \alpha_2) + (1 - \theta_{0|1}(X, A; \gamma))m_2(X, 0; \alpha_2),\\
&E[Y|Z = 0, S = 1, X, A] = m_1(X, A; \alpha_1).
\end{aligned}$$

It is easy to see that these models are compatible with the testable implications of A.1 - A.5 described in Proposition 3.4. Ordinary least squares and maximum likelihood estimation may be used for the parameter estimation. The SACE can then be estimated using (3.1),

in which $\mu^z_{LL,W}$ can be estimated using (3.11) and (3.12), $\pi_{LL|W}$ can be estimated using the product of (3.13) and (3.14), and the integration is taken with respect to the empirical distribution of W. Using standard M-estimation theory, one can show that the resulting estimate of SACE is consistent and asymptotically normally distributed.

Remark 3.5. *Our estimation method is different from the method described in Theorem 3.1. This is particularly convenient when A is continuous, in which case as detailed in Ding et al. (2011), A needs to be discretized for the estimation method in Theorem 3.1 to be applicable.*

Remark 3.6. *An alternative popular approach to estimate causal effects is based on the propensity score $e(X)$, defined as the probability of assignment to exposure conditioning on baseline covariates. However, although the propensity score is sufficient for summarizing the effect of X on Z in the sense that $Z \perp X \mid e(X)$, it is not sufficient for summarizing the effect of X on A. Thus unlike a standard causal effect estimation setting, the propensity score methods are not applicable for estimating the SACE under our identification assumptions. See Appendix B for an explanation with the DAG model in Figure 3.1(b).*

To estimate the SACE under the assumptions of Theorem 3.2, we assume M.2 and the following model:

M.5 $E[Y(0)|Z = 0, G, X, A]$, is known up to a finite-dimensional parameter α_3; that is, $E[Y(0)|Z = 0, G, X, A] = m_3(X, G; \alpha_3)$, where G takes value in $\{LL, DL\}$, $m_3(\cdot, \cdot; \alpha_3)$ is a known function and α_3 is an unknown parameter. Note that due to condition (i) of Theorem 3.2, $m_3(\cdot, \cdot; \alpha_3)$ does not depend on the value of A. Specifically, for our simulations and data examples we code LL to be 1, DL to be 0, and consider

$$m_3(X, G; \alpha_3) = \alpha_{30} + X^T \alpha_{31} + G\alpha_{33}.$$

M.6 $P(S(0) = 1|X, A)$ is known up to a finite-dimensional parameter β_0; that is $P(S(0) = 1|X, A) = \theta_0(X, A; \beta_0)$, where $\theta_0(\cdot, \cdot; \beta_0)$ is a known function and β_0 is an unknown

parameter. Specifically, for our simulations and data examples we consider

$$\theta_0(X, A; \beta_0) = \text{expit}(\beta_{00} + X^T\beta_{01} + A\beta_{02}).$$

Similarly, to estimate the SACE under the assumptions of Theorem 3.3, we assume M.3, M.4 and the following model:

M.7 $E[Y|Z, G, X, A], g = LL, LD$ is known up to a finite-dimensional parameter α_4; that is, $E[Y|Z, G, X, A] = m_4(X, A, G, Z; \alpha_4)$, where G takes value in $\{LL, LD\}$, $m_4(\cdot, \cdot, \cdot, \cdot; \alpha_4)$ is a known function and α_4 is an unknown parameter. Note that due to A.7, $m_4(X, A, G, Z; \alpha_4)$ should not contain interaction terms within the pairs (A, Z) or (A, G). Specifically, for our simulations and data examples we code LL to be 1, DL to be 0 and consider

$$m_4(X, A, G, Z; \alpha_4) = \alpha_{40} + X^T\alpha_{41} + A\alpha_{42} + G\alpha_{43} + Z\alpha_{44}.$$

Finally, if one wishes to relax both the monotonicity assumption A.1 and the exclusion restriction A.4, one may assume M.3, M.5, and the following model:

M.8 $E[Y|Z, G, X, A]$ is known up to a finite-dimensional parameter α_5; that is, $E[Y|Z, G, X, A] = m_5(X, A, G, Z; \alpha_5)$, where G takes value in $\{LL, LD, DL\}$, $m_5(\cdot, \cdot, \cdot, \cdot; \alpha_5)$ is a known function and α_5 is an unknown parameter. Note that due to A.7, $m_4(X, A, G, Z; \alpha_4)$ should not contain interaction terms within the pairs (A, Z) or (A, G). Specifically, for our simulations and data examples we code LL to be 1, DL and LD to be 0 and consider

$$E[Y|Z, G, X, A] = \alpha_{50} + X^T\alpha_{51} + A\alpha_{52} + ZG\alpha_{53} + (1 - Z)G\alpha_{55} + Z\alpha_{54}.$$

The SACE can be estimated in a similar fashion with these alternative models.

3.5 Simulation studies

In this section we examine the finite sample performance of our proposed methods. We first evaluate the performance under settings that satisfy A.1 - A.5, both in randomized studies and non-randomized studies. The proposed method is compared with the method of Ding et al. (2011) under those settings. We then evaluate the sensitivity of the proposed estimators to departure from the exclusion restriction assumption A.4.

A Settings following A.1 - A.5

We first consider simulation settings in which data were generated according to Figure 3.1(b). Specifically the baseline covariates $X = (X_1, X_2, X_3)$ are a combination of discrete and continuous variables: X_1 is a discrete variable taking value 1 or -1 with probability 1/2, and (X_2, X_3) follows a multivariate normal distribution $N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Conditional on X, the substitution variable A was generated from a Bernoulli distribution such that $P(A = 1|X) = \text{expit}(X^T\boldsymbol{u})$, where $\boldsymbol{u} = (1, 1, 1)^T$. The exposure variable Z was generated following a logistic model: $P(Z = 1|X, A) = \text{expit}(\delta_1 X\mathbf{1} + \delta_1 A)$, where δ_1 is a parameter taking value 0 or 1. The survival type G was generated from a multinomial distribution such that $P(G = LL|X, A) = \text{expit}(\gamma_0 + X^T\gamma_1 + A\gamma_2)\text{expit}(\beta_{10} + X^T\beta_{11} + A\beta_{12})$, $P(G = LD|X, A) = \left\{1 - \text{expit}(\gamma_0 + X^T\gamma_1 + A\gamma_2)\right\}\text{expit}(\beta_{10} + X^T\beta_{11} + A\beta_{12})$ and $P(G = DD|X, A) = 1 - \text{expit}(\beta_{10} + X^T\beta_{11} + A\beta_{12})$, where $(\beta_{10}, \beta_{11}, \beta_{12}) = (2, \delta_2, \delta_2, \delta_2, 1)$, $(\gamma_0, \gamma_1, \gamma_2) = (0, -3\delta_2, \delta_2, \delta_2, 1)$ and δ_2 is a parameter taking values 0 or 1. The potential outcome $Y(z)$ was generated from the following normal distributions: $Y(1)|G = LL, X, A \sim N(5 + \delta_2 X^T\boldsymbol{u}, 0.5^2)$, $Y(1)|G = LD, X, A \sim N(3 + \delta_2 X^T\boldsymbol{u}, 0.5^2)$ and $Y(0)|G = LL, X, A \sim N(4 + \delta_2 X^T\boldsymbol{u}, 0.5^2)$. The observed survival S and observed outcome Y follows from the consistency assumption. Note under our settings, A impacts the potential outcomes through

the path $A \to (S(1), S(0)) \to (Y(1), Y(0))$. Hence the treatment assignment is confounded when $\delta_1 = 1$. Similarly, X is a common cause of $A, (S(1), S(0))$ and $(Y(1), Y(0))$ when $\delta_2 = 1$. Under our simulation setting, the true value for the SACE is 1.

We compare three methods for estimating the SACE under this setting: Naive: linear regression of Y on Z, X and A among observed survivors; DGYZ: Ding et al. (2011)'s estimation method; Prop-ER: the proposed method under the exclusion restriction assumption A.4, i.e. the estimation method using M.1 - M.4. Note that Prop-ER differs from DGYZ in both the identifiability conditions and estimation method. Table 3.2 summarizes the results. The naive regression method is biased in all settings due to selection bias. Ding et al. (2011)'s method is consistent when $\delta_1 = \delta_2 = 0$, but it can be unstable when the sample size is small to moderate. To explain this, recall that Ding et al. (2011) assumes the following substitution relevance assumption:

$$P(A = 1|G = LL, Z = 1) \neq P(A = 1|G = LD, Z = 1). \tag{3.15}$$

Although (3.15) holds in our simulation setting with $\delta_1 = \delta_2 = 0$, for some simulated samples the estimate of $P(A = 1|G = LL, Z = 1)$ can be very close to the estimate of $P(A = 1|G = LD, Z = 1)$, leading to instability of the final estimates. Furthermore, one can see that Ding et al. (2011)'s method is sensitive to confounding and/or presence of a common cause of $A, (S(1), S(0))$ and $(Y(1), Y(0))$. In contrast, the proposed method performs well in all settings considered here.

B Settings with exclusion restriction violation

We now evaluate the sensitivity of the proposed method to violation of the exclusion restriction assumption A.4. We generated data in the same way as in Section A, except that the potential outcome $Y(z)$ was generated from the following normal distributions: $Y(1)|G = LL, X, A \sim N(5+\delta_2 X^T \boldsymbol{u}+A, 0.5^2)$, $Y(1)|G = LD, X, A \sim N(3+\delta_2 X^T \boldsymbol{u}+A, 0.5^2)$ and $Y(0)|G = LL, X, A \sim N(4 + \delta_2 X^T \boldsymbol{u} + A, 0.5^2)$. We also implemented the method Prop-

Table 3.2: Bias and standard deviation (in parenthesis) for various estimating methods of the SACE under settings following A.1 - A.5. $\delta_1 = 1$ and $\delta_2 = 1$ corresponds to presence of confounding and common cause of $A, (S(1), S(0))$ and $(Y(1), Y(0))$, respectively. Results are based on 1000 simulated data sets.

Sample size	δ_1	δ_2	Estimation method		
			Naive	DGYZ	Prop-ER
200	0	0	-0.73(0.14)	2.76(87.92)	0.01(1.54)
		1	-0.47(0.15)	-7.90(134.36)	-0.12(0.25)
	1	0	-0.80(0.19)	-1.29(0.52)	-0.11(2.22)
		1	-0.41(0.18)	-1.83(2.40)	-0.27(0.35)
500	0	0	-0.73(0.08)	0.34(3.73)	0.11(0.77)
		1	-0.47(0.09)	-3.43(215.02)	-0.05(0.17)
	1	0	-0.81(0.12)	-1.27(0.27)	0.11(1.19)
		1	-0.42(0.12)	-1.46(0.88)	-0.22(0.23)
1000	0	0	-0.73(0.06)	0.06(0.38)	0.05(0.37)
		1	-0.47(0.07)	-14.47(188.83)	-0.03(0.12)
	1	0	-0.81(0.08)	-1.26(0.18)	0.10(0.61)
		1	-0.41(0.08)	-1.42(0.57)	-0.11(0.23)

Table 3.3: Bias and standard deviation (in parenthesis) for various estimating methods of the SACE under settings with exclusion restriction violation. $\delta_1 = 1$ and $\delta_2 = 1$ corresponds to presence of confounding and common cause of $A, (S(1), S(0))$ and $(Y(1), Y(0))$, respectively. Results are based on 1000 simulated data sets.

Sample size	δ_1	δ_2	Estimation method		
			DGYZ	Prop-ER	Prop-NI
200	0	0	7.44(180.19)	1.48(2.53)	0.36(0.42)
		1	-11.26(247.45)	0.19(0.38)	0.04(0.20)
	1	0	0.34(0.59)	1.34(2.70)	0.47(0.54)
		1	-1.25(3.25)	0.54(0.42)	0.27(0.29)
500	0	0	2.00(6.01)	1.68(0.95)	0.14(0.33)
		1	-1.17(355.16)	0.10(0.22)	0.02(0.13)
	1	0	0.43(0.29)	1.64(1.19)	0.25(0.38)
		1	-0.71(1.35)	0.41(0.37)	0.16(0.19)
1000	0	0	1.62(0.42)	1.62(0.44)	0.08(0.25)
		1	-20.78(285.28)	0.08(0.14)	0.02(0.09)
	1	0	0.43(0.20)	1.67(0.95)	0.11(0.29)
		1	-0.62(0.84)	0.31(0.50)	0.10(0.14)

NI, the proposed method under the no-interaction assumption A.7; that is, the estimation method using M.3, M.4 and M.7. The method of Ding et al. (2011) for dealing with exclusion restriction violations is not applicable here as they require the substitution variable A to be either continuous or have at least three categories.

The simulation results are presented in Table 3.3. As expected, Prop-NI is consistent for all simulation settings considered here, whereas both DGYZ and Prop-ER are biased for estimating the SACE. However, compared with DGYZ, the inclusion of covariate information X still makes Prop-ER more robust to departure from the exclusion restriction assumption. Note that even with $\delta_1 = \delta_2 = 0$, this advantage remains under small to moderate sample sizes.

3.6 Real data analysis

A Application to a SWOG trial

We illustrate the use of the proposed estimation method using data from a randomized phase III trial to compare docetaxel plus estramustine with mitoxantrone plus prednisone in men with metastatic, hormone independent prostate cancer (Petrylak et al., 2004). For illustrative purposes, here we use the data set created by Ding et al. (2011), which contains observations on 487 men aged from 47 to 88. Of these subjects, 258 were assigned to receive docetaxel plus estramustine ($Z = 1$), and 229 were assigned to receive mitoxantrone plus prednisone ($Z = 0$). In our analysis, we are interested in comparing these two treatments in term of health related quality of life (QoL) one year after receiving the treatment.

A naive analysis shows that among patients who survived at one year after receiving the assigned treatment, the QoL for those assigned to the docetaxel plus estramustine group is higher by 2.46 units (95% CI = [-3.31,8.24]) compared to those assigned to the mitoxantrone plus prednisonegroup. However, this estimate is not causally interpretable as subjects who would survive if assigned to docetaxel plus estramustine are different from subjects who would survive if assigned to mitoxantrone plus prednisone. Moreover, as reported by Petrylak et al. (2004), docetaxel plus estramustine is beneficial for the overall survival compared to mitoxantrone plus prednisone. The direct comparison among observed survivors is hence also subject to selection bias. Instead, we apply the following three methods that take into account of truncation due to death: DGYZ, Prop-ER and Prop-NI. To account for possible common causes of the potential survivals and potential outcomes, we adjust for the following variables in the models used by Prop-ER and Prop-NI: age, race (black vs non-black), type of prognosis (PSA only vs others), bone pain (grade < 2 vs grade ≥ 2) and performance status ($0 - 1$ vs $2 - 3$). Following Ding et al. (2011), we choose the baseline QoL to be our substitution variable, and the change in QoL in the one-year period as our outcome.

We first analyze this data set under the monotonicity assumption. The point estimates as well as bootstrapped standard errors are displayed in Table 3.4. DGYZ suggested that

Table 3.4: Survivor average causal effect of docetaxel plus estramustine on health related quality of life

Estimation method	Point estimate	Bootstrapped SE	2.5%	50%	97.5%
DGYZ*	7.01	3.09	1.81	6.56	13.64
Prop-ER	3.06	11.79	-15.15	3.29	22.60
Prop-NI	2.73	3.82	-3.97	2.95	10.83

*: Results were adapted from Ding et al. (2011).

docetaxel plus estramustine had a significant causal effect on the QoL among those who would survive one year after receipt of treatment regardless of which treatment group they were assigned to. In contrast, after accounting for the baseline covariate information that might simultaneously impact the (potential) survival and the (potential) QoL, we were not able to reach such a conclusion. Both Prop-ER and Prop-NI yield a point estimate that is much closer to 0, and their 95% confidence intervals cover 0. These results show that it is important to incorporate baseline covariate information even in randomized trials.

We also note that Prop-ER is very unstable under the monotonicity assumption. In particular, the point estimate for γ is large in magnitude, suggesting that the true parameter may lie on the boundary, which indicates a monotonicity violation. Previous analyses also suggest that the monotonicity assumption might be problematic for this data set (Ding et al., 2011). Furthermore, as we discussed earlier, for randomized trials this assumption is not plausible *a priori*. For these reasons, although the overall one year survival rate in the docetaxel plus estramustine group (49.6%) is higher than that in the mitoxantrone plus prednisonegroup (38.9%), a sensitivity analysis of the monotonicity assumption is desirable. To relax the monotonicity assumption, we assume M.2, M.5 and M.6 and vary the sensitivity parameter ρ from 0 to 1. We note that under our modeling assumptions, each value of ρ corresponds to an estimate for π_{DL}, the overall fraction of patients who would die one year after receiving docetaxel plus estramustine but would survive one year after receiving mitoxantrone plus prednisone. As π_{DL} is more interpretable than ρ, in Figure 3.2, we show

the estimate of the SACE as a function of π_{DL}. The range for π_{DL} is [0.01,0.19]. The lower limit for the range of π_{DL} corresponds to $\rho = 1$. As the lower limit is greater than 0, the monotonicity assumption A.1 is not compatible with the modeling assumptions M.2, M.5 and M.6. On the other hand, the upper limit is smaller than Ding et al. (2011)'s estimate for π_{DL}, which is 0.33. This suggests that under our modeling assumptions, Ding et al. (2011)'s method requires that the potential survival under docetaxel plus estramustine is negatively correlated with the potential survival under mitoxantrone plus prednisone. We also note that Prop-ER is robust to exclusion restriction violation unless π_{DL} is smaller than 0.05. Finally, as both of the lines in Figure 3.2 cross 0, without prior knowledge on the possible values for π_{DL}, we cannot reach any definitive answer about the causal effect of docetaxel plus estramustine versus mitoxantrone plus prednisoneon the QoL one year after receiving the treatment.

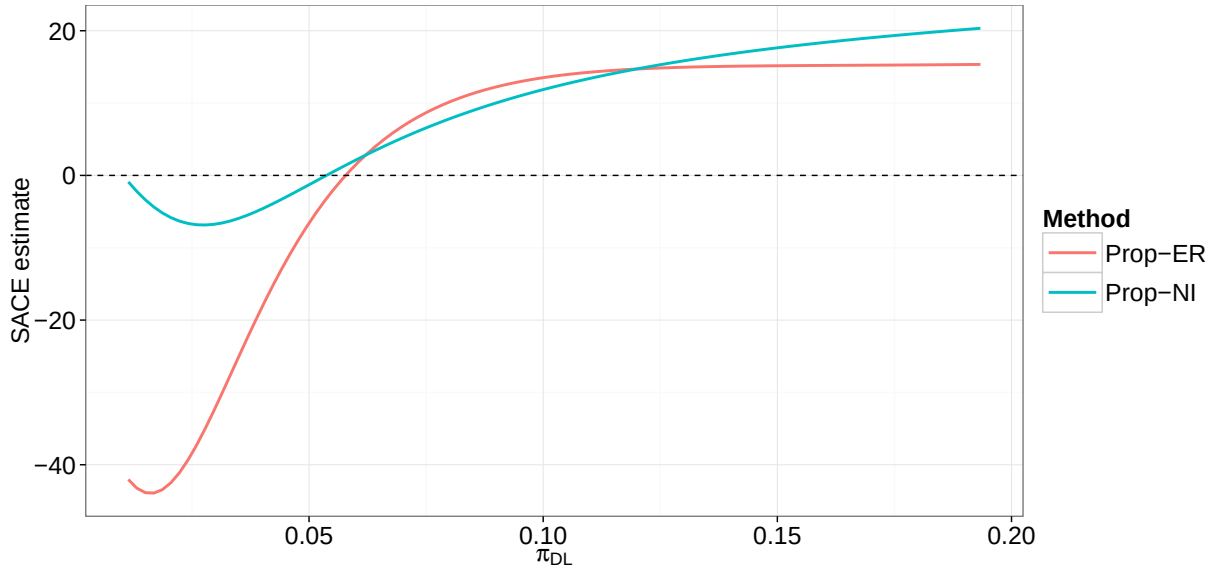

Figure 3.2: Sensitivity analysis for estimating the survivor average causal effect in the SWOG dataset.

B Application to the Health and Lifestyle Study

We now use the proposed methods to estimate the effect of smoking on memory decline using observational data from the Health and Lifestyle Study (HALS). Smoking is a leading cause of preventable disease and death. There has been consensus that smoking is harmful to the overall health of smokers, and leads to increased risks of lung cancer, heart disease and stroke. However, there have been mixed findings about effects of smoking on cognitive functioning (Anstey et al., 2007; Peters et al., 2008). Some researchers have conjectured that this inconsistency may be associated with selection bias due to censoring by death (Hernán et al., 2008).

To investigate the impact of this selection bias on estimating smoking effect on cognitive decline, we use data from the Health and Lifestyle Survey (HALS), a population-based prospective cohort study conducted in England, Scotland and Wales (Cox et al., 1987). There were two visits in the HALS. The baseline survey HALS1 was conducted in 1984-1985 and interviewed 9003 participants on their health statuses, attitudes to health and other measurements related to health and lifestyles. Among these participants, 7414 agreed to have a nurse-visit where an incidental memory test was conducted. In this test, each respondent got a score ranging from 0 to 10 based on how many items he/she correctly recalled from a memory test. The follow-up survey HALS2 was conducted in 1991-1992, at which time 808 of the 7414 original respondents to the baseline survey had died. The surviving respondents of HALS1 were re-surveyed to examine their changes on memory scores over the 7-year period.

With this data set, Whittington and Huppert (1997) examined the relationship between smoking and memory decline in the HALS. However, they restricted the analyses to survivors at the time of HALS2. To facilitate our comparison, we first reanalyzed the data ignoring the problem of truncation by death. In our analysis, we used Z to denote the smoking status, where $Z = 1$ for ever smokers. We had $S = 1$ for surviving respondents from HALS1, and $S = 0$ for respondents of HALS1 who died before HALS2 was administered. The

covariates X adjusted in our analysis included age, sex, education level, alcohol consumption, BMI, household income and family smoking history. This is a superset of the variables considered in Whittington and Huppert (1997). We use the classical linear regression to remove confounding due to observed covariates. After adjusting for covariates, smokers at the baseline visit have 0.04 more points of decline in memory score compared to non-smokers at the baseline visit (95% CI = [-0.12,0.21]). The result was non-significant, suggesting that no evidence was found for effects of smoking on memory decline. However, as noted by these authors, more regular smokers than non-smokers died in the 3 years following HALS2. This is strong evidence for the association between smoking and death. Consequently, selection bias due to truncation by death should be taken into consideration.

We then applied the proposed estimation methods to the same data set. We used family history of lung cancer as our substitution variable for the survival type as it is plausible that family history of lung cancer is not related to memory decline conditioning on the smoking behavior of the respondent and his/her parents (exclusion restriction assumption). On the other hand, as family history of lung cancer is indicative of "lung cancer genes" that are likely to be of low frequency but high penetrance (Satcher et al., 2002), it is likely to affect smoking-related mortality (substitution relevance).

In our analysis, we assumed the monotonicity assumption A.1, that is, all subjects that would survive at the time of HALS2 if they ever smoked at the time of HALS1 would also survive at the time of HALS2 if they never smoked at the time of HALS1. In other words, smoking does not save lives. To partially test this assumption, we used a logistic regression model. After adjusting for the covariates X, smoking is associated with a 94.8% (95% CI = [27.4%, 198.1%]) increase in odds of surviving between the follow-ups in the HALS. Thus the monotonicity assumption seems plausible in this case.

Table 3.5 summarizes results from the proposed methods. The point estimates with Prop-ER and Prop-NI are very close to the estimate form the naive analysis, suggesting that the selection bias due to death is minimal for this data set. However, although the estimates are close, it should be noted that their interpretations are different. The proposed methods

estimate the SACE; that is, the effect of smoking on memory decline among respondents who would survive at the time of HALS2 regardless of whether they ever smoked at the time of HALS1. As this group of people are defined at baseline, the SACE is causally interpretable. In contrast, the naive method compares the memory decline of smokers to non-smokers among respondents who were observed to survive at the time of HALS2. This comparison cannot be interpreted causally as observed survivors at the time of HALS2 represent different group of respondents at the baseline visit.

Table 3.5: Summary of results from the proposed methods

Method	Point estimate	Bootstrapped SE	95% CI
Naive	0.043	0.084	[-0.121,0.207]
Prop-ER	0.007	0.099	[-0.153,0.244]
Prop-NI	0.022	0.088	[-0.133,0.205]

3.7 Discussion

In this chapter, we have considered the identification and estimation problem of the SACE. Compared with previous works, we introduce baseline covariates X into the identification framework, which play the dual role of potential confounders and common causes of A,(S(1),S(0)) and (Y(1),Y(0)). The second role of baseline covariates is unique to our identification framework for the SACE. In particular, in contrast to the standard estimation problem of the average causal effect (ACE), inclusion of baseline covariates is crucial for obtaining unbiased estimates of the SACE even in randomized studies.

In our estimation approach, instead of imposing modeling assumptions on the observed data, we use parametric models for the potential outcomes. This not only simplifies the estimation procedure, but also makes it easier to incorporate the identifiability assumptions into our modeling assumptions. We also propose alternative models to relax the monotonicity assumption and the exclusion restriction assumption.

In the current work, we have only considered the binary exposure variable case. In practical medical studies, there may be multiple levels in the exposure variable. For example, the smoking variable may be coded as never smokers, past smokers and current smokers. In this case, it would be interesting to generalize the proposed methods to deal with ordinal exposure status.

Appendix

A Proof of Theorem 3.1

Proof. Due to A.4, we have

$$E[Y|Z=1,G=g,X,A=a_1] = E[Y|Z=z,G=g,X,A=a_0] \equiv \mu^1_{g,X}$$

for $g = LL, LD$. Equation (3.4) then follows directly from tower property of conditional expectation. Moreover, the coefficient matrix of linear system (3.4) is non-singular due to A.5, which guarantees that $p_{g|a_1,X} \neq p_{g|a_0,X}$. Therefore, $\mu^1_{LL,X}$ is identifiable from (3.4).

□

B Remarks on propensity score adjustment

Here we explain why the propensity score adjustment methods are not directly applicable to our setting. To illustrate, we use the DAG in Fig 3.1(b).

First, as shown by Rosenbaum and Rubin (1983), the propensity score is sufficient for summarizing the effect of X on Z as $Z \perp X|e(X)$. Thus we can add $e(X)$ to the DAG in Figure 3.1(b); this is shown in Figure 3.3.

Now note that a key identifying assumption in our framework is the exclusion restriction A.4. A propensity score-based estimation approach would require that we can replace X in A.4 with $e(X)$, i.e.

$$A \perp Y(1) \mid Z=1, e(X), G.$$

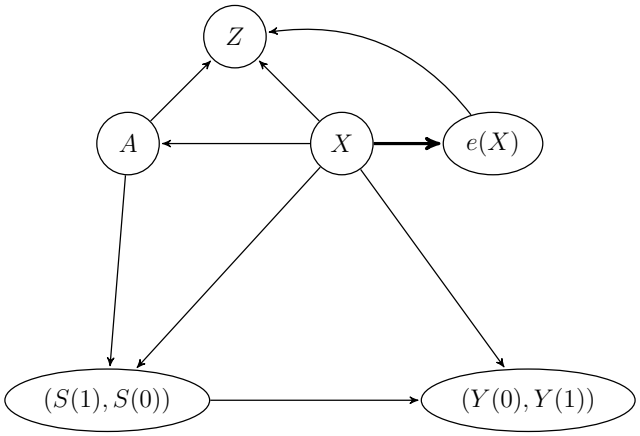

Figure 3.3: The DAG model in Figure 3.1(b) with an added node $e(X)$. The thick edge between X and $e(X)$ indicates a deterministic relationship.

However, as one can see from Figure 3.3, the backdoor path $A - X \to (Y(1), Y(0))$ is not blocked even if one conditions on $e(X)$. The underlying reason is that $e(X)$ is not sufficient for summarizing the effect of X on A; in other words,

$$A \not\perp X | e(X).$$

As a result, the path $A - X$ cannot be blocked by conditioning on $e(X)$.

C Proof of Proposition 3.4

The only non-standard constraints are (3.8) and (3.10). To prove the former, note that under A.4, (3.4) holds for all a, i.e.

$$E[Y|Z = 1, S = 1, X = x, A] = p_{LL|x,a}\mu_{1,LL,x} + p_{LD|x,a}\mu_{1,LD,x}. \tag{3.16}$$

It follows that

$$|E[Y|Z = 1, S = 1, X = x, A]| \leq \max\{|\mu_{1,LL,x}|, |\mu_{1,LD,x}|\},$$

and hence $E[Y|Z=1,S=1,X=x,A]$ is bounded for all x. On the other hand, suppose that $E[Y|Z=1,S=1,X=x,A]$ is bounded for all x. Let

$$\bar{f}(x)=\max_a E[Y|Z=1,S=1,X=x,A=a], \quad \text{and} \quad \underline{f}(x)=\min_a E[Y|Z=1,S=1,X=x,A=a].$$

Then (3.16) holds with

$$p_{LL|x,a}=\frac{\bar{f}(x)-E[Y|Z=1,S=1,X=x,A=a]}{\bar{f}(x)-\underline{f}(x)}, \quad \mu_{1,LL,x}=\underline{f}(x) \quad \text{and} \quad \mu_{1,LD,x}=\bar{f}(x).$$

Hence (3.8) summarizes all the constraints on the observed data laws derived from (3.16).

Proof of (3.10) follows similarly and is hence omitted.

Chapter 4

CAUSAL ANALYSIS OF ORDINAL TREATMENTS AND BINARY OUTCOMES UNDER TRUNCATION BY DEATH

4.1 Introduction

In multi-arm randomized trials, researchers are often interested in analyzing treatment effects on an outcome that is measured or well-defined only when an intermediate outcome takes certain values (Robins, 1986; Rubin, 2000, 2006; Egleston et al., 2007; Chiba and VanderWeele, 2011; Ding et al., 2011). For example, consider a multi-arm randomized HIV vaccine trial. Scientists might be interested in evaluating vaccine effects on HIV viral load as it correlates with infectiousness and disease progression (Hudgens et al., 2003; Gilbert et al., 2003). However, HIV viral load is typically measured only for infected individuals. Two problems occur in this case: in general, there are many potential comparisons that can be made between different vaccination groups among infected subjects; moreover, these comparisons are subject to selection bias as the vaccine may affect susceptibility to HIV infection. In the simple case of a two-arm trial, to deal with the selection bias problem, several authors have proposed to consider the vaccine effects on viral load among the always-infected stratum, the subpopulation who would become infected regardless of whether they are vaccinated or not (e.g., Hudgens et al., 2003; Gilbert et al., 2003). However, there has not been much work on analyzing this type of trial with more than two arms.

By convention, the intermediate outcome is called "survival," and we say the final outcome is truncated by death if it is only observed and/or well-defined for "survivors." Thus in the HIV vaccine example above, the always-infected stratum is referred to as the "always-survivor" stratum. The causal contrast among the always-infected subjects is hence called the (always-)survivor average causal effect (SACE) (Rubin, 2000).

In general, even in a two-arm trial, the SACE is not identifiable without strong untestable assumptions. As a result, there are no consistent tests for detecting non-null vaccine effects in the always-infected stratum. Instead, under some reasonable assumptions, Hudgens et al. (2003) tested the null hypothesis presuming the maximal degree of selection bias. Their approach is related to estimation of bounds on SACE, which has been extensively studied in literature. For example, Zhang and Rubin (2003) developed bounds on SACE under various assumptions including the monotonicity assumption and the stochastic dominance assumption. Imai (2008) provided an alternative proof that the bounds of Zhang and Rubin (2003) are sharp by formulating the truncation-by-death problem as a "contaminated data" problem. These testing and estimation methods are appealing in practice as they don't rely on strong identifiability assumptions.

However, to the best of our knowledge, there has not been much discussion on testing and estimation of SACEs in a multi-arm trial, which is fairly common in medical practice (Schulz and Grimes, 2005). Prior to our work, Lee et al. (2010) considered a sensitivity analysis approach to identify all SACEs in a three-arm trial. Their identification results rely on a strong parametric assumption and several sensitivity parameters. In this chapter, we instead propose a framework to systematically analyze SACEs in a general multi-arm trial without strong identification assumptions. To the best of our knowledge, our method is also the first that is readily applicable to randomized trials with more than three treatment arms under truncation by death.

The testing and estimation of SACEs in a multi-arm trial are more challenging compared to two-arm trials. Firstly, in general there are many different SACEs that are well-defined. As we show later in Section B, consideration of all SACEs (as in Lee et al. (2010)) can lead to paradoxical non-transitive conclusions. Hence we instead restrict our attention to comparisons within the "finest" (principal) strata, thereby avoiding this difficulty. Secondly, one needs to distinguish between an overall analysis of treatment effects and a separate analysis of each individual contrast. In the simple setting without truncation by death, it is widely known that compared to all pairwise comparisons with correction for multiple

comparisons, an overall analysis such as an ANOVA test often provides more power for testing the overall treatment effect in a multi-arm trial. When truncation by death is present, due to non-identifiability of SACEs, this advantage becomes more fundamental as it remains even when the sample size goes to infinity. In contrast to Lee et al. (2010), we distinguish between simultaneous versus marginal inference for SACEs, and argue that they should be used to answer different questions. In particular, we show that compared to marginal inference procedures, our proposed simultaneous inference procedures provide more power for testing the overall treatment effect and the advantage remains even with an infinite sample size. Thirdly, the simultaneous inference problem is unique to a multi-arm trial. Again, since SACEs are not identifiable, traditional statistical inference tools for multi-arm trials without truncation by death are not directly applicable to our setting. Instead, we develop novel simultaneous inference procedures to test an overall treatment effect, and show that they have desirable asymptotic properties. We also generalize the marginal inference procedures for a two-arm trial to get sharp bounds on SACEs for a general multi-arm trial. To focus on addressing these challenges, in this chapter, we restrict our attention to trials with ordinal treatment groups and binary outcomes.

The rest of this chapter is organized as follows. In Section 4.2, we introduce our notations, assumptions and define our causal estimands. We also address the transitivity issue and identify three specific testing and estimation questions that may arise in a general multi-arm trial with truncation by death. We then propose three novel procedures that answer these questions in Section 4.3, 4.4 and 4.5, respectively. In Section 4.3, we discuss the unique challenges for hypothesis testing with non-identifiable parameters, and develop a novel step-down testing procedure to test the overall treatment effect in this situation. In Section 4.4, we develop a linear programming algorithm to test an overall clinically relevant treatment effect. In Section 4.5, we derive the sharp marginal bounds for each causal contrast of interest. In Section 4.6, we illustrate the proposed procedure with real data analyses. Results from simulation studies can be found in the Appendix. We end with a discussion in Section 4.7.

4.2 Framework

A Data structure and assumptions

Consider a multi-arm trial with control and multiple arms of active treatment. Let Z be an ordinal treatment variable, where $Z = 0$ corresponds to the control treatment, and $Z \in \{1, \ldots, m\}$ corresponds to different arms of active treatment. In what follows, we use the terminology "treatment arms" and "treatment levels" interchangeably. We assume that each subject has $m + 1$ dichotomous potential outcomes $Y(z), z = 0, \ldots, m$, where $Y(z)$ is defined as the outcome that would have been observed if the subject was assigned to treatment arm z. Similarly, we define $S(z)$ as the potential survival status under treatment assignment z. We assume $Y(z)$ is well-defined only if $S(z) = 1$. In other words, the outcome of interest is well-defined only for subjects who survive to the follow-up visit. We also assume that the observed data $(Z_i, S_i, Y_i; i = 1, \ldots, N)$ are independently drawn from an infinite super-population.

Let $G = (S(0), \ldots, S(m))$ denotes the *basic principal stratum* (Frangakis and Rubin, 2002). If we let the letter L denote $S(z) = 1$ (meaning "live") and the letter D denote $S(z) = 0$ (meaning "die"), then G can be rewritten as a string consisting of the letters "L" and "D." For example, in a three-arm trial, $G_i = DLL$ indicates that subject i would die under control, but would survive under active treatment 1 or 2.

We make the following assumptions.

Assumption 4.1. *Stable unit treatment value assumption (SUTVA (Rubin, 1980)): there is no interference between units, and there is only one version of treatment.*

Under the SUTVA, the observed outcome equals the potential outcome under the observed treatment arm, namely $Y = Y(Z)$ and $S = S(Z)$.

Assumption 4.2. *Random treatment assignment:* $Z \perp (S(0), \ldots, S(m), Y(0), \ldots, Y(m))$.

Assumption 4.3. *Monotonicity:* $S_i(z_1) \geq S_i(z_2), i = 1, \ldots, N, z_1 \geq z_2$.

The monotonicity assumption is usually plausible in social science studies if the treatment options can be reasonably ordered. For example, in randomized experiments evaluating the effect of incentives on survey response quality, it is intuitive that higher level of incentives would not hurt survey response rates. This assumption tends to be more controversial in medical studies, in which there are often trade-offs between treatment benefits and side effects.

The only possible strata under the monotonicity assumption are strata of the form $D \cdots DL \cdots L$. To compress notation, we denote all possible principal strata as $(D^k L^{m+1-k}; k = 0, \ldots, m+1)$, where members of principal stratum $D^k L^{m+1-k}$ would die if assigned to the first k treatment arms but would survive if assigned to the remaining $m+1-k$ treatment arms.

B Causal estimands and questions

For randomized trials with two treatment arms, it is common to estimate the average causal effect in the LL stratum (Kalbfleisch and Prentice, 1980; Robins, 1986; Rubin, 2000), the only subgroup for which both of the potential outcomes are well-defined: $SACE = E[Y(1) - Y(0)|G = LL]$. In a general multi-arm trial, researchers may be interested in comparisons of potential outcomes within the same basic principal stratum. For example, in the case where we have three levels of treatment: 0, 1, 2, the target estimands are $E[Y(2) - Y(1)|G = LLL], E[Y(1) - Y(0)|G = LLL], E[Y(2) - Y(0)|G = LLL]$ and $E[Y(2) - Y(1)|G = DLL]$. These contrasts are causally meaningful as the memberships of basic principal strata are defined at baseline.

To define the causal estimands for a general multi-arm trial, we first introduce some notation. Let $\mu_g^z \equiv E[Y(z)|G = g]$ denote the mean potential outcome under treatment assignment z in basic principal stratum g. Also, let $\mathcal{M}(g)$ denote *the minimal treatment level under which members of principal stratum g can survive.* In other words, for members of principal stratum g, $S(z) = 1$ if and only if $z \geq \mathcal{M}(g)$. Consequently, μ_g^z is well-defined if and only if $z \geq \mathcal{M}(g)$. Under the monotonicity assumption, all basic principal strata take the form $g =$

$D^k L^{m+1-k}$. By definition, $\mathcal{M}(D^k L^{m+1-k}) = k$. Also let $\Omega_k = \{g : \mathcal{M}(g) \leq k\}$ denote *the collection of basic principal strata whose members would survive if assigned to treatment arm* k. The pairwise causal estimands in a multi-arm trial then take the form

$$\Delta(z_1, z_2; g) \equiv \mu_g^{z_1} - \mu_g^{z_2}, \text{ where } g \in \Omega_{m-1}, z_1 > z_2 \geq \mathcal{M}(g). \tag{4.1}$$

For notational simplicity, in this chapter, when we write the notation μ_g^z and $\Delta(z_1, z_2; g)$, we always assume that it is well-defined. We also note that the parameters involved in defining the causal contrasts $\Delta(z_1, z_2; g)$ are contained in the parameter vector $\boldsymbol{\mu}_{m-1} \equiv (\mu_g^z; g \in \Omega_{m-1}, z \geq \mathcal{M}(g))$.

There are other meaningful causal contrasts that are made within *coarsened principal strata*, defined as groups that combine several basic principal strata (Cheng and Small, 2006). For example, in the case of a three-arm trial, the contrast $E[Y(2) - Y(1)|G \in \{LLL, DLL\}]$ is also causally meaningful as memberships of the coarsened principal strata $\{LLL, DLL\}$ are also defined at baseline. Some previous researchers hence consider coarsened principal strata causal effects together with basic principal strata causal effects (Lee et al., 2010). However, as Robins (1986) noted, if one were to compare $E[Y(2)-Y(0)|G = LLL]$, $E[Y(1)-Y(0)|G = LLL]$ and $E[Y(2) - Y(1)|G \in \{LLL, DLL\}]$ simultaneously, it is possible that the last two comparisons are both positive while the first one is negative. This lack of transitivity limits the interpretability of causal effects defined within coarsened principal strata. In contrast, transitivity holds if limited to basic principal strata (e.g., LLL). Hence in this chapter, we are primarily interested in comparisons between potential outcomes in the same *basic* principal stratum.

On the other hand, as Robins et al. (2007a) noted, the size of each basic principal stratum is likely to be very small and consequently, each comparison in (4.1) only applies to a small portion of the population. Hence for randomized trials with more than three treatment arms, we may have limited power to test treatment effects for each basic principal stratum. What is more, we run into the problem of multiple comparisons as there are multiple treatment

arms and multiple basic principal strata.

Therefore, we first consider testing the global null hypothesis that the treatment is not effective in any of the basic principal strata. This question is scientifically relevant. For example, in a HIV vaccine trial, testing the global null addresses whether there exists a mechanism through which the vaccine alters viral load in infected individuals (Shepherd et al., 2006). Secondly, clinicians may also be interested in whether the overall treatment effect is clinically meaningful so that the active treatment is promising in clinical practice. For this purpose, an overall treatment effect may be declared only if it is greater than the clinical margin of relevance specified by clinicians. Finally, besides an overall treatment effect, scientists and clinicians may also be interested in isolating the non-zero/non-trivial causal contrasts. In summary, the following questions are of interest with a multi-arm trial:

1. Is there evidence of the existence of *non-zero* average treatment effects for at least one basic principal stratum between at least two treatment arms?

2. Are there *clinically relevant* average treatment effects for at least one basic principal stratum between at least two treatment arms?

3. Can we find the specific principal strata and treatment arms that correspond to the overall non-zero/clinically relevant treatment effect, if such exists?

We address these questions in Section 4.3, 4.4 and 4.5 respectively. Existing causal analysis literature in multi-arm trials with non-identifiable causal estimands (Cheng and Small, 2006; Long et al., 2010; Lee et al., 2010) focuses on answering the third question. However, as we explain later in Remark 4.8, one may be able to answer the first two questions even if there is not enough information to answer the third. Hence it is important to consider all three questions.

4.3 Testing treatment effects in a multi-arm trial

To find out if there is an overall non-zero treatment effect, it is desirable to consider the

following testing problem:

$$\mathcal{H}_0 : \Delta(z_1, z_2; g) = 0, \forall z_1, z_2, g \quad vs \quad \mathcal{H}_a : \exists z_1, z_2, g \;\; s.t. \;\; \Delta(z_1, z_2; g) \neq 0, \tag{4.2}$$

where $\forall$ means "for all," $\exists$ means "there exists" and $s.t.$ means "such that." The testing problem (4.2) is fundamentally different from (and more difficult than) a standard testing problem, in which one assumes if the observed data distribution was known, one would also know whether or not the hypothesis is true (Lehmann and Romano, 2006). The main difficulty here is that $\mathcal{H}_0$ is a statement about non-identifiable parameter vector $\boldsymbol{\mu}_{m-1}$. In other words, even if the population probabilities $P(S = 1|Z = z)$ and $P(Y = 1|S = 1, Z = z)$ were known, we could only ascertain that $\boldsymbol{\mu}_{m-1}$ resides in a region, and therefore may not know whether $\mathcal{H}_0$ is true or not.

Nevertheless, $\boldsymbol{\mu}_{m-1}$ is "partially identifiable" in the sense that the observed data distribution can narrow down the range in which $\boldsymbol{\mu}_{m-1}$ can possibly lie (Cheng and Small, 2006). For example, in a three-arm trial, the domain of $\boldsymbol{\mu}_2$ is $[0, 1]^6$. However, if the observed data distribution was known, the feasible region of $\boldsymbol{\mu}_2$ would be a subspace in $[0, 1]^6$ subject to the following constraints:

$$\begin{aligned}
&P(Y = 1|Z = 0, S = 1) = \mu^0_{LLL},\\
&P(Y = 1|Z = 1, S = 1) = p^1_{LLL}\mu^1_{LLL} + p^1_{DLL}\mu^1_{DLL},\\
&P(Y = 1|Z = 2, S = 1) = p^2_{LLL}\mu^2_{LLL} + p^2_{DLL}\mu^2_{DLL} + p^2_{DDL}\mu^2_{DDL},
\end{aligned} \tag{4.3}$$

where $p^z_g \equiv P(G = g|Z = z, S = 1)$ is identifiable (see Lemma 1 in Appendix). Figure 4.1 provides a graphical representation of the functional relations described in (4.3).

For a general multi-arm trial, if the parameter space defined by $\mathcal{H}_0$ has no intersection with the feasible region of $\boldsymbol{\mu}_{m-1}$, one would know that $\mathcal{H}_0$ is not true. In general, we introduce the following notions for hypothesis testing with non-identifiable parameters.

Definition We define a hypothesis relating to a parameter to be *compatible* with an ob-

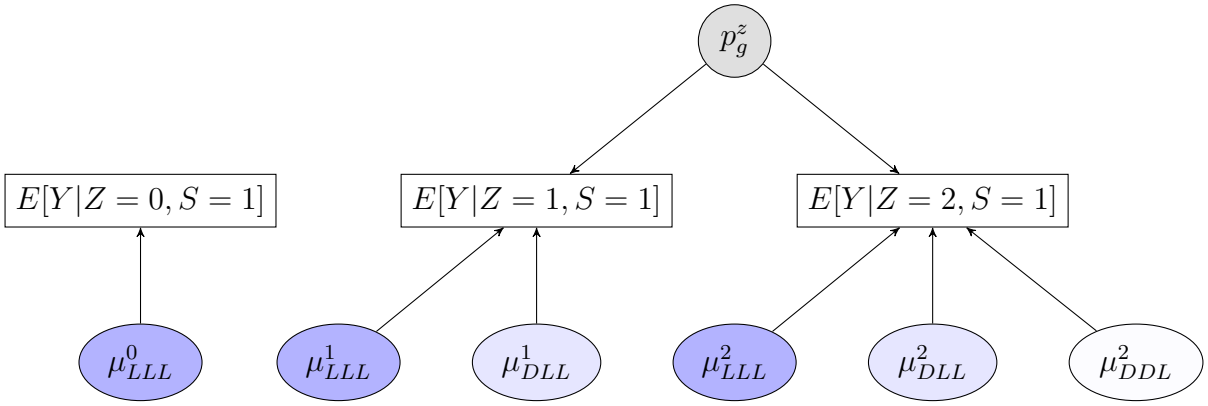

Figure 4.1: A graph representing the functional dependencies in the causal analysis of a three-arm randomized trial with truncation by death. Rectangular nodes represent observed variables; oval nodes represent unknown parameters, with different shadings corresponding to different principal strata. Under the monotonicity assumption, p_g^z can be identified from observed quantities $P(S=1|Z=z)$.

served data distribution if the parameter space defined by the hypothesis has a non-empty intersection with the feasible region of the parameter under the observed data distribution.

In particular, if a parameter is completely unidentifiable such that the observed data distribution imposes no constraints on the parameter, then all hypotheses relating to that parameter are compatible with the observed data distribution. On the other hand, if a parameter is identifiable so that its feasible region under the observed data distribution is always a single point set, then all compatible hypotheses are true.

In general, however, not all compatible hypotheses are true. Nevertheless, due to lack of identifiability, a true hypothesis may not be distinguishable from data with an untrue yet compatible hypothesis. This leads to the following notion of sharpness.

Definition We define a test to be *sharp* for testing a null hypothesis if when the null is not compatible with the observed data distribution (and is hence untrue), the power of the test tends to 1 when the sample size goes to infinity.

Intuitively, similar to consistent tests, sharp tests are those that maximize power asymptoti-

cally. The difference is that as the sample size goes to infinity, with probability tending to 1, sharp tests reject any hypotheses that are incompatible with the observed data distribution, whereas consistent tests reject any hypotheses they are untrue. In small sample settings, however, the conclusions that one would draw from a sharp test are similar to those from a consistent test. If a hypothesis is rejected, one would conclude that it is untrue (at a certain significance level); if otherwise, no claims about the correctness of the hypothesis would be made. We also note that for a standard hypothesis testing problem as described in Lehmann and Romano (2006), sharp tests are the same as consistent tests. When the null hypothesis concerns non-identifiable parameters, however, there are in general no consistent tests. Instead, sharpness plays the role of consistency in a standard hypothesis testing problem.

The notion of sharp tests is similar in spirit to the notion of sharp bounds, defined as the tightest possible bound given the observed data distribution (e.g., Imai, 2008). This notion has also been used implicitly in previous works. For example, Hudgens et al. (2003)'s test for SACE in a two-arm trial is sharp.

Below in Section A, we develop a sharp test for problem (4.2) under the presumption that the observed data distribution is known. In other words, we assume the sample size is infinite such that there is no stochastic variation in the observed data. In Section B we incorporate sampling uncertainty to our proposed test using a Bayesian method.

A A step-down procedure for testing the global null $\mathcal{H}_0$

To fix the idea, we first consider the problem for a three-arm trial, for which $\mathcal{H}_0$ holds if and only if

$$\mu^0_{LLL} = \mu^1_{LLL} = \mu^2_{LLL} \tag{4.4}$$

and

$$\mu^1_{DLL} = \mu^2_{DLL}. \tag{4.5}$$

We hence propose a two-step procedure. Firstly we test hypothesis (4.4). If (4.4) is compatible with the observed data distribution, we then test if (4.5) is compatible with the observed

data distribution conditioning on (4.4).

Specifically, one can see from Figure 4.1 that μ^0_{LLL} is identifiable from the observed data and suppose the feasible regions of μ^1_{LLL} and μ^2_{LLL} are B_{01} and B_{02}, respectively. If μ^0_{LLL} is not contained in the intersection of B_{01} and B_{02}, then (4.4) and hence $\mathcal{H}_0$ is not compatible with the observed data distribution. If otherwise, that (4.4) is compatible with the observed data distribution, we then test hypothesis (4.5) conditioning on (4.4). Note under (4.4), μ^1_{LLL} and μ^2_{LLL} are identifiable. Consequently, μ^1_{DLL} is identifiable. Suppose the feasible region of μ^2_{DLL} under the constraint (4.4) is B_{12}. If μ^1_{DLL} is not contained in B_{12}, we conclude that (4.5) is not compatible with the observed data distribution under the constraint (4.4) and hence reject $\mathcal{H}_0$. If otherwise, we conclude that $\mathcal{H}_0$ is compatible with the observed data distribution.

Algorithm 1 generalizes the procedure described above to general multi-arm trials. Theorem 4.1 states the asymptotic optimality of Algorithm 1. The proof is provided in Appendix.

Algorithm 1 A step-down algorithm for testing the global null hypothesis $\mathcal{H}_0$

1. **Set** $k = 0$

2. **For** $z = k, \ldots, m$

   obtain the feasible region B_{kz} for $\mu^z_{D^k L^{m+1-k}}$ (see Theorem 4.2)

3. **If** $\bigcap_{z=k,\ldots,m} B_{kz} = \emptyset$

   reject $\mathcal{H}_0$; report k; stop

   else

   set $\mu^k_{D^k L^{m+1-k}} = \cdots = \mu^m_{D^k L^{m+1-k}}$ (4.6)

4. **If** $k = m$

   fail to reject $\mathcal{H}_0$ and stop

   else

   set $k = k + 1$ and go to Step 2

Theorem 4.1. *The test given by Algorithm 1 is sharp for testing $\mathcal{H}_0$. In other words, it is asymptotically optimal for testing $\mathcal{H}_0$ as it maximizes power given the observed data distribution.*

To derive the feasible regions $(B_{kz}; k = 0, \ldots, m, z = k, \ldots, m)$ in Algorithm 1, we introduce notation building on Horowitz and Manski (1995). Let $Q^z_{\mathcal{G}}(\cdot)$ denote the distribution (function) of outcome Y among members of subgroup $\mathcal{G}$ who receive treatment z, and $\delta_x(\cdot)$ be a degenerate distribution function localized at x. As Y is binary, $Q^z_{\mathcal{G}}(\cdot)$ is a Bernoulli distribution with mean $m_{\mathcal{G}}(z)$: $Q^z_{\mathcal{G}}(\cdot) = (1 - m_{\mathcal{G}}(z))\delta_0(\cdot) + m_{\mathcal{G}}(z)\delta_1(\cdot)$. To compress notation, we write $Q^z_{\mathcal{G}}(\cdot)$ as $Q^z_{\mathcal{G}}$. Also let $L_\lambda(Q)$ and $U_\lambda(Q)$ be functionals that map a distribution function Q to the corresponding distributions truncated at the lower λ quantile and upper λ quantile, respectively. Theorem 4.2 gives the formula for feasible region B_{lz}. Intuitively, the bounds of B_{lz} are obtained by assigning the smallest/largest ω^z_g portion of observed outcome values in distribution $Q^z_{\boldsymbol{g}}$ to principal stratum g. The proof is in Appendix.

Theorem 4.2. *Suppose that the observed data distribution is known and* (4.6) *holds for all $k < l$. Let $g = D^l L^{m+1-l}$ and $\boldsymbol{g} = \bigcup_{\bar{g} \in \Omega_z \backslash \Omega_{l-1}} \bar{g}$ be the coarsened principal stratum whose members would survive if assigned to treatment z but would die if assigned to treatment $l-1$. The feasible region of μ^z_g is*

$$B_{lz} = \left[\int y dL_{\omega^z_g}(Q^z_{\boldsymbol{g}}), \int y dU_{\omega^z_g}(Q^z_{\boldsymbol{g}}) \right], \tag{4.7}$$

where $\omega^z_g \equiv P[G = g | G \in \Omega_z \backslash \Omega_{l-1}] = p^z_g \Big/ \left(\sum_{\bar{g} \in \Omega_z \backslash \Omega_{l-1}} p^z_{\bar{g}} \right)$ and $Q^z_{\boldsymbol{g}}$ is a Bernoulli distribution with mean

$$m_{\boldsymbol{g}}(z) = \left(m(z) - \sum_{\underline{g} \in \Omega_{l-1}} p^z_{\underline{g}} \mu^z_{\underline{g}} \right) \Big/ \left(1 - \sum_{\underline{g} \in \Omega_{l-1}} p^z_{\underline{g}} \right),$$

in which $m(z) \equiv P[Y = 1 | Z = z, S = 1]$.

Remark 4.3. *Algorithm 1 is a "step-down" procedure in the sense that the hypothesis $\mathcal{H}_0$ is decomposed into a series of hypotheses where the first hypothesis concerns the first stratum*

L^{m+1}, the second hypothesis concerns the second stratum DL^m conditioning on the first hypothesis, and so on.

B Bayesian procedures

we have so far developed a sharp test for problem (4.2). In practice, however, sampling uncertainty must be taken into account when making statistical inference. Here we introduce a Bayesian procedure to estimate the posterior probability that $\mathcal{H}_0$ is not compatible with the observed data distribution. The Bayesian method produces multiple samples of the posterior distribution, thereby reflecting randomness in observed data.

Let $p(s, y|z) = P(S = s, Y = y|Z = z)$ and $p(\cdot, \cdot|z) = (p(1, 1|z), p(1, 0|z), p(0, \uparrow |z))$, where $\uparrow$ indicates that Y is undefined when $S = 0$. Define $\boldsymbol{p} = (p(\cdot, \cdot|0), \ldots, p(\cdot, \cdot|m))$. Under independent Dirichlet priors over the *observed distributions* $p(\cdot, \cdot|z), z = 0, \ldots, m$, it is easy to sample from the posterior distribution via conjugacy. We propose to use Algorithm 2 to calculate the posterior probability that $\mathcal{H}_0$ is not compatible with the observed data distribution.

Remark 4.4. *The step-down procedure in Algorithm 1 has a similar structure to the sequential tests for nested hypotheses discussed by Rosenbaum (2008). From the Frequentist perspective, he has interesting propositions on controlling the type I error rate without resorting to multiplicity adjustment. However, with his methods one proceeds to the next step if the current hypothesis is rejected whereas in our proposal, one proceeds if the current hypothesis is* not *rejected. Moreover, in his context, the parameters of interest are identifiable. Hence his results are not directly applicable to our case.*

4.4 Testing clinically relevant treatment effects in a multi-arm trial

If a non-zero treatment effect is found using Algorithm 2, a natural question arises that whether the treatment effect is clinically meaningful. Suppose the margin of clinical relevance is Δ_0 such that a treatment effect smaller than this would not matter in practice, and also

Algorithm 2 A Bayesian procedure for testing $\mathcal{H}_0$

1. Place an independent Dirichlet prior $Dir(\alpha_{3z+1}, \alpha_{3z+2}, \alpha_{3z+3})$ on $p(\cdot, \cdot|z)$, $z = 0, \ldots, m$.

2. Simulate samples $\boldsymbol{p}^{(1)}, \cdots, \boldsymbol{p}^{(M)}$ from the posterior distributions, which are independent Dirichlet distributions

$$Dir(\alpha_{3z+1} + n_{3z+1}, \alpha_{3z+2} + n_{3z+2}, \alpha_{3z+3} + n_{3z+3}), z = 0, \ldots, m,$$

where $n_{3z+1} = \sum_{i=1}^{N} I(S_i = 1, Y_i = 1, Z_i = z), n_{3z+2} = \sum_{i=1}^{N} I(S_i = 1, Y_i = 0, Z_i = z), n_{3z+3} = \sum_{i=1}^{N} I(S_i = 0, Z_i = z)$.

3. Run Algorithm 1 with each of the posterior samples satisfying the following inequalities:

$$P(S = 1|Z = m) \geq \cdots \geq P(S = 1|Z = 1) \geq P(S = 1|Z = 0) \tag{4.8}$$

Note (4.8) characterizes the set of observed data distributions arising from the potential outcome model defined by Assumptions 4.1 - 4.3.

4. Report the proportion of posterior samples with which $\mathcal{H}_0$ is rejected.

suppose that the treamtent effect is clinically meaningful only if a higher treatment level corresponds to a higher mean potential outcome. It is desirable to consider the following testing problem:

$$\mathcal{H}_{0,c} : \Delta(z_1, z_2; g) \leq \Delta_0, \forall g, z_1 \geq z_2 \quad vs \quad \mathcal{H}_{a,c} : \exists g, z_1 \geq z_2 \;\; s.t. \;\; \Delta(z_1, z_2; g) > \Delta_0, \quad (4.9)$$

where the letter "c" in $\mathcal{H}_{0,c}$ is short for "clinical relevance." Similar to (4.2), (4.9) is a testing problem on non-identifiable parameters. However, as the null parameter space is a non-degenerate region in the domain of $\boldsymbol{\mu}_{m-1}$, the step-down procedure developed in Section 4.3 is not applicable. Instead, we define Δ_{max} to be the largest $\Delta(z_1, z_2; g)$ that appears in $\mathcal{H}_{0,c}$: $\Delta_{max} = \max_{g, z_1 \geq z_2} \Delta(z_1, z_2; g)$. (4.9) can then be rewritten in an *equivalent* form using Δ_{max}: $\mathcal{H}_{0,c} : \Delta_{max} \leq \Delta_0 \quad vs \quad \mathcal{H}_{\alpha,c} : \Delta_{max} > \Delta_0$. The following lemma says the testing problem (4.9) can be translated into the identification problem on Δ_{max}.

Lemma 4.5. *Suppose the sharp (large sample) lower bound for Δ_{max} is $\Delta_{max,slb}$. A sharp test would reject $\mathcal{H}_{0,c}$ if and only if $\Delta_{max,slb} > \Delta_0$.*

As Δ_{max} is a function of $\boldsymbol{\mu}_{m-1}$, in general, identifying $\Delta_{max,slb}$ involves minimizing Δ_{max} subject to the constraints on $\boldsymbol{\mu}_{m-1}$ imposed by the observed data distribution. Theorem 4.6 below says that the feasible region of $\boldsymbol{\mu}_{m-1}$ is a convex polytope, defined as an intersection of finitely many half spaces. Consequently, this optimization problem can be translated into a linear programming problem and efficiently solved with off-the-shelf software. See Algorithm 1 in Appendix for more details.

Theorem 4.6. *Given the observed data distribution, the feasible region of $\boldsymbol{\mu}_{m-1}$ is a subspace in $[0,1]^{dim(\boldsymbol{\mu}_{m-1})}$ subject to the following constraints:*

$$\sum_{g \in \Omega_z} p_g^z \mu_g^z = m(z), z = 0, \ldots, m-1;$$

$$\max\left(0, m(z) - p_{D^m L}^z\right) \leq \sum_{g \in \Omega_{m-1}} p_g^z \mu_g^z \leq \min\left(1 - p_{D^m L}^z, m(z)\right), z = m,$$

where p_g^z is identifiable from data (see Lemma 1 in Appendix). In particular, the feasible region of $\boldsymbol{\mu}_{m-1}$ is a convex polytope.

To incorporate statistical uncertainty, one can use Bayesian analysis methods to derive a credible interval for $\Delta_{max,slb}$. Specifically, one runs Steps 1-4 in Algorithm 2 to get multiple posterior samples that satisfy the constraint (4.8), and then produces a percentile based credible interval for $\Delta_{max,slb}$ based on the posterior samples. One may also estimate the posterior probability of rejecting $\mathcal{H}_{0,c}$ for any given positive value Δ_0 with these posterior sample draws.

4.5 Marginal credible intervals for a given contrast

If a clinically non-trivial treatment effect is found, then it is desirable to identify the principal strata and treatment arms that correspond to this treatment effect. In this case, the marginal feasible regions and associated credible intervals for $\Delta(z_1, z_2; g)$ are of interest.

If the observed data distribution was known, then the feasible region for $\Delta(z_1, z_2; g)$ can be obtained from the feasible regions for $\mu_g^{z_1}$ and $\mu_g^{z_2}$. Specifically, we have the following theorem.

Theorem 4.7. *Suppose the observed data distribution is known, and $B_{\mathcal{M}(g),z_1}$ and $B_{\mathcal{M}(g),z_2}$ are feasible regions for $\mu_g^{z_1}$ and $\mu_g^{z_2}$, respectively. Then we have the following results.*

1. *For $z = z_1, z_2$, $B_{\mathcal{M}(g),z} = \left[\int y dL_{p_g^z}(Q^z), \int y dU_{p_g^z}(Q^z)\right]$.*

2. *The feasible region of $\Delta(z_1, z_2; g)$ is $\left[\int y dL_{p_g^{z_1}}(Q^{z_1}) - \int y dU_{p_g^{z_2}}(Q^{z_2}), \int y dU_{p_g^{z_1}}(Q^{z_1}) - \int y dL_{p_g^{z_2}}(Q^{z_2})\right]$.*

In practice, credible intervals for $\Delta(z_1, z_2; g)$ can be constructed from posterior sample draws $\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(M)}$. These posterior draws may also be used to estimate the posterior probability of rejecting the null hypothesis $\mathcal{H}_{0,m} : \Delta(z_1, z_2; g) \leq \Delta_0$, where the letter "$m$" in $\mathcal{H}_{0,m}$ is short for "marginal."

Remark 4.8. *We remark that even if the observed data provide evidence for existence of non-zero/non-trivial treatment effects, it is possible that they do not contain information on the specific principal strata and treatment arms that correspond to these treatment effects. Moreover, unlike the case in multi-arm trials without truncation by death, this can happen even with an infinite sample size.*

We illustrate our point with the following numerical example. Consider a three-arm trial such that $\pi_{LLL} = \pi_{DLL} = \pi_{DDL} = 0.3, \pi_{DDD} = 0.1, m(0) = 0.3, m(1) = 0, m(2) = 0.5$, *where* $\pi_g \equiv P(G = g)$. *In this case,* $\mu^0_{LLL} = 0.3$ *and* $\mu^1_{LLL} = \mu^1_{DLL} = 0$. *It follows that* $\Delta_{max} = \max(0, \mu^2_{LLL} - \mu^1_{LLL}, \mu^2_{DLL} - \mu^1_{DLL})$. *We assume that the sample size is infinite so that we know the observed data distribution. Figure 4.2 shows the joint feasible region of* $(\mu^2_{LLL} - \mu^1_{LLL}, \mu^2_{DLL} - \mu^1_{DLL})$ *(the green shaded area). Suppose that the margin of clinical relevance* Δ_0 *is 0.1, then the acceptance region for null hypothesis* $\mathcal{H}_{0,c}$ *is the lower left area of the blue contour line. As there is no intersection between the feasible region of* $(\mu^2_{LLL} - \mu^1_{LLL}, \mu^2_{DLL} - \mu^1_{DLL})$ *and the acceptance region for* $\mathcal{H}_{0,c}$, *one may conclude that* $\mathcal{H}_{0,c}$ *should be rejected. Alternatively, one can see from the contour lines of* Δ_{max} *that the sharp lower bound of* Δ_{max} *is 0.25. As* Δ_0 *is smaller than* $\Delta_{max,slb}$, *one also rejects* $\mathcal{H}_{0,c}$.

However, by projecting the joint feasible region of $(\mu^2_{LLL} - \mu^1_{LLL}, \mu^2_{DLL} - \mu^1_{DLL})$ *onto individual axises, one concludes that the marginal feasible regions for* $\mu^2_{LLL} - \mu^1_{LLL}$ *and* $\mu^2_{DLL} - \mu^1_{DLL}$ *are both* $[0, 1]$. *As both of the marginal feasible regions contain values that are smaller than* Δ_0, *the data contain no information on the specific contrast that corresponds to the overall treatment effect.*

4.6 Data Illustrations

A Application to the HVTN 503 study

The HVTN 503 HIV vaccine study was a randomized, double-blinded, placebo-controlled Phase IIb test-of-concept clinical trial to investigate the efficacy and safety of an experimental HIV vaccine. The same vaccine was also evaluated in a different population in an earlier

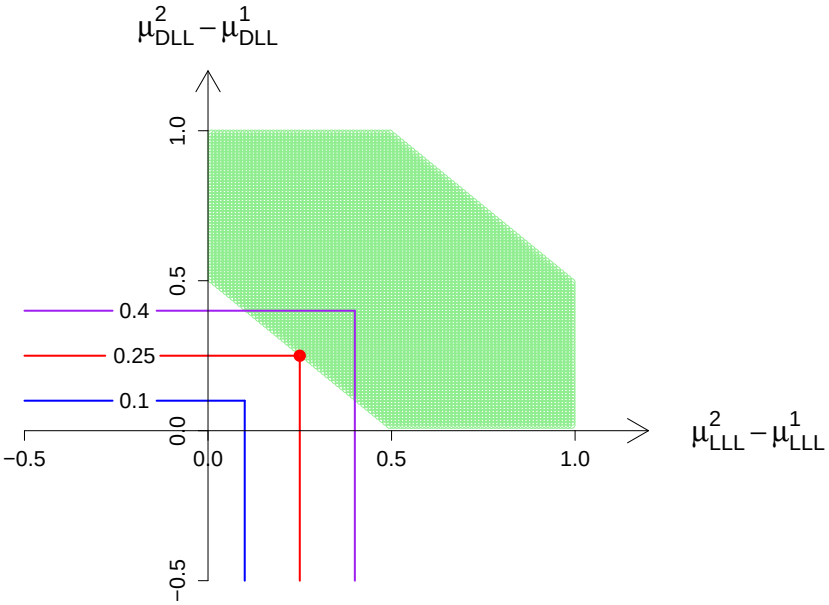

Figure 4.2: Feasible region of $(\mu^2_{LLL} - \mu^1_{LLL}, \mu^2_{DLL} - \mu^1_{DLL})$ (green shaded area). The colored lines are contour lines of Δ_{max}. The sharp lower bound of Δ_{max} is obtained at the red point.

HVTN 502/Step trial. Starting January, 2007, the HVTN 503 study enrolled 800 HIV negative subjects and randomized them to receive three doses of either the study vaccine or a placebo. The ratio of vaccine to placebo assignment was 1:1. Enrollment and vaccinations were halted in September 2007, but follow-up continued, after the HVTN 502/Step trial met its prespecified non-efficacy criteria. Details of this study can be found in Gray et al. (2011, 2014).

In our analysis, we compared CD4 counts among participants within the same principal stratum defined by their full potential infection statuses. Due to the early stopping of vaccinations of the trial, a majority of participants in the HVTN 503 trial were not fully immunized. When enrollment was stopped, 400 participants in the HVTN 503 trial were assigned to the experimental vaccine group. Of them, 112 received one injection, 259 received two injections, and only 29 received all three injections. Hence we considered the dosage of experimental vaccine as the treatment arm Z, where $Z = 0$ for all subjects in the control

group. As the trial was stopped administratively, and the time a participant entered this trial was unlikely to affect the potential outcomes of interest (CD4 count), it is reasonable to assume that the treatment arms were randomized. Furthermore, since there were only 3.6% of participants who received all three experimental vaccines, we code $Z = 2$ for all participants who receive two or more experimental vaccine injections.

A total of 100 subjects were infected during this trial. We define each subject's "median CD4 count" (the outcome of interest) as their median CD4 count measured between their confirmatory HIV testing visit and the end of follow-up or start of antiretroviral treatments. Many epidemiological studies have suggested that the risk of AIDS takes a jump up when CD4 count goes below 350 cells/mm^3, and then another jump up when CD4 count goes below 200 cells/mm^3 (World Health Organization, 2010). Based on these data, 350 cells/mm^3 and 200 cells/mm^3 have been used in previous United States Department of Health and Human Services (DHHS) guidelines for when to start antiretroviral treatment. Thus, we chose 350 and 200 as the dichotomization points for CD4 count. Note that the outcome measure is only measured for infected subjects. As 87.5% of the study subjects were uninfected, an intent-to-treat analysis with imputation for missing CD4 count values is likely to have very low power for detecting any treatment effects (Gilbert et al., 2003). Hence SACEs are of interest for analyzing this trial.

Table 4.3 in Appendix summarizes the observed data for the study participants. There were 7 infected participants who had no CD4 count measurements after their confirmatory HIV testing visit. We made the missing completely at random (MCAR) assumption and left them out of our analysis below. In treatment arm 0, 1, 2, the mean number of CD4 counts available were 5.69, 5.94 and 5.57, respectively; the mean length of time from the confirmatory HIV testing visit to the first CD4 count measure were 26 days, 25 days and 32 days, respectively, and the mean time spacing between CD4 count measurements were 127 days, 146 days and 134 days, respectively.

Presumably there was little interaction among HVTN 503 subjects so that the SUTVA was plausible. Subsequent analyses of the HVTN 502/503 data suggested that although

not possible to directly cause HIV infections itself, the investigational vaccine may increase susceptibility to HIV infection for recipients (Gray et al., 2011, 2014). Given the negative results on the primary efficacy endpoints, members of the HVTN 503 Protocol Team whom we consulted agreed that it is reasonable to make the reverse monotonicity assumption such that experimental vaccine did not help prevent HIV infection for any participant in the study population. The empirical infection rates in the $Z = 0, 1, 2$ arms were $9.25\%, 16.07\%$ and 15.63%, respectively. Thus, the reverse monotonicity assumption seemed acceptable, and we proceeded with our analysis under the reverse monotonicity assumption.

Table 4.1 summarizes the analysis results. The simultaneous testing method estimates the posterior probability of existence of an overall non-zero treatment effect, while the marginal testing method estimates the posterior probability that an overall non-zero treatment effect can be claimed along with the specific treatment arms and principal strata that correspond to this treatment effect. These posterior probabilities were high, suggesting evidence of a non-zero treatment effect on median CD4 falling below 350 or 200 cells/mm^3. The 95% credible intervals for lower bound on Δ_{max} provide information on the magnitude of vaccine effects. For example, results in Tables 4.1 show that there exists at least one basic principal stratum and treatment comparison for which the vaccine reduces the probability of median CD4 count $\leq$ 200 cells/mm^3 by 0.026, but we were not able to ascertain the specific basic principal stratum and treatment comparison that corresponds to this effect. The reason for this is two fold. Firstly, due to non-identifiability of the SACEs, if the effect size is too small, one may fail to identify the specific causal contrast that corresponds to a clinically relevant treatment effect even with an infinite sample size. Secondly, our proposed methods may deliver more conclusive results if the sample size is large enough. For example, if the sample size was 3000 (which was the estimated sample size in the HVTN 503 trial protocol) and the observed frequencies $P(S = 1|Z = z)$ and $P(Y = 1|Z = z, S = 1)$ had remained the same, then the 95% credible interval for the contrast $\mu^2_{LLL} - \mu^0_{LLL}$ would have been $[0.057, 0.186]$, which would imply that compared to the placebo, receiving two or more injections of the experimental vaccine is clinically effective for reducing the possibility of very low CD4 counts

Table 4.1: Posterior probabilities of finding a non-zero overall treatment and posterior credible intervals for lower bounds on Δ_{max} (the maximal treatment effect over all principal strata and treatment comparisons) for the HVTN 503 trial

Methods	Posterior probability of a non-zero treatment effect	95% credible interval for lower bound on Δ_{max}
Outcome:median CD4 $>$ 350		
Simultaneous	0.882	[0.000, 0.346]
Marginal	0.651	[0.000, 0.341]
Outcome:median CD4 $>$ 200		
Simultaneous	0.996	[0.026, 0.260]
Marginal	0.973	$[6 \times 10^{-4}, 0.245]$

(≤ 200 cells/mm^3) among subjects who would get infected regardless of which treatment arm they were assigned to.

We conclude this part with several caveats. First, the median CD4 count is a non-traditional endpoint for HIV vaccine efficacy trials, and it may not be completely comparable between treatment groups due to differences in the number and timing of CD4 measurements. Second, we have made the MCAR assumption for the missing values in CD4 count measures, which is hard to verify for this data set. Third, as pointed out by some authors (e.g. Pearl, 2011), under the principal stratification framework we have taken here, the vaccine effect estimates are only relevant for the subgroup of subjects who would get infected under at least one dosage level, which only constitutes a small fraction of the population. Finally, a reduction of 0.026 in the probability of median CD4 counts ≤ 200 cells/mm^3 may not be considered clinically important given the earlier finding that the vaccine increased HIV acquisition in the study population.

B Application to survey incentive trials

Faced with declining voluntary participation rates, there is now a consensus that incentives are effective for motivating response to surveys (Singer and Kulka, 2002; Singer and

Ye, 2013). There is, however, controversy on how these incentives affect the quality of data collected. Social exchange theory suggests that by establishing an explicit exchange relationship, incentives not only encourage participation in surveys, but also encourage respondents to provide more accurate and complete information (Davern et al., 2003). However, current experimental studies have mixed findings on this hypothesis. Some studies found positive effects of incentives on response quality, while others found non-significant or even negative effects (Singer and Kulka, 2002; Singer and Ye, 2013).

To the best of our knowledge, all these studies directly compare response quality in different incentive groups without accounting for the problem of truncation by response. In these studies, the treatments Z are the levels of incentive, the intermediate outcomes S are the responses to the surveys, and the final outcomes Y are measures of survey quality. Although some researchers realize that people persuaded to participate through the use of incentives will have less internal motivation for filling out the survey thoroughly (e.g. Davern et al., 2003), few, if any, of them separate this group of people in their analyses from those who would participate in the survey regardless of incentive levels, rendering their results subject to selection bias. Furthermore, arguably the response quality is *undefined* for survey non-respondents. Thus as argued by Rubin (2006), the naive comparison has no causal interpretation as it compares different groups of people at baseline. Instead, for two-arm trials, the SACE is of interest as the subgroup whose members would respond regardless of incentive level are the only group for which both of the potential outcomes are well-defined. Similarly for multi-arm trials. Moreover, it is very common that such randomized experiments have multiple incentive groups (Singer and Kulka, 2002; Singer and Ye, 2013). Hence the methodology introduced in this chapter, and more generally, identification and estimation methods for SACEs in multi-arm trials are especially relevant.

For example, Curtin et al. (2007) used data from the Survey of Consumer Attitudes (S-CA) conducted by the University of Michigan Survey Research Center to investigate whether efforts to increase the response rate jeopardize response quality. Their analysis was based on data collected between November 2003 to February 2004, during which time SCA was a

random digit dial telephone survey. In each of the four months, eligible samples were randomly assigned to one of three experimental conditions: advance letter without an incentive, advance letter plus $5 incentive and advance letter plus $10 incentive. The same follow-up procedures, including promised refusal conversion payments are used in all three groups. The measure for response quality in such studies are inevitably subjective; they can be binary (e.g., "mostly compete" vs "partially complete," or whether a particularly important question is answered) or continuous (e.g. percent of missing items). As we don't have access to this data set, below we only discuss the validity of our assumptions.

The SUTVA is reasonable as these are random digit dial samples from the coterminous United States. The monotonicity assumption is also plausible. As argued by survey sampling experts, incentives will motivate response as they compensate for the relative absence of factors that might otherwise stimulate cooperation (Singer and Kulka, 2002), so that individuals who would respond with a lower incentive would also respond if offered a higher incentive. Empirical evidence in this study also supports this assumption: the response rate for the three experimental groups were 51.7%, 63.8% and 67.7% (Curtin et al., 2007). For these and other reasons, hardly any survey methodologists today would question the value of incentives in motivating survey responses (Curtin et al., 2007).

4.7 Discussion

In randomized trials with truncation by death, the average causal effects in basic principal strata are often of interest as they provide causally meaningful and interpretable summaries of the treatment effects. However, for trials with multiple treatment arms, there are usually many such causal contrasts that are of interest to investigators. In this chapter, we consider testing and estimation problems on the basic principal stratum causal effects. Specifically, we propose three scientific questions to understand the overall treatment effect and individual principal stratum causal effects. We then develop novel inference procedures to answer these questions, and show that the proposed procedures have desirable asymptotic properties.

Compared to analyzing a multi-arm trial in a standard setting, the main difficulty in-

troduced by truncation by death is that the causal estimands are not identifiable. In this case, we show that compared to marginal methods, the (ANOVA type) simultaneous inference methods provide more power for testing the overall treatment effect, and the advantage remains even with an infinite sample size. These results demonstrate the importance of addressing both joint and marginal hypotheses in a causal analysis of multi-arm trials with truncation by death. This idea may be applied to analyze multi-arm trials in other settings in which the causal estimands are not identifiable. For example, in multi-arm trials with non-compliance, existing methods consider the causal contrasts separately (Cheng and Small, 2006; Long et al., 2010). Although results obtained with such methods are valid, they are often not informative, especially in the case where there are more than three treatment arms (Long et al., 2010). In this case, a simultaneous inference method may yield a greater posterior probability of claiming an overall treatment effect and the posterior credible intervals are less likely to contain zero.

In analyzing a multi-arm trial with truncation by death, researchers may dichotomize the treatment variable to simplify an analysis, especially in settings where the multi-arm trials consist of a placebo arm and several dosage groups for an active treatment. One such example is the HVTN 503 study, where the treatment groups 1 and 2 can be considered as different versions of the experimental vaccine. However, as noted by Hernán and VanderWeele (2011), results from these analyses may not be generalizable to other population as the causal effect of a compound treatment depends on the distribution of treatment versions in the target population. Moreover, due to the non-identifiability of SACEs, one may fail to find an overall treatment effect that could have been found by applying the proposed simultaneous inference procedure. For example, for the HVTN 503 study, if one were to collapse the active treatment groups into a single compound treatment, then the 95% credible intervals for the SACE corresponding to this compound treatment would be [0.000, 0.253], with which one could not claim any clinically relevant treatment effect.

To account for sampling uncertainty in the observed data distribution, we use Bayesian analysis methods to obtain posterior samples of identifiable quantities $\boldsymbol{p}$. An alternative

Bayesian procedure to our method involves posterior sampling on the mean potential outcomes $\boldsymbol{\mu}_{m-1}$. This alternative approach would directly yield the posterior rejection rate of $\mathcal{H}_0$ and credible intervals for $\Delta_{max,slb}$ without resorting to techniques we have introduced. However, as $\boldsymbol{\mu}_{m-1}$ is not identifiable from the observed data, it turns out that the posterior estimates of Δ_{max} are extremely sensitive to the prior specification on $\boldsymbol{\mu}_{m-1}$. We refer interested readers to Richardson et al. (2011) for a nice discussion of this issue.

The problem we consider here is similar to an instrumental variable analysis in that both problems can be analyzed under the principal stratification framework. When the exposure variable in a instrumental variable analysis is binary, the exclusion restriction assumption is closely related to the null hypothesis in the truncation by death problem, namely the causal effect in the always-survivor group is zero. Hence the approach we develop here may be used to partially test the exclusion restriction assumption of an instrumental variable model.

There are several possible extensions to our framework. For example, we have restricted our attention to binary outcomes in this chapter. We are currently exploring extensions to deal with continuous and categorical outcomes. In addition, covariate information may be employed to sharpen bounds on SACEs. Another possible extension is to introduce sensitivity parameters for better understanding of the causal effects of interest. The tests and bounds we have developed here correspond to extreme results of corresponding sensitivity analyses.

Appendix

A Algorithm for identifying $\Delta_{max,slb}$

See Algorithm 3.

B Simulation studies

We now use a hypothetical example to illustrate the advantage of the simultaneous inference procedures proposed in Section 3 and 4 in the main text for testing the overall treatment effect. Let the comparison method be the approach that considers each $\Delta(z_1, z_2; g)$ separately,

Algorithm 3 An algorithm for identifying $\Delta_{max,slb}$

1. Solve the following linear programming problem:
   minimize α subject to:

$$\sum_{g\in\Omega_z} p_g^z \mu_g^z = m(z), \quad z = 0, \ldots, m-1;$$

$$\max\left(0, m(z) - p_{D^m L}^z\right) \leq \sum_{g\in\Omega_{m-1}} p_g^z \mu_g^z \leq \min\left(1 - p_{D^m L}^z, m(z)\right), \quad z = m;$$

$$\mu_g^{z_1} - \mu_g^{z_2} \leq \alpha, \quad \forall g, z_1 \geq z_2;$$

$$0 \leq \mu_g^z \leq 1, \quad \forall g, z$$

2. Report the value of the linear programming problem above as $\Delta_{max,slb}$

and it accepts or rejects the null based on the marginal feasible regions of $\Delta(z_1, z_2; g)$. With the comparison marginal testing method, one rejects the hypothesis $\mathcal{H}_0$ only if at least one of the marginal feasible regions excludes 0. In other words, the comparison method rejects $\mathcal{H}_0$ if the observed data not only provide evidence for existence of a non-zero treatment effect, but also contain information on the specific principal strata and treatment arms that correspond to this treatment effect. As explained in Remark 3 in the main text, this generally yields a smaller posterior rejection probability. In addition, with the comparison method, one estimates the lower bound on Δ_{max} to be the maximal sharp lower bound for all $\Delta(z_1, z_2; g)$ that appear in equation (1) in the main text. We denote this lower bound as $\Delta_{max,mlb}$, where "mlb" is short for "marginal lower bound." One can see from the numerical example in Remark 3 in the main text that $\Delta_{max,mlb}$ is in general no larger than Δ_{max}. This is because the comparison marginal estimation method does not use information on the dependence among feasible regions of causal contrasts $\Delta(z_1, z_2; g)$. In the simulation studies, we empirically evaluate the difference between the proposed simultaneous inference methods and the comparison marginal inference methods for testing the overall treatment effect.

Suppose that we have a three-arm vaccine trial with two vaccine groups and one placebo

Table 4.2: Observed data counts in a hypothetical example.

Observed subgroup	Counts
$Y = 1, S = 1, Z = 0$	n_1
$Y = 0, S = 1, Z = 0$	$40 - n_1$
$S = 0, Z = 0$	360
$Y = 1, S = 1, Z = 1$	56
$Y = 0, S = 1, Z = 1$	24
$S = 0, Z = 1$	320
$Y = 1, S = 1, Z = 2$	108
$Y = 0, S = 1, Z = 2$	12
$S = 0, Z = 2$	280

group, and there are 400 subjects in each group. The hypothetical data example is listed in Table 4.2, where n_1 is a parameter taking integer values between 0 and 40. The conditional frequencies $m(0)$, $m(1)$ and $m(2)$ in this example are $0.025n_1$, 0.7 and 0.9, respectively. In our example, there are 10% of the study sample in each of the principal strata LLL, DLL, DDL, while the rest belongs to the DDD stratum.

Results in Figure 4.3 show that for some values of n_1, the simultaneous and marginal methods compared here yielded similar results. However, in some other cases, the results could be very different. For example, when $n_1 = 36$, the simultaneous testing method estimated the posterior probability of rejecting $\mathcal{H}_0$ to be 98.8%, compared to an estimate of 4.0% from the marginal testing method. When $n_1 = 20$, the simultaneous estimation method estimated the 95% credible interval for $\Delta_{max,slb}$ to be $[0.029, 0.404]$, based on which one was able to claim a clinically relevant treatment effect at margin $\Delta_0 = 0.02$. The marginal estimation method, however, estimated the 95% credible interval for $\Delta_{max,mlb}$ to be $[4 \times 10^{-4}, 0.363]$, with which one failed to claim a clinically relevant treatment effect at the same margin.

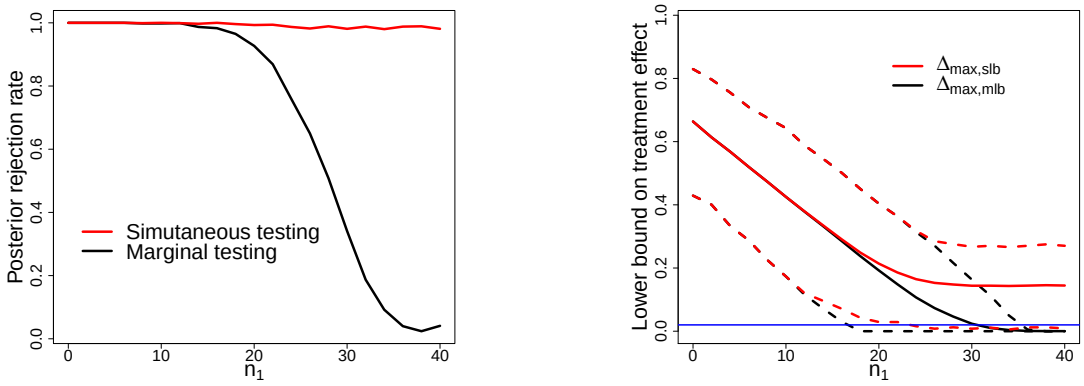

Figure 4.3: Results from analyzing the hypothetical data set in Table 4.2. The left panel shows the posterior probability of rejecting $\mathcal{H}_0$ using the proposed simultaneous testing method and the comparison marginal testing method. The right panel shows the posterior mean (solid lines) and 95% credible intervals (dashed lines) for lower bounds on Δ_{max}, the maximal treatment effect among all possible basic principal strata and treatment comparisons. The red curves correspond to sharp lower bounds obtained using the proposed simultaneous estimation method, and the black curves correspond to lower bounds obtained using the comparison marginal estimation method. The blue horizontal line corresponds to a clinically meaningful margin of 0.02.

Table 4.3: Observed data counts in the HVTN 503 trial. Z denotes the treatment arm, S denotes the infection status, and Y is the dichotomized outcome of CD4 count. $Y = *$ denotes that Y is missing.

Observed subgroup	median CD4 $>$ 350 cells/mm^3	median CD4 $>$ 200 cells/mm^3
$Y = 1, S = 1, Z = 0$	19	29
$Y = 0, S = 1, Z = 0$	14	4
$Y = *, S = 1, Z = 0$	4	4
$S = 0, Z = 0$	363	363
$Y = 1, S = 1, Z = 1$	12	16
$Y = 0, S = 1, Z = 1$	4	0
$Y = *, S = 1, Z = 1$	2	2
$S = 0, Z = 1$	94	94
$Y = 1, S = 1, Z = 2$	34	44
$Y = 0, S = 1, Z = 2$	10	0
$Y = *, S = 1, Z = 2$	1	1
$S = 0, Z = 2$	243	243

C Data Table for the HVTN 503 study

Table 4.3 gives the observed data counts for the HVTN 503 trial.

D Proofs of theorems and lemmas

Proof of Theorem 1

The proof for the general multi-arm case is very similar to the discussion for the three-arm case. The only non-trivial generalization is for Step 3 of Algorithm 1 in the main text. Instead of checking the pairwise intersections of $(B_{kz}; z = k, \ldots, m)$, we check their joint intersection. This relies on the observation that if we let $g = D^k L^{m+1-k}$, then $\omega_g^k = 1$ and the feasible region for B_{kk} is a one point set $\{\int y dQ_g^k\}$. Consequently,

$$\bigcap_{z=k,\ldots,m} B_{kz} \neq \emptyset \tag{4.10}$$

implies that

$$B_{kz_1} \cap B_{kz_2} \neq \emptyset, \forall z_1 > z_2 \geq k. \tag{4.11}$$

Note there are only $m-k$ pairs of comparisons involved in (4.10), compared to $(m+1-k)(m-k)/2$ pairs of comparisons in (4.11).

Proof of Theorem 2

To prove Theorem 2, we note that the assumptions of Theorem 2 and the observed data distribution impose the following constraints on μ_g^z:

$$Q^z = \sum_{\underline{g} \in \Omega_{l-1}} p_{\underline{g}}^z Q_{\underline{g}}^z + p_g^z Q_g^z + \sum_{\bar{g} \in \Omega_z \setminus \Omega_l} p_{\bar{g}}^z Q_{\bar{g}}^z, \tag{4.12}$$

$$\mu_{\underline{g}}^{\mathcal{M}(\underline{g})} = \cdots = \mu_{\underline{g}}^m . \forall \underline{g} \in \Omega_{l-1}, \tag{4.13}$$

where Q^z denotes the distribution of outcome Y in treatment arm z. To simplify (4.12) and (4.13), we use the following lemmas, which say that both the proportions of basic principal strata p_g^z and the means of Bernoulli distributions $(Q_{\underline{g}}^z, \underline{g} \in \Omega_{l-1}, z \geq \mathcal{M}(\underline{g}))$ are identifiable. Proofs of these lemmas are left to the end of this subsection.

Lemma 4.9. *The proportions of basic principal strata, namely* $(p_g^z; g \in \Omega_{m-1}, z \geq \mathcal{M}(g))$ *are identifiable from the observed data.*

Lemma 4.10. *Suppose that (6) in the main text holds for all* $k < l$*, then* $(\mu_{\underline{g}}^z; \underline{g} \in \Omega_{l-1}, z \geq \mathcal{M}(g))$ *are identifiable from the observed data.*

As the Bernoulli distribution $Q_{\underline{g}}^z$ is uniquely determined by its mean $\mu_{\underline{g}}^z$, the constraints on μ_g^z can be simplified as

$$Q_{\boldsymbol{g}}^z = \omega_g^z Q_g^z + \sum_{\bar{g} \in \Omega_z \setminus \Omega_l} \omega_{\bar{g}}^z Q_{\bar{g}}^z, \tag{4.14}$$

where $Q_{\boldsymbol{g}}^z$ a Bernoulli distribution with mean $m_{\boldsymbol{g}}(z)$. Applying Imai (2008)'s results to

(4.14), we have

$$B_{lz} = \left[\int y dL_{\omega_g^z}(Q_{\boldsymbol{g}}^z), \int y dU_{\omega_g^z}(Q_{\boldsymbol{g}}^z) \right].$$

This completes the proof of Theorem 2. □

Proof of Lemma 4.9

Proof. Let $\pi_g^z = P(G = g|Z = z)$. Following Assumption 2, π_g^z is independent of treatment arm z and hence can be written as π_g. Under Assumption 3, we have the following equations:

$$\begin{aligned}
P(S = 1|Z = 0) &= \pi_{L^{m+1}}, \\
P(S = 1|Z = 1) &= \pi_{L^{m+1}} + \pi_{DL^m}, \\
&\cdots \\
P(S = 1|Z = z) &= \pi_{L^{m+1}} + \ldots + \pi_{D^z L^{m+1-z}}, \\
&\cdots \\
P(S = 1|Z = m) &= \pi_{L^{m+1}} + \ldots + \pi_{D^m L}, \\
1 &= \pi_{L^{m+1}} + \ldots + \pi_{D_{m+1}}.
\end{aligned} \tag{4.15}$$

It can be shown that there exists an unique solution to equation (4.15) and hence $(\pi_g, g \in \Omega_m)$ are identifiable from equation (4.15). It then follows that $(p_g^z; g \in \Omega_{m-1}, z \geq \mathcal{M}(g))$ are also identifiable. □

Proof of Lemma 4.10

Proof. As (6) in the main text holds for all $k < l$, we only need to show that $\mu_{\underline{g}}^{\mathcal{M}(\underline{g})}$ is identifiable from the observed data. We show this by applying the induction method on $\mathcal{M}(\underline{g})$.

Base case: if $\mathcal{M}(\underline{g}) = 0$, then $\mu_{\underline{g}}^{\mathcal{M}(\underline{g})} = \mu_{L^{m+1}}^0 = P(Y = 1|Z = 0, S = 1)$ by the monotonicity assumption (Assumption 3).

Inductive step: suppose that $\mu_{\underline{g}}^{\mathcal{M}(\underline{g})}$ is identifiable from the observed data for all principle strata $\underline{g}$ such that $\mathcal{M}(\underline{g}) \leq k$. Following the monotonicity assumption (Assumption 3), we have the following identify:

$$\begin{aligned} P(Y=1|Z=k+1,S=1) &= \sum_{\underline{g}\in\Omega_k} p_{\underline{g}}^k \mu_{\underline{g}}^k + p_{D^{k+1}L^{m-k}}^{k+1} \mu_{D^{k+1}L^{m-k}}^{k+1} \\ &= \sum_{\underline{g}\in\Omega_k} p_{\underline{g}}^k \mu_{\underline{g}}^{\mathcal{M}(\underline{g})} + p_{D^{k+1}L^{m-k}}^{k+1} \mu_{D^{k+1}L^{m-k}}^{k+1}, \end{aligned} \tag{4.16}$$

where the last step in (4.16) follows from the working hypotheses.

Following Lemma 4.9, $(p_{\underline{g}}^k; \underline{g} \in \Omega_k)$ and $p_{D^{k+1}L^{m-k}}^{k+1}$ are identifiable from the observed data. Following the induction hypotheses, $(\mu_{\underline{g}}^{\mathcal{M}(g)}; \underline{g} \in \Omega_k)$ are also identifiable. Consequently, $\mu_{D^{k+1}L^{m-k}}^{k+1}$ is identifiable from (4.16). In other words, for principle strata $\underline{g}$ such that $\mathcal{M}(\underline{g}) = k+1$, $\mu_{\underline{g}}^{\mathcal{M}(\underline{g})}$ is also identifiable from the observed data.

By the induction principle, we have finished our proof. □

Proof of Theorem 4

Theorem 4 is a direct consequence of the following lemma:

Lemma 4.11. *Let h be a mixture of k Bernoulli distributions $f_1, \ldots, f_k$: $h = \sum_{j=1}^{k} \alpha_j f_j$, where the mixing proportions $\alpha_j, j = 1, \ldots, k$ are known. Let $P, P_1, \ldots, P_k$ be the probability of a positive outcome under $h, f_1, \ldots, f_k$ respectively, then*

$$\max\left(0, P - \sum_{j=l+1}^{k} \alpha_j\right) \leq \sum_{j=1}^{l} \alpha_j P_j \leq \min\left(\sum_{j=1}^{l} \alpha_j, P\right).$$

Lemma 4.11 is a generalization of Lemma 1 in Cheng and Small (2006) and can be proved by solving the linear programming problem of minimizing or maximizing $\sum_{j=1}^{l} \alpha_j f_j$ subject to constraints $P = \sum_{j=1}^{k} \alpha_j P_j$. □

REFERENCES

Andersson, S. A., Madigan, D., Perlman, M. D., et al. (1997). A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Anstey, K. J., von Sanden, C., Salim, A., and O'Kearney, R. (2007). Smoking as a risk factor for dementia and cognitive decline: a meta-analysis of prospective studies. *American Journal of Epidemiology*, 166(4):367–378.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2015). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836.

Chiba, Y. and VanderWeele, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology*, 173(7):745–751.

Cox, B., Blaxter, M., Buckle, A., Fenner, N., Golding, J., Gore, M., Huppert, F., Nickson, J., Roth, S. M., Stark, J., et al. (1987). *The Health and Lifestyle Survey. Preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults.* Health Promotion Research Trust.

Curtin, R., Singer, E., and Presser, S. (2007). Incentives in random digit dial telephone surveys: A replication and extension. *Journal of Official Statistics*, 23(1):91–105.

Davern, M., Rockwood, T. H., Sherrod, R., and Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment. *Public Opinion Quarterly*, 67(1):139–147.

Ding, P., Geng, Z., Yan, W., and Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106(496):1578–1591.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236.

Egleston, B. L., Scharfstein, D. O., Freeman, E. E., and West, S. K. (2007). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, 8(3):526–545.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Gelman, A. and Meng, X.-L. (2004). *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. John Wiley & Sons.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, 59(3):531–541.

Gray, G. E., Allen, M., Moodie, Z., Churchyard, G., Bekker, L.-G., Nchabeleng, M., Mlisana, K., Metch, B., de Bruyn, G., Latka, M. H., et al. (2011). Safety and efficacy of the HVTN 503/Phambili study of a clade-B-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. *The Lancet infectious diseases*, 11(7):507–515.

Gray, G. E., Moodie, Z., Metch, B., Gilbert, P. B., Bekker, L.-G., Churchyard, G., Nchabeleng, M., Mlisana, K., Laher, F., Roux, S., et al. (2014). Recombinant adenovirus type 5 HIV gag/pol/nef vaccine in South Africa: unblinded, long-term follow-up of the phase 2b HVTN 503/Phambili study. *The Lancet infectious diseases*, 14(5):388–396.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hájek, J. (1971). Comment on an essay on the logical foundations of survey sampling by Basu, D. *Foundations of Statistical Inference*, 236.

Hattori, S. and Henmi, M. (2014). Stratified doubly robust estimators for the average causal effect. *Biometrics*, 70(2):270–277.

Hayden, D., Pauler, D. K., and Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, 61(1):305–310.

Hernán, M. A., Alonso, A., and Logroscino, G. (2008). Cigarette smoking and dementia: potential selection bias in the elderly. *Epidemiology*, 19(3):448–450.

Hernán, M. A. and VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, 22(3):368–377.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Horowitz, J. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003). On the analysis of viral load endpoints in hiv vaccine trials. *Statistics in Medicine*, 22(14):2281–2298.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with truncation-by-death. *Statistics & Probability Letters*, 78(2):144–149.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 171(2):481–502.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kalbfleisch, J. and Prentice, R. (1980). The statistical analysis of failure time data. *Wiley series in probability and mathematical statistics*.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

Lee, K., Daniels, M. J., and Sargent, D. J. (2010). Causal effects of treatments for informative missing data due to progression/death. *Journal of the American Statistical Association*, 105(491).

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Long, Q., Little, R. J., and Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3):513–531.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

Newey, W. K. (1994). Series estimation of regression functionals. *Econometric Theory*, 10(01):1–28.

Newey, W. K., Hsieh, F., and Robins, J. (1998). Undersmoothing and bias corrected functional estimation. *Working Paper, MIT*.

Nolen, T. L. and Hudgens, M. G. (2011). Randomization-based inference within principal strata. *Journal of the American Statistical Association*, 106(494).

Paninski, L. and Yajima, M. (2008). Undersmoothed kernel entropy estimators. *Information Theory, IEEE Transactions on*, 54(9):4384–4388.

Pearl, J. (2011). Principal stratification–a goal or a tool? *The International Journal of Biostatistics*, 7(1):1–15.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Peters, R., Poulter, R., Warner, J., Beckett, N., Burch, L., and Bulpitt, C. (2008). Smoking, dementia and cognitive decline in the elderly, a systematic review. *BMC Geriatrics*, 8(1):36.

Petrylak, D. P., Tangen, C. M., Hussain, M. H., Lara Jr, P. N., Jones, J. A., Taplin, M. E., Burch, P. A., Berry, D., Moinpour, C., Kohli, M., et al. (2004). Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *New England Journal of Medicine*, 351(15):1513–1520.

Richardson, T. S., Evans, R. J., and Robins, J. M. (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J., Rotnitzky, A., and Vansteelandt, S. (2007a). Discussion of *Principal stratification designs to estimate input data missing due to death* by C.E. Frangakis, D. B. Rubin, M.-W. An & E. MacKenzie. *Biometrics*, 63(3):650–653.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007b). Comment on "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". *Statistical Science*, 22(4):544–559.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A: General*, 147:656–666.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika*, 95(1):248–252.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rotnitzky, A. and Vansteelandt, S. (2014). Double-robust methods. In Fitzmaurice, G., Kenward, M., Molenberghs, G., Tsiatis, A., and Verbeke, G., editors, *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Press.

Roy, J., Hogan, J. W., and Marcus, B. H. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, 9(2):277–289.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Rubin, D. B. (1980). Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (2000). Comment on "causal inference without counterfactuals". *Journal of the American Statistical Association*, 95(450):435–438.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science*, 21(3):299–309.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

Satcher, D., Thompson, T., and Koplan, J. P. (2002). Women and smoking: a report of the surgeon general. *Nicotine & Tobacco Research*, 4(1):7–20.

Schulz, K. F. and Grimes, D. A. (2005). Multiplicity in randomised trials I: endpoints and treatments. *The Lancet*, 365(9470):1591–1595.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*, 62(2):332–342.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. John Wiley.

Singer, E. and Kulka, R. A. (2002). Paying respondents for survey participation. In Ploeg, M. V., Moffitt, R. A., and Citro, C. F., editors, *Studies of welfare populations: Data collection and research issues*, pages 105–128. Washington, DC: National Academy Press.

Singer, E. and Ye, C. (2013). The use and effects of incentives in surveys. *The Annals of the American Academy of Political and Social Science*, 645(1):112–141.

Stallings, V. A., Suitor, C. W., Taylor, C. L., et al. (2010). *School meals: building blocks for healthy children*. National Academies Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Tchetgen Tchetgen, E. J. (2014). Identification and estimation of survivor average causal effects. *Statistics in Medicine*, 33(21):3601–3628.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75.

Whittington, J. E. and Huppert, F. A. (1997). Smoking and cognitive decline. *Human Psychopharmacology: Clinical and Experimental*, 12(5):467–480.

Williamson, E., Morley, R., Lucas, A., and Carpenter, J. (2012). Variance estimation for stratified propensity score estimators. *Statistics in medicine*, 31(15):1617–1632.

Woo Baidal, J. A. and Taveras, E. M. (2014). Protecting progress against childhood obesity – the national school lunch program. *New England Journal of Medicine*, 371(20):1862–1865.

World Health Organization (2010). Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach-2010 revision.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhao, Q. and Percival, D. (2015). Double robustness for causal effects via entropy balancing. *arXiv preprint arXiv:1501.03571*.

Znidaric, M. (2005). Asymptotic expansion for inverse moments of binomial and poisson distributions. *arXiv preprint math/0511226*.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, (just-accepted).

VITA

Linbo Wang was born in Xiapu, Fujian, CHINA in August 1991, where he spent the first 13 years of his life. He graduated from Fuzhou No.1 High School in 2007 and attended Peking University afterwards. After receiving his bachelor's degree in statistics in 2011, he continued his study in the Department of Biostatistics at University of Washington. He completed his Ph.D. in Biostatistics in March 2016 under the direction of Professor Xiao-Hua Zhou and Professor Thomas Richardson.