Semiparametric Estimation in Observational Studies and

Randomized Trials

by

Tianchen Qian

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2017

© Tianchen Qian 2017

All rights reserved
Abstract

Researchers often seek robust inference for a parameter through semiparametric estimation. Semiparametric estimation is useful, for example, for survival analysis, for estimating growth parameters in longitudinal studies, and for estimating quantities under missing data, including treatment effects based on potential outcomes. In this dissertation we study semiparametric estimation from two design perspectives: observational studies and randomized trials.

Traditional semiparametric estimation methods requires theoretical derivation of the efficient influence function, which can be a challenging and time-consuming task. To address this difficulty, we propose a new method, called “deductive estimation”, for constructing semiparametric estimators for observational studies. The method is computerizable, meaning that it does not need theoretically deriving the functional form of the efficient influence function, and is guaranteed to produce semiparametric, locally efficient estimators even for complex parameters in nonparametric models. We apply the method to two designs: the two-phase design, and the double-sampling design. We demonstrate the method with a study
ABSTRACT

on asthma care satisfaction and a study evaluating an HIV treatment in East Africa.

In randomized trials, adjusting for baseline variables and short-term outcomes can lead to increased power and reduced sample size. We investigate the strengths and limitations of a semiparametric, locally efficient estimator compared to the standard unadjusted estimator in randomized trials context. We derive formulas for the precision gain from such covariate adjustment using semiparametric estimators for the average treatment effect, and give new results on what conditions lead to substantial power gains and sample size reductions. The theory is supported by two simulation studies: simulated group sequential trials based on data from the MISTIE Phase II trial, which is a trial of a new surgical intervention for stroke, and simulated adaptive enrichment trials based on data from the Alzheimer’s Disease Neuroimaging Initiative cohort study. Our results can be used in trial planning to predict the potential precision gain from covariate adjustment, which will translate to power gain or sample size reduction.

Advisors: Constantine Frangakis, Michael Rosenblum

Committee: Nicholas Ialongo, Peng Huang

Alternates: Elizabeth Colantuoni, John Jackson
Acknowledgments

I would like to thank my advisors, Professor Constantine Frangakis and Professor Michael Rosenblum. You introduced to me the beauty, rigor, and power of biostatistics. Your advising and support helped me to grow from a amateur to a junior researcher. I aim to follow your footsteps and become a scholar with a complete profile in research, teaching, and service.

Constantine, I still remember the time in my first year, when I entered your office and asked for some papers to read. You stared outside the window for five minutes, then told me the exact title, author, and publication year of five papers. I also remember the afternoons we spent discussing on likelihood principle, and other more philosophical topics. I appreciate your wisdom, and am grateful that I get to take a sip from it during these five years.

Michael, when I look back at all the emails we sent back and forth during the years, I realize how patient and kind you are to a junior student, to someone who barely knew simulations and clinical trials. Apart from the academic things, I have also learned a lot from you on other aspects — how to talk to people, how to give
ACKNOWLEDGMENTS

presentations, and what it is like to be a researcher who is also a leader.

I wanted to thank my thesis committee, Professor Nicholas Ialongo and Professor Peng Huang. I also wanted to thank Professor Albert Wu and Professor Martin Lindquist, who were on my preliminary oral exam committee. Your suggestions have helped significantly improve my research work.

Among the many faculty members in the department who I am grateful for, I especially wanted to thank Professors Mei-Cheng Wang, Brian Caffo, Dan Scharfstein, and Hongkai Ji. You have helped and inspired me from many aspects during my graduate study.

I am lucky to have collaborated with many talented researchers and fellow classmates. Each project was a fruitful and rewarding experience, and I learned a lot from each of you — Elizabeth Colantuoni, Aaron Fisher, Yu Du, Ivan Diaz, Huitong Qiu, and Constantin Yiannoutsos.

I wanted to thank Fang Han, who opened the door to theoretical statistics for me. During our collaboration, the bar you set on learning and writing was high and motivating. I cherish the days we spent at your office, doing proofs on the board.

I wanted to thank my classmates Yuxin Zhu and Sheng Xu, who has helped me many times when I was struggling with math derivations.

Special thanks to Oliver Chen, who has encouraged and helped me on pursuing my academic career. I will always cherish the long emails you wrote to me, and
ACKNOWLEDGMENTS

the jokes we had in the office. I will continue to work hard, and when I sometimes
lost faith, I know I can always read the letters to regain confidence.

I wanted to thank Mary Joy Argo, who helped my from every aspects and made
things so easy for my graduate study. Also thank you for providing writing sug-
gestions on my draft papers.

To my wife, Manxi Yang. There are no words to describe how grateful I am to
be with you. Thank you for being supportive for every decision I make. I cannot
make it here without you.

To my parents, Jinwu Qian and Yanan Zhang. You have planted seed of academia
in my early childhood, and now it has grown to a small tree. I will keep going.

To my parents-in-law, Riqiang Yang and Lihua Liu, for your support, under-
standing, and encouragement.
Contents

Abstract ii

Acknowledgments iv

List of Tables xii

List of Figures xiii

1 Introduction 1

2 Semiparametric Estimation in Observational Studies: Deductive Estimation 6

2.1 Background ......................................................... 7

2.2 Motivation for Deductive Estimation ................................. 9

2.2.1 The Goal of a Deductive Method ................................. 9

2.2.2 Conjecturing and Functional Form as Barriers Towards a Deductive Method ................................. 10
## CONTENTS

2.3 Deductive Estimation Method for Two-phase Design .......................... 13
2.4 Extensions to More Complex Settings ........................................... 21
2.5 Remarks ...................................................................................... 24
2.6 Appendix: Algorithm Examples ..................................................... 25
   2.6.1 Estimating Mean in Two-phase Design .................................. 25
   2.6.2 Estimating Median in Two-phase Design ................................. 28

3 Extension of Deductive Estimation to Double-sampling Design .......... 32
   3.1 Background .............................................................................. 33
   3.2 Double-sampling Design and Quantity of Interest ...................... 36
   3.3 Deductive Estimation for Double-sampling Design ...................... 40
      3.3.1 General Method ................................................................. 40
      3.3.2 Specific Deductive Estimation Procedure for Double-sampling
           Design .................................................................................. 42
   3.4 Application to PEPFAR study ..................................................... 46
   3.5 Simulation Study ........................................................................ 51
   3.6 Remarks ...................................................................................... 55

4 Semiparametric Estimation in Randomized Trials: Covariate Adjustment 57
   4.1 Background .............................................................................. 58
   4.2 Notation and Assumptions .......................................................... 62
   4.3 Precision Gain When Estimating Mean Primary Outcome in One Arm 65
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Deductive estimation on asthma study</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Illustration of the Discrete Support for double-sampling designs</td>
<td>45</td>
</tr>
<tr>
<td>3.2</td>
<td>Deductive estimation with restricted double-sampling rules</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary of simulated data set</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>Simulation results for deductive estimation</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Variables in MISTIE-II dataset</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Empirical $R^2$-squared for the data generating distribution based on MISTIE-II</td>
<td>84</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation results: relative efficiency under null hypothesis</td>
<td>86</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation results: Type I error, power, and expected sample size.</td>
<td>88</td>
</tr>
<tr>
<td>4.5</td>
<td>Relative efficiency from the data generating distribution based on MISTIE-II</td>
<td>96</td>
</tr>
<tr>
<td>4.6</td>
<td>Analysis time and sample size at each interim and decision analysis for group sequential designs with $n_{\text{max}} = 480$ and $n_{\text{max}} = 300$</td>
<td>97</td>
</tr>
<tr>
<td>4.7</td>
<td>Error spending and boundaries for designs in Section 4.6.4 under prognostic setting $\text{progn}_{W,L}$</td>
<td>98</td>
</tr>
<tr>
<td>4.8</td>
<td>Simulation results: relative efficiency under alternative hypothesis</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of setups for 5 simulation studies</td>
<td>129</td>
</tr>
<tr>
<td>5.2</td>
<td>Adaptive enrichment design, and efficacy boundaries under default simulation scenario</td>
<td>133</td>
</tr>
<tr>
<td>5.3</td>
<td>Calendar time to conduct interim analysis for unadjusted and adjusted estimators under default simulation scenario</td>
<td>135</td>
</tr>
<tr>
<td>5.4</td>
<td>Type I error / power for two estimators under default simulation scenario</td>
<td>136</td>
</tr>
<tr>
<td>5.5</td>
<td>Fitted $\alpha$, $\beta$ and $\sigma^*$’s from the ADNI study data</td>
<td>151</td>
</tr>
<tr>
<td>5.6</td>
<td>Coefficients from linear regression fits using ADNI data</td>
<td>151</td>
</tr>
</tbody>
</table>
List of Figures

3.1 Characteristics for different patient types in a double-sampling design. 37
3.2 Estimated mortality rate with 95% pointwise confidence interval (shaded band). (a) Comparison of deductive estimates (with $\hat{\alpha}$ solving (3.6)) and the estimates when forcing $\hat{\alpha} = 1$, where there is no restriction on double-sampling. (b) Comparison of deductive estimates with different restrictions on double-sampling. 49
4.1 Relative efficiency when estimating mean outcome in one arm. 70
4.2 Contour plot comparing sample size reduction from adjusting for baseline variable and adjusting for short-term outcome. 72
5.1 Impact of prognostic value on trial performance. 138
5.2 Impact of delay to primary outcome on trial performance. 141
5.3 Impact of delays to short-term outcome and primary outcome on trial performance. 142
5.4 Impact of accrual rate on trial performance. 143
5.5 Information accrual rates and box-plots of estimated variance for the adjusted and unadjusted estimators under the default simulation scenario. 147
5.6 Change in power under simulation studies 1-5 using ADNI data. 156
Chapter 1

Introduction

Semiparametric models often reflect the level of available scientific knowledge more appropriately compared to parametric models. Semiparametric estimation is useful, for example, for survival analysis (Cox, 1972), for estimating growth parameters in longitudinal studies (Liang and Zeger, 1986), and for estimating quantities under missing data (Robins et al., 1994), including treatment effects based on potential outcomes (Davidian et al., 2005; Crump et al., 2009). In this dissertation we study semiparametric estimation from two design perspectives. In Chapters 2 and 3 we develop a new semiparametric estimation method for complex observational studies. In Chapters 4 and 5 we investigate the strengths and limitations of a semiparametric, locally efficient estimator compared to the standard unadjusted estimator in randomized trials.

Researchers often seek robust inference for a parameter in observational stud-
CHAPTER 1. INTRODUCTION

ies through semiparametric estimation. However, traditional semiparametric estimation methods require theoretical derivation of the efficient influence function, which can be a challenging and time-consuming task (Carone et al., 2016). To address this difficulty and make semiparametric estimation more accessible to statistical practitioners, in Chapter 2 we introduce a new method, called “deductive estimation”, for constructing semiparametric estimators. The proposed new method makes the semiparametric estimation procedure a computerizable algorithm, by utilizing numerical approximation of the efficient influence function and the concept of compatibility. It is computerizable in the sense that the method does not need theoretically deriving the functional form of the efficient influence function, and is guaranteed to produce semiparametric, locally efficient estimators even for complex parameters in nonparametric models. In Chapter 2 we formulate the general form of the algorithm, and demonstrate the method with a study on asthma as an example. We provide detailed algorithms to do semiparametric estimation for mean and median in a two-phase design.

In Chapter 3, we extend the deductive estimation method to a more complex design: the double-sampling design. Studies with long follow-up often suffer from high dropout rate. Informative dropouts (those whose dropout status depends on the outcome of interest) can make results from complete-case analysis biased. To handle such informative dropouts, double-sampling designs allocate additional resources to pursue a sample of the dropouts and find out their outcomes (An
et al., 2009). We proposed a variation of the deductive estimation method that is tailored to the specific data structure in double-sampling designs. The method involves constructing a discrete support from the observed data, which originates from the idea in Chamberlain (1987a) to approximate continuous variables with discrete ones. The discrete support makes it easier to conduct deductive estimation even for complex designs. We illustrate the estimation procedure to estimate survival probability in a double-sampling design, and demonstrate the method with a study evaluating an HIV treatment in East Africa. We also demonstrate the robustness property of the method through simulation studies: the estimator is consistent even with wrong models for survival time, as long as we have correct models for selection into the double-sampling stage. The latter model can be under control of the investigator and hence more likely to be correct.

Semiparametric estimators are also commonly used in randomized trials for covariate adjustment, because they are more robust to model misspecification than estimators based on parametric regression models (Robins and Rotnitzky, 1992; Hubbard et al., 2000; van der Laan and Robins, 2003; Stitelman et al., 2012). However, limited research has been done on quantifying the magnitude of the resulting precision and its relationship to the design characteristics. In Chapter 4 we develop theory to examine the advantages and limitations of using semiparametric estimators for covariate adjustment in group sequential designs. We focus on designs where the primary outcome is measured with a fixed the delay after enroll-
CHAPTER 1. INTRODUCTION

tment, and where there are correlated baseline variables and short-term outcomes measured prior to observing the primary outcome. We derive formulas for the precision gain (measured by the asymptotic relative efficiency against the unadjusted estimator) from adjusting for baseline variables and short-term outcomes using semiparametric estimators in randomized trials. Key components of the formulas include the variance explained by baseline variables, the residual variance explained by short-term outcomes, the proportion of pipeline participants, and the degree of treatment effect heterogeneity. The formulas can be used in trial planning to predict the potential precision gain from covariate adjustment, which will translate to reduction in sample size. The theory is supported by simulated group sequential trials based on data from the MISTIE Phase II trial, a trial of a new surgical intervention.

To further examine the sensitivity of trial performance to different trial characteristics, in Chapter 5 we conduct a comprehensive simulation study for an adaptive enrichment trial. Adaptive enrichment designs involve rules for restricting enrollment to a subset of the population during the course of an ongoing trial (Jennison and Turnbull, 2007; Stallard, 2011; Stallard et al., 2014). This can be used to target those who benefit from the experimental treatment. Through simulation studies, we assess the sensitive of the trial performance (Type I error, power, expected sample size, trial duration) is to different design characteristics. Our simulated datasets mimic features of data from the Alzheimer’s Disease Neuroimaging
CHAPTER 1. INTRODUCTION

Initiative, and involve two subpopulations of interest based on a genetic marker. We investigate the impact of the following design characteristics: the accrual rate, the delay time between enrollment and observation of the primary outcome, and the prognostic value of baseline variables and short-term outcomes. In the simulation, we use a semiparametric, locally efficient estimator, and investigate its strengths and limitations compared to standard unadjusted estimators. The simulation results are consistent with the theory presented in Chapter 4.

Chapter 6 concludes the dissertation with discussions and areas of future research.

Chapter 2 is adapted from the article “Constantine E. Frangakis, Tianchen Qian, Zhenke Wu, Ivan Diaz (2015). Deductive Derivation and Turing-computerization of Semiparametric Efficient Estimation. Biometrics, 71(4), 867-874.” Chapter 3 is adapted from the working paper “Tianchen Qian, Constantine E. Frangakis, Constantin Yiannoutsos. Deductive Semiparametric Estimation using Discrete Support: with Application to Double-sampling Designs.” Chapter 4 is adapted from the working paper “Tianchen Qian, Michael Rosenblum, Huitong Qiu. Improving Power in Group Sequential, Randomized Trials by Adjusting for Prognostic Baseline Variables and Short-Term Outcomes.” Chapter 5 is adapted from the working paper “Tianchen Qian, Elizabeth Colantuoni, Aaron Fisher, Michael Rosenblum. Sensitivity of Trial Performance to Delayed Outcomes, Accrual Rates, and Prognostic Variables Based on a Simulated Randomized Trial with Adaptive Enrichment.”
Chapter 2

Semiparametric Estimation in Observational Studies: Deductive Estimation

SUMMARY. Researchers often seek robust inference for a parameter through semiparametric estimation. Efficient semiparametric estimation currently requires theoretical derivation of the efficient influence function (EIF), which can be a challenging and time-consuming task. If this task can be computerized, it can save dramatic human effort, which can be transferred, for example, to the design of new studies. Although the EIF is, in principle, a derivative, simple numerical differ-

---

1This chapter is adapted from the following article: Constantine E. Frangakis, Tianchen Qian, Zhenke Wu, Ivan Diaz (2015). Deductive Derivation and Turing-computerization of Semiparametric Efficient Estimation. *Biometrics*, 71(4), 867-874.
CHAPTER 2. DEDUCTIVE ESTIMATION

tiation to calculate the EIF by a computer masks the EIF’s functional dependence on the parameter of interest. For this reason, the standard approach to obtaining the EIF relies on the theoretical construction of the space of scores under all possible parametric submodels. This process currently depends on the correctness of conjectures about these spaces, and the correct verification of such conjectures. The correct guessing of such conjectures, though successful in some problems, is a nondeductive process, i.e., is not guaranteed to succeed (e.g., is not computerizable), and the verification of conjectures is generally susceptible to mistakes. We propose a method that can deductively produce semiparametric locally efficient estimators. The proposed method is computerizable, meaning that it does not need either conjecturing, or otherwise theoretically deriving the functional form of the EIF, and is guaranteed to produce the desired estimates even for complex parameters. The method is demonstrated through an example.

2.1 Background

The desire for estimation that is robust to model assumptions has led to a growing literature on semiparametric estimation. Approximately efficient estimators can be obtained in general as the zeros of an approximation to the efficient influence function (EIF) (Tsiatis, 2007). Semiparametric estimation is useful, for example, for survival analysis (Cox, 1972), for estimating growth parameters in longitu-
CHAPTER 2. DEDUCTIVE ESTIMATION

dinal studies (Liang and Zeger, 1986), and for estimating quantities under missing data (Robins et al., 1994), including treatment effects based on potential outcomes (Davidian et al., 2005; Crump et al., 2009). Here, we focus on problems in which the distribution of the observed data is, in principle, unrestricted, but where estimability requires use of lower dimensional working models.

Theoretical derivation of the EIF in such problems can be challenging. If this task can be computerized, it can save dramatic human effort, which can then be transferred, for example, to designing new studies. The EIF for the unrestricted problem can be written, in general, as a Gateaux derivative (Hampel, 1974). However, if simple numerical differentiation is used to calculate the EIF by a computer to avoid theoretical derivations, then the EIF’s functional dependence on the parameter of interest is not revealed. For this reason, the derivative approach has not been generally used. Instead, the standard approach to obtaining the EIF is to construct theoretically the space of scores under all possible parametric sub-models (Begun et al., 1983). This process currently depends on the correctness of conjectures about these spaces and the correctness of their verification. The correct guessing of such conjectures can succeed in some problems, but is a nondeductive process, i.e., is not guaranteed to succeed (e.g., is not computerizable) and, as with their verification, is generally susceptible to mistakes.

We propose a method that can deductively produce semiparametric locally efficient estimators even for complex parameters. In Section 2.2, we formulate the
CHAPTER 2. DEDUCTIVE ESTIMATION

goal of a deductive method and show that it essentially requires numerical access
to the functional dependence of the EIF on the parameter of interest. Section 2.3
shows how the concept of compatibility solves the functional dependence prob-
lem, and derives a deductive method. Throughout, we use the two-phase design
as a test problem where the EIF is known theoretically, and we demonstrate our
method with a study on asthma as an example. Section 2.4 discusses extensions,
Section 2.5 concludes with remarks, and Section 2.6 gives example algorithms for
estimating mean and median in the two phase design.

2.2 Motivation for Deductive Estimation

2.2.1 The Goal of a Deductive Method

Suppose we conduct a study to measure data $D_i, i = 1, ..., n$, independent and
identically distributed (iid) from an unknown distribution $F$, in order to estimate
a root-$n$ estimable feature of the distribution $\tau(F)$. Suppose $\tau$ has a nonparametric
EIF denoted by $\phi(D_i, F - \tau, \tau)$, where $F - \tau$ denotes the remaining components
of the distribution, other than $\tau$. The goal is to find a deductive method that can
derive $\phi$ and can compute estimators $\hat{\tau}$ that solve

$$\sum_i \phi\{D_i, (F - \tau)_w, \tau\} = 0$$ (2.1)
CHAPTER 2. DEDUCTIVE ESTIMATION

after substituting for \((F - \tau)\) estimates of a working model \((F - \tau)_w\). Under some regularity conditions, estimators solving (2.1) are consistent and locally efficient if the working estimators of \((F - \tau)_w\) are consistent with convergence rates larger than \(n^{1/4}\) (van der Vaart, 2000). Our specific requirement that a method be “deductive and computerizable”, means that the method should need neither conjecturing for, nor otherwise theoretically deriving the functional form of \(\phi\), and should be guaranteed to produce an estimate in the sense of Turing (1937) (i.e., use a discrete and finite set of instructions, and, for every input, finish in discrete finite steps).

2.2.2 Conjecturing and Functional Form as Barriers Towards a Deductive Method

A test problem: estimating the mean in a two-phase design. To help make arguments concrete, we consider the following example where the EIF is well known. Suppose that in order to estimate the mean \(\tau = E(Y)\) in a population, the researcher first obtains a simple random sample of individuals and records an easily measured covariate \(X_i\). Then, the researcher is to measure the main outcome \(Y_i\) only for a subset denoted with \(R_i = 1\), where the missing data mechanism is ignorable given \(X\), i.e., \(pr(R_i = 1 \mid Y_i, X_i) = pr(R_i = 1 \mid X_i)\) (Rubin, 1976). The final data \(D_i\) are \((X_i, R_i, Y_iR_i)\), \(i = 1, \ldots, n\), iid from a distribution \(F\), and, by ignorability, the
parameter $\tau$ is identified from $F$ as

$$
\tau(F) = \int y(x)p(x)dx,
$$

(2.2)

where $p(x)$ is the density of $X_i$; and $y(x)$ is the conditional expectation $E(Y_i \mid R_i = 1, X_i = x)$. For this problem, the EIF is known (e.g., Robins and Rotnitzky (1995) and Hahn (1998)) to be

$$
\phi\{D_i, (F - \tau), \tau\} = \frac{R_i \cdot \{Y_i - y(X_i)\}}{e(X_i)} + y(X_i) - \tau,
$$

(2.3)

where $e(x)$ is the propensity score of selection into the second phase, $\text{pr}(R_i = 1 \mid X_i = x)$. The derivation has, so far, been nondeductive because it is first based on conjectures on the score space over all submodels, which are then verified to be true (e.g., Hahn (1998)).

Current estimation methods need the functional form of the EIF. Most existing approaches to using (2.1) first isolate a dependence of $\phi$ on $\tau$, then replace the remaining dependence on $F$ with a working model, and finally solve for $\tau$. For example, in the test problem above, the most common approach to using (2.3) to estimate $\tau$ first obtains working functions $y_w(X_i)$ and $e_w(X_i)$, for example using parametric MLEs,
CHAPTER 2. DEDUCTIVE ESTIMATION

and estimates \( \tau \) as the zero of the empirical sum of (2.3), to obtain:

\[
\hat{\tau}_{\text{non-deductive}} = \frac{1}{n} \sum_i R_i \cdot \{Y_i - y_w(X_i)\} e_w(X_i) + y_w(X_i); \tag{2.4}
\]

See, for example, Robins et al. (1994), Davidian et al. (2005), and Kang et al. (2007). While there also exist modified estimators like the targeted minimum loss estimator (TMLE) (van der Laan and Rubin, 2006), all methods that have been presented so far have advocated that it is critical to know the functional form dependence of \( \phi \) on \( F \), and so are nondeductive, hence, noncomputerizable without prior knowledge of the functional form.

The Gateaux derivative approach to EIF. For a general parameter \( \tau \), the EIF evaluated at an observation \( d_0 \) can be obtained as the Gateaux derivative

\[
\phi(d', F) = \lim_{\epsilon \to 0} \frac{\tau(F_{d', \epsilon}) - \tau(F)}{\epsilon}, \text{ where}
\]

\[
F_{d', \epsilon} = (1 - \epsilon) F + \epsilon \cdot 1 < d' >, \tag{2.6}
\]

where \( 1 < d' > \) denotes a point mass at \( d' \) (Hampel, 1974). Calculating this derivative at a given \( d' \) and \( F \) is a deductive and computerizable operation. To demonstrate the ease of its derivation, consider again the test problem with missing data.

Specifically, for a given observation \( d' = (x', r', y'r') \) and a distribution \( F \), it
follows from (2.2), (2.6), and Bayes rule, that

$$\tau(F_{d',e}) = \int y_{d',e}(x)p_{d',e}(x)dx,$$

where

$$p_{d',e}(x) = (1 - \epsilon)p(x) + \epsilon \cdot 1(x = x'),$$

and

$$y_{d',e}(x) = \frac{\epsilon \cdot 1(x = x', r' = 1) \cdot y' + (1 - \epsilon) \cdot p(x)e(x)y(x)}{\epsilon \cdot 1(x = x', r' = 1) + (1 - \epsilon) \cdot p(x)e(x)},$$

where $1(\cdot)$ is 1 (or 0) if the logical statement $\cdot$ is true (or false). Then, (2.5) becomes

$$\phi(d', F) = \int \left[ \frac{\partial y_{d',e}(x)}{\partial \epsilon} p_{d',e}(x) \right]_\epsilon dx + \int \left[ y_{d',e}(x) \frac{\partial p_{d',e}(x)}{\partial \epsilon} \right]_\epsilon dx. \quad (2.8)$$

The first and second terms of the above are $\frac{r'[y' - y(x')]}{e(x')}$ and $y(x') - \tau$, respectively, which is the result (2.3) above.

The problem with the derivative operation is that if simple numerical differentiation is used to calculate the EIF by a computer to avoid theoretical derivations, then the EIF’s functional dependence on the parameter of interest $\tau$ and $F$ is not revealed.

### 2.3 Deductive Estimation Method for Two-phase Design

*Method*
CHAPTER 2. DEDUCTIVE ESTIMATION

A start to finding a deductive method is to appreciate from a new perspective a problem that nondeductive estimators such as (2.4) have. Specifically, nondeductive estimators are usually constructed from a dependence of the EIF $\phi$ on $\tau$ that is different from the variation-independent partition into $[(F - \tau), \tau]$ (this is probably because of the limitations of closed-form expressions). For example, the estimator $\hat{\tau}_{\text{nonductive}}$ of (2.4) is a sample analogue of (i) the expression of the last appearance “$\tau$” in the right hand side of (2.3), using (ii) a working expectation $y_w(x)$; and (iii) the empirical estimator for $p(x)$ to average over quantities of $X_i$. However, the parameters underlying (i), (ii), and (iii) – namely, $\tau$, $y(x)$, and $p(x)$, respectively – are not variation-independent, because $\tau$ is the average of $y(x)$ over $p(x)$. This creates an incompatibility: the value of the estimator $\hat{\tau}_{\text{nonductive}}$ from this method differs (almost surely) from its defining expression $\tau(F)$ if for $F$ we use the estimates in (ii) and (iii) that are used to produce $\hat{\tau}_{\text{nonductive}}$.

The problem of incompatibility has been noted before as a nuisance (e.g., Newey (1998)) and has motivated compatible estimators like the TMLE (e.g., van der Laan and Rubin (2006)). Here, we show that, more fundamentally, the concept of incompatibility together with the Gateaux derivative create a solution to the problem of deductive estimation. In particular, the previous section noted that evaluation of the Gateaux derivative at a working distribution $F_w$ masks the dependence on $\tau$. However, the same evaluation does contain evidence that parts of the working distribution $F_w$ are misspecified, if the empirical sum of the Gateaux derivative

14
CHAPTER 2. DEDUCTIVE ESTIMATION

is not zero. This evidence of misspecified $F_w$ can be turned, by \textquotedblleft \ean\ atopon apa-\gamma\omega\gamma\eta\textquotedblright\ ("reduction to the absurd"), into estimation for $\tau$, where plausible values of $\tau$ are values $\tau(F)$ for distributions $F$ for which the empirical sum of the Gateaux derivative is zero and therefore eliminates any evidence of misspecification.

Based on the above argument, we can construct the following method that solves the deductive computerization problem by addressing the above compatibility problem.

\textbf{(step 1)}: Extend the working distribution $F_w$ to a parametric model, say, $F_w(\delta)$, around $F_w$ (i.e., so that $F_w(0) = F_w$), where $\delta$ is a finite dimensional vector. In this extension, we can keep unmodified the part of $F_w$ that is known to be most reliably estimated (e.g., a propensity score elicited by physicians).

\textbf{(step 2)}: Use the Gateaux numerical difference derivative

$$\text{Gateaux}\{\tau, F_w(\delta), D_i, \epsilon\} := \frac{\tau\{F_w(\delta, \epsilon)\} - \tau\{F_w(\delta)\}}{\epsilon}$$

for a machine-small $\epsilon$, to deduce the value of $\phi\{D_i, F_w(\delta)\}$ for arbitrary $\delta$, and find

$$\delta^{opt} \text{ that minimizes the empirical variance of } \tau\{F_w(\hat{\delta})\} \quad (2.9)$$
among all roots $\{\hat{\delta}\}$ that solve the equation

$$
\sum_i \left[ \phi[D_i, F_w(\hat{\delta})] \right] \left[ \text{Gateaux}\{\tau, F_w(\hat{\delta}), D_i, \epsilon\} \right] = 0,
$$

(2.10)

where “$\leftarrow$” means “computed as”. Property (2.10) is the empirical analogue of the central, mean-zero property if the evaluated $\phi$ at $F_w(\hat{\delta})$ is the true influence function of $\tau$. An average of the EIF at a $F_w(\delta)$ that deviates from zero is evidence that the working distribution is incorrect. This step finds a distribution $F_w(\hat{\delta})$ that eliminates such evidence. Technically, there may be no zeros, in which case \(\hat{\delta}\) can be defined as the minimizer of the absolute value of (2.10), although a better solution would be to make the model $F_w(\delta)$ more flexible (see below). More realistically, for a working model $F_w(\delta)$ there can be more than one zeros and so condition (2.9) selects the best one. Finally, although (2.9) is unambiguous if $\tau$ is a scalar, if $\tau$ is a vector then the researcher can minimize any one-dimensional criterion, such as, for example, the largest of the empirical variances of each of the components of $\tau\{F_w(\hat{\delta})\}$.

**(step 3)**: Calculate the parameter at the EIF-fitted distribution $F_w(\hat{\delta})$ as

$$
\hat{\tau}^{\text{deductive}} := \tau\{F_w(\hat{\delta}^{opt})\}.
$$

(2.11)
CHAPTER 2. DEDUCTIVE ESTIMATION

The above method is deductive because step 2 does not need the functional form of $\phi$, but deduces it by the numerical Gateaux derivative (2.5). If $\delta$ is 1-dimensional, then (2.10) is expected to have one root, and this can be found by numerical root-finding methods such as in Brent (1973) or quasi Newton-Raphson, by finding and using the numerical difference derivatives with respect to $\delta$ of the Gateaux derivative computation of $\phi$. If $\delta$ has more dimensions, then $\delta_{opt}$ can be found by either iterative quasi Newton-Raphson or by numerical Lagrange multipliers, where (2.9) can be coded as the jackknife variance. Also, the above estimates for $\tau$ and the remaining model parameters are compatible, by construction.

The deductive estimator shares useful properties of so-far known, nondeductive estimators that take $\phi$ as given. Notably, suppose the actual expectation of $\phi(D_i, F_w)$ is zero for a working distribution when, say part$_1(F_w) = \text{part}_1(F)$, or,...,or part$_K(F_w) = \text{part}_K(F)$. Then the deductive estimator above is expected to be consistent as would be usual, nondeductive estimators (e.g., Scharfstein et al. (1999)).

For example, for the two-phase design, suppose an original working function $y_w(x)$ has been obtained as the OLS fit $x'\beta_{ols}$ of a linear regression model $x'\beta$ for $E(Y \mid R = 1, X = x)$. Then, a simple model extension is to add to $x'\beta_{ols}$ a free parameter $\delta$ (this is the same as freeing-up (again) the intercept of $x'\beta_{ols}$ and let it be a parameter). The subsequent implementation steps for deriving the estimator for the mean estimand are given in Section 2.6. It is then easy to show (proof omitted) that this deductive estimator is doubly-robust (Scharfstein et al., 1999): it is
CHAPTER 2. DEDUCTIVE ESTIMATION

consistent either if the propensity score working model $e_{w}(X_i)$ (corresponding to part$_1(F_w)$ above) is correct, or if the regression working model $y_{w}(x)$ (corresponding to part$_2(F_w)$ above) is correct.

Also, the deductive estimator above shares with the TMLE the idea of extending the working model (Chaffee and van der Laan, 2011), and with other estimators the idea of empirical maximization (e.g., Rubin and van der Laan (2008)). The conditions for the deductive estimator to use the smallest empirical variance are similar to those used in Rubin and van der Laan (2008, Appendix 2) and are omitted here because of their technical nature. To our knowledge, all such existing work for local efficiency has considered it critical to have the theoretically derived form of the EIF based on the score theory. The contribution of the proposed method above is to show that this theory can be translated to estimation that can be computerized in general, by combining model extension with the Gateaux derivative.

The extension in step 1 can take different forms. For example, for the two-phase design, one can also compute an improved deductive estimator by extending $\delta$ to two dimensions (e.g., two coefficients) and minimizing the empirical variance as in step 2. If the space of distributions spanned by the one-dimensional-based extended model lies within the space spanned by the two-dimensional extended model, then the estimator based on the latter will have empirical variance at most that of the former estimator because of the larger space where minimization takes place.
CHAPTER 2. DEDUCTIVE ESTIMATION

Feasibility evaluations

To evaluate the feasibility of our method, we applied it to the study analyzed by Huang et al. (2005), as an example of the two-phase design. The goal of that study was to compare rates of patient satisfaction for asthma care as the outcome $Y$ (yes/no) among different physician groups (treatments). Physician groups differed in their distribution of patient covariates. So, in order to compare between, say, two physician groups, we set the goal to estimate the average $2.2$ of patient satisfaction for each group, standardized by the distribution of patient covariates in the combined population of the two groups. This standardization of estimands to the covariate distribution on all patients is also used in the literature, for example for point exposure studies (e.g. Rosenbaum and Rubin (1983)); and is more commonly now known as g-computation (based on Robins (1986)) also for longitudinal studies. The following covariates $X$ were considered: age, gender, race, education, health insurance, drug insurance coverage, asthma severity, number of comorbidities, and SF-36 physical and mental scores.

We tested feasibility of the above method for the comparison within two pairs of groups, denoted in Table 2.1(i) as $a_1$ vs. $b_1$ and $a_2$ vs. $b_2$ (actual names omitted). We chose $(a_1, b_1)$ as a pair for which the usual estimator $\hat{\tau}_F$ produces values diverging from the unadjusted rates for $a_1$ and $b_1$; and we chose $(a_2, b_2)$ as a pair for which the usual estimator produces values shrinking from $a_1$ and $b_1$. The nondeductive estimator used as propensity score the quintiles of the logistic regression
CHAPTER 2. DEDUCTIVE ESTIMATION

of group membership conditionally on \( X \); and a working expectation \( y_w \) as the prediction from the logistic regression of patient satisfaction conditionally on \( X \) within each group. The deductive estimator uses the same propensity score, and, for step 1 of the method, extended the working expectation \( y_w \) by including back the intercept in the logistic regression for each group as a free parameter \( \delta \). The computation of \( \phi \) for each \( \delta \) in (2.10) was obtained by straight-forward numerical differentiation for the Gateaux derivative; and the root \( \hat{\delta} \) was found by the method of Brent (1973) implemented by the function “unirout” in R. See Section 2.6 for further details.

In all cases in Table 2.1(i), the deductive estimator gives answers very close to the nondeductive estimator. This suggests that, for this problem and data, the usual doubly-robust estimator, although not derived compatibly, can be re-expressed compatibly by the set of parameter values derived by the deductive estimator. We have also studied computability of the deductive estimator for the estimand defined as the mean restricted to the patients with propensity scores in \((0.1, 0.9)\) (Table 2.1(ii)). For this estimand, for which the usual doubly-robust estimator is very close to the plain average, the deductive estimator is, again, very close to the usual non-deductive estimator. What is most important is that, although both estimators produced their answers in less than a second for each group and estimand, the deductive estimator did not need knowledge of the closed form expression (2.3) for \( \phi \), whereas the usual estimator depended critically on that
knowledge.

<table>
<thead>
<tr>
<th>(i) all patients</th>
<th></th>
<th>estimates of $\sum(F) = \int_{x \in A} y(x)p(x)dx$</th>
</tr>
</thead>
<tbody>
<tr>
<td>physician group $(g)$</td>
<td>$n$</td>
<td>unadjusted $%$</td>
</tr>
<tr>
<td>$a_1$</td>
<td>171</td>
<td>62.0</td>
</tr>
<tr>
<td>$b_1$</td>
<td>81</td>
<td>58.0</td>
</tr>
<tr>
<td>$a_2$</td>
<td>104</td>
<td>78.8</td>
</tr>
<tr>
<td>$b_2$</td>
<td>189</td>
<td>47.6</td>
</tr>
</tbody>
</table>

| (ii) patients with increased common support |  | $A : \text{patients with } \tilde{c}(x) \in (0.1, 0.9)$

| physician group $(g)$ | $n$ | unadjusted $\%$ | $\text{pr}(Y = 1 \mid G = g)$ | $\tau_{\text{non-deductive}} (%)$ | $\tau_{\text{deductive}} (%)$ |
|------------------|---|--------------------------------------------------|
| $a_1$ | 107 | 65.4 | 65.3 | 5.2 | 65.4 |
| $b_1$ | 76 | 59.2 | 59.3 | 6.9 | 59.1 |
| $a_2$ | 95 | 77.9 | 75.6 | 6.2 | 75.3 |
| $b_2$ | 154 | 46.8 | 46.2 | 5.1 | 46.3 |

*This estimand with increased "common support" (e.g., Crump et al. (2009)), excludes here 64, 5, 9 and 35 patients from $a_1, b_1, a_2, b_2$ respectively.

Table 2.1: Feasibility of the deductive method for estimating the probability of patient satisfaction adjusted for covariates for two physician group pairs using data from the asthma study of Huang et al. (2005)

2.4 Extensions to More Complex Settings

Close observation of the method for the deductive estimator for the mean in the two-phase design, as detailed in Section 2.6, actually reveals how to produce a locally semiparametric efficient estimator also for any other estimand in this design.

To see this, suppose we denote by $y_w(t; x)$ the cumulative distribution function
CHAPTER 2. DEDUCTIVE ESTIMATION

\[ pr_w(Y \leq t \mid X = x, R = 1) \] of \( Y \) for the working model. Then, by Bayes rule, we have that the cumulative distribution, say \( pr_{w(d', \epsilon)}(Y \leq t \mid X = x; R = 1) \), of \( Y \) in the perturbed distribution \( F_{d', \epsilon} \) of (2.6) at \( d' = (x', r', y'r') \), is

\[
1(y' \leq t) \frac{\epsilon \cdot 1(x = x', r' = 1)}{\epsilon \cdot 1(x = x', r' = 1) + (1 - \epsilon) \cdot p_w(x) e_w(x)} + y_w(t; x) \frac{(1 - \epsilon) \cdot p_w(x) e_w(x)}{\epsilon \cdot 1(x = x', r' = 1) + (1 - \epsilon) \cdot p_w(x) e_w(x)}. \tag{2.12}
\]

Based on this measure, implementation of steps 1-3 of Section 2.3 is relatively easy and generalizable. We have implemented this method in order to derive a locally efficient semiparametric estimator also for the median estimand in the two-phase design. This deductive estimator for the median, for which we are aware of no other implemented estimator, is given in Section 2.6. We have conducted several simulation experiments (omitted) in all of which the deductive estimator is consistent also for this estimand. A comprehensive report on the small sample properties of the deductive estimator for the median and for other more challenging estimands is of interest for future study.

In complex problems, it is possible that standard root finding methods for (2.10) are unstable. In this section we show that the Gateaux numerical derivative may still be used to construct a deductive estimation method that does not rely on solving an estimating equation.

Suppose that the parameter \( \tau(F) \) depends on \( F \) only through a set of variation
CHAPTER 2. DEDUCTIVE ESTIMATION

independent parameters $q_j(F) : j = 1, \ldots, J$. Such is the case of parameter (2.2) in our example, with $q_1(F; x) = y(x)$ and $q_2(F; x) = p(x)$. In an abuse of notation, let $\tau(q_1(F), \ldots, q_J(F)) := \tau(F)$. Since the parameters $q_j$ are variation independent, the Gateaux derivative expression of $\phi$ in (2.5) reduces to

$$
\phi(d', F) = \sum_{j=1}^J \lim_{\epsilon \to 0} \frac{\tau(q_1(F), \ldots, q_j(F_{d', \epsilon}), \ldots, q_J(F)) - \tau(F)}{\epsilon}.
$$

This expression provides the decomposition $\phi(d', F) = \sum_{j=1}^J \phi_j(d', F)$, where $\phi_j$ is the non-parametric efficient score associated to $q_j$. Once the Gateaux numerical derivatives $\phi_j$ have been computed, it is possible to implement a standard TMLE without knowledge of the functional form of $\phi$. We only provide a brief recap of the TMLE template since extensive discussions are presented elsewhere (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). For each $q_j$, consider a loss function $L_j(q_j; D)$ whose expectation is minimized at the true value of $q_j$. Consider also a working model $q_{jw}$ and a parametric extension $q_{jw}(\delta)$ satisfying

$$
\frac{d}{d\delta} L(q_{jw}(\delta); d') \bigg|_{\delta=0} = \phi_j(d').
$$

In our example, since $q_j$ are components of the likelihood, the negative log-likelihood
loss function and the exponential family may be used in this step:

\[ L(q_j; d) = - \log q_j(d), \]

\[ q_{jw}(\delta; d) \propto \exp(\delta \phi_j(d))q_{jw}(d). \]  

(2.13)

The TMLE is then defined by an iterative procedure that, at each step, estimates \( \delta \) by minimizing the expected sum of the loss functions \( L_j(q_{jw}(\delta); \cdot) \). An update of the working model is then computed as \( q_{jw} \leftarrow q_{jw}(\hat{\delta}) \), and the process is repeated until convergence. The TMLE is defined by \( \hat{\tau} = \tau(q^{*}_{1w}, \ldots, q^{*}_{Jw}) \), where \( q^{*}_{jw} \) denotes the estimate obtained in the last step of the iteration. Like the estimator presented in Section 2.3, the TMLE is a compatible estimator, and solves the EIF estimating equation. Unlike the estimator of Section 2.3, the TMLE does not require direct solution of that equation. However, the TMLE may be computationally more intensive, as it is iterative and may require numerical integration for computation of the proportionality constant in (2.13).

2.5 Remarks

We proposed a deductive method to produce semiparametric estimators that are locally efficient. The method does not rely on conjectures of tangent spaces and is not susceptible to possible errors in the verification of such conjectures. Instead,
the new method relies on computability of the estimand \( \tau \) for specified working distributions of the observed data \( F \), and on numerical methods for differentiation and for root finding.

Although we have focused on local efficiency of originally unrestricted problems, one can see a path towards finding a deductive method also for problems with restrictions set a priori. Such a path can explore, first, nesting the restricted problem within an unrestricted one, and then, making use of the proposed deductive method for the unrestricted problem, modified to impose numerically the nested restrictions. Such deductive methods can save dramatic amounts of human effort on essentially computerizable processes, and allow the transfer of that effort to other statistically demanding parts of the scientific process such as the efficient design of new studies.

### 2.6 Appendix: Algorithm Examples

#### 2.6.1 Estimating Mean in Two-phase Design

This section provides the details for how steps 1-3 of the general method of Section 2.3 are implemented in the data example given in that section.

**(Preliminaries) : Coding of functions for the estimands at working and perturbed distributions.**
CHAPTER 2. DEDUCTIVE ESTIMATION

First, a working distribution $F_w(\hat{\beta})$ was specified as follows:

(i) the working distribution, $p_w(\cdot)$, of $X$, was taken to be the empirical distribution with point-mass $1/n$ at each observed $X_i$ (one can also assign weights other than $1/n$ for standardizing to different population);

(ii) the working propensity score, $e_w(\cdot)$, was taken to be the fit from a logistic regression;

(iii) the working outcome regression, $y_w(\cdot)$ for $E(Y \mid X = \cdot, R = 1)$, was taken to be the fit from the logistic regression:

$$y_w(x, \hat{\beta}) = \expit\{\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}\},$$  \hspace{1cm} (2.14)

where $x = (x^{(1)}, \ldots, x^{(p)})$ is $p$-dimensional covariate vector and expit is the inverse logit.

Then, functions were coded for the estimands $\tau \{F_w(\beta)\}$, and $\tau \{F_{w(D_i, \epsilon)}(\beta)\}$, i.e., the perturbation at the data point $D_i = (X_i, R_i, Y_i R_i)$ and arbitrary $\beta$, $0 < \epsilon < 1$. Based on the general formula (2.7) and the above working distributions, these functions are
CHAPTER 2. DEDUCTIVE ESTIMATION

\[ \tau \{ F_w(\beta) \} = \sum_{j=1}^{n} y_w(X_j, \beta) p_w(X_j), \quad (2.15) \]

\[ \tau \{ F_w(D_i, \epsilon)(\beta) \} = \sum_{j=1}^{n} y_{w(D_i, \epsilon)}(X_j, \beta) p_{w(D_i, \epsilon)}(X_j), \quad (2.16) \]

where the components of \( F_w(D_i, \epsilon)(\beta) \) are derived using Bayes rule:

\[
p_{w(D_i, \epsilon)}(x) = (1 - \epsilon)p_w(x) + \epsilon \cdot 1(x = X_i), \]

\[
y_{w(D_i, \epsilon)}(x, \beta) = \frac{\epsilon \cdot 1(x = X_i, R_i = 1)Y_i + (1 - \epsilon) \cdot p_w(x) e_w(x) y_{w}(x, \beta)}{\epsilon \cdot 1(x = X_i, R_i = 1) + (1 - \epsilon) \cdot p_w(x) e_w(x)}.\]

Then, steps 1-3 of Section 2.3 were implemented as follows.

**step 1**: The extended working model \( F_w(\delta) \) of Section 2.3 was defined by adding to \( x'\hat{\beta} \) a free parameter \( \delta \). Specifically, for a given \( \delta \), the extended working distribution, denoted here more precisely by \( F_w(\hat{\beta}_\delta) \), takes the working distributions for the covariate and for the propensity score as in the working model (i)-(iii), but takes the working regression \( E(Y | X = x, R = 1) \) to be \( y_w(x, \hat{\beta}_\delta) \) (see (2.14)) where \( \hat{\beta}_\delta = (\delta + \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) \). Note that here \( \delta \) is 1-dim, and we have \( F_w(\hat{\beta}_\delta)|_{\delta=0} = F_w(\hat{\beta}) \).

**step 2**: The empirical influence function is numerically computed and solved for its zero. To do this, this step starts with a candidate \( \delta \) (say 0). Then,

(i) for a small \( \epsilon \), this step computes \( \tau \{ F_w(\hat{\beta}_\delta) \} \) and \( \tau \{ F_w(D_i, \epsilon)(\hat{\beta}_\delta) \} \) using the func-
CHAPTER 2. DEDUCTIVE ESTIMATION

...tions defined in (2.15)-(2.16), and hence computes the numerical derivative

\[ \phi\{D_i, F_w(\widehat{\beta}_0)\} := \frac{\tau\{F_{w(D_i, e)}(\widehat{\beta}_0)\} - \tau\{F_w(\widehat{\beta}_0)\}}{\epsilon}; \]

(ii) the sum \( \sum_{i=1}^{n} \phi\{D_i, F_w(\widehat{\beta}_0)\} \) is computed for the candidate \( \delta \);

(iii) substeps (i)-(ii) above are repeated using the bisection method to find a \( \widehat{\delta} \) such that the sum \( \sum_{i=1}^{n} \phi\{D_i, F_w(\widehat{\beta}_0)\} \) is 0 (note that because \( \delta \) has dimension 1, there is no search to optimize the empirical variance).

(Step 3) : The estimate \( \widehat{\tau}_{\text{deductive}} \) is computed using the function (2.15), giving

\[ \widehat{\tau}_{\text{deductive}} := \tau\{F_w(\widehat{\beta}_0)\}. \]

2.6.2 Estimating Median in Two-phase Design

This section describes how steps 1-3 of the general method of Section 2.3 are implemented to estimate the median outcome in the two-phase design, that is,

\[ \tau := \text{median}(F) = \inf_t \left\{ t : \int \text{pr}(Y \leq t \mid X = x, R = 1)p(x)dx \geq 0.5 \right\}, \quad (2.17) \]

where the last equality follows by ignorability in the two phase-design.

(Preliminaries) : Coding of functions for the estimands at working and per-
First, consider a working distribution \( F_w(\theta) \), with the working distribution, \( p_w(\cdot) \), of \( X \), and the working propensity score, \( e_w(\cdot) \), as (i) and (ii) in Section 2.6; and with

(iii') the working conditional distribution for the outcome given \( X \) to be the MLE fit from a normal regression \( N(\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}, \hat{\sigma}^2) \), and denote the cumulative distribution by

\[
y_w(t; x, \hat{\theta}) := \Pr \left( Y \leq t \mid X = x, R = 1, \hat{\theta} \right),
\]

(2.18)

where \( \theta = (\beta, \sigma^2) \).

Then, the median \( \tau \{ F_w(\theta) \} \), and \( \tau \{ F_{w(D_i,\epsilon)}(\theta) \} \), i.e., the perturbation at the data point \( D_i = (X_i, R_i, Y_i R_i) \) and arbitrary \( \theta, 0 < \epsilon < 1 \), can be easily derived based on the general formula (2.7) and the above working distributions, as

\[
\tau \{ F_w(\theta) \} = \inf \left\{ \sum_{j=1}^{n} y_w(t; X_j, \theta) p_w(X_j) \geq 0.5 \right\},
\]

(2.19)
CHAPTER 2. DEDUCTIVE ESTIMATION

\[
\tau \left\{ F_{w(D_i,\epsilon)} (\theta) \right\} = \\
= \inf_t \left\{ \sum_{j=1}^{n} y_{w(D_i,\epsilon)} (t; X_j, \theta) p_{w(D_i,\epsilon)} (X_j) \geq 0.5 \right\},
\]

where the components of \( F_{w(D_i,\epsilon)} (\theta) \) are derived using Bayes rule (similar argument to Section 2.6)

\[
p_{w(D_i,\epsilon)} (x) = (1 - \epsilon) p_w (x) + \epsilon \cdot 1 (x = X_i),
\]

\[
y_{w(D_i,\epsilon)} (t; x, \theta) = \\
1(Y_i \leq t) \frac{\epsilon \cdot 1 (x = X_i, R_i = 1)}{\epsilon \cdot 1 (x = X_i, R_i = 1) + (1 - \epsilon) \cdot p_w (x) e_w (x)} + y_w (t; x, \theta) \frac{(1 - \epsilon) \cdot p_w (x) e_w (x)}{\epsilon \cdot 1 (x = X_i, R_i = 1) + (1 - \epsilon) \cdot p_w (x) e_w (x)}. \tag{2.20}
\]

Then, steps 1-3 of Section 2.3 were implemented as follows.

(step 1): The extended working model \( F_w (\delta) \) of Section 2.3 was defined by freeing-up the intercept of \( \beta \). Specifically, for a given \( \delta \), the extended working distribution, denoted here more precisely by \( F_{w(\tilde{\theta}_\delta)} \), takes the working distributions for the covariate and for the propensity score as in the working model (i)-(ii), but takes the cumulative distribution \( \text{pr}(Y \leq t \mid X = x, R = 1) \) to be \( y_w (t; x, \tilde{\theta}_\delta) \) (see (2.18)) where \( \tilde{\theta}_\delta = (\delta + \tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{\sigma}^2) \).
CHAPTER 2. DEDUCTIVE ESTIMATION

(step 2): The empirical influence function is numerically computed and solved for its zero in exactly the same way as in step 2 of Section 2.6.

(step 3): The estimate \( \hat{\tau}^{\text{deductive}} \) is computed using the function (2.19), giving

\[
\hat{\tau}^{\text{deductive}} := \tau\{F_w(\hat{\theta}_\delta)\}.
\]

Note: Because (2.20) actually shows the full measure for \( Y \) under the extended working models, it can be used to compute, under these models, any estimand that can be computed based on the original working models. The above discussion, then, also serves to produce locally semiparametric efficient estimators for any other such estimand in this design.
Chapter 3

Extension of Deductive Estimation to Double-sampling Design

SUMMARY.¹ Semiparametric theory has been used to construct robust estimators in missing data problems. Efficient semiparametric estimation usually requires analytic form of the efficient influence function (EIF), the derivation of which can be difficult. Frangakis et al. [Biometrics 71, 867-874 (2015)] proposed to computerize this estimation procedure using functional derivative representation of the EIF in nonparametric models. Their approach, however, requires deriving a mixture of a continuous distribution and a point mass, which can be challenging for complicated problems. In this chapter, we propose the novel “Discrete Support” idea, which overcomes the challenge and generalizes their approach to more compli-

¹This chapter is adapted from the following working paper: Tianchen Qian, Constantine E. Frangakis, Constantin Yiannoutsos. Deductive Semiparametric Estimation using Discrete Support: with Application to Double-sampling Designs.
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

cated designs and estimands. The Discrete Support is a space constructed from the observed data, which enables i) approximation of the observed data distribution, and ii) numerical computation for constructing semiparametric estimators.

Our method is guaranteed to produce semiparametric locally efficient estimators within finite steps without knowledge of the EIF. We illustrate the construction of Discrete Support and the corresponding deductive estimation procedure in the context of double-sampling designs for handling non-ignorable dropouts. We apply the method to the estimation of mortality rate using data set from a double-sampling design evaluating an HIV treatment in East Africa, and analyze the sensitivity of the estimated mortality rate to the inclusion criteria of double-samples.

We demonstrate through simulation studies that the proposed estimator is robust to misspecified model for survival time if the propensity score model for double-sampling selection is correct.

3.1 Background

Studies with long follow-up often suffer from high dropout rate. Whether a participant drops out of a study can depend on his/her outcome of interest, even after adjusting for observed covariates. This makes the dropouts non-ignorable and the analysis based solely on the non-dropouts biased. As a way to handle non-ignorable dropouts, double-sampling designs allocate additional resources to pur-
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

sue a sample of the dropouts and find out their outcomes (Baker et al., 1993; Glynn et al., 1993; Frangakis and Rubin, 2001; Cochran, 2007). The double-sampling can be more practical or informative if it targets dropouts whose history at the dropout time has specific profiles. For example, An et al. (2009) found that such “profile designs” can save more than 35% of resources compared to the standard double-sampling design. Also, it can be more practical to double-sample relatively recent dropouts. For estimation in double-sampling designs, An et al. (2014) proposed to use parametric modeling approaches. However, analyses of such designs would be more reliable if it does not rely heavily on parametric assumptions. Robins et al. (2001) had suggested a possible way of deriving a semiparametric estimator for double-sampling designs, but that and any other existing such proposals rely on first coming up with and then verifying conjectures “by hand”. Such a process is prone to human errors and is not deductive, i.e., not generalizable.

In this chapter, we propose to conduct deductive estimation using “Discrete Support”, which overcomes the limitation in Frangakis et al. (2015). The Discrete Support is a space constructed from the observed data, which enables i) approximation of the observed data distribution, and ii) numerical computation for constructing semiparametric estimators, including augmenting fitted models and perturbing a continuous distribution by a point mass at an observed data point. This discretization technique has its root in Chamberlain (1987b). Our method is guaranteed to produce semiparametric locally efficient estimators within finite steps.
without knowledge of the EIF. This deductive and computerizable method makes semiparametric estimation more feasible to statistical practitioners.

We illustrate the construction of Discrete Support and the corresponding deductive estimation procedure in the context of double-sampling designs. We apply the method to estimating mortality rate using data from a double-sampling design evaluating an HIV treatment in East Africa. In addition, because double-sampling can be more practical if restricted to relatively recent dropouts (An et al., 2014), we explore and discuss the impact of such restrictions on the estimated mortality rate. We conduct simulation studies to demonstrate the robustness of the proposed deductive estimator for mortality rate in double-sampling designs, and find that it is consistent even if the model for survival time is misspecified, as long as the propensity score model for the double-sampling selection is correct. This robustness property is valuable in practice, because the double-sampling selection can usually be controlled by investigators and hence correctly modeled.

This chapter is organized as follows. In Section 3.2, we introduce the double-sampling design and the parameter of interest with its identifiability conditions. In Section 3.3, we give an overview of the general deductive estimation method and provide the specific algorithm for double-sampling designs, including construction of the Discrete Support. In Section 3.4, we apply the method to the HIV study and analyze the impact of restrictions on double-sampling. In Section 3.5, results from simulation studies are presented. Section 3.6 concludes the chapter.
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

with remarks.

3.2 Double-sampling Design and Quantity of Interest

For clarity, we present arguments with a double-sampling design as shown in Figure 3.1, which is modified from An et al. (2014, Figure 1). First, we describe characteristics that are inherent to a patient [i.e., potential outcomes (Rubin, 1976)] which would be realized if the program could enroll and follow a patient indefinitely with a standard effort. Then, we describe the actual design consisting of two phases.

Patient inherent characteristics and quantity of interest. For each patient, there is an enrollment time $E_i$, covariates $info_i^0$ at enrollment, and a survival time $T_i$ (time from enrollment to death). The quantity of interest is $\text{pr}(T_i > t)$ for a given $t$, which is the proportion of patients surviving beyond year $t$ for a population represented as a random sample by the enrolled patients. If the program were to follow patients indefinitely with a standard effort while they are alive, some patients would discontinue contact from the monitoring (patient types (b1), (c), (d), (e1), and (e2) in Figure 3.1). For these patients, labeled true dropouts and indicated by $R_i = 0$, denote by $L_i$ the time from enrollment to dropout and denote by $info_i^{\text{loss}}$ any in-
### Figure 3.1: Characteristics for different patient types in a double-sampling design.

Column $n$ denotes the number of patients of each type in the HIV study data set. "obs" means observed; "UD" means undefined. Quantities in brackets are unobserved. The figure is a modified version of An et al. (2014, Figure 1).
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

formation available at dropping out in addition to \(\text{info}_i^0\). For instance, \(\text{info}_i^{\text{loss}}\) may include \(L_i\) and some longitudinal measurements.

**Phase 1 of the actual design and missingness due to administrative censoring.** In the first phase, the actual design enrolls patients at times \(E_i\) and monitors them with standard effort, not indefinitely, but until “Now”—the present time. The time from enrollment \(E_i\) to “Now” is the time to administrative censoring, denoted by \(C_i\). For patient \(i\), define \(\Delta_i = 1\) if \(C_i \geq T_i\) and \(\Delta_i = 0\) otherwise. Define \(X_i = \min(C_i, T_i)\).

Simply based on Phase 1, the standard survival data \((X_i, \Delta)\) are not observed for dropouts whose time to dropout satisfies \(L_i < X_i\). Denote such observed dropout patients by \(R_i^{\text{obs}} = 0\) (patient types (c), (d), (e1), and (e2) in Figure 3.1).

**Phase 2: information after double-sampling.** Phase 2 of the design selects a subset of the observed dropouts, called a double-sample, by using characteristics of the patient known up to the dropout time \(L_i\). Additional resources are allocated for searching for and finding this double-sample at “Now”. The characteristics used in the double-sampling selection are labeled \(\text{info}_i^{\text{loss}}\). Such double-sampling is expressed as follows:

**Condition 1. Patient-dependent double-sampling.** For each observed dropout, the indicator \(S_i\) of being selected as a double-sample can depend on the patient characteristics \((\text{info}_i^0, \text{info}_i^{\text{loss}})\) up to dropout time \(L_i\); after we condition on \((\text{info}_i^0, \text{info}_i^{\text{loss}})\),
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

double-sampling $S_i$ is independent of survival and enrollment times:

\[
S_i \perp (T_i, E_i) \mid R_{i}^{\text{obs}} = 0, info_i^0, info_i^{\text{loss}}.
\]

The design must also address the information missing due to administrative censoring. This can be done in practice by limiting the design to a window period of enrollment, within which the following holds:

**Condition 2.** Exchangeability of enrollment times within a time period. Patients enrolled at different times (equivalently, having different time to administrative censoring $C_i = \text{“Now”} - E_i$) have similar survival times after conditioning on $info_i^0$:

\[
T_i \perp C_i \mid info_i^0
\]

Under Conditions 1 and 2, the estimand $P(T > t)$ is identifiable from the following components of the distribution of the observed data:

\[
\begin{align*}
\text{pr}(R_i^{\text{obs}} = 1); & \quad \text{pr}(info_i^0 \mid R_i^{\text{obs}} = 1); \quad \text{pr}\{(X_i, \Delta_i) \mid info_i^0, R_i^{\text{obs}} = 1}\}; \\
\text{pr}(R_i^{\text{obs}} = 0); & \quad \text{pr}(info_i^0, info_i^{\text{loss}} \mid R_i^{\text{obs}} = 0); \quad \text{pr}\{(X_i, \Delta_i) \mid info_i^0, info_i^{\text{loss}}, R_i^{\text{obs}} = 0, S_i = 1}\}.
\end{align*}
\]

(3.1)

(3.2)

In particular, by Condition 1 the distribution from the double-sampled individuals $\text{pr}\{(X_i, \Delta_i) \mid info_i^0, info_i^{\text{loss}}, R_i^{\text{obs}} = 0, S_i = 1\}$ is the same as the one for all dropouts,
pr\{ (X_i, \Delta_i) \mid \text{info}_i^0, \text{info}_i^{\text{loss}}, R_{i}^{\text{obs}} = 0 \}, and so, together with the second component of (3.2) gives, upon averaging over \text{info}_i^{\text{loss}}, pr\{ (X_i, \Delta_i), \text{info}_i^0 \mid R_{i}^{\text{obs}} = 0 \}. That, together with (3.1), gives pr\{ (X_i, \Delta_i), \text{info}_i^0 \}. Denote by surv(\cdot; t) the function that takes \cdot as an arbitrary distribution of (X, \Delta) from independent survival and censoring times and returns the survival probability beyond t (this function is the common probability limit of the Nelson-Aalen and Kaplan-Meier estimators). Then, by Condition 2, the estimand is calculable from the above distributions as

$$
\tau := \text{pr}(T > t) = \int \text{surv} \left[ \text{pr}\{ (X_i, \Delta_i) \mid \text{info}_i^0 \}; t \right] \text{pr}(\text{info}_i^0) d(\text{info}_i^0). \quad (3.3)
$$

The calculation described above involves regression distributions that need to be estimated with robust and deductive (i.e., easily computerizable) methods. We describe such a method next.

### 3.3 Deductive Estimation for Double-sampling Design

#### 3.3.1 General Method

Estimation of \( \tau \) requires modeling assumptions because the expression in (3.3) involves regressions that, in practice, cannot be estimated fully nonparametrically.
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

In particular, suppose the estimand \( \tau \) in (3.3) has a nonparametric efficient influence function (EIF) denoted by \( \phi(\text{Data}_i, [F - \tau, \tau]) \), where \( F - \tau \) represents the remaining components of the distribution, other than \( \tau \). We wish to find deductive estimators \( \hat{\tau} \) that solve

\[
\sum_i \phi(\text{Data}_i, [(F - \tau), \tau]) = 0 \tag{3.4}
\]

after substituting for \( (F - \tau) \) estimates of a working model. Such estimators are consistent and locally efficient if the working estimators of \( (F - \tau) \) are consistent with convergence rates larger than \( n^{1/4} \) (van der Vaart, 2000). By “deductive”, we mean that the estimator should not rely on conjectures for the functional form of \( \phi \), and should be guaranteed to produce an estimate in the sense of Turing (1937) (i.e., use a discrete and finite set of instructions, and, for every input, finish in discrete finite steps).

Recently, a deductive estimation method has been proposed that does not need the analytic form of EIF (Frangakis et al., 2015). The key idea is that, for any working distribution \( F(\alpha) \) for \( F \), we can calculate numerically the EIF as the numerical Gateaux derivative after perturbing the working distribution by a small mass at
each observed data point; i.e.,

\[ \phi\{\text{Data}_i, F(\alpha)\} \text{ gets computed as Gateaux}\{F(\alpha), \text{Data}_i, \epsilon\}, \]

(3.5)

where Gateaux\{F(\alpha), \text{Data}_i, \epsilon\} := \left[ \tau\{F(\text{Data}_i, \epsilon)(\alpha)\} - \tau\{F(\alpha)\} \right] / \epsilon,

and where \( F_{\text{Data}_i, \epsilon}(\alpha) = (1 - \epsilon)F(\alpha) + \epsilon \cdot (\text{point mass at Data}_i) \). Then, we can find the best working distribution parameter \( \hat{\alpha} \) as one that makes zero the sum of the numerical EIFs, \( \sum_i \text{Gateaux}\{F(\alpha), \text{Data}_i, \epsilon\} \). The estimator solving (3.4) approximately is then \( \tau\{F(\hat{\alpha})\} \). The standard error of the estimator can be estimated by \( n^{-1} \left[ \sum_i \text{Gateaux}\{F(\hat{\alpha}), \text{Data}_i, \epsilon\}^2 \right]^{1/2} \) with \( n \) denoting the sample size. Finding the perturbed distribution of a continuous working model at an arbitrary observed data point is easy in some designs but can be time consuming in others such as the double-sampling design. Nevertheless, this task can be rendered relatively straightforward when approximating continuous working models by discrete ones using the Discrete Support idea described in the following.

### 3.3.2 Specific Deductive Estimation Procedure for Double-sampling Design

We are interested in estimating \( \tau = \text{pr}(T_i > t) \), which is identifiable by (3.3). The observed data for patient \( i \) is \( \text{Data}_i := (R_i^{\text{obs}}, \text{info}_i^0, \text{info}_i^{\text{loss}}, S_i, X_i, \Delta_i, C_i) \), where
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

\(\text{info}_i^{\text{loss}}\) includes \(L_i\) (which is undefined or missing for anyone with \(R_i^{\text{obs}} = 1\)), \(S_i\) is undefined for anyone with \(R_i^{\text{obs}} = 1\), and \((X_i, \Delta_i)\) are missing for anyone with \(R_i^{\text{obs}} = 0\) and \(S_i = 0\). We define the Discrete Support (DS) for the double-sampling design to be the following, which depends on the observed data: \(\text{DS} := \text{DS}_{R^{\text{obs}}=0} \cup \text{DS}_{R^{\text{obs}}=1}\), where

\[
\text{DS}_{R^{\text{obs}}=0} := \{(\text{info}_i^0, \text{info}_i^{\text{loss}}, S, X, \Delta) : \text{there exists } i, j \text{ such that } (R_i^{\text{obs}}, \text{info}_i^0, \text{info}_i^{\text{loss}}) = (0, \text{info}^0, \text{info}^{\text{loss}}) \text{ and } (R_j^{\text{obs}}, S_j, X_j, \Delta_j) = (0, S, X, \Delta)\};
\]

\[
\text{DS}_{R^{\text{obs}}=1} := \{(\text{info}_i^0, X, \Delta) : \text{there exists } i, j \text{ such that } (R_i^{\text{obs}}, \text{info}_i^0) = (1, \text{info}^0) \text{ and } (R_j^{\text{obs}}, X_j, \Delta_j) = (1, X, \Delta)\}.
\]

As illustrated in Table 3.1, \(\text{DS}_{R^{\text{obs}}=0}\) is the Cartesian product space of the unique values of \((\text{info}_i^0, \text{info}_i^{\text{loss}})\) and the unique values of \((S_i, X_i, \Delta_i)\) for patients with \(R_i^{\text{obs}} = 0\), and \(\text{DS}_{R^{\text{obs}}=1}\) is the Cartesian product space of the unique values of \(\text{info}_i^0\) and the unique values of \((X_i, \Delta_i)\) for patients with \(R_i^{\text{obs}} = 1\). \(\text{DS}\) contains all observed Data\(i\) and additional points needed for the deductive estimation procedure. As we will see in the following, such construction of \(\text{DS}\) enables i) approximation of the observed data distribution, ii) augmenting fitted working models, and iii) perturbing a working distribution by a point mass at an observed data point. Note that instead of constructing an overall Cartesian product space, we keep the quantities that can be estimated reliably from the data (i.e., \(\text{pr}(R^{\text{obs}})\)), and construct product spaces
within $R_{\text{obs}} = 1$ and $R_{\text{obs}} = 0$. Denote by $F$ a (discrete) probability distribution on the space $\text{DS}$. The deductive estimation procedure for $\tau$ is as follows:

**Step 1**: Construct and code the function that takes an arbitrary probability distribution $F$ on $\text{DS}$, and outputs (through averaging and normalization) the decomposition (3.1)-(3.2) and the value $\tau(F)$ based on (3.3).

**Step 2**: Construct and code the function that takes an arbitrary probability distribution $F$ on $\text{DS}$, an index $i$, and a small $\epsilon > 0$, and outputs the perturbed mixture distribution
\[
F_{(\text{Data}_i, \epsilon)} = (1 - \epsilon)F + \epsilon \cdot (\text{point mass at Data}_i)
\]
Given $\text{Data}_i$, the perturbed probability $F_{(\text{Data}_i, \epsilon)}$ at an arbitrary point $z \in \text{DS}$ can be computed by $(1 - \epsilon)\text{pr}_F(\text{Data}_i) + \epsilon$ if $z = \text{Data}_i$, and $(1 - \epsilon)\text{pr}_F(z)$ if $z \neq \text{Data}_i$, where $\text{pr}_F(z)$ is the probability $F$ at $z$.

**Step 3**: Extend a fitted working distribution $F$ to a model $F(\alpha)$ with a tuning parameter $\alpha$, so that the left-hand side of (3.6) below explores values around 0. As a concrete example, definition of $F(\alpha)$ for the specific HIV data set is given in (3.7) in Section 3.4.

**Step 4**: Find the parameter $\hat{\alpha}$ that solves
\[
\sum_i \text{Gateaux}\{F(\hat{\alpha}), \text{Data}_i, \epsilon\} = 0, \quad (3.6)
\]
where $\text{Gateaux}\{\cdot\}$ is defined in (3.5). For computing $\text{Gateaux}\{F(\hat{\alpha}), \text{Data}_i, \epsilon\}$, $F(\alpha)$ is
computed by step 3; \( F_{\text{Data}_i, \epsilon}(\alpha) \) is computed by step 2; and \( \tau\{F(\alpha)\} \) and \( \tau\{F_{\text{Data}_i, \epsilon}(\alpha)\} \) are computed by step 1.

**(step 5):** The deductive estimator is \( \tau^{\text{deductive}} := \tau\{\hat{\alpha}\} \), computed by step 1. The standard error of \( \hat{\tau}^{\text{deductive}} \) can be estimated by \( n^{-1}\left[ \sum_i \text{Gateaux}\{F(\hat{\alpha}), \text{Data}_i, \epsilon\}^2 \right]^{1/2} \).

\[
\begin{array}{cccc}
 s & \text{(double-sampled):} & 1 & \ldots & 1 & 0 \\
 x & \text{(unique values):} & x_1 & \ldots & x_{#s=1} & \text{NA} \\
 \delta & \text{(associated with } x\text{):} & \Delta_1 & \ldots & \Delta_{#s=1} & \text{NA} \\
\end{array}
\]

<table>
<thead>
<tr>
<th>DS(_{R^{\text{obs}}=0})</th>
<th>(s)</th>
<th>(x) (unique values)</th>
<th>(\delta) (associated with (x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(x_1)</td>
<td>(\Delta_1)</td>
<td></td>
</tr>
<tr>
<td>\ldots</td>
<td>(x_{#s=1})</td>
<td>(\Delta_{#s=1})</td>
<td></td>
</tr>
</tbody>
</table>

Labels for unique values of \((\text{info}^0, \text{info}^{\text{loss}})\):

- \((1)\)
- \((2)\)
- \(\ldots\)
- \((#s = 1)\)
- \(\((#s = 1) + 1\)\)
- \(\ldots\)
- \((#R^{\text{obs}} = 0)\)

\[
\begin{array}{cccc}
 x & \text{(unique values):} & x_1 & \ldots & x_{#R^{\text{obs}}=1} \\
 \delta & \text{(associated with } x\text{):} & \Delta_1 & \ldots & \Delta_{#R^{\text{obs}}=1} \\
\end{array}
\]

<table>
<thead>
<tr>
<th>DS(_{R^{\text{obs}}=1})</th>
<th>(x) (unique values)</th>
<th>(\delta) (associated with (x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1)</td>
<td>(\Delta_1)</td>
<td></td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\Delta_{#R^{\text{obs}}=1})</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.1:** Illustration of the Discrete Support (DS) for double-sampling designs.
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

3.4 Application to PEPFAR study

We apply our method to estimate the mortality rate using the data set from a double-sampling design in East Africa evaluating the antiretroviral treatment (ART) for HIV-infected people (Geng et al., 2015). The data set consists of 1,773 HIV-infected adults from Morogoro, Tanzania, who started ART after entering the study. There are 673 dropouts during the study. Among the dropouts, 91 patients got double-sampled. We used baseline age and pre-treatment CD4 value as $i_{0}$, and the loss to follow-up time $L$ and the CD4 value measured at the last visit before dropout as $i_{\text{loss}}$. 316 out of the 1,773 enrollees (17.8%) have pre-treatment CD4 values missing, and 47 out of the 673 observed dropouts (7.0%) have CD4 value at the last visit missing. For simplicity the missing values are imputed using sample means.

Working and extended models. The estimation procedure is given in Section 3.3.2. Here we describe the working model $F$ and its extended version $F(\alpha)$ which are used in Step 3 of the estimation procedure for this specific data set. The working model $F$ on DS is factored as follows. For $\text{pr}(R_{\text{obs}}^0)$, $\text{pr}(i_{0}^0 | R_{\text{obs}}^0 = 1)$, and $\text{pr}(i_{0}^0, i_{\text{loss}}^0 | R_{\text{obs}}^0 = 0)$, we use the empirical distributions. The selection for double-sampling $\text{pr}(S = 1 | R_{\text{obs}}^0 = 0, i_{0}^0, i_{\text{loss}}^0)$ is modeled using logistic regression. The distribution $\text{pr}(X, \Delta | R_{\text{obs}}^0 = 0, i_{0}^0, i_{\text{loss}}^0)$ is modeled as the likelihood arising from independent $T, C$ given $R_{\text{obs}}^0 = 0, i_{0}^0, i_{\text{loss}}^0$. This working
independence does not need to and is not expected to hold, and only the resulting likelihood for $(X, \Delta)$ is used. The working distributions for censoring time $\Pr(C \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})$ and survival time $\Pr(T \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})$ are each taken as Cox proportional hazards model fit. An analogous but separate model is used for the distribution $\Pr(X, \Delta \mid R_{\text{obs}} = 1, \text{info}^0)$. The working distribution $F$ evaluated at an arbitrary point in $\text{DS}_{R_{\text{obs}}=0}$ or $\text{DS}_{R_{\text{obs}}=1}$ is calculated as follows (with proper normalization):

$$F(R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}, S = 1, X, \Delta) \propto \Pr(R_{\text{obs}} = 0) \Pr(\text{info}^0, \text{info}^{\text{loss}} \mid R_{\text{obs}} = 0) \times \Pr(S = 1 \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}) \Pr(X, \Delta \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}),$$

$$F(R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}, S = 0, X = \text{NA}, \Delta = \text{NA}) \propto \Pr(R_{\text{obs}} = 0) \Pr(\text{info}^0, \text{info}^{\text{loss}} \mid R_{\text{obs}} = 0) \times \Pr(S = 0 \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}),$$

$$F(R_{\text{obs}} = 1, \text{info}^0, X, \Delta) \propto \Pr(R_{\text{obs}} = 1) \Pr(\text{info}^0 \mid R_{\text{obs}} = 1) \Pr(X, \Delta \mid R_{\text{obs}} = 1, \text{info}^0).$$

To extend the working distribution $F$ to a model $F(\alpha)$ with tuning parameter $\alpha$, we keep the rest of the factors in $F$ unchanged, and only add the tuning parameter to $\Pr(X, \Delta \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})$ and $\Pr(X, \Delta \mid R_{\text{obs}} = 1, \text{info}^0)$ as follows. Denoting by $\Pr_w$ these two models for $X$ and $\Delta$, the extended models $\Pr(\cdot \mid \alpha)$ with $\alpha > 0$
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

are taken as:

\[
\begin{align*}
\text{pr}(X, \Delta \mid R^{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}; \alpha) & \propto \text{pr}_w(X, \Delta \mid R^{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})\{1 + (\alpha - 1)X/c_0\}, \\
\text{pr}(X, \Delta \mid R^{\text{obs}} = 1, \text{info}^0; \alpha) & \propto \text{pr}_w(X, \Delta \mid R^{\text{obs}} = 1, \text{info}^0)\{1 + (\alpha - 1)X/c_1\},
\end{align*}
\]

(3.7)

where \(c_r := \max\{C_i : R_i^{\text{obs}} = r\}, r \in \{0, 1\}\). The tuning parameter \(\alpha\) calibrates the working model so that (3.6) explores values around 0, by putting more weights on larger \(X\) values (when \(\alpha > 1\)) or smaller \(X\) values (when \(\alpha < 1\)).

Estimated mortality rate. The black curve in Figure 3.2a is the estimated \(t\)-year mortality rate \(\text{pr}(T < t)\) for \(0 \leq t \leq 2\) using the method in Section 3.3, along with its pointwise 95% confidence interval obtained by normal approximation. For example, the deductively estimated 1-year mortality rate is 11.4% with standard error 0.8%. As a comparison, a stratified Kaplan-Meier approach gives estimated 1-year mortality rate 14%. The yellow curve in Figure 3.2a is the estimated mortality rate when forcing \(\hat{\alpha} = 1\) in step 5 of Section 3.3, which corresponds to the estimator obtained by directly plugging in the fitted working model into the estimand, \(\tau(F)\), without the additional steps to find \(\hat{\alpha}\) that solves (3.6). This is close to the stratified Kaplan-Meier estimates. Because of the robustness property of the proposed deductive estimator, we believe it has smaller bias than the stratified Kaplan-Meier estimator.
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

(a) Deductive versus $\hat{\alpha} = 1$, no restriction

(b) Estimates with restriction on double-sampling

Figure 3.2: Estimated mortality rate with 95% pointwise confidence interval (shaded band). (a) Comparison of deductive estimates (with $\hat{\alpha}$ solving (3.6)) and the estimates when forcing $\hat{\alpha} = 1$, where there is no restriction on double-sampling. (b) Comparison of deductive estimates with different restrictions on double-sampling.

Sensitivity to double-sampling selection criteria. It can be more practical to double-sample relatively recent dropouts (An et al., 2014). We explore the impact of such selection criteria on the estimated mortality rate. In particular, for $\gamma > 0$, we consider the selection rule that only the dropouts with $C_i - L_i \leq \gamma$ are eligible to be sampled in the second phase. For different values of $\gamma$, Table 3.2 lists the proportion of double-samples in the ART data set that satisfy $C_i - L_i \leq \gamma$ and the corresponding deductively estimated 1-year mortality rate. Given $\gamma$, we obtain the estimated mortality rate by first setting $S_i = 0$ and $(X_i, \Delta_i) = (\text{NA, NA})$ for the double-samples in the data set with $C_i - L_i > \gamma$, and then using the method in Section 3.3. The estimated 1-year mortality rate and the standard error seem to be only slightly impacted by $\gamma$. When $\gamma = 262$ days (i.e., when only the most recent 19
double-samples are included), because there is too little variation in the included double-samples to fit the Cox model \( \Pr(C \mid R^{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}) \), the estimation procedure broke down and hence reports NA for the estimate and the standard error.

<table>
<thead>
<tr>
<th>( \gamma ) (days)</th>
<th>Included double-samples (%)</th>
<th>( \hat{\alpha} )</th>
<th>Estimate (%)</th>
<th>Standard error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>262</td>
<td>20</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>344</td>
<td>30</td>
<td>1.45</td>
<td>12.2</td>
<td>0.69</td>
</tr>
<tr>
<td>382</td>
<td>40</td>
<td>1.58</td>
<td>12.5</td>
<td>0.71</td>
</tr>
<tr>
<td>454</td>
<td>50</td>
<td>1.59</td>
<td>12.3</td>
<td>0.72</td>
</tr>
<tr>
<td>494</td>
<td>60</td>
<td>1.57</td>
<td>10.8</td>
<td>0.71</td>
</tr>
<tr>
<td>566</td>
<td>70</td>
<td>1.61</td>
<td>12.1</td>
<td>0.78</td>
</tr>
<tr>
<td>652</td>
<td>80</td>
<td>1.56</td>
<td>11.5</td>
<td>0.77</td>
</tr>
<tr>
<td>897</td>
<td>90</td>
<td>1.57</td>
<td>11.9</td>
<td>0.79</td>
</tr>
<tr>
<td>1061</td>
<td>100</td>
<td>1.53</td>
<td>11.7</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 3.2: Deductive estimate and standard error for the 1-year mortality rate in the ART data set when restricting double-sampling to dropouts with \( C - L \leq \gamma \) for different \( \gamma \) values.

Figure 3.2b shows the estimated \( t \)-year mortality rates without restriction on double-sampling (black curve) and when restricting the design to only double-sample the patients who dropped out within the past two years (blue curve) or the past year (yellow curve). They correspond to \( \gamma = 1061, 750, \) or 365 days; that is, when all, 85%, or 36% of the double-samples in the data set are included. For all three curves, the estimated mortality rates are similar for \( t \leq 1 \) year. The yellow curve (restricting double-sampling to only the past year dropouts) start to diverge from the other two as \( t \) grows beyond 1 year. We conclude that different selection criteria can result in quite different estimates, and evaluation and optimization of
the double-sampling selection criteria is an area of future research.

3.5 Simulation Study

We conduct a simulation study to demonstrate the robustness of the proposed estimator in a double-sampling design. Let $Z = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)$ be a vector of six independent Uniform$[0, 1]$ variables. Denote by $R$ the true dropout status of a patient, and by $R_{\text{obs}}$ the observed dropout status. Hypothetical patients in the simulation are generated as follows:

$$R \mid Z \sim \text{Bernoulli}\{(Z_1 + Z_2)/2\};$$

$$T \mid R, Z \sim \text{Exponential}\{(1.5 - R) \exp\{1 - (Z_2 + Z_3)/2\}\};$$

$$C \mid R, Z \sim \text{Exponential}(1),$$

$$L \mid R = 1, T, Z \sim \text{Uniform}(T/3, T),$$

$$S \mid R_{\text{obs}} = 1, L, Z \sim \text{Bernoulli}\{\expit(-1 + Z_4 + 2L)\}.$$ 

Descriptive statistics of a simulated data set with sample size 500 are listed in Table 3.3. The proportion of true dropouts is about 51%, and the proportion of the observed dropouts is about 42% (not all true dropouts are observed dropouts due to administrative censoring). The mean survival times for the true dropouts and non-dropouts are 1.12 and 0.41, respectively. This indicates that the dropouts have
shorter survival time on average compared to the non-dropouts, which is a pattern also observed in the HIV data set (Geng et al., 2015). We assume that the distribution of $C$ does not depend on any other variable, so that we always have a correct model for it. Among the observed dropouts, about 48% get double-sampled in the second phase.

<table>
<thead>
<tr>
<th>Variable</th>
<th>(for binary) Frequency of 1 (%)</th>
<th>(for continuous) Mean (SD) Min / Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>49.2</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>0.76 (0.97) 1e-3 / 7.07</td>
<td></td>
</tr>
<tr>
<td>(within $R = 1$)</td>
<td>1.12 (1.19) 2e-3 / 7.07</td>
<td></td>
</tr>
<tr>
<td>(within $R = 0$)</td>
<td>0.41 (0.48) 1e-3 / 2.84</td>
<td></td>
</tr>
<tr>
<td>$C$</td>
<td>0.96 (0.98) 3e-3 / 5.30</td>
<td></td>
</tr>
<tr>
<td>$L \mid R = 0$</td>
<td>0.27 (0.34) 5e-4 / 2.84</td>
<td></td>
</tr>
<tr>
<td>$R_{\text{obs}}$</td>
<td>58.2</td>
<td></td>
</tr>
<tr>
<td>$L \mid R_{\text{obs}} = 0$</td>
<td>0.19 (0.20) 5e-4 / 1.20</td>
<td></td>
</tr>
<tr>
<td>$\Delta \mid R_{\text{obs}} = 1$</td>
<td>38.8</td>
<td></td>
</tr>
<tr>
<td>$X \mid R_{\text{obs}} = 1$</td>
<td>0.46 (0.51) 2e-3 / 4.54</td>
<td></td>
</tr>
<tr>
<td>(within $\Delta = 1$)</td>
<td>0.51 (0.51) 2e-3 / 2.84</td>
<td></td>
</tr>
<tr>
<td>(within $\Delta = 0$)</td>
<td>0.29 (0.50) 3e-3 / 4.54</td>
<td></td>
</tr>
<tr>
<td>$S \mid R_{\text{obs}} = 0$</td>
<td>47.9</td>
<td></td>
</tr>
<tr>
<td>$\Delta \mid S = 1$</td>
<td>83.0</td>
<td></td>
</tr>
<tr>
<td>$X \mid S = 1$</td>
<td>0.34 (0.32) 1e-4 / 1.66</td>
<td></td>
</tr>
<tr>
<td>(within $\Delta = 1$)</td>
<td>0.31 (0.29) 1e-4 / 1.57</td>
<td></td>
</tr>
<tr>
<td>(within $\Delta = 0$)</td>
<td>0.46 (0.44) 0.07 / 1.66</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.3:** Descriptive statistics of a simulated data set with sample size 500.

The estimand we consider here is $\tau = \Pr(T > 0.15)$, which equals 0.78 under our simulation distribution. We use the same deductive estimation method as in Section 3.4. The simulation distribution makes Cox proportional hazards model correct for $\Pr(T \mid R, Z)$ and $\Pr(C \mid R, Z)$, but incorrect for $\Pr(T \mid R_{\text{obs}}, Z)$.
(Bender et al., 2005). We consider different combinations of $\text{info}^0$ and $\text{info}^{\text{loss}}$ listed in the first two columns of Table 3.4 to introduce misspecification for models of $\text{pr}(T \mid R, \text{info}^0)$ and $\text{pr}(S \mid R^{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})$. Besides the deductive estimator $\hat{\tau}^{\text{deductive}}$, we also compute three alternative estimators for comparison: an inversely-weighted Kaplan-Meier estimator (weighted by the inverse of estimated double-sampling propensity score), a stratified Kaplan-Meier estimator weighted by $\text{pr}(R^{\text{obs}})$, and a Kaplan-Meier estimator using only the first phase complete cases. Detailed implementation of the alternative estimators can be found in Geng et al. (2010).

Table 3.4 shows the Root-mean-square Error (RMSE) for the deductive estimator $\hat{\tau}^{\text{deductive}}$ and the three alternatives under different choices of $\text{info}^0$ and $\text{info}^{\text{loss}}$, using the above data generating distribution with sample size $n = 500$. The RMSEs for the stratified Kaplan-Meier and the complete cases Kaplan-Meier are constant under different $\text{info}^0, \text{info}^{\text{loss}}$, and hence are listed after the table. Our observation is three-fold. First, according to the first two rows in the table, $\hat{\tau}^{\text{deductive}}$ has the smallest RMSE among all estimators as long as the model for double-sampling selection $\text{pr}(S \mid R^{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}})$ is correct. Simulation with larger sample sizes (results omitted here) shows that the RMSE of $\hat{\tau}^{\text{deductive}}$ decreases at approximately $n^{1/2}$ rate, which means that the deductive estimator is robust to misspecification in the survival time regression model. In other words, the deductive estimator is consistent as long as the propensity score model for double-sampling selection is
correctly specified. Second, by comparing row 1 to row 2 or row 3 to row 4 of Table 3.4, we conclude that correctly specifying the model for \( \text{pr}(T \mid R, \text{info}^0) \) does not help to decrease the RMSE—instead, it is the model for \( \text{pr}(T \mid R_{\text{obs}}, \text{info}^0) \), the survival model conditional on the observed dropouts, that affects the performance of \( \hat{\tau}^{\text{deductive}} \). This is because \( \text{pr}(T \mid R_{\text{obs}}, \text{info}^0) \) is the relevant model in fitting the working distributions. Third, when the model for \( \text{pr}(T \mid R_{\text{obs}}, \text{info}^0) \) and the model for \( \text{pr}(S \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}) \) are both wrong, \( \hat{\tau}^{\text{deductive}} \) performs worse than the stratified Kaplan-Meier estimator.

<table>
<thead>
<tr>
<th>info(^0)</th>
<th>info(^{\text{loss}})</th>
<th>correct model for</th>
<th>RMSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( T \mid R )</td>
<td>( T \mid R_{\text{obs}} )</td>
<td>( S )</td>
</tr>
<tr>
<td>( Z_2, Z_3 )</td>
<td>( Z_4, L )</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>( Z_5 )</td>
<td>( Z_4, L )</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>( Z_2, Z_3 )</td>
<td>( Z_6 )</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>( Z_5 )</td>
<td>( Z_6 )</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

RMSE for K-M weighted by \( \text{pr}(R_{\text{obs}}) \) is 0.034; RMSE for K-M on the first phase complete cases is 0.153.

**Table 3.4:** Root-mean-square Error (RMSE) in the simulation study for \( \hat{\tau}^{\text{deductive}} \), inverse-weighted Kaplan-Meier (weighted by inverse of the estimated propensity score for double-sampling), Kaplan-Meier weighted by \( \text{pr}(R_{\text{obs}}) \), and Kaplan-Meier on the first phase complete cases. RMSE for the latter two estimators does not depend on the choice of \( \text{info}^0 \), \( \text{info}^{\text{loss}} \). Result is based on 1000 simulations for each scenario, with \( n = 500 \) and \( \epsilon = 1 \times 10^{-4} \).

The double-sampling selection \( \text{pr}(S \mid R_{\text{obs}} = 0, \text{info}^0, \text{info}^{\text{loss}}) \) is controlled by the study investigators; hence, it can usually be correctly modeled. When it is correctly modeled, \( \hat{\tau}^{\text{deductive}} \) is consistent even under misspecified model for the survival time. This robustness property makes the deductive estimator attractive to use in the double-sampling setting.
3.6 Remarks

We proposed a deductive method to produce semiparametric estimators for complicated problems like the double-sampling design. The method is easily computerizable by the proposed Discrete Support idea. We illustrated the specification of the method components through a double-sampling data set from an HIV study. We demonstrated through simulation studies the robustness property of the proposed estimator, in the sense that it is consistent under incorrect survival time model as long as the double-sampling selection is correctly modeled. It would also be interesting to evaluate the deductive estimator when the model for survival time is correct and the model for double-sampling selection is incorrect. However, this is less likely to happen in practice due to the inherent complexity in $\text{pr}(T \mid R_{\text{obs}}, \text{info}^0)$.

Our method applies to other estimands besides the $t$-year mortality rate. For an arbitrary estimand in the double-sampling design, the estimation procedure in Section 3.3.2 carries over directly, and the only change to make is the definition of the estimand $\tau(F)$ in Step 1.

Our method is general in that it also applies to other designs. The definition of the Discrete Support and the extended working model $F(\alpha)$ will depend on the specific design, but the estimation procedure in Section 3.3.2 still works.

In literature, discussion on robust estimators have been partly based on characterizing robust estimating functions; see, e.g., Robins et al. (2000) and Robins
CHAPTER 3. APPLICATION: DOUBLE-SAMPLING DESIGN

and Rotnitzky (2001). Although our estimator is obtained by (numerically) solving an estimation equation, it is more difficult to find analytic conditions for the estimating function to be unbiased, because our main focus is on problems with unknown efficient influence functions that are hard to derive. We conjecture that a condition for our estimator to be robust to model misspecification is when the model contains multiple working pieces orthogonal to each other. It is an area of future research to develop asymptotic theory and establish general conditions for the deductive estimator to be robust.

As mentioned at the end of Section 3.4, it would be practically valuable to develop methods for evaluating different profile double-sampling designs and determining the optimal design using the proposed estimator.
Chapter 4

Semiparametric Estimation in Randomized Trials: Covariate Adjustment

SUMMARY. In group sequential designs, adjusting for baseline variables and short-term outcomes can lead to increased power and reduced sample size. We derive formulas for the precision gain from such covariate adjustment using semiparametric estimators for the average treatment effect, and give new results on what conditions lead to substantial power gains and sample size reductions. The formulas reveal how the impact of prognostic variables on the precision gain is mod-

This chapter is adapted from the following working paper: Tianchen Qian, Michael Rosenblum, Huitong Qiu. Improving Power in Group Sequential, Randomized Trials by Adjusting for Prognostic Baseline Variables and Short-Term Outcomes.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

ified by the number of pipeline participants, analysis timing, enrollment rate, and treatment effect heterogeneity, when the semiparametric estimator uses correctly specified models. Given set prognostic value of baseline variables and short-term outcomes within each arm, the precision gain is maximal when there is no treatment effect heterogeneity. In contrast, a purely predictive baseline variable, which only explains treatment effect heterogeneity but is marginally uncorrelated with the outcome, can lead to no precision gain. The theory is supported by simulation studies based on data from a trial of a new surgical intervention for treating stroke.

4.1 Background

Group sequential designs involve interim analyses that may result in early stopping of trials for efficacy or futility. We consider designs where the primary outcome is measured at a fixed time (called the delay) after enrollment. Prior to measuring the primary outcome, practitioners may also have measured correlated short-term outcomes and baseline variables of the enrollees. In the MISTIE-II trial (Hanley et al., 2016) evaluating a surgical treatment for intracerebral hemorrhage, for example, the primary outcome is the modified Rankin Scale (a score indicating the degree of disability) at 180 days after treatment; the short-term outcomes are the modified Rankin Scales at 30 days and 90 days after treatment; and the baseline variables are a set of health-related covariates measured at enrollment.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Because of its potential to increase power and reduce sample size, research has been focused on adjusting for baseline variables and short-term outcomes in randomized trials (e.g., Yang and Tsiatis, 2001; Moore and van der Laan, 2009a; Stallard, 2010; Hampson and Jennison, 2013; Paul et al., 2014). However, according to Pocock et al. (2002), where the authors surveyed 50 clinical trial reports from major medical journals and found that 36 had used some form of covariate adjustment, “the statistical properties of covariate-adjustment are quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy.” A more recent survey also pointed out the confusing variation that exists in handling of baseline covariates in randomized trials (Austin et al., 2010). We attempt to resolve this confusion by addressing the following questions: how much precision gain can be obtained by properly adjusting for baseline variables and short-term outcomes? What factors impact such precision gain in a group sequential design?

Many variations of doubly-robust estimators have been proposed for use in randomized trials, as they are more robust to model misspecification than estimators based on parametric regression models (Robins and Rotnitzky, 1992; Hubbard et al., 2000; van der Laan and Robins, 2003; Stitelman et al., 2012). Under randomized treatment assignment and independent censoring, the doubly-robust estimators are consistent and hence guarantee Type I error control even with incorrect regression models. They are asymptotically efficient when the models are
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

correct. In particular, Lu and Tsiatis (2011) and van der Laan and Gruber (2012) give general approaches for constructing such estimators using longitudinal data. However, limited research has been done on quantifying the magnitude of the resulting precision and its relationship to the design characteristics. An exception is Moore et al. (2011), where the authors attribute the improved estimation efficiency to the correlation between the baseline variables and the primary outcome, under the setting of a single stage randomized trial.

In this chapter, we consider group sequential designs where the short-term outcome and the primary outcome can be administratively censored. We derive formulas for the precision gain (measured by the asymptotic relative efficiency against the unadjusted estimator) from adjusting for baseline variables and short-term outcomes using semiparametric estimators in randomized trials. Key components of the formulas include the variance explained by baseline variables, the residual variance explained by short-term outcomes, the proportion of pipeline participants, and the degree of treatment effect heterogeneity. The formulas can be used in trial planning to predict the potential precision gain from covariate adjustment, which will translate to reduced sample size. This theory is supported by simulation studies based on data from the MISTIE Phase II trial, a trial of a new surgical intervention.

We highlight a few implications of the theory. When estimating the mean primary outcome in one arm, adjusting for strongly prognostic baseline variables
would result in substantial precision gain, but the precision gain from adjusting for a prognostic short-term outcome can be limited. If at most $1/3$ of the enrollees are in the pipeline and have missing primary outcome, the asymptotic equivalent sample size reduction by adjusting for prognostic short-term outcomes is at most half of the reduction by adjusting for an equally prognostic baseline variable, where the sample size reduction is calculated from the benchmark of doing no covariate adjustment. If there is no missing primary outcome (e.g., at the end of a single stage trial or at the final analysis of a group sequential design with no dropouts), adjusting for prognostic short-term outcomes does not contribute precision gain at all. If there is administrative censoring or dropout (e.g., at an interim analysis of a group sequential design), adjusting for short-term outcomes can lead to precision gain, but such gain is generally smaller at later interim analyses, which are the ones that tend to impact power the most.

A conclusion for mean outcome in one arm does not necessarily imply the same conclusion for the average treatment effect. When estimating the average treatment effect, the precision gain from adjusting for baseline variables is modified by treatment effect heterogeneity. We propose a measure for the degree of treatment effect heterogeneity, defined as the standardized variation in the treatment effect across different levels of the baseline variable. Given set prognostic value of baseline variables and short-term outcomes within each arm, the precision gain is maximal when there is no treatment effect heterogeneity. In contrast, a purely pre-
dictive baseline variable, which only explains treatment effect heterogeneity but is marginally uncorrelated with the outcome, can lead to no precision gain. The precision gain from adjusting for short-term outcomes is not modified by treatment effect heterogeneity.

In Section 4.2, we introduce notation and assumptions. In Section 4.3, we establish theory to analyze the precision gain from adjusting for prognostic baseline variables and short-term outcomes using semiparametric estimators when estimating the mean primary outcome in one arm. In Section 4.4, we analyze the precision gain when estimating the average treatment effect, which is modified by treatment effect heterogeneity. We give an example of efficient semiparametric estimators in Section 4.5, which is used as the test statistic in simulation studies in Section 4.6. Section 4.7, with discussion, concludes the chapter.

### 4.2 Notation and Assumptions

For participant $i$, denote by $A_i$ the indicator of a binary treatment. Denote by $W_i$ the vector of baseline variables that are measured before the treatment assignment. Denote by $L_i$ the short-term outcome, which is observed at time $d_L$ after the treatment assignment. Denote by $Y_i$ the primary outcome of interest, observed at time $d_Y$ after the treatment assignment with $d_Y \geq d_L$. When followed up completely, participant $i$ contributes full data $(W_i, A_i, L_i, Y_i)$. The goal is to test for positive
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

average treatment effect: $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, with $\theta$ defined as

$$\theta := E(Y|A = 1) - E(Y|A = 0).$$

Denote by $n$ the sample size. Assume that $\{(W_i, A_i, L_i, Y_i) : 1 \leq i \leq n\}$ are independent and identically distributed draws from a joint distribution $P_0$ on $(W, A, L, Y)$. We assume a nonparametric model for the joint distribution $P_0$, and we assume $A$ is independent of $W$. The semiparametric results in the following sections are with respect to this model. To show the key idea, we focus on simple cases with univariate and binary $L$; the results can be easily generalized to continuous $L$.

Assume that during the course of the trial participants are continuously enrolled. Since not all participants have full data observed at an interim analysis, we use $C^L_i$ and $C^Y_i$ to denote the missingness status of $L_i$ and $Y_i$, respectively: $C^L_i = 1$ if and only if $L_i$ is observed; $C^Y_i = 1$ if and only if $Y_i$ is observed. Assume that the censoring is monotone: $C^L_i \geq C^Y_i$ (i.e., observing $Y_i$ implies observing $L_i$). At any time, an enrollee has one of the following three missingness statuses: (i) $(C^L_i, C^Y_i) = (0, 0)$: a pipeline participant with only $(W_i, A_i)$ observed. (ii) $(C^L_i, C^Y_i) = (1, 0)$: a pipeline participant with $(W_i, A_i, L_i)$ observed (also called “partially observed”). (iii) $(C^L_i, C^Y_i) = (1, 1)$: a participant with $(W_i, A_i, L_i, Y_i)$ observed (also called “fully observed”). Assume that independent censoring holds:
(C_i^L, C_i^Y) is independent of W_i, A_i, and the potential outcomes of L_i and Y_i. An example of independent censoring is when administrative censoring is the only source of missingness. See Appendix 4.8.1 in the Supplementary Material for a formulation of the assumptions within causal inference framework.

The unadjusted estimator for \( \theta \) is the difference in sample means of the two arms:

\[
\frac{\sum_i A_i C_i^Y Y_i}{\sum_i A_i C_i^Y} - \frac{\sum_i (1 - A_i) C_i^Y Y_i}{\sum_i A_i C_i^Y}.
\]

The unadjusted estimator for mean primary outcome in one arm is the sample mean of that arm. This is the benchmark against which we will compare the semi-parametric estimator.

An estimator \( \hat{\theta} \) of the parameter \( \theta \) is consistent and asymptotically normal (CAN) if \( \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2) \), where \( \xrightarrow{d} \) denotes convergence in distribution. The asymptotic variance of \( \hat{\theta} \) is \( \text{AVar}(\hat{\theta}) := \sigma^2 \). The asymptotic relative efficiency (ARE) between two estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) is the inverse of the ratio of their asymptotic variances:

\[
\text{ARE}(\hat{\theta}_1, \hat{\theta}_2) := \frac{\text{AVar}(\hat{\theta}_2)}{\text{AVar}(\hat{\theta}_1)}.
\]

Consider two Wald tests for \( H_0 : \theta = 0 \) using two CAN estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) (each test statistic being the estimator divided by its standard error). The asymptotic relative efficiency can be interpreted as the limit of the inverse of ratio of the required sample sizes for the two estimators to achieve a set power (e.g. 80%) at lo-
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

...cal alternatives (van der Vaart, 2000, Section 8.2). Hence, we define the asymptotic equivalent reduction in sample size (RedSS) of \( \hat{\theta}_1 \) compared to \( \hat{\theta}_2 \) as

\[
\text{RedSS}(\hat{\theta}_1, \hat{\theta}_2) := 1 - \text{ARE}(\hat{\theta}_1, \hat{\theta}_2)^{-1}.
\]  

(4.1)

For instance, if \( \text{ARE}(\hat{\theta}_1, \hat{\theta}_2) = 1.25 \), for both estimators to achieve the same power using the Wald test, \( \hat{\theta}_1 \) only requires \( 1/1.25 = 80\% \) sample size compared to \( \hat{\theta}_2 \) asymptotically. This implies that the asymptotic equivalent reduction in sample size is \( \text{RedSS}(\hat{\theta}_1, \hat{\theta}_2) = 20\% \).

4.3 Precision Gain When Estimating Mean Primary Outcome in One Arm

4.3.1 Theoretical Limit on the Precision Gain

In this section we focus on estimating the mean outcome in one arm at a time: \( E(Y \mid A = a) \) for \( a \in \{0, 1\} \). The estimation of the average treatment effect is addressed in Section 4.4. Given \( a \in \{0, 1\} \), let \( E_a(\cdot) \) and \( \text{Var}_a(\cdot) \) denote \( E(\cdot \mid A = a) \) and \( \text{Var}(\cdot \mid A = a) \), respectively. We first give a decomposition of \( \text{Var}_a(Y) \).
Lemma 4.3.1.1. Given $a \in \{0, 1\}$, we have

$$\text{Var}_a(Y) = \text{Var}_a\{E_a(Y \mid W)\} + \text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\} + \text{Var}_a\{Y - E_a(Y \mid L, W)\}.$$  

(4.2)

Similar to ANOVA where the total variance is decomposed into the within-group variance and the between-group variance that are orthogonal to each other, Lemma 4.3.1.1 decomposes the variance of $Y$ into three orthogonal parts: the variation explained by $W$, the additional variation explained by $L$ after adjusting for $W$, and the residual variation.

The following theorem gives the lower bound on the asymptotic variance for all regular asymptotically linear (RAL) estimators of $E(Y \mid A = a)$.

**Theorem 4.3.1.2.** Assume randomization ($A \perp W$), independent censoring on $L, Y$, and monotone censoring ($C^L \geq C^Y$). Define $p_a := P(A = a)$ for $a \in \{0, 1\}$, $p_l := P(C^L = 1)$, and $p_y := P(C^Y = 1)$. Assume $0 < p_a < 1$ and $0 < p_y \leq p_l \leq 1$.

The semiparametric lower bound on the asymptotic variance of all RAL estimators for $E(Y \mid A = a)$ is

$$\text{Var}_a\{E_a(Y \mid W)\} + \frac{1}{p_a p_l} \text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\} + \frac{1}{p_a p_y} \text{Var}_a\{Y - E_a(Y \mid L, W)\}.$$  

(4.3)

Theorem 4.3.1.2 is obtained through analyzing the efficient influence function of $E(Y \mid A = a)$. No RAL estimator can have asymptotic variance smaller than
the bound in (4.3). The estimator in van der Laan and Gruber (2012), discussed in Section 4.5, achieves this variance lower bound when all working models are correct. When the working models for this estimator are misspecified, (4.3) with conditional expectations replaced by the probability limit of the model fits may still serve as a good approximation to the asymptotic variance of the estimator, as shown in the simulation results in Section 4.6. The validity of this approximation follows from results in van der Laan and Gruber (2012, Section 4) when certain types of working models are used.

Theorem 4.3.1.2 implies that adjusting for \( W \) is more effective in reducing variance bound than adjusting for \( L \). Equation (4.3) is the weighted version of the right-hand side of (4.2), where without the weight factors \( (p_a p_t)^{-1} \) and \( (p_a p_y)^{-1} \) the three terms in (4.3) sum to \( \text{Var}_a(Y) \). Therefore, if the term with a smaller weight gets larger, the term with a larger weight will become smaller, which implies that the whole sum of (4.3) will become smaller. In a randomized trial with \( p_a = 1/2 \), the second and the third terms in (4.3) have at least twice the weight of the first term. This means that when \( \text{Var}_a(Y) \) is fixed, larger \( \text{Var}_a\{E_a(Y \mid W)\} \) (i.e., having \( W \) explain more variation in \( Y \)) reduces the variance lower bound (4.3). Similarly, when \( p_t < p_y \) (i.e., when there are partially observed participants), having \( L \) explain more variation after accounting for \( W \) also reduces the variance lower bound. Because the ordering of the weights of the first and the second terms in (4.3) is \( 1 < 1/(p_a p_t) \), making the first term larger is more effective than making the
second term larger in reducing the variance bound.

Theorem 4.3.1.2 suggests the following definition of $R$-squared, which quantifies the prognostic value of $W$ and $L$ in each arm. For $a \in \{0,1\}$, define the proportion of variation in $Y$ explained by $W$ within arm $a$ as $R^2_{W;a} := \frac{\text{Var}_a\{E_a(Y \mid W)\}}{\text{Var}_a(Y)}$; the proportion of variation in $Y$ explained by $L$ within arm $a$ as $R^2_{L;a} := \frac{\text{Var}_a\{E_a(Y \mid L)\}}{\text{Var}_a(Y)}$; the proportion of additional variation in $Y$ explained by $L$ after accounting for $W$ within arm $a$ as $R^2_{L|W;a} := \frac{\text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\}}{\text{Var}_a(Y)}$; the proportion of residual variation in $Y$ within arm $a$ as $R^2_{r;a} := \frac{\text{Var}_a\{Y - E_a(Y \mid L, W)\}}{\text{Var}_a(Y)}$. Lemma 4.3.1.1 implies the following identify:

$$R^2_{W;a} + R^2_{L|W;a} + R^2_{r;a} = 1, \quad \forall a \in \{0,1\}. \quad (4.4)$$

### 4.3.2 Baseline Variables More Useful than Equally Prognostic Short-term Outcomes When Estimating Mean Primary Outcome in One Arm

In the following, if there exists an RAL estimator that achieves the asymptotic variance lower bound, we call it the efficient RAL estimator. The following corollary of Theorem 4.3.1.2 gives the ARE between the efficient RAL estimator and the unadjusted estimator for $E(Y \mid A = a)$.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

**Corollary 4.3.2.1.** Assume all conditions in Theorem 4.3.1.2 hold. For the estimand $E(Y \mid A = a)$ with $a \in \{0, 1\}$, the asymptotic relative efficiency between the efficient RAL estimator and the unadjusted estimator is

$$\text{ARE}(\text{efficient, unadjusted}) = \frac{1}{paqR^2_{W:a} + (py/p_l)R^2_{L|W:a} + R^2_{a}}. \quad (4.5)$$

In addition, given $q \in [0, 1]$, we have the following.

(i) (Only $W$ is prognostic.) If $R^2_{W:a} = q$ and $R^2_{L|W:a} = 0$, then

$$\text{ARE}(\text{efficient, unadjusted}) = \frac{1}{paq + 1 - q}. \quad (4.6)$$

(ii) (Only $L$ is prognostic.) If $R^2_{W:a} = 0$ and $R^2_{L:a} = q$, then

$$\text{ARE}(\text{efficient, unadjusted}) = \frac{1}{(py/p_l)q + 1 - q}. \quad (4.7)$$

We plot the ARE in (4.6) against $p_y$ and the ARE in (4.7) against $p_y/p_l$ in Figure 4.1 for $q = 0.25$ and $q = 0.10$, assuming $p_a = 0.5$. The observation is threefold. First, the precision gain from adjusting for a prognostic baseline variable decreases with a smaller proportion of the pipeline participants (i.e., larger $p_y$); the precision gain from adjusting for a prognostic short-term outcome decreases with a smaller proportion of the partially observed participants (i.e., larger $p_y/p_l$).

Second, when every participant has primary outcome observed (e.g., at the fi-
Figure 4.1: Asymptotic relative efficiency between the efficient RAL estimator and the unadjusted estimator for estimating $E(Y \mid A = a)$, assuming $p_a = 1/2$. Figure (a) is when only the baseline variable $W$ is prognostic with $R^2_{W:a} = q$. Figure (b) is when only the short-term outcome $L$ is prognostic $R^2_{L:a} = q$. 
nal analysis of a group sequential design with no dropouts), adjusting for prognostic baseline variables still improves estimation precision, but adjusting for prognostic short-term outcomes does not. This is seen by letting \( p_y = p_l = 1 \) in (4.6) and (4.7).

Third, for any given \((p_y, p_l)\), adjusting for a prognostic baseline variable alone always leads to larger precision gain than adjusting for an equally prognostic short-term outcome alone. To demonstrate this in a simple example, assume \( p_a = 1/2 \).

By the definition of the asymptotic equivalent reduction in sample size (RedSS) in (4.1), adjusting for prognostic \( W \) alone with \( R_{W,a}^2 = q \) yields \( \text{RedSS} = q(1 - p_y/2) \), and adjusting for prognostic \( L \) alone with \( R_{L,a}^2 = q \) yields \( \text{RedSS} = q(1 - p_y/p_l) \).

The ratio of the two RedSSs equals (assuming \( q > 0 \))

\[
\frac{\text{RedSS from adjusting for prognostic } W \text{ with } R_{W,a}^2 = q}{\text{RedSS from adjusting for prognostic } L \text{ with } R_{L,a}^2 = q} = \frac{1 - p_y/2}{1 - p_y/p_l}. \tag{4.8}
\]

\( r > 1 \) means that the sample size reduction from adjusting for a prognostic \( W \) is larger than that from adjusting for an equally prognostic \( L \); \( r < 1 \) means the opposite. Figure 4.2 plots \( r \) against \((p_l, p_y)\) under the constraint \( 0 < p_y \leq p_l \leq 1 \). For all \((p_l, p_y)\), \( r > 1 \). If \( p_y \geq 2/3 \), \( r \geq 2 \). This means that if at most \( 1/3 \) of the enrollees are in the pipeline, the sample size reduction from adjusting for a prognostic short-term outcome alone is at most half of that from adjusting for an equally prognostic baseline variable alone. Large \( p_l \) or small \( p_y \) results in \( r \) close to 1, which means that
adjusting for $W$ or adjusting for $L$ results in similar sample size reduction when there are many participants with $L$ observed or few with $Y$ observed. This usually occurs at early stages of a group sequential design, if the delay of the primary outcome is long and the delay of the short-term outcome is relatively short.

\[
\text{Figure 4.2: Contour plot of } r(p_l, p_y), \text{ where } r \text{ is the ratio of the reductions in sample size when only } W \text{ is prognostic with } R^2_{W,a} = q \text{ and when only } L \text{ is prognostic with } R^2_{L,a} = q. \text{ The estimand is } E(Y \mid A = a). \text{ Assume } p_a = 1/2.
\]
4.4 Precision Gain When Estimating Average Treatment Effect

4.4.1 Theoretical Limit on the Precision Gain

In this section, we assume that the treatment assignment is balanced (i.e., \( p_a = 1/2 \) for \( a \in \{0, 1\} \)). The following theorem, parallel to Theorem 4.3.1.2, gives the asymptotic variance lower bound for estimating the average treatment effect: \( E(Y \mid A = 1) - E(Y \mid A = 0) \).

**Theorem 4.4.1.1.** Under the assumptions in Theorem 4.3.1.2, the semiparametric lower bound on the asymptotic variance of all RAL estimators for \( E(Y \mid A = 1) - E(Y \mid A = 0) \) is

\[
\text{Var}\{E_1(Y \mid W) - E_0(Y \mid W)\} + \sum_{a \in \{0, 1\}} \frac{2}{p_l} \text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\} + \sum_{a \in \{0, 1\}} \frac{2}{p_y} \text{Var}_a\{Y - E_a(Y \mid L, W)\} \tag{4.9}
\]

The second and third terms in (4.9) also appear in (4.3), but the first term in (4.9) is unique. \( \text{Var}\{E_1(Y \mid W) - E_0(Y \mid W)\} \) can be interpreted as the treatment effect heterogeneity, because it characterizes the variation in treatment effect across different levels of \( W \). It equals 0 when there is no treatment effect heterogeneity.
We define the degree of treatment effect heterogeneity as its standardized version:

\[
\gamma = \frac{\text{Var}\{E_1(Y \mid W) - E_0(Y \mid W)\}}{\sum_{a \in \{0, 1\}} \text{Var}_a\{Y\}}.
\] (4.10)

By definition \(0 \leq \gamma \leq 2R^2_W\); \(\gamma = 0\) corresponds to no treatment effect heterogeneity.

Similar to the \(R\)-squared for one arm defined at the end of Section 4.3.1, we define the \(R\)-squared for both arms pooled together as follows. Define the pooled proportion of variation in \(Y\) explained by \(W\) as

\[
R^2_W := \frac{\sum_{a \in \{0, 1\}} \text{Var}_a\{E_a(Y \mid W)\}}{\sum_{a \in \{0, 1\}} \text{Var}_a(Y)};
\]

the pooled proportion of variation in \(Y\) explained by \(L\) as

\[
R^2_L := \frac{\sum_{a \in \{0, 1\}} \text{Var}_a\{E_a(Y \mid L)\}}{\sum_{a \in \{0, 1\}} \text{Var}_a(Y)};
\]

the pooled proportion of additional variation in \(Y\) explained by \(L\) after accounting for \(W\) as

\[
R^2_{L \mid W} := \frac{\sum_{a \in \{0, 1\}} \text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\}}{\sum_{a \in \{0, 1\}} \text{Var}_a(Y)};
\]

the pooled proportion of residual variation in \(Y\) as

\[
R^2_r := \frac{\sum_{a \in \{0, 1\}} \text{Var}_a\{Y - E_a(Y \mid L, W)\}}{\sum_{a \in \{0, 1\}} \text{Var}_a(Y)}.
\]
Lemma 4.3.1.1 yields
\[ R_W^2 + R_{L|W}^2 + R_r^2 = 1. \]

4.4.2 Treatment Effect Heterogeneity Impacts Precision Gain When Estimating Average Treatment Effect

Moore et al. (2011) discussed the impact of prognostic baseline variables on relative efficiency when estimating the average treatment effect. In this section we generalize their result to incorporate the short-term outcomes and the proportion of pipeline participants, and analyze how the treatment effect heterogeneity modified precision gain.

Corollary 4.4.2.1. Suppose all the assumptions in Theorem 4.3.1.2 hold. For the estimand \( E(Y \mid A = 1) - E(Y \mid A = 0) \), the asymptotic relative efficiency between the efficient RAL estimator and the unadjusted estimator is

\[
\text{ARE(efficient, unadjusted)} = \frac{1}{(p_y/2)\gamma + (p_y/p_l)R_{L|W}^2 + R_r^2}. \tag{4.11}
\]

In addition, given \( q \in [0, 1] \), we have the following.

(i) (Only \( W \) is prognostic.) If \( R_W^2 = q \) and \( R_{L|W}^2 = 0 \), then

\[
\text{ARE(efficient, unadjusted)} = \frac{1}{(p_y/2)\gamma + 1 - q}. \tag{4.12}
\]
(ii) (Only $L$ is prognostic.) If $R^2_W = 0$ and $R^2_L = q$, then

$$\text{ARE(efficient, unadjusted)} = \frac{1}{(p_y/p_l)q + 1 - q}. \quad (4.13)$$

Treatment effect heterogeneity modifies the precision gain from adjusting for prognostic baseline variables when estimating the average treatment effect. By (4.12), with fixed $p_y$ and $R^2_W$, smaller treatment effect heterogeneity results in more precision gain. The gain is maximal when the treatment effect is constant across different levels of $W$, which happens, for example, under the sharp null hypothesis. By (4.13), the precision gain from adjusting for short-term outcomes is not affected by the treatment effect heterogeneity, since $L$ is measured post-treatment.

The following corollary gives the ARE when adjusting for baseline variables in two extreme cases.

**Corollary 4.4.2.2.** Assume all the conditions in Theorem 4.3.1.2 hold. Assume $R^2_W = q$ for some $q \in [0, 1]$ and $R^2_L|W = 0$. For the estimand $E(Y \mid A = 1) - E(Y \mid A = 0)$, we have the following result regarding the asymptotic relative efficiency (ARE) between the efficient RAL estimator and the unadjusted estimator.

(i) (Purely predictive baseline variable.)

If $W$ satisfies $\text{Var}\{E(Y \mid W, A = 1) - E(Y \mid W, A = 0)\} > 0$ and $\text{Var}\{E(Y \mid
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

\[ W \} = 0, \text{ then} \]

\[
\text{ARE(efficient, unadjusted)} = \frac{1}{1 - (1 - p_y)q}. \quad (4.14)
\]

(ii) (Purely prognostic baseline variable.)

If \( W \) satisfies \( \text{Var}\{E(Y \mid W, A = 1) - E(Y \mid W, A = 0)\} = 0 \) and \( \text{Var}\{E(Y \mid W)\} > 0 \), then

\[
\text{ARE(efficient, unadjusted)} = \frac{1}{1 - q}. \quad (4.15)
\]

A baseline variable \( W \) satisfying the conditions in Corollary 4.4.2.2(i) is called purely predictive. A purely predictive baseline variable indicates who benefits from the treatment (i.e., explains the treatment effect heterogeneity), but it does not explain the variation in \( Y \) marginally. For a purely predictive baseline variable, \( \text{ARE} = 1 \) in (4.14) when \( p_y = 1 \). This means that adjusting for a purely predictive baseline variable results in no precision gain when there is no missing primary outcome. In extreme situations, a baseline variable can be perfectly correlated with the outcome within each arm but still contributes nothing to the precision gain. An example is where \( W, A \) are independently distributed as Bernoulli(0.5) and \( Y = 1(A = W) \). Hence, the prognostic value of \( W \) within each arm (measured by \( R^2_{W,A} \)) does not determine the precision gain when estimating the average treatment effect; the degree of treatment effect heterogeneity is also a key factor.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

A baseline variable $W$ satisfying the conditions in Corollary 4.4.2.2(ii) is called purely prognostic, because it explains the variation in $Y$ marginally and the treatment effect is constant across strata of $W$. For a purely prognostic baseline variable, ARE in (4.15) does not depend on $p_y$. This is different from (4.6) for estimating the mean primary outcome in one arm, where $p_y$ affects ARE. Such difference is seen in the simulation results in Section 4.6.4.

4.5 Semiparametric Efficient Estimator for covariate adjustment

One semiparametric estimator that could achieve the asymptotic variance lower bounds in (4.3) and (4.9) is the Targeted Minimum Loss Estimator (TMLE) (van der Laan and Gruber, 2012). This estimator builds on the ideas of Robins (2000), Bang and Robins (2005), and van der Laan and Rubin (2006), and is implemented in the R package ltmle (Schwab et al., 2015).

Our randomization and independent censoring assumptions imply that the treatment assignment and the missingness of $L$ and $Y$ are sequentially randomized conditional on the observed history. This combined with a positivity assumption imply that TMLE is consistent no matter whether the regression models for $L$ and $Y$ is correct, an implication of TMLE’s double robustness. When the regression models are correct, TMLE achieves the asymptotic variance lower bound.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

For the binary primary outcome in the MISTIE-II data set, let \( \hat{Q}(Z_i) \) denote the prediction of \( P(Y_i = 1) \) from logistic regression model with regressor \( Z_i \). We can estimate the \( R \)-squared for one arm by their empirical versions. For example, \( R^2_{W:a} \) can be estimated by
\[
\hat{R}^2_{W:a} := \frac{\sum_{i:A_i=a} \{ \hat{Q}(A_i, W_i) - \hat{Q}(A_i) \}^2}{\sum_{i:A_i=a} \{ Y_i - \hat{Q}(A_i) \}^2}.
\]
The pooled \( R \)-squared and the degree of treatment effect heterogeneity \( \gamma \) can be estimated in a similar way.

4.6 Simulation Based on MISTIE Trial

4.6.1 Trial Example: MISTIE-II

MISTIE-II is a Phase II randomized clinical trial evaluating a new surgical treatment for intracerebral hemorrhage (ICH). The treatment is called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation, abbreviated as MISTIE (Hanley et al., 2016). In the MISTIE-II dataset, the primary and short-term outcomes of each participant correspond to the Modified Rankin Scale (mRS) on degree of disability, measured at different times after treatment. The variables used in this chapter are listed in Table 4.1. The primary outcome \( Y \) is a binary indicator of a successful outcome (mRS \( \leq 3 \)) at 180 days after treatment. The short-term outcomes \( L^{(1)} \) and \( L^{(2)} \) are the indicators of mRS no larger than 4 at 30 and 90 days after treatment, respectively. The cutoff 4 for \( L^{(1)} \) and \( L^{(2)} \) was chosen, because
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

there were relatively few participants with mRS 3 or less at 30 or 90 days after treatment. The treatment assignment indicator $A$ denotes whether a patient is assigned to the new surgical treatment. Baseline variables $W^{(1)} - W^{(4)}$ are age at enrollment (dichotomized by 65 years), baseline NIHSS (NIH Stroke Scale total score, quantifying stroke-caused impairment), ICH (Intracerebral hemorrhage volume), and GCS (Glasgow Coma Scale, a neurological scale of the conscious state).

$L^{(1)}$ and $L^{(2)}$ are both used in generating hypothetical patient data, but only $L^{(1)}$ will be used in constructing the adjusted estimator.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W^{(1)}$</td>
<td>age (years), dichotomized by 65</td>
</tr>
<tr>
<td>$W^{(2)}$</td>
<td>baseline NIHSS, continuous</td>
</tr>
<tr>
<td>$W^{(3)}$</td>
<td>baseline ICH, continuous</td>
</tr>
<tr>
<td>$W^{(4)}$</td>
<td>baseline GCS, continuous</td>
</tr>
<tr>
<td>$A$</td>
<td>treatment indicator, 1 being MISTIE</td>
</tr>
<tr>
<td>$L^{(1)}$</td>
<td>mRS at 30 days, dichotomized by 4</td>
</tr>
<tr>
<td>$L^{(2)}$</td>
<td>mRS at 90 days, dichotomized by 4</td>
</tr>
<tr>
<td>$Y$</td>
<td>mRS at 180 days, dichotomized by 3</td>
</tr>
</tbody>
</table>

Table 4.1: The variables from the MISTIE-II dataset that are used in this chapter.

4.6.2 Group Sequential Design

With the average treatment effect $\theta$ defined as $\theta := E(Y|A = 1) - E(Y|A = 0)$, we wanted to test for $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. The goal of the design was to control Type I error at 0.025 and to have at least 80% power at $\theta = 0.122$, which is the sample average treatment effect calculated from the 100 participants in the
MISTIE-II data set using the unadjusted estimator.

We used Hampson and Jennison (2013) delayed response group sequential design as the design for our simulated trials; a brief description follows. With the exception of the final stage, each stage of the design consists of two analyses: an interim analysis and a decision analysis. At each interim analysis, a test statistic is calculated using the enrollees’ data (fully observed and pipeline participants included). Whether to stop the enrollment is determined based on this test statistic. If the enrollment is not stopped, the decision analysis of that stage is skipped and the trial proceeds to the interim analysis of the next stage (or the decision analysis of the final stage, if the next stage is the final stage). Otherwise, if the enrollment is stopped, the decision analysis of that stage occurs when all the pipeline participants become fully observed. At the decision analysis, a test statistic is calculated using data of the enrollees, who are all fully observed. Whether to reject the null hypothesis is determined based on this test statistic. Details are included in Appendix 4.8.2.1 in the Supplementary Material for completeness.

Denote by $K$ the total number of stages if the trial proceeds to the final stage. We chose $K = 5$ for the design used in our simulated trials. We generated continuous participant accrual in the simulation, with a constant enrollment rate of 140 participants per year based on the projection for the enrollment rate in the MISTIE-III trial (Hanley, 2016). Given a predetermined maximum sample size $n_{\text{max}}$, the timing of interim analyses was chosen such that for $1 \leq k \leq 4$, at the
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

$k$-th interim analysis there are $(k/K)n_{\max}$ fully observed participants. The critical regions for both analyses at each stage were calculated using the error spending approach presented in Appendix 4.8.2.2 in the Supplementary Material. We used $f(t) = 0.025 \min(t^2, 1)$ for Type I error spending and $g(t) = 0.2 \min(t^2, 1)$ for Type II error spending, where $t$ denotes the information time. The variance of the test statistic at each analysis was assumed known.

4.6.3 Data Generating Distributions

The MISTIE-II data set consists of 100 participants with fully observed $(W, A, L, Y)$. If we resampled directly from the 100 participants, each simulated participant would have deterministic treatment assignment given his/her baseline variable, especially when the baseline variables are categorical with many levels or continuous, as is the case here. Such simulation fails to generate data with the randomization assumption. To overcome this, we used a resampling-based algorithm to generate hypothetical participants using the MISTIE-II data set. This algorithm generates simulated trials with the following properties:

(i) The treatment assignment is independent of baseline variables.

(ii) The average treatment effect of the simulated data equals the unadjusted estimate using the MISTIE-II data set.

(iii) The relative efficiency between TMLE and the unadjusted estimator is similar
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

to that of the MISTIE-II data set.

Details of the algorithm are in Appendix 4.8.3 in the Supplementary Material. Briefly, the algorithm ensures property (i) by adding a “twin” with identical baseline variable and opposite treatment assignment to each participant in the data set. Then it simulates $L$ and $Y$ for each “twin” using probabilistic methods to ensure properties (ii) and (iii). Properties (ii) and (iii) enable us to evaluate the power and efficiency gain from covariate adjustment in a realistic setting.

We used the above algorithm and its three variations to generate data under 4 prognostic settings. $\text{progn}_{W,L}$ is the algorithm “as-is”, where both $W$ and $L$ are prognostic. $\text{progn}_W$ is the above algorithm followed by resetting $L$ as random draws from its empirical distribution, in which case only $W$ is prognostic. Similarly, $\text{progn}_L$ resets $W$, in which case only $L$ is prognostic. $\text{progn}_\theta$ resets both $W$ and $L$, in which case none of them is prognostic.

The above algorithm generates hypothetical participants with average treatment effect $\theta = 0.122$, as desired under $H_1$. To generate participants under $H_0$, the algorithm further sets each $A_i$ to be a random draw from Bernoulli(0.5). Hence, under $H_1$ there is slight treatment effect heterogeneity with $\gamma = 0.018$, and under $H_0$ there is no treatment effect heterogeneity ($\gamma = 0$).

To assess the performance of semiparametric estimators under model misspecification, we generated the trials using all the variables listed in Table 4.1, but only used $W = (W^{(1)}, W^{(4)})$ and $L = L^{(1)}$ in fitting TMLE. Table 4.2 lists the $R$-squared
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

of the simulated data under $H_0$ and $H_1$, estimated using formula in Section 4.5.

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th></th>
<th>$H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a = 0$</td>
<td>$a = 1$</td>
<td>combined</td>
</tr>
<tr>
<td>$R^2_{W,a}$</td>
<td>0.36</td>
<td>0.34</td>
<td>0.35</td>
</tr>
<tr>
<td>$R^2_{L,a}$</td>
<td>0.31</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>$R^2_{L,W,a}$</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>$R^2_{r,a}$</td>
<td>0.57</td>
<td>0.58</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 4.2: Empirical $R$-squared for the data generating distribution based on MISTIE-II. Columns “combined” are the pooled $R$-squared.

4.6.4 Results on Relative Efficiency

We conducted simulation studies to validate our theory on asymptotic relative efficiency. To introduce model misspecification in the simulation, we only adjusted for the baseline variables $(W^{(1)}, W^{(4)})$ and the short-term outcome $L^{(1)}$. We simulated the delayed response group sequential design in Section 4.6.2 with $K = 5$, $n_{max} = 500$, enrollment rate 140 participants per year, equally spaced interim analyses, $d_L = 30$ days, and $d_Y = 180$ days. We assumed no early stopping here, so that each simulated trial proceeds to the final stage and produces the unadjusted estimator and TMLE at all 4 interim analyses and all 5 decision analyses for the three estimands: $E(Y | A = 1)$, $E(Y | A = 0)$, and $E(Y | A = 1) - E(Y | A = 0)$. We simulated 10,000 trials under each of the 4 prognostic settings.

Table 4.3 lists the theoretical relative efficiency (RE) approximated by Corollary 4.3.2.1 and Corollary 4.4.2.1, and the empirical relative efficiency from simulation
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

under $H_0$. The corresponding table for $H_1$ is included in Appendix 4.8.5 in the Supplementary Material. It is mostly similar to Table 4.3 except for some early stages, at which there is inconsistency between the theory and simulation results. This inconsistency is likely due to small sample size, misspecified models in fitting TMLE, and the fact that in the simulation we are estimating the treatment probability and the censoring model. For most settings, RE predicted by the theory is similar to that from the simulation, indicating that the asymptotic theory approximates the precision gain quite well, even in the presence of mild model misspecification.

A comparison between $\text{progn}_W$ and $\text{progn}_L$ shows that adjusting for $W$ results in much larger RE than adjusting for $L$, which is partly explained by the contrast between (4.6) and (4.7) and partly due to the larger prognostic value in $W$ than in $L$. Comparing RE under $\text{progn}_L$ at interim analysis to that at decision analysis shows that $L$ improves estimation efficiency only when there are pipeline participants. In addition, we see attenuated efficiency gain from $W$ and $L$ at later stages.

From interim analysis 1 to 4, the proportion of pipeline participants $p_y$ gets larger. As predicted by Corollary 4.4.2.2, the simulated RE under $\text{progn}_W$ for estimating $E(Y \mid A = a)$ decreases at later interim analysis, while the simulated RE under $\text{progn}_W$ for estimating $E(Y \mid A = 1) - E(Y \mid A = 0)$ is mostly constant at all interim analyses. This is because there is no treatment effect heterogeneity under $H_0$. 

85
Table 4.3: Comparison of the relative efficiency (RE) predicted by Corollary 4.3.2.1 and Corollary 4.4.2.1, and the relative efficiency from simulated trials under $H_0$. In the simulation we used a group sequential design with 5 stages, $n_{\text{max}} = 500$, enrollment rate 140 participants per year, equally spaced interim analyses, $d_L = 30$ days, and $d_Y = 180$ days. The simulated RE were based on 10,000 simulated trials.
4.6.5 Results on Sample Size Reduction

We also conducted simulation studies of the delayed response group sequential design with early stopping under the four prognostic settings using both the TMLE and the unadjusted estimator, to assess the reduction in maximum and expected sample sizes. The maximum sample size $n_{\text{max}}$ was chosen separately for each estimator under each prognostic setting, so that the design always has approximately 80% power. We adjusted for baseline variables $(W^{(1)}, W^{(4)})$ and short-term outcome $L^{(1)}$ when fitting TMLE. 10,000 trials were simulated for each estimator under each prognostic setting. The number of accrued participants, Type I and Type II errors to spend, and the critical region at each stage are listed in Appendix 4.8.4 in the Supplementary Material.

Table 4.4 lists the the $n_{\text{max}}$, Type I error, power, and the expected sample size (ESS) for $H_0$ and $H_1$ under each prognostic setting using each estimator. For each row in the table, Type I error rate is controlled at 0.025 and power is approximately 80% due to the way $n_{\text{max}}$ was chosen. ESS is calculated as the number of enrollees when the trial stops, averaging over 10,000 simulations. The performance of the unadjusted estimator is the same under all prognostic settings and is summarized in the first row.

Compared to the unadjusted estimator, TMLE reduces both the maximum sample size and the expected sample size significantly when the baseline variable is prognostic. The reduction from adjusting for prognostic short-term outcome is
limited (comparing $\text{progn}_L$ to $\text{progn}_\theta$ or $\text{progn}_{W:L}$ to $\text{progn}_W$)—only the expected sample sizes are slighted reduced. When neither $W$ nor $L$ is prognostic, the performance of the simulated trial using TMLE is comparable to the trial performance using the unadjusted estimator. Hence, TMLE is guaranteed to perform as well or better than the unadjusted estimator.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Progn. set.</th>
<th>$n_{\text{max}}$</th>
<th>Type I error</th>
<th>Power</th>
<th>ESS $H_0$</th>
<th>ESS $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>all</td>
<td>480</td>
<td>0.025</td>
<td>0.81</td>
<td>318</td>
<td>383</td>
</tr>
<tr>
<td>TMLE</td>
<td>$\text{progn}_{W:L}$</td>
<td>300</td>
<td>0.025</td>
<td>0.80</td>
<td>224</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>$\text{progn}_W$</td>
<td>300</td>
<td>0.025</td>
<td>0.80</td>
<td>228</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td>$\text{progn}_L$</td>
<td>480</td>
<td>0.025</td>
<td>0.80</td>
<td>309</td>
<td>374</td>
</tr>
<tr>
<td></td>
<td>$\text{progn}_\theta$</td>
<td>480</td>
<td>0.025</td>
<td>0.81</td>
<td>321</td>
<td>384</td>
</tr>
</tbody>
</table>

Table 4.4: The maximum sample size ($n_{\text{max}}$), Type I error, Power, and expected sample size (ESS) under $H_0$ and $H_1$ for each estimator to achieve 80% power under each prognostic setting (Progn. set.). Columns (except for $n_{\text{max}}$) are averaged from 10,000 simulated trials.

4.7 Remarks

If administrative censoring is the only source of missingness, the TMLE and the unadjusted estimator are both consistent estimators. In a real trial, however, dropouts are likely to happen, which is another source of missingness in $L$ and $Y$. If the dropout status only depends on the observed history (missing at random), TMLE is still consistent under either correctly modeled dropout distribution or correctly modeled outcome regression, but the unadjusted estimator can be biased
and have inflated Type I error in this case.

In Section 4.3 and Section 4.4 we have focused on the precision gain when using the efficient RAL estimator. In practice, with an incorrectly specified model the estimator may be inefficient. A future research area would be to establish a corresponding theory for the inefficient semiparametric estimators that use misspecified working models.

The \texttt{progn4} columns in Table 4.3 show that at early stages with small sample size, there is a non-negligible difference between the theoretical RE and the simulated RE: theoretically adjusting for noise doesn’t hurt RE asymptotically, and in the simulation it hurts slightly with finite sample size. Such a difference vanishes as the sample size gets larger. One alternative would be to use the unadjusted estimator at early analyses, and use the adjusted estimator after a relatively large number of participants has been accrued. The specific rule for when to switch the choice of estimator would need to be predetermined in the study protocol, and would need to be accounted for in estimating the covariance of test statistics at different stages.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

4.8 Appendix

4.8.1 Causal Framework

For a generic participant, define the potential outcome $Y^a$ as the primary outcome that would have been observed if the participant had been assigned treatment $A = a$ for $a \in \{0, 1\}$.

Define the average treatment effect $\theta$ to be

$$\theta := E(Y^1) - E(Y^0).$$  \hfill (4.16)

We make the following assumptions.

**Assumption 4.8.1.1** (Consistency). $Y = Y^1 A + Y^0 (1 - A)$.

**Assumption 4.8.1.2** (Randomization). $(W, Y^1, Y^0) \perp\!\!\!\perp A$.

**Assumption 4.8.1.3** (Missing completely at random). $(W, A, L, Y^1, Y^0) \perp\!\!\!\perp (C^L, C^Y)$.

**Assumption 4.8.1.4** (Monotone censoring). $C^L \geq C^Y$.

Assumption 4.8.1.1 relates the potential outcome to the observed outcome. Assumption 4.8.1.2 holds in any randomized design. Assumptions 4.8.1.3 and 4.8.1.4 hold if the only source of missingness in $L$ and $Y$ is administrative censoring.

Under Assumptions 4.8.1.1-4.8.1.4, the average treatment effect is identifiable
from the distribution of the observed data:

\[
E(Y_1) - E(Y_0) = E\{E\{E(Y \mid L, W, A = 1, C^Y = 1) \mid W, C^L = 1, A = 1\} \mid W, C^L = 1, A = 0\}
\]

(4.17)

In (4.17), for \( a \in \{0, 1\} \) the inner expectation is \( E_{Y \mid L,W,A=a,C^Y=1}(\cdot) \), the mean of \( Y \) for the fully observed participants in arm \( A = a \), within strata of \( W \) and \( L \); the middle expectation is \( E_{L \mid W,A=a,C^L=1}(\cdot) \), which integrates over the distribution of \( L \) for the partially observed participants in arm \( A = a \), within strata of \( W \); the outer expectation is \( E_W[\cdot] \), which integrates over the marginal distribution of \( W \) in arm \( A = a \). Proof for (4.17) is given as follows.

**Proof.** It suffices to show that for each \( a \in \{0, 1\} \) we have

\[
E(Y^a) = E\{E\{E(Y \mid L, W, A = a, C^Y = 1) \mid W, A = 1, C^L = 1\} \mid W, A = 1\}. \quad (4.18)
\]

Assumptions 4.8.1.1 and 4.8.1.2 imply

\[
E(Y^a) = E(Y^a \mid A = a) = E(Y \mid A = a). \quad (4.19)
\]

Using the law of iterated expectation twice, we deduce

\[
E(Y \mid A = a) = E\{E\{E(Y \mid L, W, A = a) \mid W, A = a\} \mid W, A = a\}. \quad (4.20)
\]
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Let \( f(L, W) := E(Y \mid L, W, A = a) \). By Assumption 4.8.1.3 we have

\[
E(f(L, W) \mid A = a) = E(f(L, W) \mid A = a, C^L = 1),
\]

(4.21)

and

\[
E(Y \mid L, W, A = a) = E(Y \mid L, W, A = a, C^L = 1, C^Y = 1).
\]

(4.22)

Equations (4.19)-(4.22) together yield

\[
E(Y^a) = E[E\{E(Y \mid L, W, A = a, C^L = 1, C^Y = 1) \mid W, A = 1, C^L = 1\}].
\]

(4.23)

Equation (4.23) and Assumption 4.8.1.4 yield (4.18). This completes the proof.

\[\square\]

4.8.2 Details of Group Sequential Design

4.8.2.1 Multiple Testing Procedure

In the simulation studies in Section 4.6 we use the delayed response group sequential test from Hampson and Jennison (2013) with \( K \) stages. Let \( \theta \) be the true average treatment effect. Given \( \alpha, \beta \in (0, 1) \), the design goal is to control Type I error rate at level \( \alpha \) and power \( 1 - \beta \) at \( \theta = \delta \). Such a group sequential test can terminate recruitment at an interim analysis, and if early stopping happens it
waits \( d_Y \) till all pipeline participants become fully observed before conducting a
decision analysis to reject or accept \( H_0 \). For stage \( k \), denote by \( S_k \) and \( \bar{S}_k \)
the test
statistics at interim analysis and decision analysis, respectively; denote by \( u_k \) and \( l_k \)
the efficacy and futility boundaries for terminating recruitment at interim analysis,
and \( c_k \) the critical value for the corresponding decision analysis. The following ex-
cerpt from Hampson and Jennison (2013, Figure 1) illustrates the delayed response
group sequential test:

\[
\text{“At interim analysis } k = 1, \ldots, K - 2, \\
\text{if } S_k \leq l_k \text{ or } S_k \geq u_k \text{ stop recruitment and proceed to decision analysis } k; \\
\text{otherwise continue recruitment and proceed to interim analysis } k + 1.
\]

\[
\text{At interim analysis } K - 1, \\
\text{if } S_{K-1} \leq l_{K-1} \text{ or } S_{K-1} \geq u_{K-1} \text{ stop recruitment and proceed to decision analysis } K - 1; \\
\text{otherwise complete recruitment and proceed to decision analysis } K.
\]

At decision analysis \( K \),

\[
\text{if } \bar{S}_K \geq c_k \text{ reject } H_0; \\
\text{if } \bar{S}_K < c_k \text{ accept } H_0.”
\]

4.8.2.2 Testing Boundaries

Given the Type I error spending function \( f(\cdot) \) and the Type II error spending
function \( g(\cdot) \), which are non-decreasing with \( f(0) = g(0) = 0 \) and \( f(t) = \alpha \) and
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

\[ g(t) = \beta \text{ for } t \geq 1. \] The maximum information level \( I_{\text{max}} \) is chosen depending on the power goal and the error spending functions. Denote by \( I_k \) and \( \tilde{I}_k \) the information levels at the \( k \)-th interim analysis and decision analysis, respectively. Denote by \( C_k = (l_k, u_k) \) the critical region at interim analysis \( k, 1 \leq k \leq K. \)

The testing boundaries \( u_k, l_k, \) and \( c_k \) are calculated following (12)-(15) in Hampson and Jennison (2013, Section 4.1.1). \( u_1 \) and \( l_1 \) are solutions to

\[ P(S_1 \geq u_1; \theta = 0) = f(I_1/I_{\text{max}}) \quad \text{and} \quad P(S_1 \leq l_1; \theta = \delta) = g(I_1/I_{\text{max}}). \]

For \( 2 \leq k \leq K - 1, u_k \) is the solution to

\[ P(S_1 \in C_1, \ldots, S_{k-1} \in C_{k-1}, S_k \geq u_k; \theta = 0) = f(I_k/I_{\text{max}}) - f(I_{k-1}/I_{\text{max}}), \]

and \( l_k \) is the solution to

\[ P(S_1 \in C_1, \ldots, S_{k-1} \in C_{k-1}, S_k \leq l_k; \theta = \delta) = g(I_k/I_{\text{max}}) - g(I_{k-1}/I_{\text{max}}). \]

For \( 1 \leq k \leq K - 1, \) the critical value \( c_k \) is the solution to

\[ P(S_1 \in C_1, \ldots, S_{k-1} \in C_{k-1}, S_k \geq u_k, \tilde{S}_k < c_k; \theta = 0) \]

\[ = P(S_1 \in C_1, \ldots, S_{k-1} \in C_{k-1}, S_k \leq l_k, \tilde{S}_k \geq c_k; \theta = 0). \]
And the critical value $c_K$ for the last stage is the solution to

$$P(S_1 \in C_1, \ldots, S_{K-1} \in C_{K-1}, \tilde{S}_K \geq c_K; \theta = 0) = \alpha - f(\mathcal{I}_{K-1}/\mathcal{I}_{\max}).$$

4.8.3 Resampling-based Algorithm to Simulate Trial Data

For the data generating distribution to have the three properties in Section 4.6.3, we design the algorithm to generate a simulated trial of $n$ independent and identically distributed samples from the 100 participants in MISTIE-II data set as follows.

**Step 1: Make treatment assignment independent of baseline variables.** We augment the data with a hypothetical twin for each participant. A twin has identical baseline variables as the original participant, but opposite treatment assignment. We sample $n$ independent and identically distributed participants from this augmented data set (100 original participants and 100 twins), which ensures property (i).

**Step 2: Calibrate treatment effect in simulated data.** We fit logistic regression models for $L^{(1)}$ on $(W, A)$, for $L^{(2)}$ on $(W, A, L^{(1)})$, and for $Y$ on $(W, A, L^{(1)}, L^{(2)})$, using the 100 participants in MISTIE-II data set. The preliminary $L_i$ and $Y_i$ of each newly added twin in the $n$ samples are predicted based on these logistic regression models by truncating the predicted success probability at 0.5. Then, for each newly added twin in the $n$ samples, with probability 0.03 we reset $Y_i$ to equal $A_i$. This re-
setting step increases the treatment effect of the augmented data to 0.122, matching that of the original data. Now, sampling with replacement from this augmented data set ensures properties (i) and (ii).

**Step 3: Calibrate relative efficiency of TMLE.** For \( a \in \{0, 1\} \), we empirically estimate \( p_a := P(Y = 1 \mid A = a) \) using the original data. Then, for each newly added twin in the \( n \) samples, with probability 0.164 we reset \( Y_i \) by a realization of Bernoulli random variable with success probability \( p_{A_i} \). This resetting step adds random noise to reduce the prognostic value in \( W \) and \( L \). Meanwhile it keeps the treatment effect of the augmented data unchanged. Table 4.5 displays the relative efficiency between TMLE and the unadjusted estimator in the original data, in the augmented data before Step 3, and in the augmented data after Step 3, when the trial sample size and timing are as in the first simulation study in Section 4.6.4.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Original data</th>
<th>Augmented data before step 3</th>
<th>Augmented data after step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>stage 1</td>
<td>1.54</td>
<td>1.74</td>
<td>1.52</td>
</tr>
<tr>
<td>stage 2</td>
<td>1.53</td>
<td>1.74</td>
<td>1.55</td>
</tr>
<tr>
<td>stage 3</td>
<td>1.54</td>
<td>1.75</td>
<td>1.55</td>
</tr>
<tr>
<td>stage 4</td>
<td>1.55</td>
<td>1.76</td>
<td>1.55</td>
</tr>
<tr>
<td>stage 5</td>
<td>1.55</td>
<td>1.75</td>
<td>1.55</td>
</tr>
</tbody>
</table>

**Table 4.5:** Relative efficiency between TMLE and the unadjusted estimator, where the trial sample size and timing are as in the first simulation study in Section 4.6.4.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

4.8.4 Design Example

Table 4.6 lists the sample size and analysis timing of the group sequential designs with $n_{\text{max}} = 480$ and $n_{\text{max}} = 300$ used in Section 4.6.4. For $1 \leq k \leq 4$, Analysis $k$ indicates interim analysis at stage $k$ and $k^*$ indicates the corresponding decision analysis if enrollment is early stopped at that stage. Analysis $5^*$ indicates the final decision analysis. Note that there is not interim analysis for the final stage. Fully observed participants are those with $W, L, Y$ observed; partially observed participants are those with $W, L$ but not $Y$ observed. The three groups of participants listed in Table 4.6 are inclusive of all enrollees and mutually exclusive.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>1</th>
<th>1*</th>
<th>2</th>
<th>2*</th>
<th>3</th>
<th>3*</th>
<th>4</th>
<th>4*</th>
<th>5*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_{\text{max}} = 480$</td>
<td>1.2</td>
<td>1.7</td>
<td>1.9</td>
<td>2.4</td>
<td>2.6</td>
<td>3.0</td>
<td>3.2</td>
<td>3.7</td>
<td>3.9</td>
</tr>
<tr>
<td># Fully observed (W, L, Y)</td>
<td>96</td>
<td>165</td>
<td>192</td>
<td>261</td>
<td>288</td>
<td>357</td>
<td>384</td>
<td>453</td>
<td>480</td>
</tr>
<tr>
<td># Partially observed (W, L)</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td># Pipeline with only W observed</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time (year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_{\text{max}} = 300$</td>
<td>0.9</td>
<td>1.4</td>
<td>1.4</td>
<td>1.8</td>
<td>1.8</td>
<td>2.3</td>
<td>2.2</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td># Fully observed (W, L, Y)</td>
<td>60</td>
<td>129</td>
<td>120</td>
<td>189</td>
<td>180</td>
<td>249</td>
<td>240</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td># Partially observed (W, L)</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td># Pipeline with only W observed</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.6: Analysis time and sample size at each interim and decision analysis for group sequential designs with $n_{\text{max}} = 480$ and $n_{\text{max}} = 300$. For $1 \leq k \leq 4$, Analysis $k$ indicates interim analysis and $k^*$ indicates the corresponding decision analysis if enrollment is early stopped. Analysis $5^*$ indicates the final decision analysis.

Table 4.7 lists the Type I error and Type II error spent and the testing boundaries at each stage for the designs under prognostic setting $\text{progn}_{W,L}$, where $n_{\text{max}} = 480$. 

97
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

for the unadjusted estimator and $n_{\text{max}} = 300$ for TMLE. We use error spending functions $f(t) = 0.025 \min(t^2, 1)$ and $g(t) = 0.2 \min(t^2, 1)$. Boundaries $u_k, l_k, c_k$ are computed following Section 4.8.2.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TMLE $n_{\text{max}} = 300$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I error to spend ($\times 10^{-3}$)</td>
<td>0.9</td>
<td>3.1</td>
<td>5.0</td>
<td>7.5</td>
<td>8.4</td>
</tr>
<tr>
<td>Type II error to spend ($\times 10^{-3}$)</td>
<td>7.6</td>
<td>24.6</td>
<td>40.4</td>
<td>60.3</td>
<td>67.2</td>
</tr>
<tr>
<td>Efficacy boundary at interim analysis ($u_k$)</td>
<td>3.12</td>
<td>2.72</td>
<td>2.48</td>
<td>2.26</td>
<td>-</td>
</tr>
<tr>
<td>Futility boundary at interim analysis ($l_k$)</td>
<td>-0.66</td>
<td>0.32</td>
<td>1.04</td>
<td>1.68</td>
<td>-</td>
</tr>
<tr>
<td>Critical value at decision analysis ($c_k$)</td>
<td>1.65</td>
<td>1.76</td>
<td>1.91</td>
<td>2.03</td>
<td>1.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unadjusted estimator $n_{\text{max}} = 480$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I error to spend ($\times 10^{-3}$)</td>
<td>1.0</td>
<td>3.0</td>
<td>4.9</td>
<td>7.3</td>
<td>8.8</td>
</tr>
<tr>
<td>Type II error to spend ($\times 10^{-3}$)</td>
<td>8.2</td>
<td>24.4</td>
<td>39.3</td>
<td>58.0</td>
<td>70.2</td>
</tr>
<tr>
<td>Efficacy boundary at interim analysis ($u_k$)</td>
<td>3.08</td>
<td>2.72</td>
<td>2.47</td>
<td>2.25</td>
<td>-</td>
</tr>
<tr>
<td>Futility boundary at interim analysis ($l_k$)</td>
<td>-1.08</td>
<td>-0.04</td>
<td>0.74</td>
<td>1.44</td>
<td>-</td>
</tr>
<tr>
<td>Critical value at decision analysis ($c_k$)</td>
<td>1.25</td>
<td>1.50</td>
<td>1.72</td>
<td>1.93</td>
<td>2.05</td>
</tr>
</tbody>
</table>

**Table 4.7:** Error spending and boundaries for designs in Section 4.6.4 under prognostic setting $\text{progn}_{W,L}$.

### 4.8.5 Simulation Results for Relative Efficiency Under $H_1$

See Table 4.8 for the simulation results for relative efficiency under $H_1$. This is supplementary to the simulation results presented in Table 4.3 of Section 4.6.4.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Table 4.8: Comparison of relative efficiency implied by theory in Section 4.3 and relative efficiency calculated from simulated trials, under $H_1$ of the data generating distribution based on MISTIE-II. We consider a group sequential design with 5 stages, $n_{max} = 500$, enrollment rate 140 participants per year, equally spaced interim analyses, $d_L = 30$ days, and $d_Y = 180$ days. The simulation results are based on 10,000 simulated trials.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>RE approximated by Theory</th>
<th>RE from Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E(Y</td>
<td>A = 1) - E(Y</td>
</tr>
<tr>
<td></td>
<td>$E(Y</td>
<td>A = 0)$</td>
</tr>
<tr>
<td></td>
<td>$E(Y</td>
<td>A = 1)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interim</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.71 1.58 1.12 1.00</td>
<td>2.44 1.51 1.58 0.96</td>
</tr>
<tr>
<td>2</td>
<td>1.68 1.57 1.07 1.00</td>
<td>1.96 1.55 1.26 0.98</td>
</tr>
<tr>
<td>3</td>
<td>1.67 1.57 1.05 1.00</td>
<td>1.84 1.57 1.18 0.99</td>
</tr>
<tr>
<td>4</td>
<td>1.66 1.57 1.04 1.00</td>
<td>1.74 1.57 1.14 1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.64 1.57 1.00 1.00</td>
<td>1.53 1.55 0.99 0.99</td>
</tr>
<tr>
<td>2</td>
<td>1.64 1.57 1.00 1.00</td>
<td>1.53 1.56 0.99 1.00</td>
</tr>
<tr>
<td>3</td>
<td>1.64 1.57 1.00 1.00</td>
<td>1.53 1.58 1.00 1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.64 1.57 1.00 1.00</td>
<td>1.53 1.57 1.00 1.00</td>
</tr>
<tr>
<td>5</td>
<td>1.64 1.57 1.00 1.00</td>
<td>1.52 1.56 1.00 1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decision</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.67 1.48 1.14 1.00</td>
<td>2.11 1.47 1.56 0.97</td>
</tr>
<tr>
<td>2</td>
<td>1.56 1.41 1.08 1.00</td>
<td>1.67 1.42 1.27 0.99</td>
</tr>
<tr>
<td>3</td>
<td>1.52 1.38 1.06 1.00</td>
<td>1.56 1.39 1.20 0.99</td>
</tr>
<tr>
<td>4</td>
<td>1.49 1.36 1.04 1.00</td>
<td>1.47 1.38 1.15 0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.41 1.30 1.00 1.00</td>
<td>1.29 1.32 0.99 0.99</td>
</tr>
<tr>
<td>2</td>
<td>1.41 1.30 1.00 1.00</td>
<td>1.29 1.33 1.00 0.99</td>
</tr>
<tr>
<td>3</td>
<td>1.41 1.30 1.00 1.00</td>
<td>1.29 1.32 1.00 1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.41 1.30 1.00 1.00</td>
<td>1.30 1.33 1.00 1.00</td>
</tr>
<tr>
<td>5</td>
<td>1.41 1.30 1.00 1.00</td>
<td>1.30 1.33 1.00 1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interim</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.31 1.27 1.11 1.00</td>
<td>1.83 1.22 1.56 0.97</td>
</tr>
<tr>
<td>2</td>
<td>1.26 1.23 1.06 1.00</td>
<td>1.52 1.21 1.26 0.98</td>
</tr>
<tr>
<td>3</td>
<td>1.24 1.22 1.04 1.00</td>
<td>1.40 1.20 1.18 0.99</td>
</tr>
<tr>
<td>4</td>
<td>1.22 1.21 1.03 1.00</td>
<td>1.33 1.20 1.14 0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.18 1.18 1.00 1.00</td>
<td>1.16 1.16 0.99 0.99</td>
</tr>
<tr>
<td>2</td>
<td>1.18 1.18 1.00 1.00</td>
<td>1.16 1.16 0.99 0.99</td>
</tr>
<tr>
<td>3</td>
<td>1.18 1.18 1.00 1.00</td>
<td>1.16 1.17 0.99 1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.18 1.18 1.00 1.00</td>
<td>1.15 1.16 1.00 1.00</td>
</tr>
<tr>
<td>5</td>
<td>1.18 1.18 1.00 1.00</td>
<td>1.15 1.16 1.00 1.00</td>
</tr>
</tbody>
</table>
4.8.6 Proof of Main Theorems

4.8.6.1 Proof of Lemma 4.3.1.1

Proof. Adding and subtracting terms, we have

\[ \text{Var}_a(Y) = \text{Var}_a \left\{ Y - E_a(Y \mid L, W) \right\} + \left\{ E_a(Y \mid L, W) - E_a(Y \mid W) \right\} + E_a(Y \mid W). \]

(4.24)

So for proving (4.2), it suffices to establish the following:

\[ \text{Cov}_a\{ Y - E_a(Y \mid L, W), E_a(Y \mid L, W) - E_a(Y \mid W) \} = 0, \]

(4.25)

\[ \text{Cov}_a\{ Y - E_a(Y \mid L, W), E_a(Y \mid W) \} = 0, \]

(4.26)

\[ \text{Cov}_a\{ E_a(Y \mid L, W) - E_a(Y \mid W), E_a(Y \mid W) \} = 0, \]

(4.27)

where \( \text{Cov}_a \) denotes the conditional covariance given \( A = a \).

First, we have

\[ \text{Cov}_a\{ Y - E_a(Y \mid L, W), E_a(Y \mid L, W) - E_a(Y \mid W) \} \]

\[ = E_a\{ YE_a(Y \mid L, W) - E_a(Y \mid L, W)^2 - YE_a(Y \mid W) + E_a(Y \mid L, W)E_a(Y \mid W) \}. \]

(4.28)
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

By Lemma 4.8.7.1 with $X = Z = (L, W)$, we have

$$E_a\{YE_a(Y \mid L, W)\} = E_a\{E_a(Y \mid L, W)^2\}.$$  \hfill (4.29)

By Lemma 4.8.7.1 with $X = W$ and $Z = (L, W)$, we have

$$E_a\{YE_a(Y \mid W)\} = E_a\{E_a(Y \mid L, W)E_a(Y \mid W)\}.$$  \hfill (4.30)

Equations (4.28), (4.29), and (4.30) imply (4.25).

Second, we have

$$\text{Cov}_a\{Y - E_a(Y \mid L, W), E_a(Y \mid W)\} = E_a\{YE_a(Y \mid W) - E_a(Y \mid L, W)E_a(Y \mid W)\}.$$  \hfill (4.31)

Equations (4.31) and (4.30) imply (4.26).

Third, since $E_a(Y \mid L, W) - E_a(Y \mid W)$ has expectation zero, we have

$$\text{Cov}_a\{E_a(Y \mid L, W) - E_a(Y \mid W), E_a(Y \mid W)\} = E_a\{E_a(Y \mid L, W)E_a(Y \mid W) - E_a(Y \mid W)^2\}.$$  \hfill (4.32)

In Lemma 4.8.7.1, letting $X = Z = W$ and replacing $Y$ in the lemma by $E(Y \mid L, W)$ implies

$$E_a\{E_a(Y \mid L, W)E_a(Y \mid W)\} = E_a\{E_a(Y \mid W)^2\}.$$  \hfill (4.33)

Equations (4.32) and (4.33) imply (4.27).
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

This proves (4.2). Equation (4.4) follows immediately from (4.2) and the definition of $R_W^2$ and $R_r^2$. This completes the proof for Lemma 4.3.1.1.

\[ \square \]

### 4.8.6.2 Proof of Theorem 4.3.1.2

**Proof.** Equations (4.19) and (4.23) yield

\[
E(Y \mid A = a) = E[E\{E(Y \mid L, W, A = a, C^L = 1, C^Y = 1) \mid W, A = 1, C^L = 1\}].
\] (4.34)

Treating the missingness indicators $C^L$ and $C^Y$ as binary treatments, the right-hand side of (4.34) becomes the average of outcome $Y$ under time dependent treatment assignment: $A = a, C^L = 1, C^Y = 1$. Because there is no measurement made between $A$ and $C^L$, we can combine the two as a single treatment $\tilde{A}$, with $\tilde{A} = 1$ if and only if $A = a$ and $C^L = 1$. Equation (4.34) becomes

\[
E(Y \mid A = a) = E[E\{E(Y \mid L, W, \tilde{A} = 1, C^Y = 1) \mid W, \tilde{A} = 1\}].
\] (4.35)

Using the fact that $L$ is binary-valued, by equations (24) and (28) in Rosenblum and van der Laan (2011) or Theorem 1 in van der Laan (2010), the efficient influence function for (4.35) is

\[
D(W, \tilde{A}, L, C^Y, Y) = D_0(W) + D_1(W, \tilde{A}, L) + D_2(W, \tilde{A}, L, C^Y, Y),
\] (4.36)
where

\[
D_0(W) = E(Y \mid W, \tilde{A} = 1, C^Y = 1) - E(Y \mid \tilde{A} = 1),
\]

(4.37)

and

\[
D_1(W, \tilde{A}, L) = \frac{1(\tilde{A} = 1)\{E(Y \mid L, W, \tilde{A} = 1, C^Y = 1) - E(Y \mid W, \tilde{A} = 1, C^Y = 1)\}}{P(\tilde{A} = 1 \mid W)}.
\]

(4.38)

By randomization and independent censoring assumptions, (4.37) simplifies to

\[
D_0(W) = E(Y \mid W, A = a) - E(Y \mid A = a);
\]

(4.40)

equation (4.38) simplifies to

\[
D_1(W, \tilde{A}, L) = \frac{1(C^Y = 1)1(\tilde{A} = 1)\{Y - E(Y \mid L, W, \tilde{A} = 1, C^Y = 1)\}}{P(C^Y = 1 \mid L, W, \tilde{A} = 1)P(\tilde{A} = 1 \mid W)}.
\]

(4.41)

equation (4.39) simplifies to

\[
D_2(W, \tilde{A}, L, C^Y, Y) = \frac{1(C^Y = 1)1(\tilde{A} = 1)\{Y - E(Y \mid L, W, A = a)\}}{P(C^Y = 1 \mid L, W, A = a)}.
\]

(4.42)

The following lemma states that \(D_0, D_1, \text{ and } D_2\) are pairwise uncorrelated.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Lemma 4.8.6.1. We have

\[ \text{Cov}\{D_0(W), D_1(W, \tilde{A}, L)\} = 0, \]  
(4.43)\]

\[ \text{Cov}\{D_0(W), D_2(W, \tilde{A}, L, C^Y, Y)\} = 0, \]  
(4.44)\]

\[ \text{Cov}\{D_1(W, \tilde{A}, L), D_2(W, \tilde{A}, L, C^Y, Y)\} = 0. \]  
(4.45)\]

Lemma 4.8.6.1 implies

\[ \text{Var}\{D(W, \tilde{A}, L, C^Y, Y)\} = \text{Var}\{D_0(W)\} + \text{Var}\{D_1(W, \tilde{A}, L)\} + \text{Var}\{D_2(W, \tilde{A}, L, C^Y, Y)\}. \]  
(4.46)\]

By (4.40) we have

\[ \text{Var}\{D_0(W)\} = \text{Var}\{E_a(Y \mid W)\} = \text{Var}_a\{E_a(Y \mid W)\}, \]  
(4.47)\]

where the last equality follows from randomization assumption. By (4.41) we have

\[ E\{D_1(W, \tilde{A}, L)\} = 0, \]  
so it follows from randomization and independent censoring that

\[ \text{Var}\{D_1(W, \tilde{A}, L)\} = E\left[\frac{1(\tilde{A} = 1)^2\{E(Y \mid L, W, A = a) - E(Y \mid W, A = a)\}^2}{P(A = 1)^2}\right] \]

\[ = \frac{E\{1(\tilde{A} = 1)\}}{P(\tilde{A} = 1)^2} E\left[\{E(Y \mid L, W, A = a) - E(Y \mid W, A = a)\}^2 \mid A = a\right]. \]  
(4.48)\]
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

By independent censoring we have $P(\tilde{A} = 1) = p_a p_l$. It then follows from (4.48) and randomization assumption that

$$\text{Var}\{D_1(W, \tilde{A}, L)\} = \frac{1}{p_a p_l} \text{Var}_a\{E_a(Y \mid L, W) - E_a(Y \mid W)\}. \quad (4.49)$$

Similarly, (4.42) together with randomization and monotone censoring imply

$$\text{Var}\{D_2(W, \tilde{A}, L, C^Y, Y)\} = \frac{1}{p_a p_y} \text{Var}_a\{Y^a - E_a(Y \mid L, W)\}. \quad (4.50)$$

Because the semiparametric lower bound on the asymptotic variance for an estimand equals the variance of the efficient influence function, by (4.46), (4.47), (4.49), and (4.50) we proved Theorem 4.3.1.2.

4.8.6.3 Proof of Corollary 4.3.2.1

Proof. For notation simplicity, let $C_{Y_i}$ stand for $(C^Y)_i$, the censoring indicator of $Y$ for participant $i$. The unadjusted estimator $\hat{\tau}$ for $E(Y \mid A = a)$ is defined as

$$\hat{\tau} = \frac{\sum_{i=1}^n Y_i 1(A_i = a, C_{Y_i} = 1)}{\sum_{i=1}^n 1(A_i = a, C_{Y_i} = 1)}. \quad (4.51)$$
Using the causal framework in Section 4.8.1, (4.51) becomes

\[
\hat{\tau} = \frac{\sum_{i=1}^{n} Y_i^a \mathbb{1}(A_i = a, C Y_i = 1)}{\sum_{i=1}^{n} \mathbb{1}(A_i = a, C Y_i = 1)}.
\]  (4.52)

Under Assumptions 4.8.1.2 and 4.8.1.3 in Section 4.8.1, \( \hat{\tau} \) is unbiased:

\[
E(\hat{\tau}) = E\left[ E\left\{ \frac{\sum_{i=1}^{n} Y_i^a \mathbb{1}(A_i = a, C Y_i = 1)}{\sum_{i=1}^{n} \mathbb{1}(A_i = a, C Y_i = 1)} \right| A_1, \ldots, A_n, C Y_1, \ldots, C Y_n \right]\right] = E(Y^a).
\]  (4.53)

In the following we calculate the asymptotic variance of \( \hat{\tau} \).

\[
\sqrt{n}\{\hat{\tau} - E(Y^a)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i^a \mathbb{1}(A_i = a, C Y_i = 1) - \sqrt{n}E(Y^a)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i^a - E(Y^a)\} \mathbb{1}(A_i = a, C Y_i = 1)
\]  (4.54)

By Weak Law of Large Numbers and the independent censoring assumption,

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(A_i = a, C Y_i = 1) \overset{P}{\to} p a p y,
\]  (4.55)

where \( \overset{P}{\to} \) denotes convergence in probability. By Central Limit Theorem,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i^a - E(Y^a)\} \mathbb{1}(A_i = a, C Y_i = 1) \overset{d}{\to} N(0, \sigma^2),
\]  (4.56)
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

where by randomization and independent censoring we have

\[
\sigma^2 = \text{Var}\left[\{Y_a - E(Y^a)\} \mathbb{1}(A = a, C^Y = 1)\right] = E\left[\{Y_a - E(Y^a)\}^2 \mathbb{1}(A = a, C^Y = 1)^2\right] \\
= p_a p_y \text{Var}(Y^a). \tag{4.57}
\]

Combining (4.54)-(4.57), it follows from Slutsky’s theorem that

\[
\sqrt{n} \{\hat{\tau} - E(Y^a)\} \xrightarrow{d} N(0, (p_a p_y)^{-1} \text{Var}(Y^a)).
\]

So the asymptotic variance of \(\hat{\tau}\) is \((p_a p_y)^{-1} \text{Var}(Y^a)\), which by randomization yields

\[
\text{AVar(unadjusted)} = \frac{1}{p_a p_y} \text{Var}_a(Y). \tag{4.58}
\]

Equations (4.58) and (4.4) imply

\[
\text{AVar(unadjusted)} = \frac{1}{p_a p_y} \text{Var}_a(Y)(R_{W,a}^2 + R_{L|W,a}^2 + R_{r,a}^2). \tag{4.59}
\]

On the other hand, Theorem 4.3.1.2 and the definition of \(R_{W}^2\) and \(R_{r}^2\) imply

\[
\text{AVar(efficient)} = \text{Var}_a(Y)(R_{W,a}^2 + \frac{1}{p_a p_t} R_{L|W,a}^2 + \frac{1}{p_a p_y} R_{r,a}^2). \tag{4.60}
\]

Equations (4.59) and (4.60) yield (4.5).

Equation (4.6) follows immediately from (4.5). For (4.7), note that if \(R_{W,a}^2 = 0\),
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

by definition we have $R^2_{L|W,a} = R^2_{L,a}$. So (4.7) follows from (4.5). The proof is thus finished. \hfill \square

4.8.6.4 Proof of Theorem 4.4.1.1

For notation simplicity, denote by $Q(\cdot)$ the conditional expectation $E(Y \mid \cdot)$. Using the derivation in (4.34)-(4.42) twice for $A = 1$ and $A = 0$, we get the efficient influence function $D(W, A, C^L, L, C^Y, Y)$ for $E(Y \mid A = 1) - E(Y \mid A = 0)$:


where

$$D_0(W) = \{Q(W, A = 1) - Q(A = 1)\} - \{Q(W, A = 0) - Q(A = 0)\}, \quad (4.62)$$

$$D_1(W, A, C^L, L) = \frac{AC^L}{p_1p_l} \left\{Q(W, L, A = 1) - Q(W, A = 1)\right\}$$

$$- \frac{(1 - A)C^L}{p_0p_l} \left\{Q(W, L, A = 0) - Q(W, A = 0)\right\}, \quad (4.63)$$

and

$$D_2(W, A, C^L, L, C^Y, Y) = \frac{AC^Y}{p_1p_y} \left\{Y - Q(W, L, A = 1)\right\} - \frac{(1 - A)C^Y}{p_0p_y} \left\{Y - Q(W, L, A = 0)\right\}. \quad (4.64)$$
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

The following lemma states that $D_0$, $D_1$, and $D_2$ are pairwise uncorrelated.

**Lemma 4.8.6.2.** We have

\[
\text{Cov}\{D_0(W), D_1(W, A, C^L, L)\} = 0, \tag{4.65}
\]

\[
\text{Cov}\{D_0(W), D_2(W, A, C^L, L, C^Y, Y)\} = 0, \tag{4.66}
\]

\[
\text{Cov}\{D_1(W, A, C^L, L), D_2(W, A, C^L, L, C^Y, Y)\} = 0. \tag{4.67}
\]

Lemma 4.8.6.2 implies

\[
\text{Var}\{D(W, A, C^L, L, C^Y, Y)\} = \text{Var}\{D_0(W)\} + \text{Var}\{D_1(W, A, C^L, L)\}
+ \text{Var}\{D_2(W, A, C^L, L, C^Y, Y)\}. \tag{4.68}
\]

By 4.62 we have

\[
\text{Var}\{D_0(W)\} = \text{Var}\{Q(W, A = 1) - Q(W, A = 0)\}. \tag{4.69}
\]
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

By (4.63) we have

\[
\text{Var}\{D_1(W, A, C, L)\} = \frac{AC^L}{p_1p_0^2} \left[ Q(W, L, A = 1) - Q(W, A = 1) \right]^2 + \frac{(1 - A)C^L}{p_0^2p_1^2} \left[ Q(W, L, A = 0) - Q(W, A = 0) \right]^2 \\
= \sum_{a \in \{0,1\}} \frac{1}{p_ap_l} E \left[ \left( Q(W, L, A = a) - Q(W, A = a) \right)^2 \mid A = a \right] \\
= \sum_{a \in \{0,1\}} \frac{1}{p_ap_l} \text{Var}\{Q(W, L, A = a) - Q(W, A = a) \mid A = a\}. \tag{4.70}
\]

The step from (4.70) to (4.71) utilizes the independent censoring and randomization assumptions. Similarly, (4.64) together with randomization and independent censoring imply

\[
\text{Var}\{D_2(W, A, C, L, C^Y, Y)\} = \sum_{a \in \{0,1\}} \frac{1}{p_ap_l} \text{Var}\{Y - Q(W, L, A = a) \mid A = a\}. \tag{4.73}
\]

Because the semiparametric lower bound on the asymptotic variance for an estimand equals the variance of the efficient influence function, by (4.68), (4.69), (4.72), and (4.73) we proved Theorem 4.4.1.1.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

4.8.6.5 Proof of Corollary 4.4.2.1

The unadjusted estimator for the average treatment effect is

\[
\hat{\tau} = \frac{\sum_{i=1}^{n} Y_i 1(A_i = 1, C_i = 1)}{\sum_{i=1}^{n} 1(A_i = 1, C_i = 1)} - \frac{\sum_{i=1}^{n} Y_i 1(A_i = 0, C_i = 1)}{\sum_{i=1}^{n} 1(A_i = 0, C_i = 1)}.
\]

Similar to the derivation from (4.51) to (4.58), when estimating the average treatment effect with \(P(A = 1) = P(A = 0) = 1/2\), we have

\[
\text{AVar(adjusted)} = \frac{1}{0.5 p_y} \sum_{a \in \{0, 1\}} \text{Var}_a(Y). \tag{4.74}
\]

The result in Corollary 4.4.2.1 then follows immediately from Theorem 4.4.1.1.

4.8.6.6 Proof of Corollary 4.4.2.2

(i) Purely predictive baseline variable. Because

\[
\text{Var}\{E(Y \mid W)\} = \text{Var}\{0.5E(Y \mid W, A = 1) + 0.5E(Y \mid W, A = 1)\},
\]

\[
\text{Var}\{E(Y \mid W)\} = 0 \text{ implies}
\]

\[
\text{Var}\{E(Y \mid W, A = 1)\} + \text{Var}\{E(Y \mid W, A = 0)\} = -2\text{Cov}\{E(Y \mid W, A = 1), E(Y \mid W, A = 0)\}. \tag{4.75}
\]
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

By (4.75) and the definition of \( \gamma \) and \( R_{W}^{2} \), we have \( \gamma = 2R_{W}^{2} \). This combined with (4.12) imply (4.14).

(ii) Purely prognostic baseline variable. By the definition of \( \gamma \) and \( \text{Var}\{E(Y \mid W, A = 1) - E(Y \mid W, A = 0)\} = 0 \), (4.15) follows immediately from (4.12).

4.8.7 Proof of Auxiliary Lemmas

4.8.7.1 Additional Supporting Lemmas

Lemma 4.8.7.1. Consider three random variables \( X, Y, \) and \( Z \). Denote by \( \sigma(Z) \) the \( \sigma \)-field generated by \( Z \). If \( X \in \sigma(Z) \), then

\[
E\{YE(Y \mid X)\} = E\{E(Y \mid Z)E(Y \mid X)\}. \tag{4.76}
\]

Proof. By the law of iterated expectation, we have

\[
E\{YE(Y \mid X)\} = E[E\{YE(Y \mid X) \mid Z\}]. \tag{4.77}
\]

Because \( X \in \sigma(Z) \), we have \( E(Y \mid X) \in \sigma(X) \subset \sigma(Z) \). This implies

\[
E\{YE(Y \mid X) \mid Z\} = E(Y \mid X)E(Y \mid Z). \tag{4.78}
\]

Equations (4.77) and (4.78) imply (4.76). This completes the proof.
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Lemma 4.8.7.2. Consider three random variables \( W, L, \) and \( Y \). For any measurable functions \( f(W) \) and \( g(W, L) \), we have

\[
\begin{align*}
E[f(W) \{E(Y \mid W, L) - E(Y \mid W)\}] &= 0, \quad (4.79) \\
E[g(W, L) \{Y - E(Y \mid W, L)\}] &= 0. \quad (4.80)
\end{align*}
\]

Proof. By the law of iterated expectation, we have

\[
\begin{align*}
E\{g(W)E(Y \mid W, L)\} &= E[E\{g(W)E(Y \mid W, L) \mid W\}] \\
&= E[g(W)E\{E(Y \mid W, L) \mid W\}] \\
&= E\{g(W)E(Y \mid W)\},
\end{align*}
\]

which proves (4.79).

Similarly, we have

\[
E\{g(W, L)Y\} = E[E\{g(W, L)Y \mid W, L\}] = E\{g(W, L)E(Y \mid W, L)\},
\]

which proves (4.80).
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

4.8.7.2 Proof of Lemma 4.8.6.1

Proof. By (4.40) and (4.41) we have

\[ \text{Cov}(D_0, D_1) \propto E[1(\tilde{A} = 1)\{E_a(Y \mid W) - E_a(Y)\}\{E_a(Y \mid L, W) - E_a(Y \mid W)\}] . \]

(4.81)

Combining (4.27), (4.81), and Assumptions 4.8.1.2-4.8.1.3, we derive (4.43).

By (4.40) and (4.42) we have

\[ \text{Cov}(D_0, D_2) \propto E[1(\tilde{A} = 1, C^Y = 1)\{E_a(Y \mid W) - E_a(Y)\}\{Y - E_a(Y \mid L, W)\}] . \]

(4.82)

Combining (4.26), (4.82), and Assumptions 4.8.1.2-4.8.1.3, we derive (4.44).

By (4.41) and (4.42) we have

\[ \text{Cov}(D_1, D_2) \propto E[1(\tilde{A} = 1, C^Y = 1)\{E_a(Y \mid L, W) - E_a(Y \mid W)\}\{Y - E_a(Y \mid L, W)\}] . \]

(4.83)

Combining (4.25), (4.83), and Assumptions 4.8.1.2-4.8.1.3, we derive (4.45).

This completes the proof.
4.8.7.3 Proof of Lemma 4.8.6.2

For notation simplicity, we use $E_1(\cdot)$ and $E_0(\cdot)$ to denote $E(\cdot \mid A = 1)$ and $E(\cdot \mid A = 0)$, respectively. By (4.62) and (4.63) we have

\[
\text{Cov}(D_0, D_1) \propto E\left[A\{Q(W, A = 1) - Q(W, A = 0)\}\{Q(W, L, A = 1) - Q(W, A = 1)\}\right] \\
- E\left[(1 - A)\{Q(W, A = 1) - Q(W, A = 0)\}\{Q(W, L, A = 0) - Q(W, A = 0)\}\right] \\
= \sum_{a \in \{0,1\}} E_a[\{E_1(Y \mid W) - E_0(Y \mid W)\}\{E_a(Y \mid W, L) - E_a(Y \mid W)\}] \times P(A = a).
\]

(4.84)

Both terms in (4.84) equals 0 by (4.79) in Lemma 4.8.7.2 with $f(W) = E_1(Y \mid W) - E_0(Y \mid W)$. This yields (4.65).

By (4.62) and (4.64) we have

\[
\text{Cov}(D_0, D_2) \propto E\left[A\{Q(W, A = 1) - Q(W, A = 0)\}\{Y - Q(W, L, A = 1)\}\right] \\
- E\left[(1 - A)\{Q(W, A = 1) - Q(W, A = 0)\}\{Y - Q(W, L, A = 0)\}\right] \\
= \sum_{a \in \{0,1\}} E_a[\{E_1(Y \mid W) - E_0(Y \mid W)\}\{Y - E_a(Y \mid W, L)\}] \times P(A = a).
\]

(4.85)

Both terms in (4.85) equals 0 by (4.80) in Lemma 4.8.7.2 with $g(W, L) = E_1(Y \mid W) - E_0(Y \mid W)$. This yields (4.66).
CHAPTER 4. COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

By (4.63) and (4.64) we have

\[
\text{Cov}(D_1, D_2) \propto E \left[ a \{Q(W, L, A = 1) - Q(W, A = 1)\} \{Y - Q(W, L, A = 1)\} \right]
- E \left[ (1 - a)\{Q(W, L, A = 0) - Q(W, A = 0)\}\{Y - Q(W, L, A = 0)\} \right]
= \sum_{a \in \{0, 1\}} E_a \left[ \{E_a(Y \mid W, L) - E_a(Y \mid W)\} \{Y - E_a(Y \mid W, L)\} \right] \times P(A = a).
\]

(4.86)

Both terms in (4.86) equals 0 by (4.80) in Lemma 4.8.7.2 with \( g(W, L) = E_a(Y \mid W, L) - E_a(Y \mid W) \). This yields (4.67).

This completes the proof.
Chapter 5

Simulation Study for Adaptive Enrichment Trials

Summary. Adaptive enrichment designs involve rules for restricting enrollment to a subset of the population during the course of an ongoing trial. This can be used to target those who benefit from the experimental treatment. Trial characteristics such as the accrual rate and the prognostic value of baseline variables are typically unknown when a trial is being planned; these values are typically assumed based on information available before the trial starts. Because of the added complexity in adaptive enrichment designs compared to standard designs, it may be of concern how sensitive the trial performance is to deviations from assumptions. Through

---

1This chapter is adapted from the following working paper: Tianchen Qian, Elizabeth Colantuoni, Aaron Fisher, Michael Rosenblum. Sensitivity of Trial Performance to Delayed Outcomes, Accrual Rates, and Prognostic Variables Based on a Simulated Randomized Trial with Adaptive Enrichment.
simulation studies, we evaluate the sensitivity of Type I error, power, expected sample size, and trial duration to different design characteristics. Our simulation distributions mimic features of data from the Alzheimer’s Disease Neuroimaging Initiative cohort study, and involve two subpopulations based on a genetic marker. We investigate the impact of the following design characteristics: the accrual rate, the time from enrollment to measurement of a short-term outcome and the primary outcome, and the prognostic value of baseline variables and short-term outcomes. To leverage prognostic information in baseline variables and short-term outcomes, we use a semiparametric, locally efficient estimator, and investigate its strengths and limitations compared to standard estimators. We apply information-based monitoring, and evaluate how accurately information can be estimated in an ongoing trial.

5.1 Background

Adaptive enrichment designs involve pre-planned rules for restricting enrollment based on accrued data in an ongoing trial (Wang et al., 2007). If, for example, a subpopulation shows evidence of no benefit of treatment, its enrollment could be stopped while the complementary subpopulation continues to be enrolled. Stallard et al. (2014) give an overview of statistical methods for adaptive enrichment designs, including the p-value combination approach (Bretz et al., 2006; Schmidli
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

et al., 2006; Jennison and Turnbull, 2007; Brannath et al., 2009; the conditional error function approach (Friede et al., 2012); and approaches using group sequential computations (Stallard, 2011; Magnusson and Turnbull, 2013). We use an adaptive enrichment design from the general class of Rosenblum et al. (2016), which is based on the group sequential computation approach.

We consider trials where the primary outcome is observed a fixed amount of time after enrollment (called the delay); we refer to such outcomes as delayed responses. To illustrate, we use data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort study. We set the primary outcome to be a measure of change in severity of dementia symptoms from baseline to 2 year of follow-up described below; this is similar to the primary outcome in an ongoing, phase 3 clinical trial of a drug to slow cognitive and functional decline from early Alzheimer’s Disease (Biogen, 2016). Also recorded are baseline variables and the short-term outcome of change in severity of dementia symptoms measured at 1 year of follow-up.

To leverage prognostic information in baseline variables and the short-term outcome, we use a semiparametric, locally efficient estimator (called the adjusted estimator, for conciseness) from van der Laan and Gruber (2012). The adjusted estimator in a randomized trial is consistent under mild regularity conditions without requiring any parametric model assumptions. It has potential to improve precision, power, expected sample size, and trial duration when variables are sufficiently prognostic for the outcome. In trials with delayed responses, the adjusted
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

estimator uses information from pipeline participants, i.e., enrollees whose primary outcome has not yet been observed.

Trial characteristics like the accrual rate and the prognostic value in covariates are typically unknown and assumed based on historical information before a trial begins. Deviation from these assumptions may result in different information accrual rates in each subpopulation, and may change the covariance matrix of the test statistics. Because of the intrinsic complication in adaptive enrichment designs including the interdependence of the stopping rules of multiple subpopulations, it is crucial and difficult to assess the sensitivity of the trial performance to such deviations. In this chapter, we evaluate the sensitivity of Type I error, power, expected sample size, and trial duration to different design characteristics through simulation studies. Our simulation distributions mimic features of data from the Alzheimer’s Disease Neuroimaging Initiative, and involve two subpopulations of interest based on a genetic marker. We investigate the impact of the following design characteristics: the accrual rate, the delay time of the short-term outcome and the primary outcome, and the prognostic value of baseline variables and short-term outcomes.

The simulated trials involve multiple stages, and information-based monitoring is used to determine the time of interim analyses. We evaluate the accuracy of information estimates when using the adjusted estimator versus the unadjusted estimator, which is critical in order that the familywise Type I error be controlled.
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

In Section 5.2 we describe the ADNI study. In Section 5.3 we present notation. The simulation setup is given in Section 5.4. Section 5.5 presents simulation results, including the impact of prognostic baseline variables and a short-term outcome (Section 5.5.1), the impact of varying delay time (Section 5.5.2), and the impact of varying the accrual rates (Section 5.5.3) on the performance of the adaptive design. In Section 5.6 we discuss information accrual rates and how accurately these can be estimated in an ongoing trial. Section 5.7 concludes with discussions and future research directions.

5.2 Data Example

Our simulations are based on distributions that mimic features of the data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), an observational longitudinal study of cognitive impairment and progression to Alzheimer’s disease. The ADNI was initiated in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of the study has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer’s disease.\(^2\) The Clinical Dementia Rating (CDR) scale is used to assess the severity of dementia symptoms and provides both a numeric global score

\(^2\)For up-to-date information, see www.adni-info.org.
ranging from 0 to 3, and a sum of boxes (SOB) score ranging from 0 to 18.

Our data come from 286 patients who entered the ADNI study with mild cognitive impairment (CDR 0.5 with a SOB score 2.5 or less) and who remained in the study for the full 12 months of follow-up. For conciseness, we refer to the CDR sum of boxes score as the CDR score. We define the primary outcome $Y$ as the difference between the CDR score at baseline and at 2 years. We define the short-term outcome $L$ as the difference between the CDR score at baseline and at 1 year. Let $W$ denote the following five prognostic baseline variables: CDR score at baseline; age; $A_{42}$ (a type of amyloid plaque involved in Alzheimer’s disease progression); Alzheimer’s Disease Association (ADA, 13 items) scale; and the Mini Mental State Examination (MMSE) score. We consider two distinct subpopulations defined by apolipoprotein E (APOE) $e4$ carrier status. Subpopulation 1 consists of those with no $e4$ alleles, and subpopulation 2 consists of those with at least one $e4$ allele. Among the 286 patients, 47% carry no APOE $e4$ alleles. We consider a hypothetical treatment whose goal is to delay the progression of disease.

5.3 Notation

When followed up completely, each participant $i$ in the trial has full data vector $D_i = (S_i, W_i, A_i, L_i, Y_i)$. We use the vector $D = (S, W, A, L, Y)$ when referring to a generic participant. The variable $S_i \in \{1, 2\}$ denotes the subpopulation that
participant \( i \) belongs to; \( W_i \) denotes a vector of baseline variables; \( A_i \) denotes the treatment assignment indicator; \( L_i \) denotes the short-term outcome; and \( Y_i \) denotes the primary outcome. We assume that \((S_i, W_i, A_i)\) are observed when participant \( i \) is enrolled, and that \( L_i \) and \( Y_i \) are observed at time \( d_L \) and \( d_Y \), respectively, from the time of enrollment. Assume \( d_L \leq d_Y \). Each vector \( D \) is assumed to be an independent, identically distributed draw from an unknown distribution \( Q \), with the only restriction being that \( A \) is randomized by design with equal probability of being 0 or 1, independent of \( S, W \). The short-term outcome \( L \) can be any predefined measurement made after randomization. No assumptions on its relationship to \( Y \) are needed in order that our estimators (adjusted and unadjusted) are consistent and asymptotically normal (Gruber and van der Laan, 2012).

For a given population, the average treatment effect is defined to be the difference between the population mean of the primary outcome under treatment \((A = 1)\) versus under control \((A = 0)\). Denote the average treatment effect in subpopulation 1, subpopulation 2, and the combined population by \( \Delta_1 \), \( \Delta_2 \), and \( \Delta_0 \), respectively, where \( \Delta_0 = E(Y|A = 1) - E(Y|A = 0) \) and for each subpopulation \( s \in \{1, 2\} \), \( \Delta_s = E(Y|A = 1, S = s) - E(Y|A = 0, S = s) \). Let \( p_s \) denote the proportion of subpopulation \( s \) in the combined population, and we have \( \Delta_0 = p_1 \Delta_1 + p_2 \Delta_2 \). Define the null hypotheses

\[
H_{01} : \Delta_1 \leq 0; \quad H_{02} : \Delta_2 \leq 0; \quad H_{00} : \Delta_0 \leq 0,
\]
which represent no average treatment benefit in subpopulation 1, subpopulation 2, and the combined population, respectively.

We quantify the prognostic value of \( W \) and \( L \) for explaining variance in the primary outcome \( Y \) for the combined population. Define the \( R \)-squared of \( W \) and \( R \)-squared of \( L \) as

\[
R^2_W = \frac{\text{Var}\{E(Y \mid W)\}}{\text{Var}(Y)} , \quad R^2_L = \frac{\text{Var}\{E(Y \mid L)\}}{\text{Var}(Y)} .
\]  

(5.1)

\( R^2_W \) represents the fraction of variance in \( Y \) explained by \( W \). \( R^2_L \) represents the fraction of variance in \( Y \) explained by \( L \).

Using the ADNI study data, we approximated (5.1) to roughly determine how much of the variance of the outcome \( Y \) is explained by \( W \) or \( L \). The empirical \( R^2_W \) is computed as in (5.1), with \( E(Y \mid W) \) estimated by a linear model with intercept and main terms \( W_3, W_4 \), and the variances are estimated by the empirical variance. (We use only \( W_3, W_4 \) in the working model for constructing the adjusted estimator; see Section 5.4.2.) A similar computation was done to obtain the empirical \( R^2_L \). The resulting values are 0.20 and 0.48 for \( R^2_W \) and \( R^2_L \), respectively, for the combined population.

We also estimated \( R^2_W \) and \( R^2_L \) within each subpopulation, and found the prognostic values differ by subpopulation. The corresponding empirical \( R^2_W \) is 0.30 for subpopulation 1 and 0.14 for subpopulation 2; the empirical \( R^2_L \) is 0.44 for sub-
population 1 and 0.50 for subpopulation 2. This differential prognostic value by subpopulation impacts information accrual and power for the adjusted estimator as described in Section 5.5. In what follows, $R^2_W$ and $R^2_L$ refer to (5.1) for the combined population.

5.4 Simulation Setup

5.4.1 Overview

Our goal is to evaluate the performance of an adaptive enrichment design with a delayed response when we vary the prognostic values in baseline variables and short-term outcome, accrual rates, delay time, and estimator used. The performance is evaluated based on Type I error, power, expected sample size and average duration of the trial, and is based on two estimators: the unadjusted estimator (the difference between the sample means of the primary outcome between the two study arms), and an adjusted estimator that leverages baseline variables and the short-term outcome. The latter is a targeted maximum likelihood estimator (TMLE) of van der Laan and Gruber (2012) implemented in the R package ltmle (Schwab et al., 2015). The R code we used for the adjusted estimator is provided in the Supplementary Materials. Other candidate adjusted estimators include those of Lu and Tsiatis (2011), Rotnitzky et al. (2012), and Gruber and van der Laan
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

(2012). Colantuoni and Rosenblum (2015) showed that all these adjusted estimators have similar performance in the context of randomized trials. Both the unadjusted and adjusted estimators are consistent and asymptotically normal under mild regularity conditions (van der Laan and Gruber, 2012).

We vary the following in our simulation studies: the prognostic value of baseline variables $W$ and short-term outcome $L$ represented by the $R$-squared formulas in (5.1); the delay time $d_L$ of the short-term outcome; the delay time $d_Y$ of the final outcome; and the accrual rate.

5.4.2 Data Generating Distributions Based on ADNI data

Hypothetical trials are populated with participants, each of whose data vector $D$ is drawn independently from a data generating distribution $Q$, which differs by simulation study. We construct each $Q$ to mimic certain observed relationships between $W$, $L$ and $Y$ within each subpopulation $s \in \{1, 2\}$ in the ADNI study. For simplicity, we center $W$ within each subpopulation.

There is no treatment in the ADNI study. For simulation purpose we assign the treatment variable $A$ independent of $S, W$ as a Bernoulli random variable with success probability 0.5, and having a relationship with $Y$ as described next. The minimum, clinically meaningful, average treatment effect for our hypothetical tri-
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

als is $\delta_{\text{min}} = 0.42$, which corresponds to a 30% relative improvement in mean CDR score change, i.e., a 30% reduction in disease progression. Within each of our five simulation studies (described below in Table 5.1), we generate data under four treatment effect settings (abbreviated as “effect setting” hereafter): (a) treatment benefits neither subpopulation ($\Delta_1 = \Delta_2 = 0$); (b) treatment benefits subpopulation 1 only ($\Delta_1 = \delta_{\text{min}}, \Delta_2 = 0$); (c) treatment benefits subpopulation 2 only ($\Delta_1 = 0, \Delta_2 = \delta_{\text{min}}$); and (d) treatment benefits both subpopulations ($\Delta_1 = \Delta_2 = \delta_{\text{min}}$). Effect settings (b) and (c) involve treatment effect heterogeneity.

The data generating distribution is denoted by

$$Q = Q (\Delta_1, \Delta_2, R_{W}, R_{L}, d_L, d_Y, \text{accrual rate}),$$

which is determined by the following: the pair of treatment effects for each subpopulation ($\Delta_1, \Delta_2$), the prognostic value of the baseline covariates $R_{W}$, the prognostic value of the short-term outcome $R_{L}$, the delay between enrollment and the short-term outcome $d_L$, the delay between enrollment and the primary outcome $d_Y$, and the accrual rate. We set the enrollment process to be random, where the enrollment time of the patients follows a homogeneous Poisson process with intensity equal to the accrual rate. We assume that each subpopulation’s accrual rate is proportional to its prevalence in the combined population. In each simulation study, we vary one or several of the above at a time to assess the impact on trial
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

performance.

Within each subpopulation $S = s$, $W$ is randomly sampled from the observed data, and $Y$ and $L$ are generated from the linear models:

$$L = \alpha_0^s + \alpha_W^s W + \alpha_A^s A + \epsilon_L, \quad \epsilon_L \sim N \left(0, (\sigma_L^s)^2\right) (5.2)$$

$$Y = \beta_0^s + \beta_W^s W + \beta_A^s A + \beta_L^s L + \epsilon_Y, \quad \epsilon_Y \sim N \left(0, (\sigma_Y^s)^2\right) (5.3)$$

with $\epsilon_Y$ and $\epsilon_L$ independent of $(W, A)$ and of each other. The values of $\beta_0^s$, $\beta_W^s$, $\beta_L^s$, $\sigma_Y^s$, $\sigma_W^s$, $\alpha_0^s$, $\alpha_W^s$ and $\sigma_L^s$ are based on the above models fit to the ADNI study data separately within each stratum $S = s$ and leaving out $A$. We set $\alpha_A^s = 0.5\Delta_s$ and $\beta_A^s = \Delta_s - \alpha_A^s \beta_L^s$, where $\Delta_s$ is the desired treatment effect of the corresponding effect setting. This makes the treatment effect on the short-term outcome half of that on the final outcome, which we believe is plausible.

We construct simulation distributions with a range of $R_W^2$ and $R_L^2$ values by varying $\beta$’s and $\sigma$’s. We do so in such a way that the average treatment effect within each subpopulation is unchanged, and the variance of $Y$ within each subpopulation and each treatment arm is unchanged. This is to ensure that the (asymptotic) performance of the unadjusted estimator is unchanged, providing a benchmark to compare against. To obtain different values of $R_W^2$ and $R_L^2$, we multiply the original fits $\beta_W^s$ and $\beta_L^s$ from the ADNI data set by a tuning parameter $p$, and change $\beta_0^s$, $\beta_A^s$, $\sigma_Y^s$ and $\sigma_L^s$ accordingly to ensure that the variance of $Y$ given $A$, $S$ and the
average treatment effect given $S$ are unchanged. Details are given in Section 5.5).

Let the default simulation scenario be the one with design characteristics corresponding to the empirical distribution of the ADNI study data: $R^2_W = 0.20$, $R^2_L = 0.48$, $d_L = 1$ year, $d_Y = 2$ years, and the accrual rate for the combined population 167 patients/year. We conduct 5 sets of simulations with various design characteristics that are summarized in Table 5.1. Each combination of $(R^2_W, R^2_L, d_L, d_Y, \text{accrual rate})$ is referred to as a simulation scenario. For example, in simulation study 1 (row 1 in Table 5.1), $R^2_W$ is varied from 0 to 0.6 and all other characteristics are the default value.

<table>
<thead>
<tr>
<th>Simulation study</th>
<th>$R^2_W$ (years)</th>
<th>$d_L$ (years)</th>
<th>$d_Y$ (years)</th>
<th>accrual rate (patients/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 – 0.6</td>
<td>default</td>
<td>default</td>
<td>default</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 – 0.6</td>
<td>default</td>
<td>default</td>
</tr>
<tr>
<td>3</td>
<td>default</td>
<td>0</td>
<td>0 – 4</td>
<td>default</td>
</tr>
<tr>
<td>4</td>
<td>default</td>
<td>default</td>
<td>0 – $d_Y$</td>
<td>0.1, 1, 2, 3, 4, default</td>
</tr>
<tr>
<td>5</td>
<td>default</td>
<td>default</td>
<td>default</td>
<td>50 – 500</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of setups for 5 simulation studies. Default value of parameter: $R^2_W = 0.20$, $R^2_L = 0.48$, $d_L = 1$ year, $d_Y = 2$ year, accrual rate 167 patients/year. Ranges of values $x – y$ indicate the design characteristic(s) varied in the corresponding simulation study.

In all simulations, we use the full set of baseline covariates $(W_1, W_2, W_3, W_4, W_5)$ in the data generating distributions (5.2) and (5.3) for $L$ and $Y$, but we only include baseline variables $W_3$, $W_4$ ($A\beta_{42}$ and ADA) in the working models used by the adjusted estimator. We intentionally induced such model misspecification, since in practice the working models used by the adjusted estimator will generally be
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

misspecified. In addition, the TMLE estimator uses logistic regression working models (by first scaling the outcome to the interval \([0, 1]\)) rather than linear models, which can lead to additional misspecification. (The usage of logistic regression on bounded continuous variables is justified in Gruber and van der Laan (2010, Lemma 1) because it preserves the bounds on the outcome.) Though the adjusted estimator is robust to the above model misspecification in that it is still consistent and asymptotically normal, the misspecification may reduce its precision (Gruber and van der Laan, 2012, Section 4).

5.4.3 Adaptive Enrichment Design

We define a new adaptive enrichment design using the general framework developed by Rosenblum et al. (2016). We consider two subpopulations denoted by \(S\): \(S = 1\) if the patient has no APOE \(\varepsilon 4\) allele, and \(S = 2\) if the patient has one or more APOE \(\varepsilon 4\) allele. Denote by \(S = 0\) the combined population. We consider an adaptive enrichment design with maximum number of stages \(K = 5\). At each analysis \(k \leq K\), denote by \(Z_{s,k}\) the Wald statistic (estimator divided by its standard error) for null hypothesis \(H_{0s}\) \((s \in \{0, 1, 2\})\). For each population \(s\) and stage \(k \leq K\), let \(u_{s,k}\) denote the efficacy boundary for the null hypothesis \(H_{0s}\) \((s \in \{0, 1, 2\})\), and let \(l_{s,k}\) denote the futility stopping boundary \((s \in \{1, 2\})\). Below are the steps that are followed at each analysis \(k \leq K\) to determine the continuation (or stop) of enrollment and the results of hypothesis testing.
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

1. For each \( s \in \{1, 2\} \), if subpopulation \( s \) has not had enrollment stopped at a previous analysis, and if \( Z_{s,k} > u_{s,k} \), reject \( H_{0_s} \).

2. For each \( s \in \{1, 2\} \), if \( H_{0_s} \) is rejected or \( Z_{s,k} < l_{s,k} \), stop subpopulation \( s \) enrollment.

3. If both \( H_{0_1} \) and \( H_{0_2} \) are rejected, or (if both subpopulations have not had enrollment stopped at a previous analysis and \( Z_{0,k} > u_{0,k} \)), reject \( H_{00} \).

The trial continues until both subpopulations terminate enrollment or the final analysis \( K \) is reached. For \( s \in \{0, 1, 2\} \), if \( H_{0_s} \) is not rejected in the above steps, we fail to reject it.

Define the power of \( H_{0_1} \) to be the probability to reject \( H_{0_1} \) under effect setting (b), power of \( H_{0_2} \) to be the probability to reject \( H_{0_2} \) under effect setting (c), and power of \( H_{00} \) to be the probability to reject \( H_{00} \) under effect setting (d). The design’s goals are to achieve at least 80% power to reject the corresponding null hypothesis under each effect setting (b), (c), and (d), and to strongly control the familywise Type I error rate at level 0.025, asymptotically. For example, the requirement under effect setting (b) is 80% power for \( H_{0_1} \).

The Type I error spent at each stage, futility boundaries \( l_{s,k}, s \in \{1, 2\}, 1 \leq k \leq K \) and the information level (inverse of the estimator’s variance) used for analysis timing are in Table 5.2. They were constructed by approximately solving the following optimization problem: for the unadjusted estimator under the default sim-
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

ulation scenario, minimize the expected sample size averaged over effect settings (a)-(d), subject to the Type I error and power constraints in the previous paragraph. The optimization was solved using an approach from Fisher and Rosenblum (2016), and does not necessarily equal the true optimum solution (which is currently an open research question). The asymmetry in the solution is because the proportion of subpopulation 1 $p_1 = 0.47$ and the variances differ by subpopulation.

In determining the values of efficacy boundaries $u_{s,k}, s \in \{0, 1, 2\}, 1 \leq k \leq K$, we use the error spending approach as described in Rosenblum et al. (2016, Section 3.2), which extends the approach of Slud and Wei (1982); Lan and DeMets (1983) to multiple populations. The boundaries are numerically calculated to ensure that the test at each stage maintains its pre-specified Type I error, by assuming joint normal distribution of the test statistics; see the Supplementary Material for details. These efficacy boundaries depend on the covariance matrix of the estimator being used. As shown in Rosenblum et al. (2016), the design is guaranteed to strongly control the familywise Type I error rate at level 0.025, asymptotically, for Wald statistics based on either the unadjusted or adjusted estimators.

5.4.4 Analysis Timing and Information Accrual

We present our method to determine the time of each analysis based on information monitoring. Consider either the adjusted or the unadjusted estimator. There are 3 populations of interest (the two subpopulations and the combined pop-
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

<table>
<thead>
<tr>
<th>Analysis ($k$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I error spent for Subpop. 1</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0028</td>
<td>0.0015</td>
<td>0.0038</td>
</tr>
<tr>
<td>Type I error spent for Subpop. 2</td>
<td>0.0001</td>
<td>0.0023</td>
<td>0.0012</td>
<td>0.0026</td>
<td>0.0027</td>
</tr>
<tr>
<td>Type I error spent for Comb. Pop.</td>
<td>0.0028</td>
<td>0.0006</td>
<td>0.0009</td>
<td>0.0013</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

Futility boundary ($l_{1,k}$)  
Futility boundary ($l_{2,k}$)  

Information threshold for Subpop. 1 | 13.0 | 20.2 | 24.9 | 40.1 | 69.1 |
Information threshold for Subpop. 2 | 13.4 | 20.2 | 25.7 | 41.1 | 69.6 |
Information threshold for Comb. Pop. | 27.1 | 40.8 | 50.1 | 80.3 | 138.5 |

| Efficacy boundaries for the unadjusted estimator under default simulation scenario |       |       |       |       |       |
| Efficacy boundary ($u_{1,k}$) | 3.12 | 3.06 | 2.64 | 2.77 | 2.53 |
| Efficacy boundary ($u_{2,k}$) | 3.52 | 2.76 | 2.78 | 2.63 | 2.62 |
| Efficacy boundary ($u_{0,k}$) | 2.78 | 3.08 | 2.92 | 2.86 | 2.89 |

Table 5.2: Adaptive enrichment design, and efficacy boundaries under default simulation scenario.

ulation) in our design. For each population there is a treatment effect estimator whose variance changes over time as patients are continuously enrolled. We define the information accrued for each population as the reciprocal of the corresponding estimator’s variance. The $k$th analysis occurs at the earliest time when the information accrued for every population is above its corresponding, preset threshold (which is a preset function of the Type I error allocated at that stage, i.e., part of the trial design). Information thresholds in the design, shown in Table 5.2, were set such that for the unadjusted estimator in the default simulation scenario, the information accrual for each population crosses its threshold at the same calendar time. Information can accrue at different rates depending on whether the unadjusted or adjusted estimator is used, as shown in our simulations (Section 5.6). Faster
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

information accrual can lead to earlier analyses in calendar time.

Since in a real trial the variance of each estimator is unknown, one could use a variance estimator that is updated whenever new data accrues. (See Section 5.6 where we investigate the accuracy of information estimation at given time points.) However, it is not computationally feasible to implement this in our simulations where each data generating distribution is used to simulate 50,000 trials. Instead, we set analysis timing once for each simulation scenario and estimator type, using an approximation described in the Supplementary Material.

Table 5.3 shows the calendar times of each analysis for the unadjusted and the adjusted estimators under the default simulation scenario. The cumulative sample size at each analysis time is random due to the random accrual process; Table 5.3 is an example realization. Time of analysis and sample sizes are substantially smaller for the adjusted estimator compared to the unadjusted estimator due to the former having a faster information accrual rate.

5.5 Results

We simulated 50,000 trials for each simulation scenario and effect setting combination. Table 5.4 shows the empirical probability of rejecting each hypothesis under the four effect settings in the default simulation scenario. The numbers with * indicate Type I error, i.e., rejecting at least one true null hypothesis. Under effect
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

<table>
<thead>
<tr>
<th>Analysis ($k$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>stop enroll</th>
<th>5 (final)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unadjusted estimator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time (years)</td>
<td>4.4</td>
<td>5.6</td>
<td>6.4</td>
<td>9.1</td>
<td>12.3</td>
<td>14.3</td>
</tr>
<tr>
<td>CSS (Subpop. 1)</td>
<td>202 (+148)</td>
<td>299 (+135)</td>
<td>353 (+149)</td>
<td>544 (+157)</td>
<td>928 (+138)</td>
<td>1066 (+0)</td>
</tr>
<tr>
<td>CSS (Subpop. 2)</td>
<td>211 (+190)</td>
<td>329 (+175)</td>
<td>405 (+158)</td>
<td>620 (+170)</td>
<td>1040 (+183)</td>
<td>1223 (+0)</td>
</tr>
<tr>
<td>CSS (Comb. Pop.)</td>
<td>413 (+338)</td>
<td>628 (+310)</td>
<td>758 (+307)</td>
<td>1164 (+327)</td>
<td>1968 (+321)</td>
<td>2289 (+0)</td>
</tr>
<tr>
<td><strong>Adjusted estimator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time (years)</td>
<td>3.7</td>
<td>4.8</td>
<td>5.5</td>
<td>7.7</td>
<td>10.2</td>
<td>12.2</td>
</tr>
<tr>
<td>CSS (Subpop. 1)</td>
<td>138 (+159)</td>
<td>219 (+158)</td>
<td>278 (+156)</td>
<td>453 (+158)</td>
<td>824 (+171)</td>
<td>995 (+0)</td>
</tr>
<tr>
<td>CSS (Subpop. 2)</td>
<td>150 (+164)</td>
<td>236 (+169)</td>
<td>295 (+182)</td>
<td>500 (+196)</td>
<td>916 (+186)</td>
<td>1102 (+0)</td>
</tr>
<tr>
<td>CSS (Comb. Pop.)</td>
<td>288 (+323)</td>
<td>455 (+327)</td>
<td>573 (+338)</td>
<td>953 (+354)</td>
<td>1740 (+357)</td>
<td>2097 (+0)</td>
</tr>
</tbody>
</table>

**Table 5.3**: Calendar time to conduct interim analysis for unadjusted and adjusted estimators under default simulation scenario. For one realization of the trial we show the cumulative sample size (CSS) with the format: number of participants with $Y$ observed (+ number of pipeline participants). If no early stop occurs, “stop enroll” column shows the time of last participant enrolled, and we wait until all participants have $Y$ observed then conduct the final analysis (analysis 5).

setting (a), all null hypotheses are true; under effect setting (b) (or (c)), only $H_{01}$ (or $H_{02}$) is true; under effect setting (d), none of the null hypotheses are true.

Across all the simulation scenarios we considered, the familywise Type I error rate was always controlled at 0.025 for both adjusted and unadjusted estimators. All the power goals in Section 5.4.3 are met. For the unadjusted estimator, the powers of $H_{00}$, $H_{01}$ and $H_{02}$ are all about 80% under different simulation scenarios. This is as expected due to our method of determining the analysis timing described in Section 5.4.4. For the adjusted estimator, the power of $H_{02}$ also stays near 80% under different simulation scenarios, whereas the power of $H_{00}$ and $H_{01}$ under certain simulation scenarios can be much higher than 80%. For example, when the prognostic value in $W$ ($R^2_W$) is over 0.3, the power of $H_{01}$ can exceed 90%. This is because when adjusting for baseline variables, the ratio of information accrual rate
between the two subpopulations is different than when the unadjusted estimator is used, which changes the covariance matrix of the test statistics. If one intended to have exactly 80% power for all three hypotheses for the adjusted estimator, we could have optimized a separate adaptive design for the adjusted estimator to incorporate the different $R^2_W$ in two subpopulations. However, this would make it harder to do a head-to-head comparison of the unadjusted and the adjusted estimators, so we believe the current simulation setup makes more sense.

<table>
<thead>
<tr>
<th>Effect setting</th>
<th>Percent probability to reject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_{00}$</td>
</tr>
<tr>
<td>Adjusted estimator</td>
<td></td>
</tr>
<tr>
<td>(a) $\Delta_1 = \Delta_2 = 0$</td>
<td>0.7*</td>
</tr>
<tr>
<td>(b) $\Delta_1 = \delta_{\text{min}}, \Delta_2 = 0$</td>
<td>12</td>
</tr>
<tr>
<td>(c) $\Delta_1 = 0, \Delta_2 = \delta_{\text{min}}$</td>
<td>16</td>
</tr>
<tr>
<td>(d) $\Delta_1 = \Delta_2 = \delta_{\text{min}}$</td>
<td>83</td>
</tr>
<tr>
<td>Unadjusted estimator</td>
<td></td>
</tr>
<tr>
<td>(a) $\Delta_1 = \Delta_2 = 0$</td>
<td>0.6*</td>
</tr>
<tr>
<td>(b) $\Delta_1 = \delta_{\text{min}}, \Delta_2 = 0$</td>
<td>12</td>
</tr>
<tr>
<td>(c) $\Delta_1 = 0, \Delta_2 = \delta_{\text{min}}$</td>
<td>15</td>
</tr>
<tr>
<td>(d) $\Delta_1 = \Delta_2 = \delta_{\text{min}}$</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 5.4: Type I error / power for two estimators under default simulation scenario. Type I errors (numbers with *) are computed assuming nonbinding futility boundaries; powers are computed assuming binding futility boundaries. In “Percent probability to reject”, to reject an individual hypothesis means to reject at least that hypothesis; All/Any means to reject all/any of the three hypotheses. The empirical values corresponding to the power requirements are in bold for each scenario (b)-(d).

In what follows, we focus on comparing the expected sample size (ESS) and the expected duration (ED) as summaries of trial performance under different simulation scenarios and between the two estimators.
5.5.1 Simulation Studies 1-2: Effect of Prognostic Value of Baseline Variables and Short-term Outcome

Figure 5.1 illustrates how ESS and ED are affected when one of $R^2_W$ or $R^2_L$ varies. The performance of the unadjusted estimator remains the same when the prognostic value in $W$ and $L$ changes, providing a benchmark to compare against. The adjusted estimator performs similar to the unadjusted when there is no prognostic value in $W$ or $L$, i.e. $R^2_W = R^2_L = 0$. As $R^2_W$ or $R^2_L$ increases, the adjusted estimator leverages this to achieve faster information accrual and fewer participants per stage, which leads to smaller ESS and ED. In simulation study 1, $R^2_W$ is varied from 0 to 0.6; in simulation study 2, $R^2_L$ is varied from 0 to 0.6 (Table 5.1).

Our results indicate that for the adjusted estimator, a prognostic baseline variable is more valuable than an equally prognostic short-term outcome in terms of reducing ESS and ED. For instance, under effect setting (d), increasing $R^2_W$ from 0 to 0.25 results in a 19% drop in ESS (1618 to 1314), whereas increasing $R^2_L$ from 0 to 0.25 only renders a 1% drop (1618 to 1608). This is because all enrolled patients’ baseline variables contribute to the precision of the adjusted estimator; however, although the short-term outcome of every participant is used, the efficiency gain from adjusting for $L$ is proportional to the number of participants in the pipeline (i.e., those who have $L$ but not $Y$ observed). Moreover, a participant’s baseline variables potentially improve precision for estimation of both $E(Y|A = 1)$ and
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

\[ E(Y|A = 0) \], while a participant’s short-term outcome is only used toward improving precision for one of these, corresponding to the treatment that participant received.

Figure 5.1: Left: impact of \( R^2_W \) on ESS and ED in simulation study 1. Right: impact of \( R^2_L \) on ESS and ED in simulation study 2. Since the results corresponding to unadjusted estimator do not change as \( R^2_W \) and \( R^2_L \) are varied, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols. \( \delta \) refers to \( \delta_{\min} \).
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

5.5.2 Simulation Studies 3-4: Effect of Delay Times $d_Y$ and $d_L$

We assess the impact of delay times $d_Y$ and $d_L$ on the performance of the design. In simulation study 3, we vary $d_Y$ from 0 years (immediate $Y$) to 4 years with $L$ being not prognostic at all, in order to separate the impact of $d_Y$. In simulation study 4, we set $d_Y$ to several levels, and in each case vary $d_L$ from 0 (immediate $L$) to $d_Y$; in this situation we set the prognostic value of $L$ to default (same as in the ADNI data set).

Figure 5.2 shows the comparison under simulation study 3. ESS and ED increase with longer $d_Y$ for both estimators. This is intuitive: the longer it takes to observe the primary outcome, the more time is needed to accumulate the necessary information. The adjusted estimator leads to smaller ESS and ED than the unadjusted estimator uniformly over all values of $d_Y$ because of gains from adjusting for baseline variables $W$. In addition, ESS and ED for both estimators are approximately linear in $d_Y$.

Figure 5.3 shows the comparison under simulation study 4. When $d_Y$ is fixed, the performance of the unadjusted estimator remains the same regardless of the length of $d_L$, because $L$ is not used in the unadjusted estimator. For the adjusted estimator, a longer $d_L$ results in a smaller proportion of pipeline participants who have $L$ observed—hence, slower information accrual and larger ESS and ED. Such
impact of $d_L$ is modified by $d_Y$ in that having a quickly-observable short-term outcome (i.e., smaller $d_L$) is slightly more beneficial when the delay of the final outcome is longer. For example, when $d_Y = 4$ years, decreasing $d_L$ from $d_Y$ to 0 results in a 2% drop in average duration (8.9 years to 8.7 years); when $d_Y = 0.1$ years, changes in $d_L$ have almost no effect on the trial. Of course, this is also because we are considering $d_L$ on a relative scale of $d_Y$. Finally, note that even when $d_L = d_Y$, which implies no asymptotic precision gain from adjusting for $L$, the adjusted estimator still gains efficiency from adjusting for prognostic $W$.

5.5.3 Simulation Study 5: Effect of Accrual Rate

Figure 5.4 illustrates how the ESS and ED are affected by accrual rate when the outcomes are observed with delay. Because the information depends either entirely (for the unadjusted estimator) or largely (for the adjusted estimator) on the number of participants who have the delayed response $Y$ observed, with faster accrual there will generally be more pipeline participants at interim analyses. These additional pipeline participants make ESS larger. Therefore, having fast accrual can have the negative consequence of increasing the overall study size when the primary outcome is measured with delay. For ED the result is intuitive: the duration of the trial gets shorter with faster accrual. We observe similar trends for both estimators.
Figure 5.2: Impact of $d_Y$ on ESS and ED in simulation study 3. Different line types indicate the ESS and ED under four effect settings. For the unadjusted estimator, the lines for ED under effect settings (b) and (c) are clustered together. For the adjusted estimator, the lines for ED under effect settings (b)-(d) are clustered together. $\delta$ refers to $\delta_{\text{min}}$. 

\[\text{Effect Setting}\]

- a) $\Delta_1=0, \Delta_2=0$
- b) $\Delta_1=\delta, \Delta_2=0$
- c) $\Delta_1=0, \Delta_2=\delta$
- d) $\Delta_1=\delta, \Delta_2=\delta$
Figure 5.3: Effect of $d_Y$ and $d_L$ on ESS and ED in simulation study 4. Since the results corresponding to unadjusted estimator do not change when $d_L$ varies as long as $d_Y$ is fixed, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols. $\delta$ refers to $\delta_{\text{min}}$. 
Figure 5.4: Effect of accrual rate on ESS and ED. Different line types indicate the ESS and ED under four effect settings. For each estimator, the curves for ED under effect settings (b)-(d) are clustered together. $\delta$ refers to $\delta_{\text{min}}$. 
5.6 Information Accrual Rates and Estimating Information Levels

In Section 5.4.4 we presented our approach for determining the time for analyses based on information monitoring. Here we explore information accrual more thoroughly and discuss how accurately information can be estimated in an ongoing trial. At time $t$, we are interested in two types of information level: the current information, i.e., the inverse of variance of the estimator computed using available data at time $t$, and the wait-for-pipeline information, i.e., the inverse of variance of the estimator using available data at time $t$ plus the not yet observed $L$ and $Y$ of the pipeline participants at time $t$. In order words, the wait-for-pipeline information for time $t$ is computed as if enrollment were stopped at time $t$ and we wait till all pipeline participants finish the trial before calculating the estimator. The current information is used for determining time for interim analyses, and the wait-for-pipeline information is used for determining time for the final analysis where we wait until all pipeline participants finish the trial and then test hypotheses.

Figure 5.5(a) shows how the two types of information accrue over time for the two estimators under the default simulation scenario when enrollment for both subpopulations continues. For the unadjusted estimator, the information at a given time is proportional to the number of patients with $Y$ observed; for the adjusted estimator, such proportionality is only approximate because the pipeline participants...
also contribute information. There is an approximately constant gap between the current information and the wait-for-pipeline information for each estimator, because the extra information in the not yet observed outcomes from the pipeline participants stays roughly constant over time. The adjusted estimator results in a faster information accrual compared to the unadjusted estimator, which is consistent with better trial performance (as shown in Section 5.5). The information accrual rates do not depend on $\Delta_1$ and $\Delta_2$ since in our setup these do not impact the estimator’s variance.

In practice, one needs a reliable method for estimating the information level using data from the ongoing trial in order to determine information-based timing for interim and final analyses. The sample variance is used to estimate the true variance of the unadjusted estimator. For the adjusted estimator, its variance can be estimated using the nonparametric bootstrap or by the influence curve. The \texttt{ltmle} package computes an influence-curve-based variance estimate (ICVE) for the TMLE estimator. In theory ICVE can be conservative in the sense that it may overestimate the variance (van der Laan and Gruber, 2012); in our simulation it approximates the variance quite well.

Figure 5.5(b) summarizes the performance of the variance estimators under the default simulation scenario. The solid red line connects the true information levels over time, and the box-plots represent the distribution of inverse of variance estimator at 5 analyses assuming no early stopping (sample variance estimator for
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

the unadjusted, ICVE for the adjusted). For the information of the adjusted estimator, the mean and the spread of the distribution increase with time (and hence with sample size $n$), because the information level is approximately $n$ times the reciprocal of the variance of the estimator’s influence curve, and the latter is estimated with standard error proportional to $n^{-1/2}$ asymptotically. Therefore, the spread in the box plots representing the approximate interquartile range grows at rate $n^{1/2}$. A similar observation applies to the sample variance estimate for the unadjusted estimator. Estimation accuracy for information accrual is similar for the two estimators.

5.7 Remarks

In this chapter, we conducted extensive simulation studies to examine the sensitivity of trial performance (measured by Type I error, power, expected sample size, and average duration) to different trial characteristics, including prognostic value of the baseline variable $W$ and the short-term outcome $L$, delay time to observe the short-term outcome ($d_L$) and the primary outcome ($d_Y$), and the accrual rate. We constructed simulation distributions to mimic features of the ADNI data set. We used the full set of baseline variables in generating data, and only used a subset in the adjusted estimator to incorporate model misspecification in our simulation study. Throughout this chapter, we do not assume that the short-term
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

(a) Information accrual under the default simulation scenario. Yellow corresponds to unadjusted estimator and black to adjusted estimator.

(b) Box-plots of estimated information level for adjusted estimator (using influence-curve-based method) and unadjusted estimator (using sample variance) at each of the five analyses assuming no early stopping of enrollment (so that enrollment stops \( d_Y = 1 \) year before the final analysis; see Table 5.3). The red solid line connects the true information levels, and each box-plot shows the spread of the estimated information level.

Figure 5.5: Information accrual rates and box-plots of estimated variance for the adjusted and unadjusted estimators under the default simulation scenario.
outcome $L$ is a surrogate. That is, we are not using $L$ as basis for stopping rules. $L$ is only used for improving estimation precision of the treatment effect, due to its correlation with the primary outcome.

For both the unadjusted and the adjusted estimator, expected sample size and trial duration increases with longer delay of the primary outcome. Faster patient accrual results in shorter trial duration, but can have the negative consequence of increasing the overall study size when the primary outcome is measured with delay.

For trials using the adjusted estimator, with more prognostic $W$ or $L$ the power increases, and the expected sample size and average duration decrease. A prognostic $W$ results in better trial efficiency compared to an equally prognostic $L$ (measured in terms of $R^2$). Shorter $d_L$ helps to slightly reduce expected sample size and average duration.

For trials using the unadjusted estimator, because it only uses information in the observed primary outcome, the performance is not affected by the prognostic value of $W$ or $L$, or the delay to the short-term outcome $d_L$.

We conclude that the adjusted estimator is especially useful when there are strongly prognostic baseline variables or short-term outcome available, or when the primary outcome is measured with considerable delay while the short-term outcome is observed relatively fast. Our simulation results can inform trial planning that involves delayed response.
The adaptive enrichment design involves multiple subpopulations and the hypothesis tests are interdependent. As prognostic value, delay time or patient accrual rate changes, ratio of information accrual rate in the two subpopulations changes, which in turn changes the covariance structure of the test statistics. This is reflected in higher than 80% power for certain hypothesis tests using the adjusted estimator. Familywise Type I error is always controlled in our simulation studies when using information monitoring.

In simulation studies 3 and 4 in Section 5.5.2, we set constant prognostic values $R_{W}^{2}$ and $R_{L}^{2}$, while varying $d_{L}$ and $d_{Y}$. It may also be of interest to consider a range of simulation scenarios where the prognostic value changes with delay. For example, it is possible that with longer $d_{Y}$, the baseline variables $W$ become less correlated with the final outcome $Y$, e.g., if these variables measure the same quantity at different time points. In addition, if $d_{L}$ is closer to $d_{Y}$ then the correlation between $L$ and $Y$ may become stronger. It is an area of future research to explore such simulation scenarios, in which there is a trade-off that shorter $d_{L}$ results in more pipeline participants with $L$ observed, but such $L$ is less prognostic for $Y$.

Open research problems include investigating the impact of subpopulation proportion, and generalizing the findings to other designs and data generating mechanisms. Another problem is to evaluate the impact of dropout in the simulation. The adjusted estimator can provide advantages over the unadjusted estimator for handling dropout under the missing at random assumption, in which case the un-
adjusted estimator will typically be inconsistent (van der Laan and Gruber, 2012).

As a final note, throughout this chapter we assumed the primary outcome $Y$ to be continuous. For binary outcome, the adjusted estimator in Moore and van der Laan (2009b) can be used instead, in which case $R^2$ is still the correct measure of prognostic value. We conjecture that the impact of $R^2$ on trial performance be similar to our findings here. For censored, time-to-event outcome, we can use the adjusted estimator in Brooks et al. (2013). It is an area of future research on how to properly characterize the prognostic value of $W$ and $L$ for time-to-event $Y$.

5.8 Appendix

5.8.1 Detail for Varying Prognostic Values in the Data Generating Distribution

The fitted $\alpha, \beta$ and $\sigma$'s from the ADNI study data are shown in Table 5.6.

In varying $R^2_W$ and $R^2_L$ as described in Section 5.4.2, it is desired that the average treatment effect (on both $L$ and $Y$) within each subpopulation remains unchanged, and the variance of $Y$ given $A, S$ remains unchanged. This implies that $E(Y \mid A = 1, S = s) - E(Y \mid A = 0, S = s)$ ($s \in \{1, 2\}$) and $\text{Var}(Y \mid A = a, S = s)$ ($a \in \{1, 2\}, s \in \{1, 2\}$) need to be unchanged. Throughout this subsection we omit the superscript $s$, because the following procedures will be conducted separately.
within each subpopulation. We first centered the baseline variables $W$ within each subpopulation in the ADNI data set.

Equations (5.2) and (5.3) imply

$$Y = (\beta_0 + \beta_L \alpha_0) + (\beta_W + \beta_L \alpha_W)W + (\beta_A + \beta_L \alpha_A)A + (\beta_L \epsilon_L + \epsilon_Y). \quad (5.4)$$

We aim to keep the following terms unchanged when we vary $R^2_W$ and $R^2_L$:

(i) The average treatment effect on $Y$: $\beta_A + \beta_L \alpha_A$.

(ii) The average outcome $Y$ in control arm: $\beta_0 + \beta_L \alpha_0$.

(iii) The average treatment effect on $L$: $\alpha_A$.

(iv) The average short-term outcome $L$ in control arm: $\alpha_0$.

(v) The variance of $Y$ within each arm:

$$((\beta_W + \beta_L \alpha_W)^T \text{Var}(W)(\beta_W + \beta_L \alpha_W) + \beta_L^2 \sigma_L^2 + \sigma_Y^2.$$
terms of the coefficients $\alpha, \beta$ and $\sigma$ in (5.2) and (5.3) as follows:

\begin{align*}
\beta_W(p) & := p\beta_W, \\
\beta_L(p) & := p\beta_L, \\
\beta_0(p) & := \beta_0 + (1 - p)\alpha_0\beta_L, \\
\beta_A(p) & := \beta_A + (1 - p)\alpha_A\beta_L, \\
\alpha_0(p) & := \alpha_0, \\
\alpha_W(p) & := \alpha_W, \\
\alpha_A(p) & := \alpha_A.
\end{align*}

This ensures that (i) – (iv) above are unchanged. We numerically solve for $\sigma_L(p)$ and $\sigma_Y(p)$ such that (v) is also unchanged. Because there are two degrees of freedom to solve for, we add an additional constraint that $\sigma_L(p)/\sigma_Y(p) = \sigma_L/\sigma_Y$ (i.e., they are changing proportionally).

For each value of $p$ such that positive solutions to $\sigma_L(p)$ and $\sigma_Y(p)$ exist, we simulate a data set with $1 \times 10^6$ patients and calculate $R^2_W$ and $R^2_L$ numerically for that $p$. In computing the adjusted estimator where we want $R^2_W$ or $R^2_L$ to be zero, we simply replace $W$ or $L$ with random noise after simulating the data set but before computing the estimator.
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

5.8.2 Algorithm to Compute Efficacy Boundaries $u_{s,k}$

At stage $k$ for the null hypothesis $H_{0s}$, $1 \leq k \leq K$, $s \in \{0, 1, 2\}$, denote $\alpha_{s,k}$ as the Type I error to be spent, and $u_{s,k}$ the efficacy boundary for the Wald statistic $Z_{s,k}$ (estimator divided by its standard deviation). Define the ordering $(s', k') \prec (s, k)$ if and only if $k' < k$ or ($k' = k$ and $s' < s$). Define $u_{s,k}$ to be the solution to

$$P \{ Z_{s',k'} \leq u_{s',k'} \text{ for all } (s', k') \prec (s, k), \text{ and } Z_{s,k} > u_{s,k} | \Delta_1 = \Delta_2 = 0 \} = \alpha_{s,k},$$

where the joint distribution of $Z_{s,k}$ is approximated by a normal distribution, with variance-covariance matrix estimated from 10,000 simulated trials.

5.8.3 Detail for Analysis Timing

For each fixed data generating distribution

$$Q = Q (\Delta_1, \Delta_2, R_W^2, R_L^2, d_L, d_Y, \text{accrual rate})$$

and each estimator, our method to determine time of analyses consists of four steps:

Step 1: Generate 10,000 pilot simulated trials where interim analyses are conducted at 25 pre-selected calendar time points $t_1, \ldots, t_{25}$, such that approximately 50 patients from subpopulation 1 are enrolled between $t_j$ and $t_{j+1}$. For each $t_j$,
we record the estimated treatment effect \(\tau_j\) at that time and the “wait-for-pipeline”
treatment effect \(\tilde{\tau}_j\) that is obtained by assuming enrollment is stopped at \(t_j\) and
estimating the treatment effect after \(Y\) is measured for all pipeline participants.

**Step 2:** Compute the variances of \(\tau_j\) and \(\tilde{\tau}_j\) from the 10,000 pilot simulated trials,
the inverse of which are the current information and wait-for-pipeline information
at \(t_j\), respectively.

**Step 3:** For interim analyses 1 - 4, linearly interpolate to find calendar time \(T_k\) of
which the current information equals that listed in Table 5.2 for the corresponding
\(k \in \{1, 2, 3, 4\}\). For the final analysis, linearly interpolate to find the calendar time
\(T_5\) of which the wait-for-pipeline information equals that listed in Table 5.2 for
\(k = 5\).

**Step 4:** In the simulated trials, interim analyses 1 - 4 are conducted at calendar times \(T_1, \ldots, T_4\), enrollment stops at \(T_5\) (if no early stopping occurs), and final
analysis is conducted at calendar time \(T_5 + d_Y\).

In step 3, for each \(k\) we identify the calendar time such that the information
accrued for subpopulation 1, subpopulation 2 and the combined population all
exceed the corresponding threshold in Table 5.2. The design is optimized for the
unadjusted estimator, so that when using the unadjusted estimator, the thresholds
for the three populations are crossed at almost the same time. However, for the
adjusted estimator, since \(W\) is more prognostic in subpopulation 1 than in subpop-
ulation 2 in the ADNI study data, by the time the information accrued for subpop-
CHAPTER 5. APPLICATION: ADAPTIVE TRIAL

ulation 2 reaches the threshold, information accrued for subpopulation 1 and the combined population already exceed their corresponding thresholds. Thus, for the adjusted estimator at each interim analysis the information for subpopulation 2 is exactly as in Table 5.2, whereas the information for subpopulation 1 and the combined exceeds the thresholds in Table 5.2. This makes the power for \( H_{01} \) and \( H_{00} \) higher than 80% for the adjusted estimator (as presented in Section 5.8.4.1, Figure 5.6).

5.8.4 Additional Simulation Results

5.8.4.1 Impact of Prognostic Value, Delay Time and Accrual Rate on Power

Figure 5.6 shows that for the adjusted estimator power of \( H_{01} \) and \( H_{00} \) increases with larger \( R^2_W \), whereas the power of \( H_{02} \) remains roughly constant. Change in \( R^2_L, d_Y, d_L \) or accrual rate does not substantially affect power. For the unadjusted estimator power is always constant.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01

155
Figure 5.6: Change in power under simulation studies 1-5. The power of \( H_{00} / H_{01} / H_{02} \) is the probability to reject at least \( H_{00} / H_{01} / H_{02} \) under effect setting (d)/(b)/(c). Since the results corresponding to unadjusted estimator do not change as the design characteristics are varied, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols.
AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
Chapter 6

Discussion

In this dissertation we studied semiparametric estimation for observational studies and for randomized trials. We proposed the “deductive estimation” method for constructing semiparametric estimators in observational studies. The new method overcomes the challenges in traditional semiparametric estimation, i.e., theoretical derivation of the analytic form of the efficient influence function. We presented the general framework of the proposed estimation procedure, and illustrated with detailed estimation algorithms for two-phase designs and double-sampling designs. The robustness property of the proposed estimators was demonstrated through simulation studies.

We examined the advantages and limitations of semiparametric estimators to adjust for baseline variables and short-term outcomes in randomized trials. We provided formulas to quantify the asymptotic relative efficiency when using a
semiparametric locally efficient estimator compared to the standard unadjusted estimator. Applicability of the theory to finite sample situation is validated through simulation studies, which are based on datasets from real trials. The formulas and the simulation results can guide trial planning where adjusted estimators are used.

In Chapters 2 and 3 we have presented the deductive estimation method for constructing semiparametric locally efficient estimators. Currently the method only works for pathwise differentiable parameters in nonparametric models, because semiparametric theory only justifies the equivalent of efficient influence function and Gateaux derivative under such models. It is an area of future research to extend deductive estimation to other restricted models where the efficient influence function no longer equals the Gateaux derivative.

In Chapter 3 we demonstrated the robustness of the deductive estimator in the double-sampling design through simulation studies. It is of both practical and theoretical interest to establish asymptotic theory of such estimators, for better understanding of the applicability and limitations of the deductive estimation method.

Through the simulation studies in Chapters 4 and 5, we found that the variance estimation of both the adjusted estimator and the unadjusted estimator can be anti-conservative when the sample size is small. This could lead to inflated Type I error. It is an area of future research to find approaches to correct the biased variance estimation in order to preserve trial Type I error.
Bibliography


BIBLIOGRAPHY


Friede, T., N. Parsons, and N. Stallard (2012). A conditional error func-


BIBLIOGRAPHY


Robins, J. (1986). A new approach to causal inference in mortality studies with sus-


BIBLIOGRAPHY


BIBLIOGRAPHY


Vita

Tianchen Qian was born on October 6th, 1989, in Shaoxing, Zhejiang, China. He got his Bachelor’s degree in Mathematics from Tsinghua University, Beijing. He then came to Johns Hopkins University pursuing PhD degree in Biostatistics. He has mainly worked in the area of causal inference and semiparametric estimation. His work involves estimation methods for observational studies and randomized trials. He has also worked on adaptive trial designs. He was named United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar in 2015 and 2016. He was awarded the Margaret Merrell Award for Excellence in Research from Johns Hopkins University in 2017. He was inducted to Phi Beta Kappa in 2017.
EDUCATION AND TRAINING

2017  
**Johns Hopkins University**, Baltimore, MD  
Ph.D. in Biostatistics  
Thesis title: *Semiparametric Estimation in Observational Studies and Randomized Trials*  
Advisors: Constantine Frangakis and Michael Rosenblum

2012  
**Tsinghua University**, Beijing  
B.S. in Mathematics  
Thesis title: *Statistical Methods for Algae Bloom Analysis and Prediction*  
Advisor: Kai-Fun Yu

PUBLICATIONS AND MANUSCRIPTS

PUBLISHED


UNDER REVIEW / IN PREPARATION


7. **Qian T., Frangakis C., and Yiannoutsos C.** Deductive Semiparametric Estimation in Double-Sampling Designs.

8. **Qian T., Rosenblum M., and Qiu H.** Improving Power in Group Sequential, Randomized Trials by Adjusting for Prognostic Baseline Variables and Short-Term Outcomes.

**Presentations**

2017 Robust Methods for Improving Power in Group Sequential Randomized Trial Designs, by Leveraging Prognostic Baseline Variables and Shortterm Outcomes. *ENAR*, Washington, DC

2016 Sensitivity of Trial Performance to Delayed Outcomes, Accrual Rates, and Prognostic Variables based on a Simulated Randomized Trial with Adaptive Enrichment. *Clinical Trials on Alzheimer’s Disease*, San Diego, CA


2015 Impact of Delayed Outcomes and Prognostic Variables on Simulated Randomized Trial Design with Adaptive Enrichment. *FDA Office of Regulatory Science and Innovation (ORSI) Science Symposium*, White Oak, MD

2015 Impact of Delayed Outcomes and Prognostic Variables on Simulated Randomized Trial Design with Adaptive Enrichment. *Johns Hopkins University*, Baltimore, MD

2014 Causal Inference on the Difference of the Restricted Mean Lifetime between Two Groups. *Johns Hopkins University*, Baltimore, MD
Awards and Honors

1. Phi Beta Kappa, 2017
2. Margaret Merrell Award for Excellence in Research, Johns Hopkins University, 2017
3. United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar, 2016
4. United States Food and Drug Administration (FDA) and Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI) Scholar, 2015
5. Tsinghua University Academic Award (Dean’s List equivalent), 2011
6. Fang-xiong Zhao Award for Excellence in Mathematical Programming, 2011
7. Tsinghua University Academic Scholarship (Dean’s List equivalent), 2010

Teaching Experience

Invited Lecturer

2015 Two Phase Design and Efficient Influence Function. 140.734 (Statistical Theory IV). Department of Biostatistics, Johns Hopkins University.


Teaching Assistant


2015 140.741 (Survival Analysis II). Johns Hopkins Bloomberg School of Public Health.

2015 140.641 (Survival Analysis I). Johns Hopkins Bloomberg School of Public Health.

2015 140.734 (Statistical Theory IV). Johns Hopkins Bloomberg School of Public Health.

2015 140.733 (Statistical Theory III). Johns Hopkins Bloomberg School of Public Health.


