Distribution Agreement

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

Signature:

_____ _____

Li Li Date

Semiparametric efficient and robust estimation of treatment effects from observational data

By

Li Li

Doctor of Philosophy

Biostatistics

Brent A. Johnson, Ph.D.
Advisor

John Hanfelt, Ph.D.
Committee Member

Qi Long, Ph.D.
Committee Member

Patrick S. Sullivan, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Semiparametric efficient and robust estimation of treatment effects from observational data

By

Li Li
B.S., University of Science and Technology of China, 2002
M.S., University of Science and Technology of China, 2005

Advisor: Brent A. Johnson, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

Semiparametric efficient and robust estimation of treatment effects from observational data

By Li Li

This dissertation aims to solve two problems. One is to evaluate the effect of different treatment switching strategies in HIV studies and the second is to evaluate the effect of treatment duration in infusion studies.

The current goal of initial antiretroviral (ARV) therapy is suppression of plasma HIV-1 RNA levels to below the detection limits of currently available assays. A substantial proportion HIV-infected patients who initiate antiretroviral therapy in clinical practice or antiretroviral clinical trials either fail to suppress HIV RNA or have HIV RNA levels rebound on therapy. In some clinical trials, such as the AIDS Clinical Trials Group (ACTG) Study A5095, patients randomized to initial antiretroviral treatment combinations but fail to suppress HIV RNA or have a rebound of HIV RNA on therapy are allowed to switch to second-line regimen subject to provider-specific and patient-specific information. The optimal timing of switching ARV therapy to ensure sustained virologic suppression and prolonged clinical stability in patients who have rebound in their HIV RNA is not known. Randomized clinical trials to compare early versus delayed switching have been difficult to design and even more difficult to enroll. Here, we provide a statistical framework to compare early versus late regimen change using observed data from the ACTG A5095 study. Using efficient and doubly-robust estimators for the average causal effect, we conclude that patients who follow treatment strategies that switch within eight weeks of confirmed virologic failure have significantly better health outcomes, on average, than patients following strategies that do not switch within eight weeks.

The second topic is motivated by a treatment duration-response study, ESPRIT (Enhanced Suppression of the Platelet IIb/IIa receptor with Integrilin Therapy) trial.

The ESPRIT trial targeted patients with coronary artery disease scheduled to undergo percutaneous coronary intervention (PCI) with stent implantation in a native coronary artery. The experimental treatment regimen consisted of an eptifibatide bolus and a continuous eptifibatide infusion for 18-24 hours, with a similar regimen for the placebo group. The study protocol also required that patients experiencing serious complications immediately discontinue the infusion process to receive appropriate medical attention; we define these protocol-defined adverse events as infusion-terminating events. Once treatment is found to be effective, attention often focuses on optimum treatment delivery. A treatment duration policy for t unites of time is defined as a recommendation to treat for t units of time or until a treatment-terminating event occurs, whichever comes first. Johnson and Tsiatis (2004) have shown how to consistently estimate the population mean response for the treatment policy by considering propensity score weight, when treatment duration can take on only a finite number of values, $t_1, \cdots, t_m$. However, the estimator proposed by Johnson and Tsiatis (2004) is consistent only when the model for propensity score is correctly specified. We propose a doubly robust estimator to protect again model misspecification using semiparametric theory by defining potential outcomes and regarding observed data as the coarsened full data. In addition, the new estimator is locally efficient when all the models in the estimator are correctly specified, so is more efficient than Johnson and Tsiatis' estimator.

In the end, we propose a nonparametric method to estimate the mean outcome corresponding to Definition 1 for HIV-1 infected patients as an alternative method to the semiparametric method proposed in the first topic when semiparametric method does not perform very well at small sample size, having high dimensional confounding and a highly skewed outcome. Simulation studies showed that nonparametric methods had smaller MSE than semiparametric method for the cases mentioned above.

Semiparametric efficient and robust estimation of treatment effects from observational data

By

Li Li
B.S., University of Science and Technology of China, 2002
M.S., University of Science and Technology of China, 2005

Advisor: Brent A. Johnson, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Acknowledgement

I would like to thank and acknowledge several people who helped me along my path to today. First, I'd specially thank my advisor, Professor Johnson, who always encourages me to try challenging topics, inspires me with creative ideas and gives me helpful advice over the past three years. I could not have the dissertation done without his unconditional support. I'd like to thank Professor Hanfelt for providing me many insightful suggestions since I started my study at Emory University. I am also very grateful to Professor Long and Professor Sullivan for their prompt response all the time and constructive comments. I also would like to thank Anita Chen, my supervisor at Duke Clinical Research Institute, for her encouragement and support. In addition, I wish to extend my sincere thanks to all those who helped me with my work in the Department of Biostatistics and Bioinformatics. Next, I would like to thank all of my dear friends for their warm friendship and continuous encouragement. Finally, I would like to thank my parents for their selfless and endless caring and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

The methods proposed in this dissertation are motivated by two clinical trials. The first one is to compare the effect of different treatments related to switching strategies for HIV-1 infected patients and the second one is to evaluate the effect of different treatment durations in infusion study.

1.1.1 The ACTG A5095 Data

One of the most threatening infectious diseases presently facing global public health is the human immunodeficiency virus (HIV). By the end of the year of 2008, an estimated 33.4 million people worldwide were living with HIV/AIDS (UNAIDS, 2009). At this time, there is no cure for AIDS, but medications are effective in fighting HIV and its complications. Treatments are designed to reduce HIV virus level in patients' body, keep immune system as healthy as possible and decrease the complications. Antiretroviral drug treatment is the most current treatment for HIV or AIDS. The immediate goal of antiretroviral therapy is to reduce plasma HIV-1 RNA levels to below detectable limits. Standard antiretroviral therapy consists of the use of at least

three antiretroviral (ARV) drugs to maximally suppress the HIV virus and stop the progression of HIV disease. A sequential antiretroviral treatment regime (SATR) is the most current antiretroviral method for treating HIV-1 infected patients in clinical practice and in many clinical studies. In general, a sequential antiretroviral treatment regime (SATR) is defined by first-line ARV regime followed by a switch to second-line ARV regime if virologic failure on the first; third-line ARV regime follows if failure on the second-line regime and so on. Figure 1.1.1 described an episode of SATR.

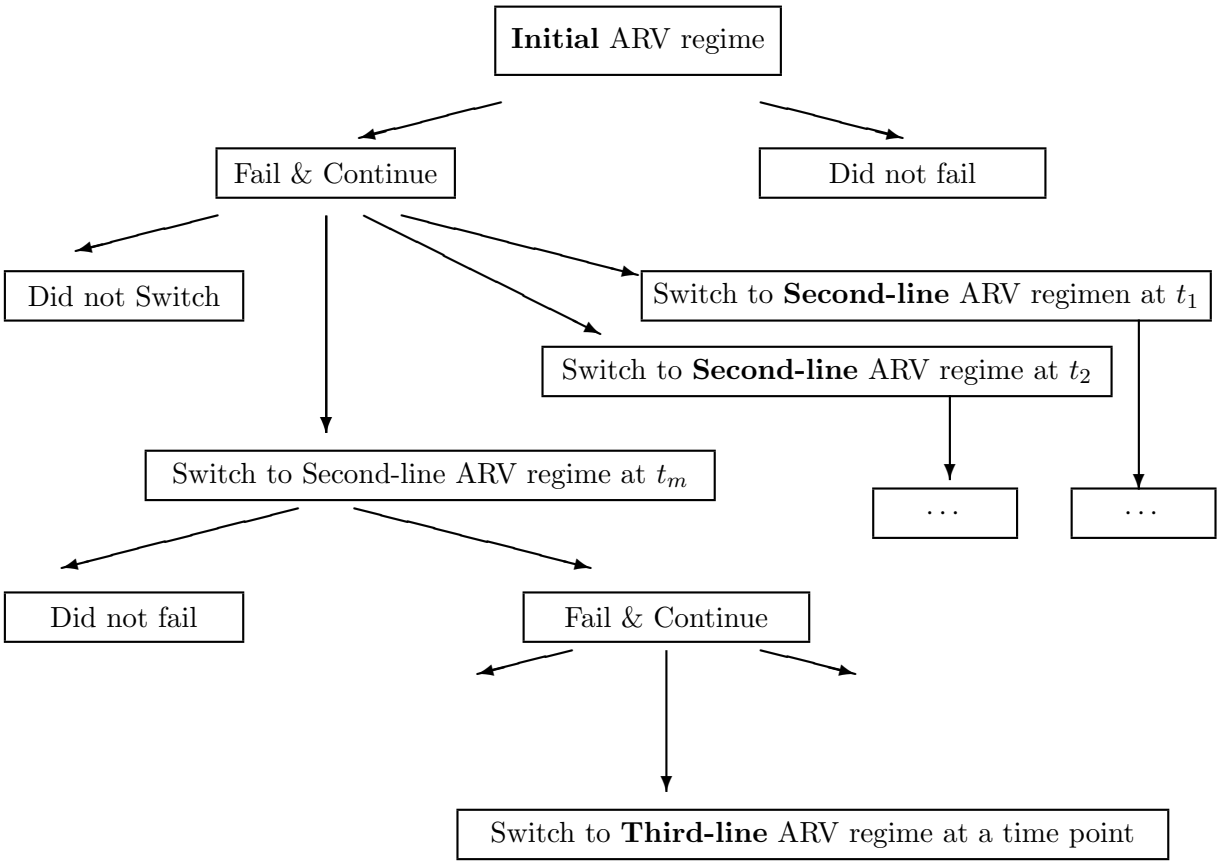


Figure 1.1: An Episode of SATR Procedure.

A regime switch includes any and all drug augmentations and/or substitutions if and only if a patient has failed on the previous regime. The current standard in clinical practice and in many clinical studies is to allow for patient/provider-initiated

treatment options or decisions after the initial treatment assignment. In antiretroviral studies, patients have historically been given treatment choices in an effort to maintain high adherence levels and avoid opportunistic infections. Now, patients are encouraged to switch regimes after (confirmed) virologic failure since increased viral replication on the same regime may compromise therapy, increase the possibility of cross resistance to other agents in the same class, and decrease the likelihood that the patient responds to subsequent antiretroviral therapy. The negative aspects of staying on a failing regime are counterbalanced by a patient who is clinically stable and tolerating their initial regime despite virologic failure. Thus, some patients will not switch ARV regimes within the course of the study while others may fail and switch regimes repeatedly. The natural question then arises, when should patients or their attending physicians switch ARVs after experienced the virologic failure on the previous treatment in order to prolong the duration of HIV RNA suppression?

ACTG A5095 was a randomized, multi-center clinical trial designed to compare three antiretroviral regimens Abacavir (ABC)/ Lamivudine (3TC)/ Zidovudine (ZDV), 3TC/ZDV/ Efarirenz (EFV), and ABC/3TC/ZDV/EFV in HIV-infected, antiretroviral therapy-naive patients with HIV-RNA levels $\leq$ 400 copies/mL. The goal of the study was to suppress and maintain HIV-1 RNA levels $<$ 200 copies/mL. The primary efficacy endpoint was time to first virologic failure, defined as two subsequent assays where HIV-1 RNA levels $\geq$ 200 copies/mL. The study was designed to last 96 weeks. After 32 weeks of follow-up, 82 of 382 patients (21 percent) in the triple NRTI group versus 85 of 765 patients (11 percent) in the combined efavirenz group experienced virologic failure; hence, the triple nucleoside reverse-transciptase inhibitor (NRTI) regimen (ABC/3TC/ZDV) appeared inferior when compared to the combined efavirenz-containing regimens. Moreover, the time to virologic failure was significantly shorter in the combined efavirenz arm. The data safety and monitoring board recommended that the triple NRTI arm be discontinued but to follow the other

two arms for the clinical endpoints for the remaining duration of the study. The initial study report (Gulick et al., 2004) details all differences between the triple NRTI and combined efavirenz groups.

All study patients who failed on the initial antiretroviral treatment regimen had the opportunity to switch ARV regimen in favor of another ARV regimen, subject to standard pharmacological restrictions (i.e. all known adverse drug interactions were disallowed). The second-line ARV regimens could include on- or off-study medications. Once virologic failure was confirmed, the decision to switch ARVs was left to the discretion of the patient and his or her attending physician. A caveat to this open invitation to switch ARV regimens post-virologic failure was that a patient was *not required* to switch regimens even after failing on the initial ARV regimen. (ACTG) study A5095 allows that the decision to switch initial ARV regimen after virologic failure is left to the patient and his or her primary-care provider. Decisions to stop, continue, or switch treatments, even in the face of incomplete suppression of HIV-1 replication, often depend on multiple other factors including a patient's medical history of ARVs, immunologic and clinical response to those ARVs or alternatively the desire to limit resistance emergence. Because the same factors that affect a patient's treatment decision and assignment may subsequently affect response, we have a classic case of confounding.

The issue when to switch antiretroviral therapy has been discussed by physicians or clinicians. The weight of evidence(Cozzi-Lepri et al. (2007), Tozzi et al. (2006), GoetzMB et al. (2006)) suggests that continued exposure to failing ARV regimes will rapidly lead to the development of HIV strains that are resistant to drugs in the failing regime and possibly to those that may be required in the future. Consequently, most HIV treatment guidelines recommend that ARV regimes are changed rapidly in patients experiencing virological failure (Gazzard et al., (2006), Hammer et al.(2006)). However, Tenorio et al. (2009) argued that delaying a treat-

ment switch in antiretroviral-treatment HIV-1 infected patients with detectable drug-resistant viremia does not have a profound effect on immune parameters using the AIDS clinical trials group study A5115. Therefore, as Dr. Deeks (2003) mentioned, the best approach remains unclear for patients who have failed multiple treatment regimes. The management of such patients requires a careful understanding of the pathogenesis of drug-resistant HIV-1, the clinical consequences of virological failure, the potential benefits and limitations of diagnostic assays, and the likelihood that agents in development will be effective. This project is to propose statistical methods to address the scientific question in HIV/AIDS research where there is an abundance of conjecture and speculation but only limited information: when to switch from a failing ARV regime?

1.1.2 ESPRIT infusion trial

The ESPRIT (Enhanced Suppression of the Platelet IIb/IIa receptor with Integrilin Therapy) trial targeted patients with coronary artery disease scheduled to undergo percutaneous coronary intervention (PCI) with stent implantation in a native coronary artery. The main objective of ESPRIT was to compare eptifibatide (Integrilin) therapy to placebo on the basis of the composite binary endpoint of death, myocardial infarction (MI), or urgent target vessel revascularization within 30 days. The study enrolled 2064 eligible patients who were randomized to either study drug (1040) or palcebo (1024) regimen. The experimental treatment regimen consisted of an eptifibatide bolus and a continuous eptifibatide infusion for 18-24 hours, with a similar regimen for the placebo group. The study protocol also required that patients experiencing serious complications, such as abrupt closure, no reflow, or coronary thrombosis immediately discontinue the infusion process to receive appropriate medical attention; we define these protocol-defined adverse events as infusion-terminating events, or more generally as treatment-terminating events.

Once treatment has been proven effective, study investigators are often interested in the best treatment duration which optimizes the response. As Johnson and Tsiatis(2004) argued that because infusion can not continue after a treatment-terminating event, a recommendation to infuse for t units of time necessarily implies that treatment would be discontinued either after drug was administered for t units of time or when a treatment-terminating event occurs. Thus, censoring is as an essential part of treatment policy and a treatment duration policy for t unites of time of interest is defined as "a recommendation to treat for t units of time or until a treatment-terminating event occurs, whichever comes first".

Johnson and Tsiatis (2004; subsequently referred to as JT) have shown how to estimate consistently the population mean response for the treatment duration policy by incorporating propensity score in the estimator without modeling outcome regression on covariates. However, the JT estimator is neither the most efficient nor doubly robust; that is, it does not remain consistent and asymptotically normal if either the propensity score model or the outcome regression model is correct. Considerable recent interest has focused on doubly robust estimators for a population mean response in the presence of incomplete data, which involve models for both the propensity score and the regression of outcome on covariates. Given the protection afforded by the property being doubly robust, these estimators have been advocated for routine use (Bang and Robins, 2005). In this paper we will propose a double robust estimator which is more efficient than the JT estimator to estimate the mean outcome corresponding to the treatment duration policy defined above.

1.2 Outline

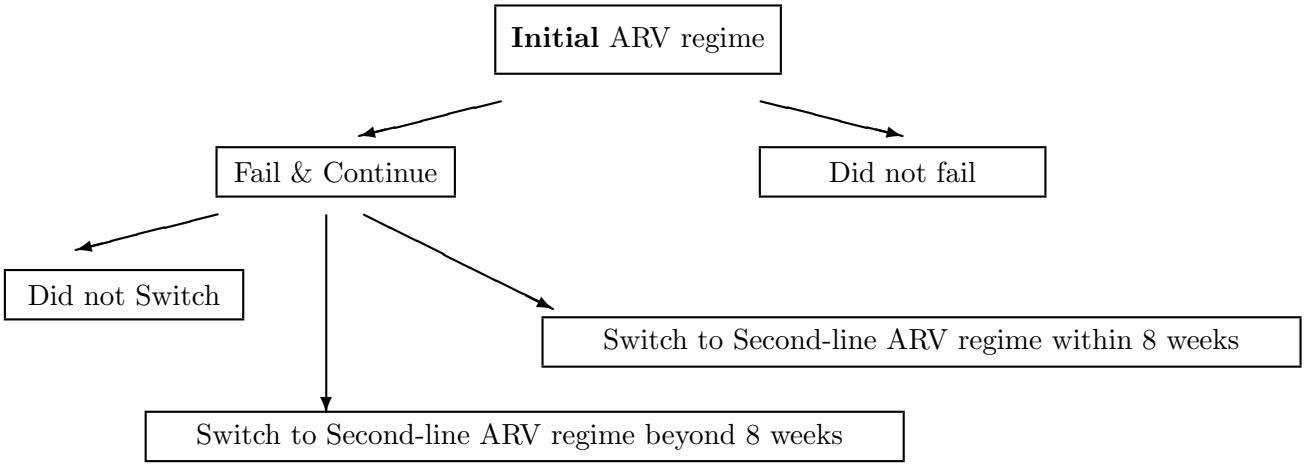
Throughout this method, we adopt the point of view proposed by Neyman (1923) and Rubin (1974), where casual effects are defined through potential outcomes or

counter-factual random variables. The organization of the dissertation is as follows:

In chapter 2, we simplify k-stage ARV regime in HIV treatment study to 2-stage regime allowing patients switch to the second-line regime only once, after they experienced virologic failure, and we consider the case where patients either switched to the second-line treatment within 8 weeks after they failed on the initial treatment or switch after 8 weeks, as in Figure 1.2. We estimate the mean outcome for the two strategies. We describe our methodology for extension of randomized two-stage design to observational study to specifically compare two policies:

- "Switching Early" Policy . A patient had started an initial regimen, he would switch early to next regimen, if virologic failure on the initial regimen
- "Switching Late" Policy . A patient had started an initial regimen, he would switch late to next regimen, if virologic failure on the initial regimen

A through case study report on ACTG5095 data including sensitivity analysis by applying the method proposed in chapter 2 is provided, alongside simulations that highlight interesting features of the methods such as increased power, improved efficiency, etc.



In Chapter 3, we propose how to develop a doubly robust and efficient semiparametric estimator for mean outcome corresponding to the policy, which is "A treatment

duration policy for t units of time as a recommendation to treat for t units of time or until a treatment-terminating event occurs, which comes first", in infusion study, followed by the discussion on properties of proposed estimator and simulation studies comparing double robustness and efficiency with the estimator proposed by Johnson and Tsiatis (2004).

In Chapter 4, we propose a nonparametric method to estimate the mean outcome corresponding to Definition 1 for HIV-1 infected patients as an alternative method to the semiparametric method proposed in Chapter 2 when semiparametric method does not perform very well at small sample size and high dimensional confounding. Simulation studies to compare semiparametric and nonparametric methods are provided.

1.3 Contribution

Estimating the mean clinical endpoint on a given ARV treatment policy and comparing policy changes are significant and important practical problems in current HIV/AIDS research. Our methods proposed in the dissertation specifically addressed the following two policies related to the issue of when to modify treatment.

Definition 1 (Dual-stage ARV Regimen Policy)**.** *An initial ARV regimen "a" followed by a switch at time "s" to any second-line ARV regimen, if virologic failure on the initial regimen*

Definition 2 (Duration Stopping Rule)**.** *A treatment duration policy for t units of time as a recommendation to treat for t units of time or until a treatment-terminating event occurs, which comes first.*

The first method would estimate average causal effect for two regime policies: a policy that switches to new ARV regime soon after confirmed virologic compared to delayed ARV regime change. The final clause in Definition 2, *if virologic failure on*

the initial regimen, is rather important because it underscores the practical relevance of our policy by reflecting standard clinical practice. The clause in Definition 2 also necessarily implies that patients assigned to second-line treatment are not a random sample of the study population. We developed our estimator by augmenting IPW estimator incorporating with outcome regression to improve efficiency and provide double robustness under model misspecification for the average causal effect of the policy change via Definition 2. Moreover, our method is an extension of Wahed and Tsiatis (2004) to a two-stage design where treatment assignment at the second stage is confounded.

Another contribution is we provide a thorough case study of the ACTG A5095 data. We proposed unusually different endpoints which can be computed even when a mortality outcome is not available and we found that for patients on a efavirenz-based ART , regimen changes made within 8 weeks of confirmed virologic failure on initial ARV regimen were associated with lower cumulative HIV RNA level, higher cumulative CD4 cell counts, and spent a larger proportion of the follow-up period with suppressed HIV RNA levels, on average. To the best of our knowledge, this is the first paper to report such findings.

We proposed an estimator having improved efficiency and double robustness for the treatment duration policy in the infusion study, compared to the estimator proposed by Johnson and Tsiatis (2004). The method can be applied to estimate consistently the population mean response for policy in Definition 1 in any observational duration-response studies with duration possibly right-censored.

Nonparametric analysis using boosting algorithm proposed in the third topic is to alleviate the impact of assumption of wrong working model on the bad performance of semiparametric estimators. When semiparametric estimator we proposed in the previous chapters does not perform very well because of small sample size, the large number of confounding, or incorrect assumption of working models, non-parametric

methods provides an alternative choice to analyze data with those features by avoiding the hypothesis that the regression function belongs to a certain finite-dimensional parametric family (Gonzalez-Manteiga and Perez-Gpnzalez, 2004).

Chapter 2

Optimal Estimation of Mean Endpoint on Two-stage Sequential Antiretroviral Treatment Regimen Using Observational HIV Data

2.1 Introduction and Background

In HIV treatments, a sequential antiretroviral treatment regimen (STAR) is defined by first-line initial ARV regimen followed by a switch to second-line ARV regimen if virologic failure on the first-line regimen; third-line ARV regimen follows if failure on the second-line regimen and so on. When to switch to the next line regimen after experiencing virologic failure has caught clinicians and statisticians' attention in HIV-1 studies recently. For the time being we focus on two-stage sequential antiretroviral treatment regimen where patients are only allowed to switch once and we will extend to k-stage sequential antiretroviral treatment regimen where multiple switches are allowed during the whole clinical practice in the future. Because only those patients who

failed on the initial treatment have chance to enter the second-line ARV regimen, most of current methods addressed the issue of when to switch by retrospectively identifying all subjects who experienced virologic failure on the initial treatment, such as the application of history-adjusted marginal structural model(Petersen et al, 2008) and the application of semiparametric method proposed by Johnson and Tsiatis (2004). However, the endpoint is not only the result of the second-line regimen but also the consequences of combination of initial treatment and second-line regimen. Therefore, estimating mean endpoint as the consequence of combination of sequential treatment regimens is more attractive and reasonable than merely focusing on the second-line regimen, that is, only including patients who experienced virologic failure in the analysis. When randomization up to front is not available in reality due to the fact that decision when to switch to the second-line regimen after virologic failure is left to patients themselves or physicians, two-stage randomized design which allowing patients who meet the criteria at the end of first stage enter the second stage and randomly receive treatments at the beginning of second stage becomes promising. Lunceford et al.(2002) proposed inverse probability weighting (IPW) methods for 2-stage design. In this chapter, we would extend his method to 2-stage SATR where at the beginning of the second stage patients received one of two treatments not randomly using observational data and two treatments options are defined as switch to the second-line treatment within or after a specific period. We propose a doubly robust estimator which Lunceford's estimator is not and the proposed estimator has an improved efficiency over IPW estimators whenever propensity score model is correctly specified by borrowing Tan (2006)'s idea. In addition, we apply our method to ACTG5095 data and propose alternative endpoints which is computed when endpoints related to mortality are not available.

Figure 2.1: Antiretroviral treatment strategy in ACTG A5095.

2.2 Methods

In this section, we describe the work in details that forms the basis of our proposed research.

2.2.1 Causal Model: Notation and Assumptions

2.2.1.1 Potential outcome framework

As in previous chapter , throughout this chapter we adopt Rubin's causal model (1974) and the ideas of potential random variables (Neyman, 1923; Rubin, 1974). We define Ya, s as potential outcomes for which a patient would have started initial treatment a and then second-line treatment s if he experienced virologic failure on the initial regimen. For simplicity, consider the two potential outcomes, $Y(a, 0)$ and $Y(a, 1)$, where $Y(a, s)$ is the outcome that one would observe if a patient were assigned to policy (a, s), where $a \in \{0, 1\}$ corresponds to combined efavirenz (0) or triple-nucleoside (1) therapies and $s \in \{0, 1\}$ corresponds to switch early (0) versus switch late (1) after virologic failure. We also define the potential random variable $R(a)$ as the failure status indicator on the initial treatment $a \in \{0, 1\}$. Hence, the full set of potential random variables for a randomly selected patient from the population is $\{R(0), R(1), Y(0, 0), Y(0, 1), Y(1, 0), Y(1, 1)\}$.

2.2.1.2 Identifiability and Consistency

We assume that potential outcomes for patients who do not fail on initial ARV treatment would be the same whether they were assigned to a policy s which switched early or late after virologic failure. That is, we assume that the distribution of

$\{R(a), Y(a,0), Y(a,1)\}$ obeys the constraint

$$Y(a,0) = Y(a,1) \text{ if } R(a) = 0\ , \quad a = 0, 1. \tag{2.1}$$

In the observational study, one only observes the random variables $\{A, R, RS, Y, \mathbf{X}\}$. Here, A is the initial ARV assignment indicator (1 for $a = 1$), R is an indicator of virologic failure (1 if failed virologically on initial regimen), S is an indicator of switching early (0) versus late (1) and is only observed for those patients who failed on the initial ARV regimen (i.e. $R = 1$), Y is the observed outcome, and $\mathbf{X}$ are patient characteristics. We assume that the potential failure indicators, $R(0)$ and $R(1)$, are related to the observed failure indicator R through $R = (1-A) \times R(0) + A \times R(1)$. Furthermore, using, we assume that potential outcomes $\{Y(a,0), Y(a,1)\}$ and observed outcome Y are related through

$$Y = (1-R) \times Y(a,0) + R(1-S) \times Y(a,0) + RS \times Y(a,1) \quad \text{on the event } \{A = a\}. \tag{2.2}$$

Incidentally, because we assume (2.1), then we also have that (2.2) $= (1 - R) \times Y(a,1) + R(1 - S) \times Y(a,0) + RS \times Y(a,1)$.

2.2.1.3 Treatment assignment

In ACTG A5095, subject to exclusion criteria, patients were randomly assigned to initial ARV treatment independent of any patient characteristics. Hence, we have that $P(A = 1|Y(a,0), Y(a,1), \mathbf{X}) = P(A = 1)$, because of randomization. Similarly, it is necessary to consider how patients are assigned to switch ARVs early versus late post-virologica failure. Let the random variable S denote whether a patient is assigned to the switch early (0) group versus switch late (1) group. Here, patients were not assigned to treatment independent of individuals characteristics; rather patients and their attending physicians intentionally chose to switch early versus late based

on individual cases. Hence, we have $P(S = 1|R = 1, Y(a,0), Y(a,1), \mathbf{X}) \neq P(S = 1|R = 1)$. However, we do assume

$$P(S = 1|R = 1, Y(a,0), Y(a,1), \mathbf{X}) = P(S = 1|R = 1, \mathbf{X}) \quad \text{on the event } \{A = a\}. \tag{2.3}$$

Expression (2.3) is often referred to as the no unmeasured confounders assumption or sequential randomization. We interpret (2.3) as the assumption that a decision to switch early versus late depends on auxiliary variables $\mathbf{X}$ up to that point at which a switch is made but not dictated by future events $Y(a,0)$ and $Y(a,1)$.

2.2.2 Estimation in the observational study

Statistical methods for estimation and testing in two-stage randomization designs (Corimer et al.; Stone et al., 1995; Lunceford et al., 2002) are now well developed and are distinguished from standard two-sample problems by the failure indicator R. Because not every patient fails on initial ARV treatment regimen, we only see a sicker subpopulation on switching to second-line ARV regimens. In order to estimate $E\{Y(a,s)\}$ consistently, a proper adjustment for patients who fail and switch versus those who do not fail is required.

2.2.2.1 Hypothetical two-stage trial

To motivate our estimator, consider the hypothetical two-stage randomization trial where treatment assignment at the second-stage is randomized so $P(S = 1|A = a, R = 1, Y(a,0), Y(a,1), \mathbf{X}) = P(S = 1|A = a, R = 1)$ for each $a \in \{0,1\}$. This hypothetical trial is exactly the scenario considered by Lunceford et al. (2002). Using inverse weighting methods and standard conditioning arguments via (2.1) and (2.2),

Lunceford et al. (2002) show that for each (a, s) combination, we have

$$\mu_{as} = E_{\bar{d}_2}\left[Y\mathbf{1}(A=a)\times\left\{(1-R)+\frac{R\mathbf{1}(S=s)}{P_{\bar{d}_2}(S=s|A=a,R=1)}\right\}\right] \qquad (2.4)$$

where $E_{\bar{d}_2}$ and $P_{\bar{d}_2}$ are expectation and probability in the hypothetical two-stage trial, respectively. A sample average of the random variables on the right-hand side of (2.4) will be a consistent and asymptotically normal estimator for the estimand μ_{as} under (2.1) and (2.2). The interpretation of the weighting scheme in (2.4) is as follows: if a patient does not fail on the initial ARV regimen a, then that patient represents him/herself and hence receives a weight of one; if a patient fails on initial ARV regimen, however, then that patient represents $\{P_{\bar{d}_2}(S=s|A=a,R=1)\}^{-1}$ similar patients who could have potentially been assigned to combined first- and second-line ARV regimens (a, s). In the following subsection, we show how to extend the expression (2.4) for the analysis of the ACTG A5095 data.

2.2.2.2 The Radon-Nikodym derivative

Murphy, van der Laan, and Robins (2001) give an elegant result for estimating the marginal means of potential random variables when treatment assignment is sequential and depends on patient characteristics. Under regularity conditions, including $P(S=s|A=a,R=1,\mathbf{X})>0$, Lemma 4.1 of Murphy et al. (2001) asserts that the distribution of (Y, A, R, S) under $P_{\bar{d}_2}$ is absolutely continuous with respect to the distribution of (Y, A, R, S) under P, and a version of the Radon-Nikodym derivative is $E\left\{W_{\bar{d}_2}(A,R,S,\mathbf{X})|Y=y,A=a,R=r,S=s\right\}$, where

$$W_{\bar{d}_2}(a,r,s,\mathbf{x}) = \mathbf{1}(A=a)\times\left\{\mathbf{1}(R=0)+\mathbf{1}(R=1,S=s)\frac{P_{\bar{d}_2}(S=s|A=a,R=1)}{P(S=s|A=a,R=1,\mathbf{X}=\mathbf{x})}\right\}.$$

The Radon-Nikodym derivative can be applied directly to the expression in (2.4) to show that

$$\begin{aligned} & E_{\bar{d}_2}\left[Y\mathbf{1}(A=a)\times\left\{(1-R)+\frac{R\mathbf{1}(S=s)}{P_{\bar{d}_2}(S=s|A=a,R=1)}\right\}\right] \\ = \;& E\left[W_{\bar{d}_2}(A,R,S,\mathbf{X})\times Y\mathbf{1}(A=a)\times\left\{(1-R)+\frac{R\mathbf{1}(S=s)}{P_{\bar{d}_2}(S=s|A=a,R=1)}\right\}\right] \\ = \;& E\left[Y\mathbf{1}(A=a)\times\left\{(1-R)+\frac{R\mathbf{1}(S=s)}{P(S=s|A=a,R=1,\mathbf{X})}\right\}\right]. \end{aligned} \tag{2.5}$$

The last expression in (2.5) is a function of the observed data and a sample average yields a consistent estimator for μ_{as} if the propensity score $P(S=s|A=a,R=1,\mathbf{X})$ were known.

By definition, the propensity score $P(S=s|A=a,R=1,\mathbf{X})$ is identified by those patients who failed on their initial ARV regimen (i.e. $R=1$) and not by the entire sample. Hence, in small samples, estimation and inference for the causal estimand μ_{as} may be sensitive to the overall marginal probability of failing on initial ARV regimen. Nevertheless, the propensity score may be modeled parametrically, semiparametrically, or nonparametrically as a function of $(A,\mathbf{X})$. The most common approach is to model $P(S=s|A=a,R=1,\mathbf{X})$ using maximum likelihood via generalized linear models (e.g. probit or logistic regression) and separately for each initial ARV regimen ($a=0,1$). Substituting the fitted propensity score $\widehat{P}(S=s|A=a,R=1,\mathbf{X})$ in expression (2.5) and taking a sample average leads to the inverse-probability weighted (IPW) estimator

$$\widehat{\mu}_{as}=\mathcal{E}_n\left[Y\mathbf{1}(A=a)\times\left\{(1-R)+\frac{R\mathbf{1}(S=s)}{\widehat{P}(S=s|A=a,R=1,\mathbf{X})}\right\}\right],$$

where $\mathcal{E}_n$ denotes sample average (i.e. $\mathcal{E}_n f(Z)=\sum_{i=1}^n f(Z_i)/n$). Assuming that the propensity model is correctly specified and under standard regularity conditions, $\sqrt{n}(\widehat{\mu}_{as}-\mu_{as})$ converges in distribution to a mean-zero normal random variable with

asymptotic variance that can be derived using standard arguments (e.g. Tsiatis, 2006) and consistently estimated from the data (e.g. Lunceford et al., 2002; Tsiatis, 2006). A consistent estimator for the average causal effect (ACE) for switching ARV regimens early versus late on the initial ARV regimen $A = a$ on the combined efavirenz arm is given by

$$\text{ACE} = \widehat{\mu}_{01} - \widehat{\mu}_{00},$$

and similarly for the triple nucleoside arm. Unfortunately, it is well-known that IPW estimators are inefficient. In the following subsection, we show to improve the precision and robustness of IPW estimators through adaptation, maximum likelihood and the theory of control variates.

2.2.3 Doubly-Robust, Locally Efficient, and Optimal Estimation

Semiparametric efficient estimation has received a significant amount of attention in the statistical literature over the past two decades beginning with two comprehensive accounts by Newey (1990) and (Bickel, Klassen, Ritov, and Weller, 1993). Newey's methods were applied to a general class of missing data problems by Robins, Rotnizky, and Zhao (1994). Because our two-stage estimation problem may be considered as type of missing data problem, the Robins et al. (1994) theory applies here as well. In fact, Wahed and Tsiatis (2004) have considered different efficient and optimal (in the sense of Robins et al., 1995) estimators for the two-stage randomization design. Through control variates and Monte Carlo integration, Tan (2006) recently offered a novel estimator which offers some finite sample advantages over that proposed by Robins et al. (1995). Our contribution is to extend Tan's estimator to our estimation problem and subsequently apply it to the ACTG A5095 data. When the decision to switch to second-line ARV regimens is independent of patient characteristics $\mathbf{X}$,

our two-stage estimation problem reduces to the two-stage randomization design considered by Wahed and Tsiatis (2004). In this case, our estimator improves on the estimator by Wahed and Tsiatis (2004) in the same way that Tan's (2006) estimator improves on the estimator by Robins et al. (1995) for the simple one- and two-sample causal estimands.

For the purposes of robust and efficient estimation, we drop the indicator $\mathbf{1}(A = a)$ from our estimators with the understanding that estimators are calculated separately for each initial ARV regimen $a = 0, 1$. Also, we will restrict our attention to the estimator where patients "switch early" to second-line ARV regimens, i.e. $\{S = 1\}$; that is, we restrict our attention to the following estimand and IPW estimator

$$\mu = E\left[Y\left\{(1-R) + \frac{RS}{\pi(\mathbf{X})}\right\}\right] \quad \text{and} \quad \widehat{\mu}_{\text{IPW}} = \mathcal{E}_n\left[Y\left\{(1-R) + \frac{RS}{\widehat{\pi}(\mathbf{X})}\right\}\right], \quad (2.6)$$

with propensity score $P(S = 1|R = 1, \mathbf{X}) = \pi(\mathbf{X})$. The IPW estimator for "switch late" to second-line ARV regimen is given by replacing $S/\widehat{\pi}(\mathbf{X})$ in (2.6) with $(1 - S)/\{1 - \widehat{\pi}(\mathbf{X})\}$. We now show how to construct improved estimators by considering their influence functions.

2.2.3.1 Semiparametric AIPW class of estimators

For regular and asymptotically linear (RAL) estimators, we know their influence function satisfies $(\widehat{\mu} - \mu_0) \cong \mathcal{E}_n\varphi(\mathbf{Z})$, where $\mathbf{Z}$ are the observed data $(Y, R, RS, \mathbf{X})$ and "$\cong$" denotes a difference of the order $o_p(n^{-1/2})$. A straightforward application of the Robins et al. (1994) theory suggests that all RAL estimators $\widehat{\mu}$ have influence function belonging to the class

$$\Phi = \left\{\varphi \;\middle|\; \varphi = \left[\left\{(1-R) + \frac{RS}{\pi(\mathbf{X})}\right\}Y - \mu + R\left(\frac{S}{\pi(\mathbf{X})} - 1\right)h(\mathbf{X})\right], h \in \mathcal{H}\right\}, \quad (2.7)$$

where $\mathcal{H}$ consists of all arbitrary functions of $\mathbf{X}$. In the missing data literature,

the last expression $R\{S/\pi(\mathbf{X}, \boldsymbol{\psi}) - 1\}h(\mathbf{X})$ is called the augmentation term. The IPW influence function φ_{IPW} is found when $h \equiv 0$; hence, $\varphi_{\text{IPW}} \in \Phi$ trivially. The semiparametric efficient estimator (i.e. the one with smallest asymptotic variance in the class Φ) is found when $h_{\text{eff}} = E(Y|R = 1, \mathbf{X})$ and its influence function is denoted by φ_{eff}. Because the true conditional mean model $E(Y|R = 1, \mathbf{X})$ is unknown, we posit a statistical model model for it, say $E(Y|R = 1, \mathbf{X}) = m(\mathbf{X})$. Then, replacing $h(\mathbf{X})$ with the posited model $m(\mathbf{X})$ leads to the augmented inverse-probability weighted (AIPW) class of estimators:

$$\mathcal{C}_{\text{AIPW}} = \left\{ \widehat{\mu} \;\middle|\; \widehat{\mu} = \widehat{\mu}_{\text{IPW}} - \mathcal{E}_n \left[R \left\{ \frac{S}{\widehat{\pi}(\mathbf{X})} - 1 \right\} \widehat{m}(\mathbf{X}) \right], m \in \mathcal{M} \right\},$$

where $\mathcal{M} \subseteq \mathcal{H}$ and defines a subset by modeling the first conditional moment $m(\mathbf{X}) = E(Y|R = 1, \mathbf{X})$. Clearly, the influence function φ_{AIPW} belongs to the class Φ with $h(\mathbf{X}) = m(\mathbf{X})$. In addition, we know that (a) the AIPW estimator is consistent if either $\pi(\mathbf{X})$ or $m(\mathbf{X})$ is correctly specified (i.e. double robustness), and (b) the AIPW estimator is semiparametric efficient if $\pi(\mathbf{X})$ and $m(\mathbf{X})$ are correctly specified (i.e. local efficiency). Unfortunately, if the conditional mean model $m(\mathbf{X})$ is incorrectly specified, there is no guarantee that the AIPW estimator is more efficient than the IPW estimator. Hence, the estimator class $\mathcal{C}_{\text{AIPW}}$ will not include the efficient estimator if $m(\mathbf{X})$ is misspecified even though it does include the efficient estimator when $m(\mathbf{X})$ is correctly specified. These considerations lead to other classes of estimators which may be motivated through maximum likelihood (Tan, 2006) or the theory of control variates (Hammersley and Handscomb, 1964).

2.2.3.2 The regression estimator

Tan (2006, Theorem 2) proposed "regression" or "tilde" estimators and belong to a larger family of optimal control variate (CV) estimators (Hammersley and Hand-

scomb, 1964). Now, we describe this family of estimators and detail how they are implemented in practice for our estimation problem. First, we model the propensity score

$$\text{logit } \pi(\mathbf{X}, \boldsymbol{\psi}) = \psi_0 + \psi_1 X_1 + \cdots + \psi_{q-1} X_{q-1}, \tag{2.8}$$

where $\boldsymbol{\psi} = (\psi_0, \ldots, \psi_{q-1})^{\mathrm{T}}$ is a q-vector of unknown parameters for predictors including baseline CD4 and two nadir RNA covariables (See Section 2.3). Second, we model the conditional mean of Y given $\mathbf{X}$ *for patients who failed on initial ARV regimen* linearly as

$$E(Y|R=1, \mathbf{X}) = m(\mathbf{X}, \boldsymbol{\xi}) = \xi_0 + \xi_1 X_1 + \cdots + \xi_{r-1} X_{r-1}, \tag{2.9}$$

where $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_{r-1})^{\mathrm{T}}$ is an r-vector of unknown parameters for covariables including baseline CD4 and nadir RNA levels. Now, define the following class of estimators

$$\mathcal{C}_{\mathrm{CV}} = \left\{ \widehat{\mu} \;\middle|\; \widehat{\mu} = \widehat{\mu}_{\mathrm{IPW}} - \kappa \times \mathcal{E}_n \left[R \left\{ \frac{S}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})} - 1 \right\} m(\mathbf{X}, \widehat{\boldsymbol{\xi}}) \right], \kappa \in \Re \right\}.$$

A first-order approximation to $\widehat{\mu}_{\mathrm{CML}}$ leads to the following optimal CV estimator,

$$\widehat{\mu}_{\mathrm{OPT}} = \widehat{\mu}_{\mathrm{IPW}} - \widetilde{\kappa} \times \mathcal{E}_n \left[R \left\{ \frac{S}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})} - 1 \right\} m(\mathbf{X}, \widehat{\boldsymbol{\xi}}) \right],$$

where $\widetilde{\kappa}$ is the first element of $\{\mathcal{E}_n(VW^{\mathrm{T}})\}^{-1} \mathcal{E}_n(VU)$,

$$U = \frac{RSY}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})}, \qquad V = R \left(\frac{S}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})} - 1 \right) \times G, \quad W = R \left(\frac{S}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})} \right) \times G,$$

$$G = \begin{pmatrix} m(\mathbf{X}, \widehat{\boldsymbol{\xi}}) \\ \left\{ (\partial / \partial \boldsymbol{\psi}) \pi(\mathbf{X}, \widehat{\boldsymbol{\psi}}) \right\} / \left\{ 1 - \pi(\mathbf{X}, \widehat{\boldsymbol{\psi}}) \right\} \end{pmatrix}.$$

Alternatively, Robins et al. (1995) defined $\widehat{\kappa}$ as the first element to the classic multiple linear regression $\{\mathcal{E}_n(VV^{\mathrm{T}})\}^{-1}\mathcal{E}_n(VU)$. Such use of control variates lead to the estimator ,

$$\widehat{\mu}_{\mathrm{RRZ}} = \widehat{\mu}_{\mathrm{IPW}} - \widehat{\kappa} \times \mathcal{E}_n \left[R \left\{ \frac{S}{\pi(\mathbf{X}, \widehat{\boldsymbol{\psi}})} - 1 \right\} m(\mathbf{X}, \widehat{\boldsymbol{\xi}}) \right].$$

Note that both $\widehat{\mu}_{\mathrm{RRZ}}$ and $\widehat{\mu}_{\mathrm{OPT}}$ belong to the class $\mathcal{C}_{\mathrm{CV}}$. A key difference between $\widehat{\mu}_{\mathrm{OPT}}$ and $\widehat{\mu}_{\mathrm{RRZ}}$ is that the former is doubly robust while the latter is not (Tan, 2008, Proposition 4). Standard errors may be estimated using usual sandwich formulae via M-estimation theory (cf. van der Vaart, 1998, ch. 5).

2.2.4 Estimating Equations and Asymptotic Variance

Under standard regularity conditions, our estimator behaves asymptotically as if the parameter κ_j, $j = 0, 1$ were known *a priori* and defined in 3.3.3. For completeness, we define the whole system of estimating equations, $0 = \mathcal{E}_n \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{Z}, \boldsymbol{\theta})$, where $\boldsymbol{\phi}_{\boldsymbol{\theta}} = (\phi_{\mu_1}, \phi_{\mu_0}, \boldsymbol{\phi}_{\boldsymbol{\psi}}^{\mathrm{T}}, \boldsymbol{\phi}_{\boldsymbol{\xi}_1}^{\mathrm{T}}, \boldsymbol{\phi}_{\boldsymbol{\xi}_0}^{\mathrm{T}})^{\mathrm{T}}$, and,

$$\begin{aligned}
\phi_{\mu_1} &= Y \left\{ (1-R) + \frac{RS}{\pi(\mathbf{X}, \boldsymbol{\psi})} \right\} - \widetilde{\kappa}_1 \times R \left\{ \frac{S}{\pi(\mathbf{X}, \boldsymbol{\psi})} - 1 \right\} (1, m(\mathbf{X}, \boldsymbol{\xi}_1))^{\mathrm{T}} - \mu_1, \\
\phi_{\mu_0} &= Y \left\{ (1-R) + \frac{R(1-S)}{1-\pi(\mathbf{X}, \boldsymbol{\psi})} \right\} - \widetilde{\kappa}_0 \times R \left\{ \frac{1-S}{1-\pi(\mathbf{X}, \boldsymbol{\psi})} - 1 \right\} (1, m(\mathbf{X}, \boldsymbol{\xi}_0))^{\mathrm{T}} - \mu_0, \\
\boldsymbol{\phi}_{\boldsymbol{\psi}} &= R\{S - \pi(\mathbf{X}, \boldsymbol{\psi})\}\mathbf{X}, \\
\boldsymbol{\phi}_{\boldsymbol{\xi}_1} &= RS\{Y - m(\mathbf{X}, \boldsymbol{\xi}_1)\}\mathbf{X}, \\
\boldsymbol{\phi}_{\boldsymbol{\xi}_0} &= R(1-S)\{Y - m(\mathbf{X}, \boldsymbol{\xi}_0)\}\mathbf{X}.
\end{aligned}$$

Standard arguments lead to the usual sandwich formula for the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ and a consistent estimator given by $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}(\widehat{\mathbf{A}}^{-1})^T$, The asymptotic covariance estimator for $\boldsymbol{\mu} = (\mu_1, \mu_0)^{\mathrm{T}}$, i.e. $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \widehat{\mathrm{cov}}(\widehat{\boldsymbol{\mu}})$, is the upper right 2×2 matrix of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Our Wald test statistic for the difference between early vs. late switch $T = (C\mu)^{\mathrm{T}}(C\boldsymbol{\Sigma}_{\boldsymbol{\mu}}C^{\mathrm{T}})^{-1}(C\mu)$, where C=(1, -1) is asymptotically distributed as χ_1^2

under the null hypothesis of no difference.

2.2.5 Length-adjusted Area Under the Curve

In two-stage analyses, the endpoint Y must be well-defined for all treatment combinations. Perhaps the most natural endpoint is death or time-to-death. Fortunately, mortality is no longer considered a primary endpoint in many HIV studies because ARV regimens and our ability to treat HIV has significantly improved. Also, because some HIV studies do not extend beyond 2-3 years of follow-up, few HIV-related deaths are actually recorded. In A5095, for example, only 24 patients died during 5 years of follow-up. In the sequel, we use length-adjusted area-under-the-curve (AUC) endpoints.

Mean area-under-the-curve (AUC) of the outcome (or the outcome minus its baseline value) over time are often compared among treatment groups in many clinical trials where subjects are evaluated for a continuous outcome (e.g. drug concentration, HIV-1 viral load, CD4 T-Cell count) at multiple fixed study time points (Spritzler, 2008). When the length of follow up varies by subject, this procedure is commonly modified by defining time-averaged AUC as the area under the curve from the first to the last observed evaluation, divided by the time from the first to the last observed evaluation, which is called length-adjusted area under the curve. AUC analysis provide an obvious way to combine measurements across timepoints, even if data may be missing at certain time points. For example in the briefing document produced by Gilead for the NDA review of tenofovir, the mean AUC of HIV-1 RNA at 24 weeks adjusted for baseline was compared between patients receiving tenofovir and those receiving placebo within subgroups defined by baseline resistance mutations (FDA, 2001). The co-primary endpoint of the randomized placebo-controlled clinical trial of a Merck therapeutic vaccine for HIV, A5197, in the AIDS Clinical Trial Group, is the HIV-1 RNA AUC during a sixteen week analytical treatment interruption phase

(AACTG, 2007). In a clinical trial of colloids versus crystalloids for fluid resuscitation in critically ill patients a secondary endpoint was the AUC of mean arterial pressure over 24 hours (NCT00318942, 2007). The secondary endpoints of the HEGPOL randomized placebo-controlled trial of glycine in the postoperative phase of liver transplantation included the AUCs of AST, ALT and bilirubin serum levels over the first eight days after transplantation (HEGPOL, 2005). In ACTG A5095 analysis where mortality related outcomes are not available we will use length-adjusted area-under-the-curve related endpoints to evaluate the performance of different regimens.

2.3 Analysis of the ACTG A5095 data

2.3.1 The study sample

A total of 1147 subjects enrolled in the ACTG A5095 between March 2001 and November 2002. Of the original 1147 patients, 12 patients never started their initial treatments. These twelve patients were removed from our analysis.

As mentioned briefly in the previous section, the data safety and monitoring board discontinued the triple-nucleoside arm at the second annual review in February 2003. For ease of exposition, our analysis is restricted to the combined efavirenz arm. Thus, our effective sample size is 758 patients in the combined efavirenz groups, where 146 (19.3% of 758) patients experienced virologic failure on their initial ARV treatment regimens. An additional 50 (6.6% of 758) patients experienced virologic failure but only after a protocol-approved substitution; hence, they were not following their initial ARV at the time of failure. These 50 patients are assumed to not switch within 8 weeks after the first failure in the main analysis. The robustness and sensitivity of our analytic results excluding these 50 patients is considered below.

2.3.2 Treatment and endpoint definitions

The results of our analyses depend critically on the definitions of virologic failure (R), switching off a failing regimen early versus late (S), and the endpoint (Y). Our definition of confirmed virologic failure follows one defined in the ACTG A5095 protocol: lab readings from two consecutive visits where HIV-1 RNA $\geq$ 200 copies/mL. We define "failure" as "confirmed virologic failure on the first-line treatment regimen" and first-line regimen include initial ARV regimen plus any protocol-approved substitutions. Our definitions for "early" versus "late" ARV regimen switch and three different outcomes are described in the paragraphs below.

Our preferred definition for switching ARV regimens "early" was switching regimens less than eight weeks after confirmed virologic failure. In Figure 1, we see that the proportion of patients switching ARV regimen less than eight weeks was 16.8% (31 of 196) patients on the combined efavirenz arm.

Now we are defining our endpoints using length-adjusted area under the curve as talked in section (2.2.5). Let $H(t)$ be the HIV viral load or CD4 cell count at time t and $\alpha(t)$ a non-negative weight function. A patient's AUC is defined by the Riemann-Stieltjes integral

$$\text{AUC} = \int_0^L H(t)\alpha(t)\, dt,$$

where L is the patient's length of follow-up and follow-up is defined as length from the first drawing viral load(or CD4) date to the off study date. Our endpoints are defined $Y = \text{AUC}/L$ and interpreted as the length-adjusted AUC, introduced by Spritzler et al. (2008). We adopt length-adjusted AUC instead of original AUC to adjust for difference in follow-up time. For example, suppose patient 1 has HIV= 100 copies/mL for 1-year follow-up while patient 2 has HIV=100 copies/mL for 2-years follow-up. Apparently patient 2 has better performance on sustaining viral load below a limit

copies than patient 1. Without adjusting follow-up length, AUC of viral load for patient 2 is twice that of patient 1; however, length-adjusted AUC for patient 1 and patient 2 are the same. In practice, the AUC is approximated through the Riemann sum $\sum_{j=1}^{J} H_j \Delta\alpha_j$, where $H_j = H(t_j)$ for $j = 1, \ldots, J$ and $\Delta\alpha_j = \alpha(t_j) - \alpha(t_{j-1})$. One conventional definition uses a constant weight $\alpha(t) = 1$ which implies $\Delta\alpha_j = (t_j - t_{j-1})$ while the modified definition $\Delta\alpha_j = [(t_{j+1} - t_j) + (t_j - t_{j-1})]/2$ leads to the linear trapezoidal rule (Yeh and Kwan, 1978); we report results using the latter definition. Missing data was dealt with using the principle of "last value carry forward". Here, we present a simple example to show how we dealt with the missing value at certain time point. If VL is 1000 at t1 and 6 month later is 2000 at t2, then AUC of VL for this 6 month period is $\frac{1}{2} \times (1000 + 2000) \times 6 = 9000$. If a patient has VL 1000 at t1, and goes off with no value after t1, AUC is $\frac{1}{2} \times (1000 + 1000) \times 6 = 9000$ for time period from t1 and t2. We consider three specific endpoints: (i) $H(t)$ is HIV-1 RNA (copies/mL) at time t and $\alpha(t) = 1$ for all t; (ii) $H(t) = 1$ and $\alpha(t) = \mathbf{1}\{H(t) \leq 200\}$; (iii) $H(t)$ is CD4 cell count at time t with $\alpha(t) = 1$. The first endpoint is interpreted as cumulative HIV with large values suggesting sicker patients. We interpret the second endpoint as proportion of time with non-detectable viral load or time below a limit of detection adjusted for lengths of follow-up. The third endpoint is the same as the first but with CD4 cell count replacing viral load. Length-adjusted AUC of viral load and CD4 counts are transformed on a natural logarithmic scale(see also, subsection 4.4). The mean and standard deviation of our endpoints are included as part of our analytic results in Table 4.3.

Because both treatment policy and endpoint may depend on viral load levels, we must ensure, for example, that an early switch to second-line regimen does not necessarily imply a smaller endpoint. We explore and explain the concepts through two exemplary HIV trajectories Figure 2 illustrates. Patient 1 and patient 2 have the same trajectory of viral load over time before week 6 after confirmed failure.

In the left panel, patient 1 switched to the second-line regimen within 8 weeks after confirmed failure, and then viral load dropped below 200 copies/mL quickly. Patient 2 switched to the second-line regimen at week 10 and then viral load dropped below 200 copies/mL immediately. We assume they have the same follow-up length. Therefore, cumulative viral load for patient 1 (AUC of purple line with squares Y_1) is less than cumulative viral load for patient 2 (AUC of blue line with dots Y_2). On the other hand, $Y_1 > Y_2$ is in the right panel because HIV for patient 1 does not drop significantly after switching to second-line regimen. A similar phenomenon occurs for the rate of time of suppression endpoint. Because both panels in Figure 2 are scientifically plausible, we argue our endpoints are not determined necessarily by definition of treatment policy.

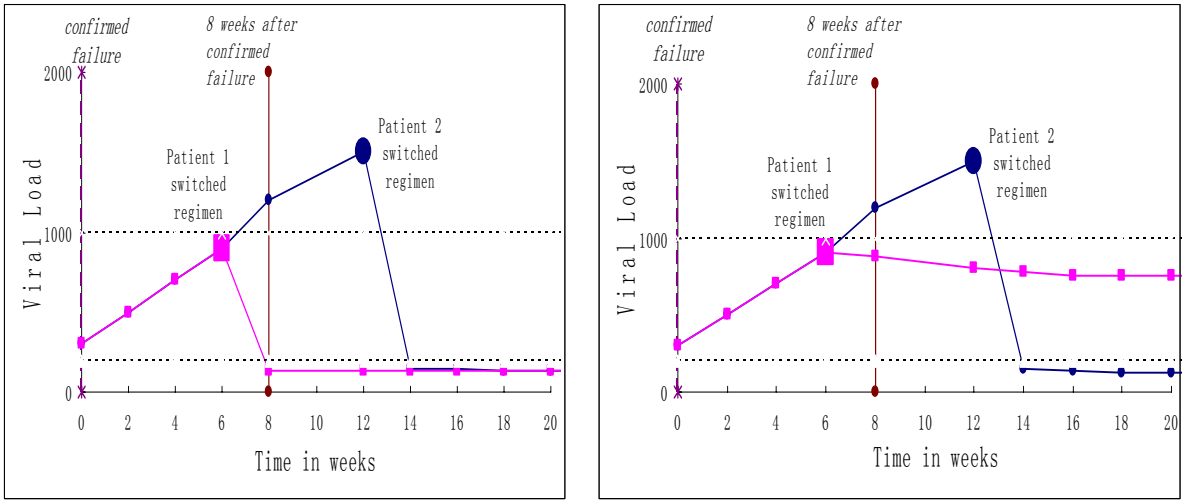


Figure 2.2: Two exemplary HIV trajectories. Patient 1 has smaller AUC and longer time of suppression than Patient 2 in the left panel. Right panel shows the opposite phenomenon.

We included 11 covariates as auxiliary variables. Patients' age ranges from 18 to 77 years old; height has a mean of 173cm with standard deviation of 9cm; weight has a mean of 76kg with standard deviation of 16kg. Sex is an indicator variable, being 1 if male and 0 if female; race is a three-level variable, being 0 or 1 corresponding

to white or black, respectively, and 2 otherwise; drug history indicates whether a patient ever previously used IV drug, being 1 if ever using IV drug and 0 otherwise. Baseline RNA and maximum viral load prior to treatment switch are presented on the logarithmic scale; baseline CD4, baseline CD8, and time (in weeks) from baseline to first virologic failure remain their original scales. The characteristics of auxiliary covariates are presented in Table 2.1.

Table 2.1: Descriptive statistics of auxiliary covariates

Covariates	Overall (758) Mean (SD.)	Switch Early (31) Mean (SD.)	Switch Late (165) Mean (SD.)
Baseline RNA(log_{10})	4.86 (0.73)	5.29 (0.70)	4.84 (0.71)
Maximum RNA† (log_e)	9.06 (2.23)	10.57 (2.03)	8.78 (2.16)
Baseline CD4 Counts	239.74 (191.95)	183.05 (165.21)	250.95 (208.02)
Baseline CD8 Counts	846.15 (505.91)	882.61 (598.55)	848.03 (582.74)
Time to first failure	57.88 (39.75)	57.08 (39.49)	58.03 (39.92)
Height	173.54 (9.04)	171.68 (9.38)	173.68 (9.87)
Body Weight	76.14 (16.20)	70.97 (14.15)	75.01 (14.71)
Age	37.52 (9.32)	38.16 (8.88)	37.64 (9.62)
Sex	80.00	80.65	83.03
Drug history	10.82	19.35	14.55
Race			
Black	35.49	41.94	44.24
Hispanic or others	23.48	29.03	19.39

†Maximum RNA is calculated on the time interval from first failure up to the minimum of switching time and 8 weeks after failure.

2.3.3 Main Analysis

The first step in our statistical analysis involves estimating auxiliary parameters in the propensity score and conditional mean models. We first performed multiple univariate logistic regression analyses using the binary switch indicator (1 if switched early within 8 weeks after failure and 0 otherwise) and covariables including age, weight, height, race, sex, time from baseline to the first virologic failure, baseline CD4, baseline CD8, baseline viral load, and maximum viral load prior to treatment switch, drug history.

Maximum viral load prior to treatment switch as auxiliary variables is not present when endpoint is adjusted cumulative viral load. Each of the variables (age, height, weight, baseline CD4, baseline CD8 and time to first failure) are normalized before entering models. We fit similar univariate models for each of our three endpoints via simple linear regression. The estimated regression coefficients in the propensity score model and the outcome regression model are described in Tables 2.2 and 2.3, respectively.

Table 2.2: Estimates in propensity score model for switching to second-line ARV regimens on the combined Efavirenz arm

Covariate	Est. (SE)[1]	Est.(SE)[2]
Intercept	-8.77 (2.18)	-6.32 (1.93)
Baseline viral load	0.69 (0.36)	0.81(0.34)
Maximum viral load†	0.34 (0.10)	-
Time to 1^{st} failure	0.11 (0.22)	-0.04(0.21)
Baseline CD4	-0.16 (0.32)	-0.12(0.29)
Baseline CD8	0.01 (0.19)	0.10(0.17)
Height	-0.13 (0.31)	-0.16(0.29)
Body Weight	-0.05 (0.28)	-0.11(0.28)
Age	-0.01 (0.22)	-0.03(0.20)
Sex	0.04 (0.78)	0.21(0.74)
Drug History	0.48 (0.58)	0.41(0.55)
Race		
Black	0.17 (0.64)	0.25(0.51)
Hispanic or other	0.37 (0.57)	0.55(0.60)

†Maximum viral load is calculated on the time interval from first failure up to the minimum of switching time and 8 weeks after failure.

[1] Model includes maximum viral load.

[2] Model does not include maximum viral load.

Probability modeled is switch early.

In Table 2.2, we found that patients with high baseline Viral load levels were more likely to switch earlier rather than later. Also, maximum HIV-viral load prior to switching had a profound effect on the probability of switching early on the combined efavirenz arm with higher maximum HIV-viral load values more likely to switch early when it was included in the propensity model. The remaining covariates are not

significantly associated with switching to second-line ARV among those who failed on first-line regimen. The outcome regression model results are presented in Table 2.3. First, controlled for other covariates, baseline viral load is highly correlated with cumulative viral load. Time to virological failure on initial regimen is negatively associated with cumulative viral load. Second, maximum viral load prior to switching is negatively associated with proportion of time with non-detectable viral load. Patients with larger rate of suppression time who switched to second-line regimens at least 8 weeks after first failure had longer time to the first failure, less baseline CD4 counts. Third, cumulative CD4 counts has a negative association with maximum viral load. Time to failure and baseline CD4 counts are positively correlated with cumulative CD4 counts among patients switching to the second-line regimen after 8 weeks.

Table 4.3 presents the analytic results for three primary endpoints length-adjusted HIV-viral load AUC, length-adjusted time-of-viral load supression, and length-adjusted CD4 AUC using auxiliary covariates in Table 1. Our tables include the optimal (OPT) regression, inverse-probability weighted (IPW), and naive parameter estimates, the last of which is defined as the empirical average of endpoints for the subset of patients who failed on first-line ARV regimen. We conduct formal hypothesis tests that the average causal effect equals zero, that is, no difference between early ($s = 1$) versus late ($s = 0$) ARV regimen switch on the initial ARV regimen. In the case of naive estimates, we report the squared two-sample t-test statistic so that it is asymptotically distributed χ^2_1 under the null hypothesis. The Wald test statistics for OPT and IPW estimates are described in Appendix A.

In Table 4.3, we note that naive estimators did not detect any significant difference for any endpoint and IPW estimators only showed significance for difference in cumulative viral load. However, with the auxiliaries of covariates, our new results suggest there are significant differences in cumulative viral load, proportion of time with non-detectable viral load and cumulative CD4 cell counts between patients

Table 2.3: Estimates for conditional mean model on the combined Efavirenz arm

	Cumulative Virus Load[1]		Time of Viral Load Suppression[2]		Cumulative CD4 Cell Count[3]	
Covariate	Early	Late	Early	Late	Early	Late
Intercept	-1.06 (2.40)	2.27 (0.86)	1.74 (0.51)	0.72 (0.18)	7.59 (0.92)	5.96 (0.36)
Baseline viral load	1.84 (0.41)	1.34 (0.16)	0.02 (0.09)	-0.01 (0.03)	0.22 (0.15)	0.03 (0.06)
Maximum viral load†	-	-	-0.08 (0.03)	-0.02 (0.01)	-0.19 (0.05)	-0.03 (0.02)
Time to failure	-0.35 (0.36)	-0.24 (0.10)	-0.05 (0.07)	0.12 (0.02)	-0.05 (0.13)	0.20 (0.04)
Baseline CD4	0.11 (0.40)	0.17 (0.11)	-0.05 (0.08)	-0.06 (0.02)	0.24 (0.15)	0.47 (0.04)
Baseline CD8	0.39 (0.25)	0.15 (0.09)	0.03 (0.05)	-0.01 (0.02)	0.05 (0.10)	0.00 (0.04)
Height	-0.16 (0.37)	-0.23 (0.14)	0.06 (0.08)	-0.01 (0.03)	0.11 (0.13)	-0.01 (0.05)
Body Weight	-0.16 (0.34)	-0.11 (0.13)	0.02 (0.07)	0.04 (0.02)	-0.01 (0.12)	0.01 (0.05)
Age	-0.38 (0.27)	-0.15 (0.10)	0.06 (0.05)	0.04 (0.02)	0.03 (0.10)	-0.02 (0.04)
Sex	0.31 (1.01)	0.36 (0.38)	-0.38 (0.20)	0.05 (0.07)	-0.78 (0.36)	-0.03 (0.14)
Drug Use	0.01 (0.69)	0.27 (0.29)	-0.09 (0.14)	-0.07 (0.05)	-0.42 (0.25)	-0.04 (0.11)
Race						
Black	1.05 (0.57)	0.40 (0.23)	-0.30 (0.15)	-0.11 (0.04)	-0.55 (0.27)	-0.12 (0.09)
Hispanic or others	0.47 (0.85)	0.25 (0.30)	-0.06 (0.17)	-0.00 (0.06)	-0.29 (0.30)	-0.03 (0.11)

[1] Virus Load: Length-adjusted AUC of Virus Load, logarithm scale;

[2] Rate of Suppression Time: Rate of Time of Suppression of Virus;

[3] CD4 Counts: Length-adjusted AUC of CD4 Counts, logarithm scale

switching early versus late to second-line ARV regimens on the combined efavirenz arm. In particular, cumulative viral load is generally smaller while cumulative CD4 cell counts is larger for those patients switching earlier rather than later on the combined efavirenz arm. Our findings also suggest that patients who switch within 8 weeks after confirmed virologic failure tend to have larger proportion of time with non-detectable viral load, on average. For example, after 1 year follow-up, patients following a treatment policy which switched to second-line regimen within 8 weeks after virologic failure suppressed viral load levels below 200 copies/mL for an average $365 \times 0.80 \simeq 293(\pm 4)$ days, compared to $365 \times 0.76 \simeq 277(\pm 4)$ days for switching to second-line regimen beyond 8 weeks after failure. Hence, on average, patients spend about three more weeks with viral load levels below 200 copies/mL if they switched prior to 8 weeks. In conclusion, when we define failure as "confirmed failure on the first-line regimen which is initial ARV regimen plus any protocol-approved substitutions", we find some evidence to suggest a smaller cumulative viral load and higher proportion of time with non-detectable viral load and cumulative CD4 counts for patients that switch off a failing ARV regimen within 8 weeks.

Table 2.4: Estimates of mean outcomes, 758 patients, full model

		OPT		IPW		Naive	
Endpoint	Switch	Est.(SE)	T	Est.(SE)	T	Est.(SE)	T
Virus [1]	Early	7.93 (0.08)	14.89	7.91 (0.10)	5.47	9.56 (0.32)	0.72
	Late	8.12 (0.07)		8.12 (0.07)		9.33 (0.12)	
Rate [2]	Early	0.80(0.01)	19.38	0.81 (0.03)	3.13	0.52 (0.05)	1.00
	Late	0.76(0.01)		0.76 (0.01)		0.46 (0.02)	
CD4 [3]	Early	5.96(0.02)	16.65	6.01 (0.21)	0.33	5.69(0.12)	0.40
	Late	5.89(0.02)		5.89 (0.02)		5.74(0.06)	

NOTE: The estimated endpoint is reported for combination of initial ARV treatment regimen (A=Combined Efavirenz) and switching status (S);

Report the Wald test statistic for a test of the null hypothesis of no average causal effect (ACE);

[1] Virus: Length-adjusted AUC of Virus Load, logarithm scale;

[2] Rate: Rate of Time Suppression of Virus;

[3] CD4: Length-adjusted AUC of CD4 Counts, logarithm scale

2.3.4 Sensitivity analyses

The sensitivity our analytic results depend on many assumptions, some of which are identified and others of which are not identified by the observable data. Because these statistical assumptions play no small role in the analysis of observational data, many authors have proposed a wide range of tools for model diagnostics and sensitivity analyses (cf. Rosenbaum, 1983; Robins, 1999; Robins, Rotnizky, and Scharfstein, 1999; Rotnizky, Scharfstein, Su, and Robins, 2001). Our sensitivity analyses included, but not limited to, comparing the effect on our parameter estimates when weak *observed* confounders were removed and when all the confounders were included in the models. In addition, we will conduct analysis to look into the influence of those 50 patients who experienced virologic failure but only after a protocol-approved substitution. Investigating the sensitivity of our analytic results to nonidentifiable assumptions is beyond the scope of the current paper. Hence, our results rest on the validity of the "no unmeasured confounders" assumption. However, this assumption is ubiquitous in the literature and a well-know limitation of causal inference.

In our main analysis, we include all potential confounders which were significantly or mildly related to treatment switching, endpoints, or both, and used the same set of variables throughout. Different endpoints have different important covariables sets. Baseline viral load, baseline CD4 cell counts, time to viral failure on initial regimen, race and body weight are found important for cumulative viral load. Baseline viral load, maximum viral load before switching, baseline CD4 cell counts, baseline CD8 cell counts, time to viral failure on initial regimen, sex and race have important effects on proportion of time with non-detectable viral load. Baseline viral load, maximum viral load before switching, baseline CD4 cell counts, baseline CD8 cell counts, time to viral failure on initial regimen, body weight, sex and race are significantly associated with cumulative viral load. In Table 1, we report the point estimates of mean outcomes on treatment policies, their standard error estimates, and ACE. Compared to the

main analysis, we found that point estimates and standard error estimates changed little when unimportant covariates were removed from the models and the difference of mean endpoints are still significant between patients who switch to a second-line regimen within 8 weeks after failure on the first-line regimen than those patients who switch late. The findings are summarized in Table 2.5.

Table 2.5: Analytic results after removing weak confounders

		OPT		IPW		Naive	
Endpoint	Switch	Est.(SE)	T	Est.(SE)	T	Est.(SE)	T
Virus [1]	Early	7.93 (0.08)	11.79	7.92 (0.10)	5.61	9.56 (0.32)	0.72
	Late	8.12 (0.07)		8.11 (0.07)		9.33 (0.12)	
Rate [2]	Early	0.80 (0.01)	18.59	0.82 (0.03)	3.98	0.52 (0.05)	1.00
	Late	0.76 (0.01)		0.76 (0.01)		0.46 (0.02)	
CD4 [3]	Early	5.96 (0.02)	16.17	6.07 (0.23)	0.63	5.69 (0.12)	0.40
	Late	5.89 (0.02)		5.89 (0.02)		5.74 (0.06)	

NOTE: The estimated endpoint is reported for combination of initial ARV treatment regimen (A=Combined Efavirenz) and switching status (S);
The Wald test statistic for a test of the null hypothesis of no average causal effect (ACE);
[1] Virus: Length-adjusted AUC of Virus Load, logarithm scale;
[2] Rate: Rate of Time Suppression of Virus;
[3] CD4: Length-adjusted AUC of CD4 Counts, logarithm scale

We repeat the analysis for our three endpoints in Table 2.6, excluding 50 patients from our analysis. This way accounts for defining "failure" as "confirmed failure on the first-line regimen" and first-line regimen only includes initial treatment. As in the main analysis, we use all potential confounders. Conclusion did not change from main analysis(Table 2.6).

In the main analysis, length-adjusted AUC of viral load and CD4 cell counts are computed on the original scale, then we transform to the natural logarithmic scale. In this section, length-adjusted AUC is calculated on a natural logarithmic scale of viral load and CD4 cell counts. Although point estimates for viral load are different from main analysis, the conclusions are exactly the same. That is, patients that switch off a failing ARV regimen within 8 weeks have a smaller cumulative viral load and

Table 2.6: Analytic results after excluding 50 patients who were not following initial ARV regimen at first virologic failure

		OPT		IPW		Naive	
Endpoint	Switch	Est.(SE)	T	Est.(SE)	T	Est.(SE)	T
Virus [1]	Early	7.91 (0.08)	3.58	7.90 (0.10)	1.23	9.56 (0.32)	1.23
	Late	8.10 (0.07)		7.99 (0.08)		9.15 (0.15)	
Rate [2]	Early	0.82 (0.01)	17.47	0.82 (0.03)	1.77	0.52 (0.05)	0.73
	Late	0.78 (0.01)		0.78 (0.01)		0.48 (0.03)	
CD4 [3]	Early	5.96 (0.02)	10.80	5.98 (0.22)	0.15	5.67 (0.11)	0.42
	Late	5.90 (0.03)		5.89 (0.03)		5.75 (0.07)	

NOTE: The estimated endpoint is reported for combination of initial ARV treatment regimen (A=Combined Efavirenz) and switching status (S);
Wald test statistic for a test of the null hypothesis of no average causal effect (ACE);
[1] Virus: Length-adjusted AUC of Virus Load, logarithm scale;
[2] Rate: Length-adjusted Time of Suppression of Virus;
[3] CD4: Length-adjusted AUC of CD4 Counts, logarithm scale

higher cumulative CD4 cell counts. Point estimates and standard error estimates are displayed in Table 2.7.

2.4 Simulation Study

In this section, we carry out simulation studies to investigate the performance of proposed estimator and how unknown parameters affect the power of our test statistics to detect significant differences between policies that switch early versus late.

We firstly conducted simulation studies to examine the operating characteristics of several estimators. We consider a special case where all the patients experienced virologic failure and switched to the second-line treatment either early or late. For the true propensity score models and outcome regression models, we follow simulation scenarios similar to those in Cao et al. (2009). For each i, $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ was generated as standard multivariate normal, and the elements of $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ were defined as $X_{i1} = \exp(Z_{i1}/2), X_{i2} = Z_{i2}/(1 + \exp(Z_{i1})) + 10, X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$ and $X_{i4} = (Z_{i1} + Z_{i2})^2$, so that Z_i may be expressed in terms of X_i. The true

Table 2.7: Analytic results when outcomes are length-adjusted AUC of logarithm of original scale

		OPT		IPW		Naive	
Endpoint	Switch	Est.(SE)	T	Est.(SE)	T	Est.(SE)	T
Virus [1]	Early	4.23 (0.09)	4.95	4.22 (0.09)	4.97	5.98 (0.41)	0.51
	Late	4.41 (0.07)		4.41 (0.07)		6.18 (0.16)	
Rate [2]	Early	0.80 (0.01)	19.38	0.81 (0.03)	3.13	0.52 (0.05)	1.00
	Late	0.76 (0.01)		0.76 (0.01)		0.46 (0.02)	
CD4 [3]	Early	5.91 (0.03)	18.35	5.96 (0.21)	0.37	5.60 (0.13)	0.48
	Late	5.83 (0.03)		5.83 (0.03)		5.67 (0.06)	

NOTE: The estimated endpoint is reported for combination of initial ARV treatment regimen (A=Combined Efavirenz) and switching status (S);

Wald test statistic for a test of the null hypothesis of no average causal effect (ACE);

[1] Virus: Length-adjusted AUC of Virus Load, logarithm scale;

[2] Rate: Length-adjusted Time of Suppression of Virus;

[3] CD4: Length-adjusted AUC of CD4 Counts, logarithm scale

propensity score model is $\pi_0 = \text{expit}(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$. We consider two types of true outcome regression models. The first type is an additive linear regression of endpoint on Z. The second setting is partly motivated by our observation that AUC outcomes have long right-tails; hence, we assume endpoints are exponentially distributed. We carry out the following simulations: (1) true PS models and normal distribution for endpoints, and correctly specified posited models; (2) true PS models and normal distribution for endpoints, and correctly specified posited models with more insignificant covariates; (3) true PS models and normal distribution for endpoints, and true OR models are on X, but posited OR models are on Z; (4) true PS models and exponential distribution for endpoints, but posited model are linear regression of endpoints with normal standard errors. For each scenario of n=1000 and n=100, we generate 1000 Monte Carlo data sets. For each estimator, sandwich standard errors are calculated as described in the Appendix. Results for all simulation scenarios are presented in Table 2.4.

When sample size is as large as 1000 and PS model is correctly specified, IPW, AIPW, RRZ and Tan estimators all showed negligible Monte Carlo bias. In addition,

AIPW, RRZ and Tan estimator showed improved efficiencies than IPW estimators, as expected. Moreover, Tan estimator exhibit the best performance in terms of efficiency. However, when sample size is 100 and OR model is misspecified, RRZ estimator shows significant small sample bias. In most cases, Tan's estimator was competitive with other estimators in the scenarios we considered. However, based on the simulation results, we see that even Tan's optimal estimators can perform poorly in small samples, even when the PS model is correctly specified.

The second set of simulations is used to evaluate the influence of coefficient in the outcome regression model on the power. we are interested in the probability of $T > 3.84$ where T is statistics calculated according to the formula $T = (C\mu)^{\mathrm{T}}(C\boldsymbol{\Sigma_\mu}C^{\mathrm{T}})^{-1}(C\mu)$, where C=(1, -1) is asymptotically distributed as χ_1^2 under the null hypothesis of no difference, and we estimate it based on 1000 MCMC simulations.

We consider combination of the following conditions:(a) switching rate $\approx (0.1, 0.5)$ by adjusting intercept(β) in the propensity score, and (b) coefficients in the outcome regression $\gamma = (-0.1, -0.3, -0.5)$. Failure rate was fixed at 0.3, sample size n=700 and other parameters are chosen to be close to the values in the A5095.

The simulation results were presented in the Figures below. The power versus rate of difference of mean over its standard deviation are plotted. First, an apparent trend was found that power will increase with the increase of standardized difference of means in two groups given the effect of covariate on outcome in outcome regressions. Second, we noticed that for given standardized difference of mean, power increases with the increase of influence of covariate on outcome. The findings are consistent no matter how much the switching rate is, 0.1 or 0.5. In short, the more important covariates in the propensity score model or conditional mean models, the more powerful the proposed estimator.

The last simulations is intended to show how the switching rate affects the power

Table 2.8: Simulation results based on 1000 Monte Carlo replications.

	n=1000				n=100			
Method	Bias	MCSD	AAVE	MAVE	Bias	MCSD	AAVE	MAVE
				PS correct OR correct				
IPW	0.011	4.685	4.453	3.926	1.782	24.274	13.978	22.370
AIPW	0.029	1.149	1.147	1.147	0.127	3.658	3.605	3.596
RRZ	-0.312	1.300	1.172	1.164	5.937	9.130	5.348	4.520
Tan	-0.029	1.149	1.147	1.147	0.123	3.652	3.602	3.598
				PS correct, OR incorrect [1]				
IPW	0.040	4.869	4.488	3.975	2.660	24.154	20.533	14.165
AIPW	0.029	1.150	1.149	1.148	0.128	3.664	3.612	3.609
RRZ	-0.323	1.306	1.175	1.167	8.410	10.997	7.211	6.198
Tan	0.030	1.150	1.149	1.148	0.127	3.664	3.607	3.605
				PS correct, OR incorrect [2]				
IPW	-0.497	8.324	7.685	6.757	2.676	41.472	24.535	18.256
AIPW	-0.496	2.617	2.381	2.176	0.500	7.813	6.263	5.506
RRZ	0.177	2.619	2.227	2.118	10.303	16.368	9.495	7.989
Tan	-0.082	2.386	2.201	2.058	1.454	7.655	6.644	5.219
				PS correct, OR incorrect [3]				
IPW	0.373	12.647	12.722	12.067	0.759	50.897	37.851	32.091
AIPW	0.464	12.080	12.113	11.592	3.064	42.294	35.611	32.585
RRZ	0.183	12.071	11.745	11.346	3.583	40.884	33.476	30.598
Tan	0.619	12.149	11.929	11.415	3.571	45.991	42.915	33.171

[1] Posited OR models are linear, having more insignificant covariates than true models

[2] Posited OR models are linear, having different significant covariates from true models

[3] Posited OR models are linear, while true distribution of endpoint are exponential

Bias, Monte Carlo bias; MCSD, Monte Carlo standard deviation;

AAVE, average of sandwich standard errors; MAVE, median of sandwich standard errors

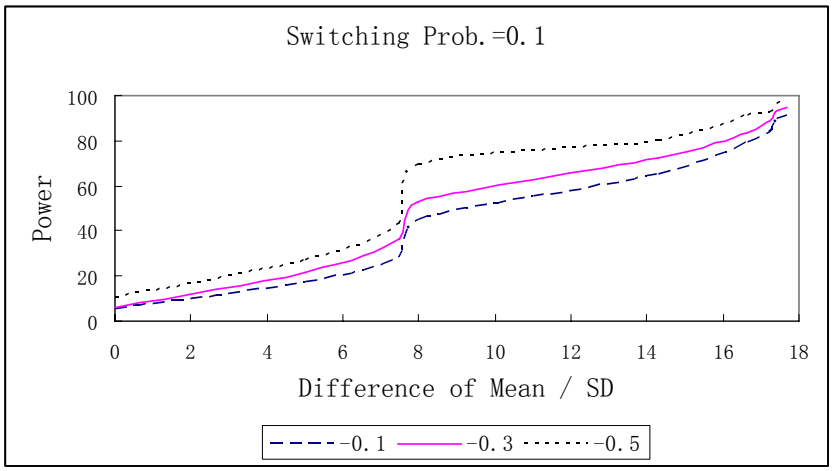


Figure 1: Sample size=700, failure rate=0.3, switching probability among failed patients is about 0.1, coefficients for Cd4 in the linear regression are -0.1, -0.3, -0.5.

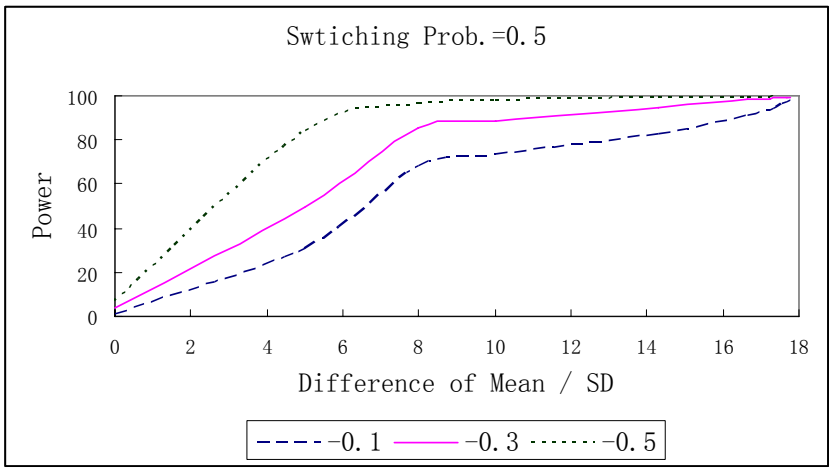


Figure 2: Sample size=700, failure rate=0.3, switching probability among failed patiets is about 0.5, coefficients for Cd4 in the linear regression are -0.1, -0.3, -0.5.

of our estimators. In this simulation, we mange to set all the parameters close to the values in the data of A5095, except the intercept in the propensity score which adjusts the switching rate. In order to mimic the real data, we choose 6 covariates as in real data analysis, 4 continuous covariates following standard normal distribution and 2 categorical covariates following binary distribution with probability of 0.35 and 0.23, respectively. The propensity score and outcome regressions are assumed as follows:

(1) logit $P(S=1|X,R=1) = \beta + 0.82X_1 - 0.04X_2 - 0.05X_3 - 0.18X_4 + 0.22X_5 + 0.61X_6$

(2) $E(Y_1|R=1,X) = -1.18 + 1.91X_1 - 0.47X_2 + 0.25X_3 - 0.20X_4 + 0.98X_5 + 0.81X_6$

$E(Y_0|R=1,X) = 2.61 + 1.33X_1 - 0.26X_2 + 0.23X_3 - 0.17X_4 + 0.41X_5 + 0.43X_6X$

Except β, all the parameters in the models above are chosen to reflect aspects of the dataset in A5095. In the simulation scenarios to follow, failing rate $R = 0.26$ and sample size $n = 758$. The simulation results are presented in the Table 2.9 below. The results in the Table 2.9 showed that Larger switching rate would result in larger power, and if we simulate an environment which have the same parameters with the dataset in A5095, our estimators would have a power of more than 80% when switching rate reachs 16%.

Table 2.9: Power under different switching rates

Switching Rate	0.07	0.16	0.19	0.24	0.33	0.54
$\hat{\mu}_1$(se.)	7.99 (0.62)	7.96 (0.09)	7.96 (0.09)	7.95 (0.08)	7.95 (0.08)	7.94 (0.07)
$\hat{\mu}_0$(se.)	8.13 (0.07)	8.13 (0.07)	8.13 (0.07)	8.13 (0.07)	8.13 (0.07)	8.12 (0.07)
Power	0.425	0.837	0.894	0.945	0.994	0.999

True values: μ_1=7.92; μ_0=8.13

2.5 Discussion

We extended LDT's estimators proposed in the two-stage randomization setups to observational data by introducing a hypothetical randomization design and applying a result about Radon-Nikodym derivative presented by Murphy, van der Laan, and Robins(2001). Moreover, by adopting the idea of Tan(2006), our new estimator gained more efficiency and robustness.

The results gained by applying our method to ACTG A5095 data answered a scientific question of interest of the physicians in ACTG team: Should a patient switched to a second-line regimen within 8 weeks after he/she had been confirmed to be failed on the first-line regimen? Specifically, our findings support the statement that patients would have smaller cumulative virus load, higher rate of time of suppressing viral load below 200mL/copies and larger cumulative CD4 counts if they had failed on the first-line regimen, then switched to second-line regimen within 8 weeks after failure.

As in most observational studies, a key assumption to our estimator is no unmeasured confounding. This assumption is, unfortunately, also the most difficult to verify. What we did was to include all the covariates that we can collect from the existing datasets in our method to acquire some confidence that we had all the observed important factors. In sensitivity analysis, we removed those covariates which had no effect on the propensity score and endpoints. It turns out that results did not change very much and the same conclusion remained. However, if we removed strong covariates related to propensity score and endpoints, the significance lost or decreased. Such findings justified the way we selected the auxiliary variables and support our conclusions that switching to 8 weeks within 8 weeks after failure is better than switching late. Another special assumption for ACTG A5095 data is we assumed 50 patients who experienced virologic failure but only after a protocol-approved substitution did not switch to second-line regimen within 8 weeks and included them

in the main analysis. Sensitivity analysis showed that including these 50 patients or not did not affect our previous findings. This does not mean our method is robust to drug-substitution happening to HIV patients. The findings about this are only restricted to ACTG A5095 data.

In addition to additional analyses listed in Subsection 2.3.4, we also considered different thresholds for an early versus delayed regimen change. The decision to define switching early as switching within 8 weeks after failing on the first-line regimen was made by investigators in the ACTG A5095 team and physicians who actually participated the study. Nevertheless, we tried different thresholds including 90, 120 and 150 days. In general, we found that significance differences between early and late regimen change disappeared as timing of the threshold increases. Robust methods that do not require thresholds are desirable but we conjecture would not make a significant impact in the current analysis due to limited numbers of patients who actually changed regimens within several months of confirmed virlogic failure.

The ACTG A5095 study is an excellent example of a clinical trial that was designed to test a particular hypothesis on the efficacy of initial regimens but we used it for a particularly intriguing secondary analysis of regimen change. If ACTG A5115 is any indication of the future, it will be difficult to design and enroll a completely randomized study of regimen change. In this case, data like that from ACTG A5095 and the framework employed here will be germane for evaluating the effect of early regimen change. As with many clinical studies, however, there are always twists and caveats that make secondary analyses tricky. A limitation of our analysis is that patients who switched to second-line regimen within 8 weeks of initial virologic failure may have more follow-up data post-second-line treatment than do patients who delay switching. We attempted to adjust for the discrepancy by adjusting for length of follow-up but it would have been preferable to measure and analyze an endpoint that was exogenous to the timing of the second-line ARV treatment decision. Our first

response to this criticism is that it is not evident, based on the literature or the clinical experience of the co-authors, whether more follow-up necessarily leads to better or worse outcomes and such assumption was one of our working hypotheses when we started this project. A second response to this criticism is consider other analyses and other larger data sets. For example, augmenting the A5095 data with data from other ACTG studies will allow us to better address the long-term effect of regimen changes and is also the subject of ongoing research.

Although dissemination of our findings from ACTG A5095 data may not be available, our method is anticipated to prove useful in multiple fields of applied research. Further developments will focus on the method to choose the optimal waiting time to switch to the second-line regimen and waiting time will remain continuous instead of dichotomy on a prefixed cutting point.

Finally, the findings about influence of coefficients in the conditional mean models are interesting in the simulation. Whether it is a general phenomenon or a special case depending on the data setting needs further exploration.

Chapter 3

Locally efficient and Double Robust Semiparametric Estimator for the Treatment Duration, with Duration Possibly Right-censored

3.1 Introduction

Once treatment has been proven effective, study investigators are often interested in the best treatment duration which optimizes the response. As Johnson and Tsiatis(2004) argued that because infusion can not continue after a treatment-terminating event, a recommendation to infuse for t units of time necessarily implies that treatment would be discontinued either after drug was administered for t units of time or when a treatment-terminating event occurs. Thus, censoring is as an essential part of treatment policy and a treatment duration policy for t unites of time of interest is defined as "a recommendation to treat for t units of time or until a treatment-terminating event occurs, whichever comes first".

Johnson and Tsiatis (2004; subsequently referred to as JT) have shown how to estimate consistently the population mean response for the treatment duration policy by incorporating propensity score in the estimator without modeling outcome regression on covariates. However, the JT estimator is neither the most efficient nor doubly robust; that is, it does not remain consistent and asymptotically normal if either the propensity score model or the outcome regression model is correct. Considerable recent interest has focused on doubly robust estimators for a population mean response in the presence of incomplete data, which involves models for both the propensity score and the regression of outcome on covariates. Given the protection afforded by the property being doubly robust, these estimators have been advocated for routine use (Bang and Robins, 2005). In this chapter we use Robins, Rotnitzky, and Zhao (1994)'s theory to identify a class of augmented inverse probability weighted estimators and show how to derive a doubly robust estimator which is more efficient than the JT estimator.

3.2 Method

As in Johnson and Tsiatis (2004), throughout our analysis, we adopt the point of view proposed by Neyman (1923) and Rubin (1974), where casual effects are defined through potential outcomes or counter-factual random variables. Specially, for each level of the treatment T having m discrete values: $t_1, \cdots, t_m$, we assume that there exists a potential outcome Y_t^*, where Y_t^* denotes the response of a randomly selected individual had, possibly contrary to fact, been given treatment $T = t$. If a patient experienced a treatment-terminating event at time C, he would have potential outcome Y_C^*. The parameter of interest would be mean outcome, $E(Y_{t\wedge C}^*)$, where $t \wedge C$ denotes the minimum of t and C and $Y_{t\wedge C}^*$ is the response if a patient would have been treated for t units of time or until a treatment-terminating event occurs, whichever

comes first. This may also be written as

$$\mu_r = E(Y^*_{t_r \wedge C}) = E\{Y^*_{t_r} I(C > t_r) + Y^*_C I(t_r \geq C)\}.$$

The random variables defined above are referred to as potential random variables, or counterfactuals, because, contrary to the fact, they may not actually be observed. In contrast, for a randomly selected individual from our population, the observable random variables are subsets of potential outcomes. We regard them as incomplete data of potential outcomes. In this chapter, instead of talking about missing data, we refer to a more general notion of "coarsening" of data. The concept of coarsened data was first introduced by Heitjan and Rubin (1991) and studied more extensively by Heitjan (1993) and Gill, van der Laan, and Robins (1996). In a number of common situations, data are neither entirely missing nor perfectly present. Instead, we observe only a subset of the complete-data sample space in which the true and unobservable data lie (Heitjan and Rubin (1991)). This kind of incomplete data is referred to coarse data. The purpose we introduce coarsening data is not to investigate its feature or explore more perspectives about coarsening data, instead, we are more interested in the application of coarsening data as a tool in our problem to derive an efficiently doubly robust semiparametric estimator. There are several concepts related to coarsened data, which are full data, coarsening variable and observed data, and like function making connection between full data and observed data through coarsening variables. We aim to make inference on the parameter of interest of full data through observed data.

3.2.1 Full Data and Observed Data

Let T denote treatment duration, having m discrete values: $t_1, \cdots, t_m$, and C is the termination event time, and C^* is the coarsened variable. $\{Y^*_1, ..., Y^*_m, Y^*_C\}$ are

potential outcomes while Y is the observed outcome. Let $X^H(C)$ denote all the history information collected up to treatment-terminating event time .

In an imaginarily ideal world, we are able to observe full data $Z = \{Y_1^*, ..., Y_m^*, Y_C^*, C, X^H(C)\}$. Parameters of interest are

$$\mu_r = E\left[Y_r^* I(C > t_r) + Y_C^* I(C \leq t_r)\right], r = 1, ..., m.$$

With full data we can make inference on $\mu_r (r = 1, \cdots, m)$ using standard statistical theory.

However, in a real world, we can only observe $\{Y, U, \Delta, X^H(U)\}$, where $U = \min(T, C), \Delta = I(T < C)$ and a key assumption is

$$Y = Y_1^* I(T = t_1, C > t_1) + ... + Y_m^* I(T = t_m, C > t_m) + Y_C^* I(T \geq C)$$

$$= Y_1^* I(U = t_1, \Delta = 1) + ... + Y_m^* I(U = t_k, \Delta = 1) + Y_C^* I(\Delta = 0).$$

Remark. *The assumption above is referred as Stable Unit Treatment Value Assumption (SUTVA) referred by Rubin(1978a). SUTVA implies that there must not be any interference in the response from other subjects and the plausibility of this assumption needs to be evaluated on a case-by-case basis (Tsiatis, 2006). Since the disease in our study is not infectious through common contacts and patients were not collected by families, we have reasons to believe the SUTVA assumption holds.*

Remark. *Full data, latent vectors and observed data: We assume that underlying any problem related to potential outcomes there are unobservable latent variables*

$$Z^* = (Y_1^*, ..., Y_m^*, Y_C^*, C, X^H(C), T).$$

The joint distribution of $p(t, c, y_1^*, ..., y_m^*, y_c^*, x)$ *can be written as*

$$p(t|c, y_1^*, ..., y_m^*, y_c^*, x)p(c, y_1^*, ...y_m^*, y_c^*, x),$$

where $p(C, y_1^*, ...y_m^*, y_c^*, x)$ *denotes the density of the full data had we been able to observe them. Therefore, full data which we need to make statistical inference for parameter of interest is*

$$Z = \{Y_1^*, ..., Y_m^*, Y_C^*, C, X^H(C)\}.$$

If we define C^* *is the coarsened variable, both* (C^*, Z) *and* Z *are referred to full data in the coarsening problem (Tsiastis, P156, 2006). To avoid confusion, we call* Z *is full data and* (C^*, Z) *are latent random vectors.*

In survival analysis, death or failure, or other competing risk is considered an "event". Usually, survival time T is either recorded completely or censored by some termination event. However, here we treat time to terminating event as being censored by treatment T, instead of T being censored by C. The reason is due to the following logic. Because potential outcomes $Y_1^*, \cdots, Y_m^*$ have been assumed to be obtained in the full data, so T does not play any role for the potential outcomes in the full data. Plus, in full data each patient is assumed to be observed on his/her terminating event time C and the corresponding potential outcome Y_C^*. However, T plays an important role in deciding whether C could be observed in reality. Therefore, instead of regarding T is censored by C, this problem could be viewed in this way that C is censored by T and T plays role as coarsening variable or part of coarsening variable. If we believe latent variables data is $(C^*, Z) = (T, Y_1^*, \cdots, Y_m^*, Y_C^*, C, X^H(C))$(we never observed latent variables (C^*, Z) or full data Z), the observed data are given as the transformation of the latent variables, namely $\{Y, U, \Delta, X^H(U)\}$, where $U = \min(T, C), \Delta = I(T < C)$ and $Y = Y_1^* I(U = t_1, \Delta = 1) + ... + Y_m^* I(U = t_m, \Delta =$

$1) + Y_C^* I(\Delta = 0) = Y_1^* I(T = t_1, C > t_1) + ... + Y_m^* I(T = t_m, C > t_k) + Y_C^* I(T \geq C)$, and $X^H(U)$is the history information collected up to observed time U.

3.2.2 Coarsening Variable and Link Functions

To make the connection between the full data and the observed data, we define coarsening variable as below,

$$(C^* = r) = (T = t_r, C > t_r), r = 1, 2, ..., m.$$
$$(C^* = \infty) = (C \leq T).$$

Thus, $r+1$ types of coarsening exist. Note there is no possibility to observe the full data in reality. We need to point out that $(C^* = \infty)$ does not mean we have observed the full data, only denotes one type of coarsening where terminating event can be observed. When $C^* = r$, we observe $G_r(Z) = [Y_r^*, CI(C \leq t_r)], X^H(\min(t_r, C))]$ for $r \leq m$ and $G_\infty(Z) = [Y_C^*, C, X^H(C)]$ for $r = \infty$. $\{G_r(Z), r = 1, \cdots, m, \infty\}$ are called link functions.

Obviously $G_r(Z)$ is not a function of $G_{r+1}(Z)$ if we do not put any assumption such as monotone on the relationship for potential outcomes $Y_r^*, r = 1, ..., m$. However, we would like to express $G_r(Z) = [Y_r^*, CI(C \leq t_r), X(\min(t_r, C))] = [Y_r^*, G_r^1(C, X)]$ for $r \leq m$ and $G_\infty(Z) = [Y_C^*, C, X^H(C)] = [Y_C^*, G_\infty^1(C, X)]$ for $r = \infty$. Thus, although $G_r(Z)$ is not a function of $G_{r+1}(Z)$, $G_r^1(C, X)$ is a function of $G_{r+1}^1(C, X)$. Furthermore, we assume $\{(Y_1^*, ..., Y_m^*, Y_C^*) \perp T | X^H(C)\}$. This reminds us it is convenient to consider coarsening models expressed by the discrete hazard function, and we will discuss it later.

The observed data can be expressed as

$$(C^*, G_{C^*}(Z)) = \{C^*, G_{C^*}(Y_1^*, \cdots, Y_m^*, Y_C^*, C, X^H(C))\},$$

where

$$\begin{aligned}\{C^* = r, G_r(Y_1^*, ..., Y_m^*, Y_C^*, C, X^H(C))\} &= \{T = t_r, C > t_r, Y_r^*, X^H(t_r)\}, r = 1, \cdots, m,\\ \{C^* = \infty, G_\infty(Y_1^*, ...Y_m^*, Y_C^*, C, X^H(C))\} &= \{T \geq C, C, Y_C^*, X^H(C)\}.\end{aligned}$$

Important assumptions not only include SUTVA assumption but also include:

(1) Coarsening at random, i.e., $P(C^* = r|Z) = \omega(r, G_r(Z))$; (3.1)

(2) No unmeasured confounding, i.e., $(Y_1^*, ..., Y_m^*, Y_C^*) \perp T|X$. (3.2)

Assumption (3.1) is a common assumption in the coarsening problem and Assumption (3.2) is an important assumption on which inference on causal effect necessarily relies. Assumption (3.2) implies that $P(C^* = r|Z) = \omega(r, G_r^1(C, X))$, allowing to model coarsening probability by borrowing idea when coarsening is monotone. We will introduce "partially-monotone coarsening" in the next section. We simplify the symbol of $X^H(U)$ as X from now on.

3.2.3 Partially-monotone Coarsening

When data are coarsening at random, we consider models for the coarsening probabilities, which, in general, are denoted by

$$P(C^* = r|Z = z, \gamma) = \omega(r, G_r(Z), \gamma),$$

in terms of the unknown parameters γ. Tsiatis(2005) elaborated how to make inference when coarsening is monotone, when the link function $G_r(Z)$ is a many-to-one function of $G_{r+1}(Z)$. Although the assumption of monotone coarsening does not hold in our case, we can borrow the idea of dealing with monotone coarsening to handle

the coarsening mechanism in our case. Specifically, for the rth link function

$$G_r(Z) = \{Y_r^*, CI(C \leq t_r), X(\min(t_r, C))\}, r \leq m;$$
$$G_\infty(Z) = \{Y_C^*, C, X^H(C)\}, r = \infty.$$

we define another set of functions as below:

$$G_r^1(C, X) = [CI(C \leq t_r), X\{\min(t_r, C)\}], r \leq m;$$
$$G_\infty^1(C, X) = [C, X^H(C)], r = \infty.$$

Thus, although $G_r(Z)$ is not a function of $G_{r+1}(Z)$, $G_r^1(C, X)$ is a function of $G_{r+1}^1(C, X)$, and we call such type of coarsening "Partially-monotone Coarsening".

With partially-monotone coarsening, it is convenient to consider models for the discrete hazard function, defined as

$$\lambda_r(G_r^1(C, X)) = \begin{cases} P(C^* = r | C^* \geq r, Z) & r = 1, 2, \cdots, m \\ 1 & r = \infty. \end{cases} \tag{3.3}$$

That $\lambda_r(\cdot)$ is a function of $G_r^1(C, X)$ follows by noting that the right-hand side of equation (3.3) equals

$$P(C^* = r | C^* \geq r, Z) = \frac{P(C = r|Z)}{P(C^* \geq r|Z)}$$
$$= \frac{\omega(r, G_r^1(C, X))}{1 - P(C^* < r|Z)} = \frac{\omega(r, G_r^1(C, X))}{1 - \sum_{r' \leq r-1} \omega(r', G^1(C, X))},$$

and where $G_{r'}^1(C, X)$ is a function of $G_r^1(C, X)$ for all $r' < r$.

We also define

$$K_r\{G_r^1(C, X)\} = P(C^* > r|Z) = \prod_{r'=1}^{r} [1 - \lambda\{G_{r'}^1(C, X)\}], r = 1, \cdots, m-1.$$

Consequently, we can equivalently express the coarsening probabilities in terms of the discrete hazard functions; namely,

$$\omega(r, G_r^1(C,X), \gamma) = \begin{cases} \lambda_1\{G_1^1(C,X)\} & r = 1, \\ \prod_{r'=1}^{r-1}[1-\lambda_{r'}\{G_{r'}^1(C,X)\}]\lambda_r\{G_r^1(C,X)\} & r = 2, \cdots, m, \\ \prod_{r'=1}^{m}[1-\lambda_{r'}\{G_{r'}^1(C,X)\}] & r = \infty. \end{cases}$$

In addition, we define rth discrete cause-specific hazard functions only specified by observed data X as follows:

$$\widetilde{\lambda}_r(X) = P(U = t_r, \Delta = 1 | U \geq t_r, X), r = 1, 2, \cdots, m.$$

Note one of the special features in ESPRIT study is if a patient has not experienced any terminating event before time t_m which is maximum treatment length, this patient would be forced to receive treatment with length t_m. Consequently, when $r = m, \widetilde{\lambda}_r(X) = 1$. This feature leads to the fact that

$$K_m(G_m(Z)) = P(C^* > m | Z) = 0, \text{if } C > t_m.$$

We show relationship between discrete hazard function $\lambda\{G_r^1(C,X)\}$ and discrete cause-specific hazard function $\widetilde{\lambda}(X)$ in the next Lemma.

Lemma 1. *Under the CAR assumption* (3.1) *and no unmeasured confounding assumption* (3.2) *,*

$$\lambda_r\{G_r^1(C,X)\} = \widetilde{\lambda}_r(X) I(C > t_r), r = 1, \cdots, m,$$

where $\lambda_r\{G_r^1(C,X)\}$ *denotes discrete hazard function* $P(C^* = r | C^* \geq r, Z)$ *and* $\widetilde{\lambda}_r(X)$ *denotes discrete cause-specific hazard function* $P(U = t_r, \Delta = 1 | U \geq t_r, X)$.

Proof:

The event $C^* \geq r$, which includes $C^* = \infty$, is equal to

$$C^* \geq r = (T \geq t_r, C > T) \cup (C \leq T).$$

Therefore,

$$\begin{aligned} & \lambda_r(G_r^1(C, X)) = P(C^* = r | C^* \geq r, Z) \\ = \;& P(T = t_r, C > t_r | (T \geq t_r, C > T) \cup (C \leq T), C, X). \end{aligned}$$

If $C \leq t_r$, then $\lambda_r(G_r(Z)) = 0$, whereas if $C > t_r$, then

$$\begin{aligned} & (T \geq t_r, C > T) \cup (C \leq T)\} \cap (C > t_r) \\ = \;& (T \geq t_r, C > T) \cap (C > t_r)\} \cup \{(C \leq T) \cap (C > t_r)\} \\ = \;& (T \geq t_r, C > T) \cup (C > t_r, C \leq T) \\ = \;& (T \geq t_r, C > T, C > t_r) \cup (C > t_r, C \leq T, T \geq t_r) \\ = \;& (T \geq t_r, C > t_r). \end{aligned}$$

$$\begin{aligned} & \lambda_r(G_r^1(C, X)) = P(C^* = r | C^* \geq r, Z) \\ = \;& P(C^* = r | (T \geq t_r, C > T) \cup (C \leq T), C, X) \\ = \;& P(C^* = r | (T \geq t_r, C > T) \cup (C \leq T), C, X) I(C > t_r) \\ = \;& P(T = t_r, C > t_r | T \geq t_r, C, X) I(C > t_r) \\ = \;& P(T = t_r, C > t_r | T \geq t_r, C \geq t_r, X) I(C > t_r) \\ = \;& P(U = t_r, \Delta = 1 | U \geq t_r, X) I(C > t_r) \\ = \;& \widetilde{\lambda}_r(X) I(C > t_r), \; r = 1, \cdots, m. \end{aligned}$$

□

Lemma 1 builds up the fact that $\omega(r, G_r^1(C,X))$ can be written as a function of X and indicator of $C > t_r$, specifically,

$$\omega(r, G_r^1(C,X)) = \begin{cases} \widetilde{\lambda}_1(X) I(C > t_1) & r = 1 \\ \prod_{r'=1}^{r-1} \left[1 - \widetilde{\lambda}_{r'}(X)\right] \widetilde{\lambda}_r(X) I(C > t_r) & r = 2, \cdots, m \\ \prod_{r'=1}^{m-1} \left[1 - \widetilde{\lambda}_{r'}(X) I(C > t'_r)\right] I(C \le t_m) & r = \infty. \end{cases}$$

To simplify notation, we will use the following symbols from now on.

$$\begin{aligned} &\lambda_r = \lambda_r(G_r^1(C,X)), \\ &\widetilde{\lambda}_r = \widetilde{\lambda}_r(X), \text{ so } \lambda_r = \widetilde{\lambda}_r I(C > t_r); \\ &K_r = K_r\{G_r^1(C,X)\} = \prod_{r'=1}^{r} [1 - \lambda\{G_{r'}^1(C,X)\}], \\ &\widetilde{K}_r = \prod_{r'=1}^{r} (1 - \widetilde{\lambda}_{r'}); \\ &\omega_r = \omega(r, G_r^1(C,X)), \\ &\widetilde{\omega}_r = \begin{cases} \widetilde{\lambda}_1 & r = 1 \\ \prod_{r'=1}^{r-1} (1 - \widetilde{\lambda}_{r'}) \widetilde{\lambda}_r & r = 2, \cdots, m \end{cases} \end{aligned}$$

Any symbol with a "$\sim$ " on it denotes a function with history information X collected up to observed time only.

3.2.4 Influence Functions of Full Data and Observed Data

In order to derive the observed-data regular asymptotic linear (RAL) estimator for $\mu_r = E[Y_r^* I(C > t_r) + Y_C^* I(C \le t_r)], r = 1, ..., m$, we need to know the influence function of full data then that of the observed data.

Influence Functions of Full Data

A full-data M-estimator $\hat{\mu}_r (r = 1, \cdots, m)$ can be derived by solving the estimating

equations

$$\sum_{i=1}^{n}\{Y_{ri}^{*}I(C_i > t_r) + Y_{Ci}^{*}I(C_i \leq t_r)\} - \mu_r = 0, r = 1, \cdots, m.$$

The influence function of $\hat{\mu}_r$ is given by (P.31, Tsiatis 2005)

$$-\left[E\frac{\partial m_r(Z,\mu_{0r})}{\partial \mu_{0r}}\right]^{-1} m_r(Z,\mu_{0r}),$$

where μ_{0r} is the true value and $m_r(Z,\mu_{0r}) = Y_r^{*}I(C > t_r) + Y_C^{*}I(C \leq t_r) - \mu_{0r}$. Since $-E\{\frac{\partial m_r(Z,\mu_{0r})}{\partial \mu_{0r}}\} = 1$, we immediately deduce that the influence function of $\hat{\mu}_r$ is

$$m_r(Z,\mu_{0r}) = \varphi_r^{F}(Z) = Y_r^{*}I(C > t_r) + Y_C^{*}I(C \leq t_r) - \mu_r, \quad r = 1, ..., m.$$

General Form of Influence Function for Observed Data

Definition. A mapping, also sometimes referred to as an operator, $\mathcal{K}$, is a function that maps each element of some linear space to an element of another linear space. In all of our applications, the linear spaces are well-defined Hilbert spaces. We define the many-to-one mapping

$$\mathcal{K} : \mathcal{H} \to \mathcal{H}^{F}$$

to be

$$\mathcal{K}(h) = E[h\{h(C^{*}, G_C(Z))\}|Z],$$

for $h \in \mathcal{H}$. We define an inverse operator, for any element $h^{F} \in (H)^{F}, \mathcal{K}^{-1}(h^{F})$ corresponds to the set of all elements $h \in \mathcal{H}$ such that $\mathcal{K}(h) = h^{F}$.

Parameter of interest is a vector with m dimensions, i.e., $\mu = (\mu_1, \cdots, \mu_m)$, so we discuss estimating approach and make inference for a given μ_r, finally stack them up.

If $\varphi^{F}(Z)$ is the full-data influence function for μ_r, then the observed-data influence function corresponds to functions $K^{-1}(\varphi^{F}(Z))$, $(K : H \to H^{F})$ given by definition

above, is a function belong to the class of

$$\{\varphi(Z)\} = h(Y, U, \Delta, X) + \Lambda_2,$$

where

(Condition 1.) $h(Y, U, \Delta, X)$ is any function that satisfies the relationship

$$E\{h(Y, U, \Delta, X)|Y_1^*, \cdots, Y_m^*, Y_C^*, C, X\} = \varphi^F(Z) = Y_r^* I(C > t_r) + Y_C^* I(C \leq t_r) - \mu_r;$$

(Condition 2.) Λ_2 is the linear subspace in Hilbert space $\mathcal{H}$ consisting of elements $L_2(Y, U, \Delta, X)$ such that

$$E\{L_2(Y, U, \Delta, X)|Y_1^*, \cdots, Y_m^*, Y_C^*, C, X\} = 0.$$

Theorem 1. *When the coarsening probability is known to us, say γ_0, the class of observed-data influence functions is denoted as:*

$$\{\varphi(Z)\} = h(Y, U, \Delta, X, \gamma_0) + L_2,$$

where

$$h(Y, U, \Delta, X) = (Y - \mu_r)\left\{\frac{I(U = t_r, \Delta = 1)}{\omega_k} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty}\right\}. \tag{3.4}$$

$$L_2 = \sum_{k=1}^{m-1}\left\{I(C^* = k) - \frac{\omega_k I(C^* = m)}{\omega_m} - \frac{\omega_k I(C^* = \infty, t_k < C \leq t_m)}{\omega_\infty}\right\} l_k(X) \tag{3.5}$$

$l_k(X)(k = 1, \cdots, m-1)$ *are any functions of X.*

Before we prove Theorem 1, we first introduce two Lemmas, each of which shows that Condition 1 and Condition 2 are satisfied by defining function h and L_2 as in (3.4) and (3.5).

Lemma 2.

$E\{h(Y,U,\Delta,X)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\} = Y_r^* I(C>t_r)+Y_C^* I(C<t_r)-\mu_r, r=1,\cdots,m.$

where $h(Y,U,\Delta,X)$ *is defined in the* (3.4).

Proof:

$$\begin{aligned}
& E\{h(Y,U,\Delta,X)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\} \\
= \; & E\left\{\frac{YI(U=t_r,\Delta=1)}{\omega_r}|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\} \\
& +E\left\{\frac{YI(U\le t_r,\Delta=0)}{\omega_\infty}|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}-\mu_r \\
= \; & \frac{E\left\{Y_r^* I(T=t_r,C>t_r)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_r} \\
& +\frac{E\left\{Y_C^* I(C\le t_r,C\le T)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_\infty}-\mu_r.
\end{aligned}$$

Conditioning on full data, the last equation leads to the following equation:

$$\begin{aligned}
= \; & \frac{Y_r^* I(C>t_r)E\left\{I(T=t_r)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_r} \\
& +\frac{Y_C^* I(C\le t_r)E\left\{I(C<T)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_\infty}-\mu_r \\
= \; & \frac{Y_r^* I(C>t_r)P\left\{(T=t_r)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_r} \\
& +\frac{Y_C^* I(C\le t_r)P\left\{I(C<T)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\right\}}{\omega_\infty}-\mu_r \\
= \; & \frac{Y_r^* I(C>t_r)P\left\{T=t_r,C>t_r|C,X\right\}}{\omega_r}+\frac{Y_C^* I(C\le t_r)P\left\{C<T|C,X\right\}}{\omega_\infty}-\mu_r
\end{aligned}$$

The last equation holds because of assumption (3.2) and $P\left\{T=t_r,C>t_r|C<t_r,X\right\} =$

0. Therefore,

$$\begin{aligned} & E\{h(Y,U,\Delta,X)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\} \\ = & \frac{Y_r^* I(C>t_r)\omega_r}{\omega_r} + \frac{Y_C^* I(C<t_r)\omega_\infty}{\omega_\infty} - \mu_r \\ = & Y_r^* I(C>t_r) + Y_C^* I(C<t_1) - \mu_r \\ = & \varphi^F(Z). \end{aligned}$$

□

Lemma 3.

$$E\{L_2(Y,U,\Delta,X)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\} = 0,$$

where

$$L_2 = \sum_{k=1}^{m-1} \left\{ I(C^*=k) - \frac{\omega_k I(C^*=\infty, t_k < C \le t_m)}{\omega_\infty} - \frac{\omega_k I(C^*=m)}{\omega_m} \right\} l_k(X),$$

$l_k(X)$ *is any function of* X *only,* $k = 1,\cdots,m-1$.

Proof:

Since the coarsening variable C^* is discrete, we can express any function,

$$L_2(C^*, G_{C^*}(Y_1^*,\cdots,Y_m^*,Y_C^*,C,X)) \in H,$$

as

$$\sum_{k=1}^{\infty} I(C^*=k) L_{2k}(G_k(Z)).$$

So, the space of functions $L_2 \in \Lambda_2 \subset H$ must satisfy

$$E(\sum_{k=1}^{\infty} I(C^*=k) L_{2k}(G_k(Z))|Z) = 0.$$

Hence,

$$\sum_{k=1}^{\infty} E(I(C^* = k)L_{2k}(G_k(Z))|Z) = \sum_{k=1}^{\infty} L_{2k}(G_k(Z))\omega(k, G_k(Z)) = 0. \tag{3.6}$$

The form of $\omega_k, k = 1, \cdots, m, \infty$, and their properties discussed in the previous section are applied to (3.6), then we have the following results.

When $C \leq t_1$, $\omega_1 = \omega_2 = ...\omega_m = 0$ and $\omega_\infty = 1$. Plug them into equation (3.6), equation (3.6) will be reduced to

$$L_{2\infty}(G_\infty(Z)) = 0,$$

Therefore, to make the equation above hold, it only requires that $L_{2\infty} = 0$, but $L_{2k}, k = 1, ..., m$, can be any function when $C \leq t_1$.

When $t_1 < C \leq t_2$, $\omega_k = 0$ for $k = 2, ..., m$, and $\omega_1 = \widetilde{\lambda}_1$, $\omega_\infty = 1 - \widetilde{\lambda}_1$. Equation (3.6) will be reduced to

$$\omega_1 L_{21} + \omega_\infty L_{2\infty} = 0.$$

Because both ω_1 and ω_∞ are functions of X only, then the equation above can only be true if both L_{21} and $L_{2\infty}$ are function of X when $t_1 < C \leq t_2$. Thus,

$$\Rightarrow L_{2\infty} = -\frac{\omega_1}{\omega_\infty} L_{21},$$

and $L_{2k}, k = 2, ..., m$ can be any function.

Generally, when $t_{r-1} < C \leq t_r (r = 2, \cdots, m)$, we have $\omega_1 = \widetilde{\lambda}_1$, $\omega_k = \prod_{r'=1}^{k-1} [1 - \widetilde{\lambda}'_r] \widetilde{\lambda}_k I(C > t_k)$ for $k = 2, \cdots, r$, $\omega_k = 0$ for $k = r+1, ..., m$ and $\omega_\infty = \prod_{r'=1}^{m} [1 - \widetilde{\lambda}_{r'} I(C > t'_r)]$. Thus, equation (3.6) is reduced to

$$\omega_\infty L_{2\infty} + \sum_{k=1}^{r} \omega_i L_{2k} = 0 \Rightarrow L_{2\infty} = -\frac{\sum_{k=1}^{r} \omega_k L_{2k}}{\omega_\infty},$$

and $L_{2k}, k = r+1, ..., m$, can be any function.

Finally, when $C > t_m$, $\omega_\infty = 0$. We have

$$\sum_{k=1}^{m} \omega_k L_{2k} = 0 \Rightarrow L_{2m} = -\frac{\sum_{k=1}^{m-1} \omega_k L_{2k}}{\omega_m},$$

and $L_{2\infty}$ can be any function.

Therefore, if we choose any $m-1$ functions of X, $l_{2k} = l_{2k}(X), k = 1, \cdots, m-1$, and define

$$\begin{aligned}
L_{2k} &= l_{2k}, k = 1, \cdots, m-1, \\
L_{2m} &= \frac{-\sum_{k=1}^{m-1} \omega_k l_{2k}}{\omega_m}, \\
L_{2\infty} &= -\frac{1}{\omega_\infty}\{I(t_1 < C \le t_2)\omega_1 l_{21} + I(t_2 < C \le t_3)(\omega_1 l_{21} + \omega_2 l_{22}) + \cdots \\
&\quad + I(t_{m-1} < C \le t_m)(\omega_1 l_{21} + \omega_2 l_{22} + ... + \omega_{m-1} l_{2,m-1}, \\
&= -\frac{1}{\omega_\infty} \sum_{k=1}^{m-1} \sum_{i=1}^{k} I(t_i < C \le t_{i+1}) \omega_i l_{2i}.
\end{aligned}$$

Rearrange the terms in the last line, we obtain that

$$L_{2\infty} = -\frac{I(C \le t_m)}{\omega_\infty} \sum_{k=1}^{m-1} I(C > t_k) \omega_k l_{2k}.$$

The choice of $l_{2k}, k = 1, \cdots, m-1$, and definition for $L_{2k}, k = 1, \cdots, m, \infty$, satisfy

the constraints we discussed previously. Therefore,

$$\begin{aligned} L_2 &= \sum_{k=1}^{\infty} I(C^*=k) L_{2k}(G_k(Z)) \\ &= \sum_{k=1}^{m-1} I(C^*=k) l_{2k} - I(C^*=t_m)\frac{\sum_{k=1}^{m-1}\omega_k l_{2k}}{\omega_m} - I(C^*=\infty)\frac{I(C\le t_m)}{\omega(\infty)}\sum_{k=1}^{m-1} I(C>t_k)\omega_k l_{2k} \\ &= \sum_{k=1}^{m-1}\left\{ I(C^*=k) - \frac{\omega_k I(C^*=m)}{\omega_m} - \frac{\omega_k I(C^*=\infty, t_k < C \le t_m)}{\omega(\infty)} \right\} l_{2k}. \end{aligned}$$

The procedure to construct L_2 ensures that $E\{L_2(Y,U,\Delta)|Y_1^*,\cdots,Y_m^*,Y_C^*,C,X\}=0$. □

Lemma 2 and Lemma 3 together support the statement of Theorem 1.

The Augmentation Space Λ_2 with Partially-monotone Coarsening

The advantage of defining partially-monotone coarsening is we can derive another equivalent representation for L_2, the element of space Λ_2, which helps to derive the most efficient influence function of μ_r. This representation takes advantage of cause-specific hazards as we demonstrate in the following Lemma.

Lemma 4. *Under partially-monotone coarsening and assumption* $\widetilde{\lambda}_m(X) = P(U = t_m, \Delta = 1 | U \ge t_m, X) = 1$, *a typical element of* Λ_2 *can be expressed as*

$$\sum_{k=1}^{m-1}\left\{ \frac{I(C^*=r) - \lambda_r I(C^*\ge r)}{K_k} \right\} l_k(X) \tag{3.7}$$

$$= \sum_{k=1}^{m-1}\left\{ \frac{I(U=t_k,\Delta=1) - \widetilde{\lambda}_r I(U\ge t_k)}{\widetilde{K}_k} \right\} l_k(X), \tag{3.8}$$

Where $K_k = \prod_{i=1}^{k}[1-\lambda_i]$ *and* $\widetilde{K}_k(X) = \prod_{i=1}^{k}[1-\widetilde{\lambda}_i]$.

Let's derive an equation needed for proving Lemma 4.

Preposition 1. *For a fixed i,* $i = 0, \cdots, m-2$,

$$\sum_{j=i+1}^{m-1} \left\{ \frac{I(C^* = j) - \lambda_j I(C^* \geq j)}{K_j} \right\} = \frac{I(C^* \geq i+1)}{K_i} - \frac{I(C^* = m)}{K_{m-1}} - \frac{I(C^* = \infty, C \leq t_m)}{K_{m-1}},$$

where $K_j = P(C^* = j|Z) = \prod_{j' \leq j} [1 - \lambda_{j'}]$ *and* $\lambda_j = P(C^* = j | C^* \geq j, Z)$.

$$\begin{aligned}
& \sum_{j=i+1}^{m-1} \left\{ \frac{I(C^* = j) - \lambda_j I(C^* \geq j)}{K_j} \right\} \\
= & \sum_{j=i+1}^{m-1} \left\{ \frac{I(C^* = j)}{K_j} - \frac{\lambda_j I(C^* \geq j)}{K_j} \right\} \\
= & \sum_{j=i+1}^{m-1} \left\{ \frac{I(C^* \geq j)}{K_j} - \frac{I(C^* \geq j+1)}{K_j} - \frac{\lambda_j I(C^* \geq j)}{K_j} \right\}.
\end{aligned}$$

The equation above holds because of the fact that $I(C^* = j) = I(C^* \geq j) - I(C^* \geq j+1)$. The fact that $K_j = \prod_{j' \leq j} [1 - \lambda_{j'}] = K_{j-1}(1 - \lambda_j)$ results in the following procedure:

$$\begin{aligned}
= & \sum_{j=i+1}^{m-1} \left\{ \frac{I(C^* \geq j)}{K_{j-1}} - \frac{\lambda_j I(C^* \geq j+1)}{K_j} \right\} \\
= & \frac{I(C^* \geq i+1)}{K_i} - \frac{I(C^* \geq m)}{K_{m-1}} \\
= & \frac{I(C^* \geq i+1)}{K_i} - \frac{I(C^* = m)}{K_{m-1}} - \frac{I(C^* = \infty)}{K_{m-1}} \\
= & \frac{I(C^* \geq i+1)}{K_i} - \frac{I(C^* = m)}{K_{m-1}} - \frac{I(C^* = \infty, C \leq t_m)}{K_{m-1}}.
\end{aligned}$$

The last equation holds because the special mechanism in our case to assign the treatment to a patient, which if a patient has not experience any termination event at time t_m, which is the last treatment length, the the patient would receive the

treatment length of t_m for sure. Therefore, event $\{C^* = \infty\}$ only happened when $C \leq t_m$. $\square$

Now, we start proving Lemma 4.

Proof for Lemma 4: In Lemma 3, we have defined a typical element of Λ_2 as

$$\sum_{k=1}^{m-1} \left\{ I(C^* = k) - \frac{I(C^* = m)\omega_k}{\omega_m} - \frac{\omega_k I(C^* = \infty, t_k < C \leq t_m)}{\omega(\infty)} \right\} l_{2k}. \tag{3.9}$$

We also have facts that (f1) $K_{i-1} - \omega_i = K_i$ and (f2) $I(C^* = k) = I(C^* \geq k) - I(C^* \geq k+1)$. As proved by Lemma 2,

$$\begin{aligned} L_2 &= \sum_{k=1}^{m-1} \left\{ I(C^* = k) - \frac{\omega_k I(C^* = \infty, t_k < C \leq t_m)}{\omega(\infty)} - \frac{I(C^* = m)\omega_k}{\omega_m} \right\} l_{2k}(x) \\ &= \sum_{k=1}^{m-1} \left\{ \frac{I(C^* = k)K_{i-1}}{K_i} - \frac{I(C^* = k)\omega_k}{K_k} - \frac{\omega_k I(C^* = \infty, t_k < C \leq t_m)}{\omega_\infty} - \frac{I(C^* = m)\omega_k}{\omega_m} \right\} l_{2k}. \end{aligned}$$

The last equation holds because of the fact (f1), and the fact (f2) leads to next equation.

$$\begin{aligned} = \sum_{k=1}^{m-1} \Bigg\{ & \frac{I(C^* = k)K_{k-1}}{K_k} - \frac{I(C^* \geq k)\omega_k}{K_k} + \frac{I(C^* \geq k+1)\omega_k}{K_k} \\ & - \frac{I(C^* = \infty, t_k < C \leq t_m)\omega_k}{\omega_\infty} - \frac{I(C^* = m)\omega_k}{\omega_m} \Bigg\} l_{2k}(x). \end{aligned}$$

Applying the result in Lemma 3 here, we have

$$
\begin{aligned}
L_2 &= \sum_{k=1}^{m-1} \Bigg\{ \frac{I(C^*=k) - I(C^* \geq k)\lambda_k}{K_k} + \lambda_k \sum_{j=k+1}^{m-1} \frac{I(C^* \geq j) - I(C^* \geq j)\lambda_j}{K_j} \\
&\quad + \frac{I(C^* = \infty, C \leq t_m)\lambda_k}{K_{m-1}} - \frac{I(C^* = \infty, t_k < C \leq t_m)\lambda_k}{\omega_\infty} \Bigg\} l_{2k}(x) K_{k-1} \\
&= \sum_{k=1}^{m-1} \Bigg\{ \frac{I(C^*=k) - I(C^* \geq k)\lambda_k}{K_k} + \lambda_k \sum_{j=k+1}^{m-1} \frac{I(C^* \geq j) - \lambda_j I(C^* \geq j)}{K_j} \\
&\quad + I(C^* = \infty, t_k < C \leq t_m)\lambda_k \left(\frac{1}{K_{m-1}} - \frac{1}{\omega_\infty} \right) \qquad (3.10) \\
&\quad + \frac{I(C^* = \infty, C \leq t_k)\lambda_k}{K_{m-1}} \Bigg\} l_{2k}(x) K_{k-1}. \qquad (3.11)
\end{aligned}
$$

Note:

$$
\begin{aligned}
K_{m-1} &= \prod_{i=1}^{m-1} (1 - \lambda_i(x) I(C > t_i)), \\
\omega_\infty &= \prod_{i=1}^{m-1} (1 - \lambda_i(x)) I(C \leq t_m).
\end{aligned}
$$

When $C \leq t_m$, $K_{m-1} = \omega_\infty$,(3.10)=0. When $C < t_k$, (3.11)=0. In addition, $I(C \leq t_i)\lambda_i$=0.

So,

$$
\begin{aligned}
L_2 &= \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-I(C^*\geq k)\lambda_k}{K_k}+\lambda_k\sum_{j=k+1}^{m-1}\frac{I(C^*=j)-I(C^*\geq j)}{K_j}\right\}K_{k-1}l_{2k} \\
&= \sum_{k=1}^{m-1}\frac{I(C^*=k)-I(C^*\geq k)\lambda_k}{K_k}K_{k-1}l_{2k}+\sum_{k=1}^{m-1}\omega_k l_{2k}\sum_{j=k+1}^{m-1}\frac{I(C^*=j)-\lambda_j I(C^*\geq j)}{K_j} \\
&= \sum_{k=1}^{m-1}\frac{I(C^*=k)-I(C^*\geq k)\lambda_k}{K_k}K_{k-1}l_{2k}+\sum_{j=2}^{m-1}\frac{I(C^*=j)-\lambda_j I(C^*\geq j)}{K_j}(\sum_{k=1}^{j-1}\omega_k l_{2k}) \\
&= \sum_{k=1}^{m-1}\frac{I(C^*=k)-I(C^*\geq k)\lambda_k}{K_k}K_{k-1}l_{2k}+\sum_{k=2}^{m-1}\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}(\sum_{j=1}^{k-1}\omega_j l_{2j}) \\
&= \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\}l_k.
\end{aligned}
$$

where$l_1 = l_{21}$ and $l_k = K_{k-1}l_{2k}+\sum_{j=1}^{k-1}\omega_j l_{2j}, k=2,\cdots,m-1$, and it is a function of X only. Equation (3.7) is proved. Plug $\lambda_r = \widetilde{\lambda}_r I(C>t_r)$ into (3.7), we have equation (3.8). □

Optimal Influence Function

So far, we have demonstrated that a class of the influence functions of RAL estimators for μ_r, when the parameters in the coarsening probability models are known to us, is given by

$$
\{\varphi(Z)\} = (Y-\mu_r)\left\{\frac{I(U=t_r,\Delta=1)}{\omega_k}+\frac{I(U\leq t_r,\Delta=0)}{\omega_\infty}\right\}+\Lambda_2,
$$

where Λ_2 is the linear subspace consisting of elements L_2 and L_2 is defined by Lemma 4. The optimal influence function in the class of $\{\varphi(Z)\}$ is the one with the smallest variance or, equivalently, the element with the smallest norm. This is obtained by choosing L_2 as the projection of $h(Y,U,\Delta,X,\psi_0)$ onto Λ_2, $\prod(h|\Lambda_2)$, in which case

the optimal influence function is given by

$$h(Y, U, \Delta, X, \psi_0) - \prod(h|\Lambda_2),$$

for a fixed full-data influence function, where ψ_0 is the nuisance parameter. Theoretical proof can be found on P.223, Tsiatis (2006).

However, if the coarsening probability were not known and had to be modeled using the unknown ψ, then Theorem 9.1 (Tsiatis, 2005) showed that if the coarsening process follows a parametric model, and if the nuisance parameter is estimated using the maximum likelihood method, or any other efficient method, then the solution to the estimating equation

$$\sum_{i=1}^{n}[h(Y, U, \Delta, X, \hat{\psi}) + L_0(\hat{\psi})] = 0, \tag{3.12}$$

will be an estimator whose influence function is $h(Y, U, \Delta, X) - \prod(h|\Lambda_2)$ and $L_0 = \prod(\{h\}|\Lambda_2)$ is defined as the projection of $h(Y, U, \Delta, \psi_0)$ on the space of Λ_2.

Thus far, we defined partially-monotone coarsening and explored features under this setting. We also described the class of observed-data influence functions when data are partially-monotone coarsened at random by taking advantage of results obtained for a full-data semiparametric model and showed the special form of influence functions under this setting. We provided the tool to derive an efficient estimator by solving the estimating equation defined in (3.12). Ultimately, the goal is to derive as efficient an estimator for μ_r as is possible using partially monotone coarsened data. In order to fulfill the task, all the work narrows down finding the projection of $h(Y, U, \Delta, X)$ on Λ_2. Now, we show how to derive the projection of $h(Y, U, \Delta, X)$ on Λ_2.

3.2.5 Projection of $h(Y, U, \Delta, X)$ on Λ_2

In order to improve efficiency, we need to find the projection of $h(Y, U, \Delta, X)$ on Λ_2, where

$$\begin{aligned} & h(Y,U,\Delta,X) = h_1(Y,U,\Delta,X) + h_2(Y,U,\Delta,X) \\ = \;& (Y-\mu_r)\frac{I(U=t_r,\Delta=1)}{\omega_r} + (Y-\mu_r)\frac{I(U<t_r,\Delta=0)}{\omega_\infty}, \end{aligned} \tag{3.13}$$

in which $\omega_r = P(C^* = r|Z) = \prod_{r'=1}^{r}(1-\lambda_{r'})\lambda_r$ and $\lambda_r = P(C^* = r|C^* \geq r, Z)$.

By the linearity of projections,

$$\prod(\{h(Y,U,\Delta,X)\}|\Lambda_2) = \prod(\{h_1(Y,U,\Delta,X)\}|\Lambda_2) + \prod(\{h_2(Y,U,\Delta,X)\}|\Lambda_2),$$

so that we can derive the projections of h_1 and h_2 onto Λ_2 separately.

Since $\prod(h_1|\Lambda_2)$ and $\prod(h_2|\Lambda_2) \in \Lambda_2$, with Lemma 4 they can be written as

$$\sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_{0k},$$

where $K_k = P(C^* > k|Z) = \prod_{r'=1}^{r}(1-\lambda_{r'})$, such that,

$$\begin{aligned} & E\left(\left[h_1 - \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_{0k}\right]\right. \\ & \left.\times \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_k\right) = 0, \end{aligned} \tag{3.14}$$

and

$$E\left(\left[h_2 - \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_{0k}\right] \times \sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_k\right) = 0, \tag{3.15}$$

for all $l_k, k = 1, \cdots, m-1$.

We demonstrate the specific form of projection of h_1 and h_2 in the next two prepositions.

Theorem 2. *Let r be given.*

$$L_{01} = \prod h_1(Y,U,\Delta,X)|\Lambda_2 = \sum_{k=1}^{r}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_{0k}, \tag{3.16}$$

where

$$l_{0k} = \begin{cases} -\frac{p_r E(Y_r^*|X)}{p_k} & k = 1, \cdots, r-1 \\ \frac{(1-\widetilde{\lambda}_r)E(Y_r^*|X)}{\widetilde{\lambda}_r} & k = r \end{cases}$$

, and $p_k = Prob.(C > t_k|X)$, and $\prod h_1(Y,U,\Delta,X)|\Lambda_2$ denotes the projection of $h_1(Y,U,\Delta,X)$ onto Λ_2.

Theorem 3. *Let r be given.*

$$L_{02} = \prod h_2(Y,U,\Delta,X)|\Lambda_2 = \sum_{k=1}^{r-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\} l_{0k}, \tag{3.17}$$

where

$$l_{0k} = -\frac{1}{p_k}\sum_{j=k}^{r-1} p_{j,j+1} E(Y_C^*|X, t_j < C \leq t_{j+1}), k = 1, \cdots, r-1.$$

$p_{j,j+1} = Prob.(t_j < C \leq t_{j+1}|X)$ *and* $p_k = Prob.(C > t_k|X)$.

We first derive some relationships that will simplify the calculations in two the-

orems above in the following lemmas. The first two prepositions have been proved Tsiatis (2006), so here we just listed them. We prove the remainings.

Lemma 5. *For $k \neq k'$,*

$$\begin{aligned} E & \left(\left\{ \frac{I(C^* = k) - \lambda_k I(C^* \geq k)}{K_k} \right\} l_{0k} \right) \\ \times & \left(\left\{ \frac{I(C^* = k') - \lambda_{k'} I(C^* \geq k')}{K_{k'}} \right\} l_{0k'} \right) = 0. \end{aligned} \tag{3.18}$$

Lemma 6. *For $k = k'$,*

$$\begin{aligned} E & \left(\left\{ \frac{I(C^* = k) - \lambda_k I(C^* \geq k)}{K_k} \right\} l_{0k} \right) \\ \times & \left(\left\{ \frac{I(C^* = k') - \lambda_{k'} I(C^* \geq k')}{K_{k'}} \right\} l_{0k'} \right) \\ = & \; E \left[\frac{\lambda_k}{K_k} l_{0k} \right]. \end{aligned} \tag{3.19}$$

Note $\lambda_k = \widetilde{\lambda}_k(X) I(C > t_k)$ and when $C > t_k$ $K_k = \widetilde{K}_k$, we continue (3.19) by conditioning on X, then we can get the following result.

$$(3.19) \;\; = \;\; E \left[\frac{p_k \widetilde{\lambda}_k}{\widetilde{K}_k} l_{0k} \right]. \tag{3.20}$$

Lemma 7. *Let r be given, for $k < r$,*

$$\begin{aligned} & E \left(Y \frac{I(U = t_r, \Delta = 1)}{\omega_r} \right) \times \left(\left\{ \frac{I(C^* = k) - \lambda_k I(C^* \geq k)}{K_k} \right\} l_k \right) \\ = & \; -E \left[\frac{\widetilde{\lambda}_k p_r E(Y_r^* | X)}{\widetilde{K}_k} l_k \right]. \end{aligned} \tag{3.21}$$

where $p_r = P(C > t_r | X)$.

Proof: Because of equivalence of event $\{U = t_r, \Delta = 1\}$ and event $\{C^* = r\}$,

$$\begin{aligned} & E\left\{Y\frac{I(U=t_r,\Delta=1)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\} \\ = \; & E\left\{Y\frac{I(C^*=r)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\} \\ = \; & E\left[-\left\{Y\frac{I(C^*=r)\lambda_k l_k Y}{\omega_r K_k}\right\}\right]. \end{aligned}$$

When $I(C > t_k) = 1$, $\omega_k = \widetilde{\omega}_k$ and $K_k = \widetilde{K}_k$. Therefore, combined with the SUTVA assumption, the equation above is equal to

$$\begin{aligned} & -E\left\{Y_r^*\frac{I(C^*=r)\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{\omega}_r\widetilde{K}_k}\right\} \\ = \; & -E\left\{E\left[Y_r^*\frac{I(C^*=r)\widetilde{\lambda}_k I(C>t_k)l_k Y_r^*}{\widetilde{\omega}_r\widetilde{K}_k}|C^*,X\right]\right\} \\ = \; & -E\left\{I(C^*=r)\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{\omega}_r\widetilde{K}_k}\left[EY_r^*|C^*=r,X\right]\right\} \\ = \; & -E\left\{E\left[I(C^*=r)\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{\omega}_r\widetilde{K}_k}\left[EY_r^*|C^*=r,X\right]\right]|C,X\right\} \\ = \; & -E\left\{\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{\omega}_r\widetilde{K}_k}\left[EY|C^*=r,X\right]E\left[I(C^*=r)|C,X\right]\right\} \\ = \; & -E\left\{\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{\omega}_r\widetilde{K}_k}\left[EY|C^*=r,X\right]\omega_r\right\} \\ = \; & -E\left\{\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{K}_k}\left[EY_r^*|C^*=r,X\right]\right\} \\ = \; & -E\left\{E\left\{\frac{\widetilde{\lambda}_k I(C>t_k)l_k}{\widetilde{K}_k}\left[EY_r^*|C^*=r,X\right]\right\}|X\right\} \\ = \; & -E\left[\frac{\widetilde{\lambda}_k p_r E(Y_r^*|X)}{\widetilde{K}_k}l_k\right], \end{aligned}$$

where $p_r = P(C > t_r|X)$. □

Lemma 8. *Let r be given, if $k = r$,*

$$\begin{aligned} & E\left\{Y\frac{I(U=t_r,\Delta=1)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\} \\ = \; & E\left\{\frac{(1-\widetilde{\lambda}_k)\left[EY_r^*|X\right]p_r}{\widetilde{K}_k}l_k\right\}, \end{aligned} \tag{3.22}$$

where $p_r = P(C > t_r|X)$.

Proof:

Note that

$$\begin{aligned} & E\left\{Y\frac{I(U=t_r,\Delta=1)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=r)-\lambda_r I(C^*\geq r)}{K_r}\right]l_r\right\} \\ = \; & E\left\{Y\frac{I(C^*=r)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=r)-\lambda_r I(C^*\geq k)}{K_r}\right]l_r\right\} \\ = \; & E\left\{Y\frac{I(C^*=r)(1-\lambda_r)l_r}{\omega_r}\right\}. \end{aligned}$$

The rest of proof is similar to the proof for the Lemma 7. □

Lemma 9. *Let r be given, for $k > r$,*

$$E\left\{Y\frac{I(U=t_r,\Delta=1)}{\omega_r}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\} = 0. \tag{3.23}$$

The Proof is easily obtained by noticing that the fact $I(C^* = r) \times I(C^* \geq k) = 0$ if $k > r$.□

Lemma 10. *Let r be given, for $k < r$,*

$$\begin{aligned} & E\left\{Y\frac{I(U\leq t_r,\Delta=0)}{\omega_\infty}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\} \\ = \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1}p_{j,j+1}E(Y_C^*|C^*=\infty,t_j<C\leq t_{j+1},X)\right\}, \end{aligned} \tag{3.24}$$

where $p_{j,j+1} = P(t_j < C \leq t_{j+1}|X)$.

Proof:

$$\begin{aligned}
& E\left\{Y\frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty}\right\} \times \left\{\left[\frac{I(C^* = k) - \lambda_k I(C^* \geq k)}{K_k}\right] l_k\right\} \\
= \; & E\left\{Y\frac{I(C^* = \infty, C \leq t_r)}{\omega_\infty}\right\} \times \left\{\left[\frac{I(C^* = k) - \lambda_k I(C^* \geq k)}{K_k}\right] l_k\right\} \\
= \; & -E\left\{Y\left[\frac{\lambda_k I(C^* = \infty, C \leq t_r)}{K_k \omega_\infty}\right] l_k\right\} \\
= \; & -E\left\{Y\left[\frac{\widetilde{\lambda}_k I(C > t_k) I(C^* = \infty, C \leq t_r)}{K_k \omega_\infty}\right] l_k\right\} \\
= \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1}\frac{I(C^* = \infty, t_j < C \leq t_{j+1})Y}{\widetilde{K}_j}\right\}.
\end{aligned}$$

The last equation holds because when $I(C^* = \infty, t_j < C \leq t_{j+1})$=1 at $j > k$, $\omega_\infty = \prod_{i=1}^{j}(1 - \widetilde{\lambda}(X)) = \widetilde{K}_j$ and $K_k = \widetilde{K}_k$. Therefore, after conditioning on $I(C^* = \infty, t_j < C \leq t_{j+1})$ and X, we have

$$-E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1}\frac{I(C^* = \infty, t_j < C \leq t_{j+1})E(Y_C^*|C^* = \infty, t_j < C \leq t_{j+1}, X)}{\widetilde{K}_j}\right\}.$$

Continue conditioning on C and X, then conditioning on X, we get

$$\begin{aligned}
= \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1}\frac{I(t_j < C \leq t_{j+1})E(Y_C^*|C^* = \infty, t_j < C \leq t_{j+1}, X)\omega_\infty}{\widetilde{K}_j}\right\} \\
= \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1} I(t_j < C \leq t_{j+1})E(Y_C^*|C^* = \infty, t_j < C \leq t_{j+1}, X)\right\} \\
= \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1} P(t_j < C \leq t_{j+1}|X)E(Y_C^*|C^* = \infty, t_j < C \leq t_{j+1}, X)\right\} \\
= \; & -E\left\{\frac{\widetilde{\lambda}_k l_k}{\widetilde{K}_k}\sum_{j=k}^{r-1} p_{j,j+1}E(Y_C^*|C^* = \infty, t_j < C \leq t_{j+1}, X)\right\}. \quad \square
\end{aligned}$$

Lemma 11. *Let r be given, for $k \geq r$*

$$E\left\{Y\frac{I(U<t_r,\Delta=0)}{\omega_\infty}\right\}\times\left\{\left[\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right]l_k\right\}=0. \quad (3.25)$$

Lemma 11 is valid because of the fact that $\{C^* = k\}$ implies $\{U > t_k\}$ for $k = 1, \cdots, m$. □

Using the results of Lemma 5 to Lemma 11, now we start proving Theorem 2 and Theorem 3.

Proof for Theorem 2 and Theorem 3:

Plug results (3.21)-(3.23) into (3.14), we have

$$\begin{aligned}
& E\left\{\left[Y\frac{I(U=t_r,\Delta=1)}{\omega_r}-\sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\}l_{0k}\right]\right. \\
\times\quad & \left.\sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\geq k)}{K_k}\right\}l_k\right\} \\
=\quad & E\left\{\sum_{k=1}^{r-1}\left[\frac{\widetilde{\lambda}_k}{\widetilde{K}_k}(p_k l_{0k}+p_r E(Y_r^*|X))l_k\right]+\frac{\widetilde{\lambda}_r l_{0r}-(1-\widetilde{\lambda}_r)E(Y_r^*|X)}{\widetilde{K}_r}p_r l_r+\sum_{k=r+1}^{m-1}\left[\frac{\widetilde{\lambda}_k}{\widetilde{K}_k}l_{0k}l_k\right]\right\}=0.
\end{aligned}$$

In order to make the equation above hold for any function $l_k, k = 1, \cdots, m-1$, if and only if

$$l_{0k}=\begin{cases}-\frac{p_r E(Y_r^*|X)}{p_k} & k=1,\cdots,r-1;\\ \frac{(1-\widetilde{\lambda}_r)E(Y_r^*|X)}{\widetilde{\lambda}_r} & k=r;\\ 0 & k>r,\end{cases}$$

where $p_k = P(C > t_k|X)$.

Similarly, after plugging results of (3.24)-(3.25) into equation (3.23), we have

$$
\begin{aligned}
& E\left\{\left[Y\frac{I(U\le t_r,\Delta=1)}{\omega_r}-\sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\ge k)}{K_k}\right\}l_{0k}\right]\right. \\
\times\ & \left.\sum_{k=1}^{m-1}\left\{\frac{I(C^*=k)-\lambda_k I(C^*\ge k)}{K_k}\right\}l_k\right\} \\
=\ & E\left\{\sum_{k=1}^{r-1}-\frac{\widetilde{\lambda}_k}{\widetilde{K}_k}\left[\sum_{j=k}^{r-1}p_{k,k+1}E(Y_C^*|C^*=\infty,t_j<C<t_{j+1},X)+p_k l_{0k}\right]l_k+\sum_{k=r}^{m-1}\frac{\widetilde{\lambda}_k}{\widetilde{K}_k}l_{0k}l_k\right\}=0.
\end{aligned}
$$

In order to make the equation above hold for any function $l_k, k=1,\cdots,m-1$, if and only if

$$
l_{0k}=\begin{cases}
-\frac{1}{p_k}\sum_{j=k}^{r-1}p_{j,j+1}E(Y_C^*|C^*=\infty,t_j<C\le t_{j+1},X) & k=1,\cdots,r-1; \\
0 & k\ge r,
\end{cases}
$$

where $p_{j,j+1}=P(t_j<C\le t_{j+1}|X)$. □

If the observed data were coarsened by design, we could use results above to generate the following estimating equation for μ_r:

$$
\begin{aligned}
& h_1(Y,U,\Delta,X,\gamma_0)-\prod h_1(Y,U,\Delta,X,\gamma_0)|\Lambda_2 \\
+\ & h_2(Y,U,\Delta,X,\gamma_0))-\prod h_2(Y,U,\Delta,X,\gamma_0)|\Lambda_2=0,
\end{aligned}
$$

Where h_1 and h_2 are defined as (3.13) and $\prod(h_1|\Lambda_2)$ and $\prod(h_2|\Lambda_2)$ are defined as in Theorem 2 and Theorem 3.

If the coarsening probabilities were not known and had to be modeled through the unknown parameter γ, then we would derive an estimator for μ_r by solving the

estimating equation

$$\begin{aligned} & h_1(Y,U,\Delta,X,\hat{\gamma}) + \prod h_1(Y,U,\Delta,X,\hat{\gamma})|\Lambda_2 \\ + \quad & h_2(Y,U,\Delta,X,\hat{\gamma}) + \prod h_2(Y,U,\Delta,X,\hat{\gamma})|\Lambda_2 = 0, \end{aligned} \tag{3.26}$$

where $\hat{\gamma}$ is the MLE for the parameter γ in the model of cause-specific hazard and h_1, h_2, $\prod(h_1|\Lambda_2)$ and $\prod(h_2|\Lambda_2)$ are defined as in Theorem 2 and Theorem 3.

In order to solve the estimating equation (3.26) successfully, we need to estimate γ and compute conditional means $E(Y_r^*|X), r = 1, \cdots, m$, and $E(Y_C^*|t_{j-1} < C \leq t_j), j = 1, \cdots, m-1\}$, and $P(C > c|X)$.

3.2.6 MLE Approach to Estimate the Parameters in the Cause-specific Hazard Function $\widetilde{\lambda}$

We posit a model for the cause-specific hazard functions $\widetilde{\lambda}_r$ through parameter γ, $r = 1, \cdots, m-1$. (Note: $\widetilde{\lambda}_m = 1$). We have showed that the coarsening probability can be deduced through the cause-specific discrete hazard leading to the model

$$\omega(r, G_r^1(Z)) = \begin{cases} \widetilde{\lambda}_1(X)I(C > t_1) & r = 1; \\ \prod_{r'=1}^{r-1}[1 - \widetilde{\lambda}_{r'}(X)I(C > t_r')]\widetilde{\lambda}_r(X)I(C > t_r) & r = 2, \cdots, m-1; \\ \prod_{r'=1}^{m-1}[1 - \widetilde{\lambda}r'(X)I(C > t_r')]I(C > t_m) & r = m; \\ \prod_{r'=1}^{m-1}[1 - \widetilde{\lambda}r'(X)I(C > t_r')]I(C \leq t_m) & r = \infty. \end{cases}$$

We use maximum likelihood estimation (MLE) to estimate the parameters γ in the models above based on observed data. Specifically, the maximum likelihood estimator $\hat{\gamma}_n$ for γ is obtained by maximizing

$$\prod_{i=1}^{n} \omega(r, G_r^1(Z), \gamma). \tag{3.27}$$

Substituting the right-hand side $\omega(r, G_r^1(Z))$ into (3.27) and rearranging terms, we obtain that the likelihood can be expressed as

$$
\begin{aligned}
& \prod_{r=1}^{m-1} \prod_{i:C^* \geq r} [\lambda_r]^{I(C^*=r)} [1-\lambda_r]^{I(C^*>r)} \\
= & \prod_{r=1}^{m-1} \prod_{i:C^* \geq r} \left[\widetilde{\lambda}_r I(C > t_r)\right]^{I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(U>t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(\Delta=0)} \\
= & \prod_{r=1}^{m-1} \prod_{i:C^* \geq r} \left[\widetilde{\lambda}_r\right]^{I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(U>t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(U>t_r,\Delta=0)} \\
= & \prod_{r=1}^{m-1} \prod_{i:U \geq t_r} \left[\widetilde{\lambda}_r\right]^{I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(U>t_r)} \\
= & \prod_{r=1}^{m-1} \prod_{i:U \geq t_r} \left\{ \left[\widetilde{\lambda}_r\right]^{I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{I(U\geq t_r)} \right. \\
& \left. \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{-I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{-I(U=t_r,\Delta=0)} \right\}.
\end{aligned}
$$

Notice the fact that $\{U = t_r, \Delta = 0\}$ implies $\{C \leq T, C = t_r\}$, so

$$\left[1-\widetilde{\lambda}_r I(C > t_r)\right]^{-I(U=t_r,\Delta=0)} = 1.$$

Therefore, likelihood is expressed as

$$\prod_{r=1}^{m-1} \prod_{i:U \geq t_r} \left[\frac{\widetilde{\lambda}_r}{1-\widetilde{\lambda}_r}\right]^{I(U=t_r,\Delta=1)} \left[1-\widetilde{\lambda}_r\right]^{I(U\geq t_r)}.$$

We use the same generalized linear model for $\widetilde{\lambda}_r, r = 1, \cdots, m-1$, as Johnson and Tsiatis(2002) discussed, where

$$\widetilde{\lambda}_r = F(\gamma_r^T X),$$

$r = 1, \cdots, m-1$, and $F(t)$ is the logistic function,i.e., $e^t/(1+e^t)$. This is similar to the continuation ratio logit model(Agresti, 1990, p.319), in which case likelihood

becomes

$$\prod_{r=1}^{m-1} \prod_{i:U \geq t_r} \frac{\exp(\gamma_r^T X_i) I(U_i = t_r, \Delta_i = 1))}{1 + \exp(\gamma_r^T X_i)}.$$

The parameterizations above allows for different parameters associated with each time interval t_k. For parsimony, when applicable, we may assume that some of these parameters are the same across the different time intervals.

3.2.7 Adaptive Estimation to Estimate Conditional Means and Distribution of Terminating Event

We adopt adaptive method proposed by Tsiatis (2006) to find improved estimators, for the purpose of approximating the conditional expectations of potential outcomes and conditional probability of terminating event. That is, We first posit models for conditional probability $P(C|X)$ in terms of the parameter ξ_1 and conditional means $E(Y_r^*|X), r = 1, \cdots, m$, and $E(Y_C^*|X)$ in terms of the parameter ξ_2. Substitute $\hat{\xi_1}$ and $\hat{\xi_2}$ to compute conditional means and conditional probabilities appearing in the estimating equation (3.26). The estimator for μ_r obtained by the following estimating equation is denoted as $\hat{\mu}_r$.

$$\sum_{i=1}^{n} \left[(Y - \mu_r) \left\{ \frac{I(U = t_r, \Delta = 1)}{\omega_r(\hat{\gamma})} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty(\hat{\gamma})} \right\} - L_0(U, \Delta, X, Y, \hat{\gamma}, \hat{\xi_1}, \hat{\xi_2}, \mu_r) \right]. \tag{3.28}$$

Remark: Under suitable regularity conditions, the estimator $\hat{\xi}$ will converge in probability to a constant ξ^*. If models are correctly specifically, ξ^* is right ξ_0 where ξ_0 is true value to describe the models. Even though the posited models may not be correctly specified, $n^{1/2}(\hat{\xi} - \xi^*)$ is bounded in probability. Also, even if the posited model is incorrect, the function $L_0(U, \Delta, X, Y, \hat{\gamma}, \hat{\xi_1}, \hat{\xi_2})$ and $L_0(U, \Delta, X, Y, \hat{\gamma}, \xi^*)$ still

belong to Λ_2. We denote the solution to the estimating equation

$$\sum_{i=1}^{n}\left[(Y-\mu_r)\left\{\frac{I(U=t_r,\Delta=1)}{\omega_r(\hat{\gamma})}+\frac{I(U\leq t_r,\Delta=0)}{\omega_\infty(\hat{\gamma})}\right\}-L_0(U,\Delta,X,Y,\hat{\gamma},\xi_1^*,\xi_2^*,\mu_r^0)\right] \tag{3.29}$$

with μ_r set to the true value μ_r^0 and ξ_1^*,ξ_2^* fixed in $L_0(\cdot)$, is an estimator for μ_r, denoted as $\hat{\mu_r^*}$.

3.2.8 Estimating the Asymptotic Variance

We denote the asymptotic variance of the RAL estimator $\hat{\mu}_r$ by Σ. An estimator for the asymptotic variance, $\hat{\Sigma}$, can be obtained using a sandwich variance estimator(Tsiatis, 2005). Tsiatis(2005, p.207) described how to construct this estimator:

$$\begin{aligned}\hat{\Sigma} &= \hat{E}\left[\left\{\frac{\partial m(Z,\hat{\mu})}{\partial\mu^T}\right\}\right]^{-1}\\ &\times \left[n^{-1}\sum_{i=1}^{n}g(C_i^*,G_{C_i^*}(Z_i),\hat{\gamma},\hat{\mu})g^T(C_i^*,G_{C^*i}(Z_i),\hat{\gamma},\hat{\mu})\right]\\ &\times \hat{E}\left[\left\{\frac{\partial m(Z,\hat{\mu})}{\partial\mu^T}\right\}\right]^{-T},\end{aligned}$$

where $\hat{E}$ denotes sample average,

$$\begin{aligned}&\frac{\partial m(Z,\hat{\mu})}{\partial\mu^T}=\frac{\partial q(C_i^*,G_{C^*i}(Z_i),\hat{\gamma},\hat{\mu})}{\partial\mu_r}\\ &g(C_i^*,G_{C^*i}(Z_i),\hat{\gamma},\hat{\mu})=q(C_i^*,G_{C_i^*}(Z_i),\hat{\gamma},\hat{\mu_r})-\hat{E}(qS_\gamma^T)\{\hat{E}(S_\gamma S_\gamma^T)\}^{-1}S_\gamma,\\ &q(C_i^*,G_{C^*i}(Z_i),\hat{\gamma},\hat{\mu})=(Y-u_r)\left\{\frac{I(U=t_r,\Delta=1)}{\omega_r(\hat{\gamma})}+\frac{I(U\leq t_r,\Delta=0)}{\omega_\infty(\hat{\gamma})}\right\}\\ &-L_0(C^*,G_{C^*}(Z),\hat{\mu}_r,\hat{\gamma},\hat{\xi_1},\hat{\xi_2}),\end{aligned}$$

where,

$$\hat{E}(qS_\gamma^T) = n^{-1}\sum_{i=1}^{n} q(C_i^*, G_{C^*i}(Z_i), \hat{\gamma}, \hat{\mu})S_\gamma,$$
$$\hat{E}(S_\gamma S_\gamma^T) = n^{-1}\sum_{i=1}^{n} S_\gamma S_\gamma^T,$$
$$S_\gamma = \frac{\partial log[\omega(C^*, G_{C^*}(Z), \gamma)]}{\partial\gamma} \mid_{\gamma=\hat{\gamma}} .$$

□

3.3 Properties of Proposed Estimator

3.3.1 Double Robustness of Proposed Estimator

The major advantages of the propose estimator, $\hat{\mu}_r$, are that it is not only efficient but also double robust. Double robustness means that it is a consistent estimator if either the model for $\widetilde{\lambda}_r, r = 1, \cdots, m-1$, or the posited models for both $P(C > c|X)$ and the set of conditional means, $E(Y_r^*|X), r = 1, \cdots, m$, and $E(Y_C^*|t_{j-1} < C \leq t_j), j = 1, \cdots, m-1$, are correctly specified. We firstly show $\hat{\mu}_r^*$ is double robust , then show $\hat{\mu}_r^*$ is equivalent to $\hat{\mu}_r$ in probability. Therefore, $\hat{\mu}_r$ is double robust.

Using standard asymptotic arguments, the estimator $\hat{\mu}_r^*$ will be consistent and asymptotically normal if we can show that

Preposition 2. *At each true value of μ_r, γ and ξ,*

$$E\left\{\left[\frac{I(U = t_r, \Delta = 1)}{\omega_r(\gamma^*)} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty(\gamma^*)}\right](Y - \mu_r^0) - L_{01} - L_{02}\right\} = 0, \quad (3.30)$$

where μ_r^0 *is true value of* μ_r *and*

$$\begin{aligned} L_{01} &= \sum_{k=1}^{r-1}\left\{\frac{I(U=t_r,\Delta=1)-\widetilde{\lambda}_r(\gamma^*)I(U\geq t_r)}{\widetilde{K}_r(\gamma^*)}\right\}p_r(\xi_{1r}^*)E(Y_r^*-\mu_r^0|X,\xi_{2r}^*), \\ &- \frac{I(U=t_r,\Delta=1)-\widetilde{\lambda}_r(\gamma^*)I(U\geq t_r)}{\widetilde{\omega}_r(\gamma^*)}E(Y_r^*-\mu_r^0|X,\xi_{1r}^*), \\ L_{02} &= \sum_{k=1}^{r-1}\left\{\frac{I(U=t_r,\Delta=1)-\widetilde{\lambda}(\gamma^*)_rI(U\geq t_r)}{\widetilde{K}_r(\gamma^*)}\right\}\sum_{j=1}^{k-1}p_j(\xi_{1r}^*)E(Y_C^*-\mu_r^0|t_{j-1}<C\leq t_j,X,\xi_{2r}^*). \end{aligned}$$

if either $\widetilde{\lambda}_r(r=1,\cdots,m-1)$ *or the posited models for both* $P(C>c|X)$ *and the set of conditional mean* $\{E(Y_r^*|X),r=1,\cdots,m$ *and* $E(Y_C^*|t_{j-1}<C\leq t_j),j=1,\cdots,m-1\}$ *are correctly specified.*

Before showing Preposition 2 is true, we first derive two facts.

Lemma 12.

$$\frac{I(C^*=r)}{\widetilde{\omega}_r}-1=-\sum_{k=1}^{r-1}\left\{\frac{I(C^*=k)-\widetilde{\lambda}_kI(C^*\geq k)}{\widetilde{K}_k}\right\}+\frac{I(C^*=r)-\widetilde{\lambda}_rI(C^*\geq r)}{\widetilde{\omega}_r}.$$

Proof:

We start proof with expressing the first term in the right hand as follows:

$$\begin{aligned} &\sum_{k=1}^{r-1}\left\{\frac{I(C^*=k)-\widetilde{\lambda}_kI(C^*\geq k)}{\widetilde{K}_k}\right\} \\ =\ &\sum_{k=1}^{r-1}\left\{\frac{I(C^*=k)}{\widetilde{K}_k}\right\}-\sum_{k=1}^{r-1}\left\{\frac{\widetilde{\lambda}_kI(C^*\geq k)}{\widetilde{K}_k}\right\}. \end{aligned} \tag{3.31}$$

Because of the discreteness of C^*, we can write the second term

$$\sum_{k=1}^{r-1}\left\{\frac{I(C^*=k)}{\widetilde{K}_k}\right\}=\frac{I(C^*\leq r-1)}{\widetilde{K}_{C^*}}. \tag{3.32}$$

By the definitions of $\widetilde{\lambda}_r$ and $\widetilde{K}_r$, we obtain that

$$\frac{\widetilde{\lambda}_r}{\widetilde{K}_r} = \frac{1}{\widetilde{K}_r} - \frac{1}{\widetilde{K}_{r-1}} \text{ and } \sum_{k=1}^{j} \frac{\widetilde{\lambda}_k}{\widetilde{K}_k} = 1 - \frac{1}{\widetilde{K}_j} \text{ any j} > k. \tag{3.33}$$

Therefore, we can write the second term in (3.31) as follows,

$$\begin{aligned} & \sum_{k=1}^{r-1} \left\{ \frac{\widetilde{\lambda}_k I(C^* \geq k)}{\widetilde{K}_k} \right\} = \sum_{k=1}^{r-1} \left\{ \frac{\widetilde{\lambda}_k I(C^* \geq r)}{\widetilde{K}_k} \right\} + \sum_{k=1}^{r-1} \left\{ \frac{\widetilde{\lambda}_k I(k \leq C^* \leq r-1)}{\widetilde{K}_k} \right\} \\ = \; & -I(C^* \geq r) \left(1 - \frac{1}{\widetilde{K}_{r-1}} \right) + \sum_{k=1}^{r-1} \sum_{j=1}^{r-1} \left\{ \frac{\widetilde{\lambda}_k I(C^* = j)}{\widetilde{K}_k} \right\}. \end{aligned}$$

Exchange summation order of the second term in the last line, we can have

$$\begin{aligned} = \; & -I(C^* \geq r) \left(1 - \frac{1}{\widetilde{K}_{r-1}} \right) + \sum_{j=1}^{r-1} \sum_{k=1}^{j} \left\{ \frac{\widetilde{\lambda}_k I(C^* = j)}{\widetilde{K}_k} \right\} \\ = \; & -I(C^* \geq r) \left(1 - \frac{1}{\widetilde{K}_{r-1}} \right) + I(C^* \leq r-1) \sum_{k=1}^{C^*} \frac{\widetilde{\lambda}_k}{\widetilde{K}_k} \\ = \; & -I(C^* \geq r) \left(1 - \frac{1}{\widetilde{K}_{r-1}} \right) - I(C^* \leq r-1) \left(1 - \frac{1}{\widetilde{K}_{C^*}} \right). \end{aligned} \tag{3.34}$$

The last equation holds because of (3.33). Then, plug (3.32) and (3.34) into (3.31), we can obtain the following equation,

$$\begin{aligned} & \sum_{k=1}^{r-1} \left\{ \frac{I(C^* = k) - \widetilde{\lambda}_k(\gamma) I(C^* \geq k)}{\widetilde{K}_k(\gamma)} \right\} - \frac{I(C^* = r) - \widetilde{\lambda}_r(\gamma) I(C^* \geq r)}{\widetilde{\omega}_r(\gamma)} \\ = \; & 1 - \frac{I(C^* > r-1)}{\widetilde{K}_{r-1}} - \frac{I(C^* = r)}{\widetilde{\omega}_r} + \frac{I(C^* > r-1)}{\widetilde{K}_{r-1}} \\ = \; & 1 - \frac{I(C^* = r)}{\widetilde{\omega}_r}. \end{aligned}$$

□

Using similar procedure we can prove the Lemma below.

Lemma 13.

$$\frac{I(C^*=\infty, C\le t_m)}{\omega_\infty} - 1 = -\sum_{k=1}^{m-1}\left\{\frac{I(C^*=r)-\widetilde{\lambda}_k I(U\ge t_k)}{\widetilde{K}_k}\right\} - \frac{I(C^*=m)}{\widetilde{K}_{m-1}}.$$

□

Proof for Theorem 2:

(1) If models for $\widetilde{\lambda}_r(X,\gamma)(r=1,\cdots,m-1)$ are all correctly specified, i.e., $\gamma_r^* = \gamma_r^0$.

We have proved $E(h) = E(E(h|Z)) = E(\varphi^F(Z) = 0$ at the true value of μ_r^0 and γ_r^0. In addition, we have known $E(L_0) = E(E(L_0)|Z) = 0$, because $L_0(\xi^*) \in \Lambda_2$. Thus, Equation (3.31) holds.

(2) If models for the conditional means $\{E(Y_r^*|X), r=1,\cdots,m,\ E(Y_C^*|t_{j-1} < C \le t_j), j=1,\cdots,m-1\}$ and $P(C>c|X)$ are correctly specified.

By the SUTVA assumption,

$$\begin{aligned}
&\frac{I(U=t_r,\Delta=1)}{\omega_r(\gamma_r^0)}(Y-\mu_r^0) = \frac{I(C^*=r)}{\omega_r(\gamma_r^0)}(Y_r^*-\mu_r^0)\\
&\text{Note that } I(C^*=r) = I(C^*=r) * I(C>t_r)\\
=\ & I(C>t_r)(Y_r^*-\mu_r^0) + \left\{\frac{I(C^*=r)}{\omega_r}-1\right\} I(C>t_r)(Y_r^*-\mu_r^0), \qquad (3.35)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{I(U\le t_r,\Delta=0)}{\omega_r(\gamma_r^0)}(Y-\mu_r^0) = \frac{I(C^*=\infty, C\le t_r)}{\omega_r(\gamma_r^0)}(Y_c^*-\mu_r^0)\\
=\ & I(C\le t_r)(Y_c^*-\mu_r^0) + \left\{\frac{I(C^*=\infty, C\le t_r)}{\omega_r}-1\right\} I(C\le t_r)(Y_r^*-\mu_r^0) \qquad (3.36)
\end{aligned}$$

Take the summation of the two equations above, (3.31) becomes

$$\begin{aligned}
&E\varphi^F(Z) + E\left[\left\{\frac{I(C^*=r)}{\omega_r}-1\right\} I(C>t_r)(Y_r^*-\mu_r^0) - L_{01}\right] \qquad (3.37)\\
+\ & E\left[\left\{\frac{I(C^*=\infty, C\le t_r)}{\omega_r}-1\right\} I(C\le t_r)(Y_r^*-\mu_r^0) - L_{02}\right]. \qquad (3.38)
\end{aligned}$$

The first term is equal to 0. We would show the other two terms are equal to 0. Plug result in Lemma 12 and specific forms of L_{01} and L_{02} into (3.37), (3.37) can be written as

$$\begin{aligned} & E \left[-\sum_{k=1}^{r-1} \left\{ \frac{I(C^* = r) - \widetilde{\lambda}_k(\gamma^*) I(C^* \geq k)}{\widetilde{K}_k(\gamma^*)} \right\} \right. \\ & \left. \times \left(\frac{-p_r(\xi_{1r}^0) E(Y_r^* - \mu_r^0 | X, \xi_{2r}^0)}{p_k(\xi_1^0)} + I(C > t_r)(Y_r^* - \mu_r^0) \right) \right] \\ - \quad & E \left[\frac{I(C^* = r) - \widetilde{\lambda}_r(\gamma^*) I(C^* \geq r)}{\widetilde{\omega}_r(\gamma^*)} \left\{ E(Y_r^* - \mu_r^0 | X, \xi_{2r}^0) - (Y_r^* - \mu_r^0) \right\} \right]. \end{aligned}$$

Conditional on full data Z, then conditional on C and X, the last equation becomes

$$\begin{aligned} = \quad & E \left[-\sum_{k=1}^{r-1} \left\{ \frac{P(C^* = r | Z) - \widetilde{\lambda}_k(\gamma^*) P(C^* \geq k | Z)}{\widetilde{K}_k(\gamma^*)} \right\} \right. \\ & \left. \times \left\{ \frac{-p_r(\xi_{1r}^0) E(Y_r^* - \mu_r^0 | X, \xi_{2r}^0)}{p_k(\xi_1^0)} + I(C > t_r) E(Y_r^* - \mu_r^0 | Z) \right\} \right] \\ & -E \left[\frac{P(C^* = r | Z) - \widetilde{\lambda}_r(\gamma^*) P(C^* \geq r) | Z}{\widetilde{\omega}_r(\gamma^*)} \left(E(Y_r^* - \mu_r^0 | X, \xi_{2r}^0) - E(Y_r^* - \mu_r^0 | C, X) \right) \right]. \end{aligned} \tag{3.39}$$

Because of assumption that $\{Y_r^* \perp C | X\}$ and assumption of coarsening at random, the second expectation in equation (3.39) is 0 and the first expectation is in equation

(3.39) equal to

$$
\begin{aligned}
& E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})I(C>t_k)}{\widetilde{K}_k(\gamma^*)}\right\}\right. \\
& \left.\left\{\frac{-p_r(\xi_{1r}^0)E(Y_r^*-\mu_r^0|X,\xi_{2r}^0)}{p_k(\xi_1^0)}+I(C>t_r)E(Y_r^*-\mu_r^0|X)\right\}\right] \\
= \; & E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})I(C>t_k)}{\widetilde{K}_k(\gamma^*)}\right\}\frac{-p_r(\xi_{1r}^0)E(Y_r^*-\mu_r^0|X,\xi_{2r}^0)}{p_k(\xi_1^0)}\right] \\
& +E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})}{\widetilde{K}_k(\gamma^*)}\right\}I(C>t_r)E(Y_r^*-\mu_r^0|X)\right].
\end{aligned}
$$

Conditional on X,

$$
\begin{aligned}
= \; & E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})}{\widetilde{K}_k(\gamma^*)}\right\}P(C>t_k|X)\frac{-p_r(\xi_{1r}^0)E(Y_r^*-\mu_r^0|X,\xi_{2r}^0)}{p_k(\xi_1^0)}\right] \\
& +E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})}{\widetilde{K}_k(\gamma^*)}\right\}P(C>t_r|X)(E(Y_r^*-\mu_r^0|X)\right] \\
= \; & E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})}{\widetilde{K}_k(\gamma^*)}\right\}\left\{-p_r(\xi_{1r}^0)E(Y_r^*-\mu_r^0|X,\xi_{2r}^0)\right\}\right] \\
& +E\left[-\sum_{k=1}^{r-1}\left\{\frac{(\widetilde{\omega}_r(\gamma^0)-\widetilde{\lambda}_k(\gamma^*)\widetilde{K}_{k-1})}{\widetilde{K}_k(\gamma^*)}\right\}\left\{p_r(\xi_{1r}^0)E(Y_r^*-\mu_r^0|X)\right\}\right] \\
= \; & 0.
\end{aligned}
$$

Similar procedure can show (3.38) is 0. Therefore, Preposition 2 has been proved and it provides the guarantee of double robustness of $\hat{\mu}_r^*$. □

In the next preposition, we would like to prove $\hat{\mu}_r$ is equivalent to $\hat{\mu}_r^*$ in probability.

Preposition 3. *Assuming coefficients in the propensity score models, $\hat{\gamma}$, and parameters in the augmentation part, $\hat{\xi}$, are estimated consistently. In addition, $n^{1/2}(\hat{\xi}-\xi_r^0)$ are assumed to be bounded in probability. Under the CAR assumption* (3.1) *and no unmeasured confounding assumption* (3.2)*, if coarsening models are correctly speci-*

fied, that is $\gamma^* = \gamma_r^0$*, then*

$$n^{1/2}(\hat{\mu}_r - \hat{\mu}_r^*) \xrightarrow{P} 0. \tag{3.40}$$

If coarsening models are not correctly specified, then

$$(\hat{\mu}_r - \hat{\mu}_r^*) \xrightarrow{P} 0. \tag{3.41}$$

Proof:

We note that the expansion of (3.29) about μ_r, but keeping $\hat{\gamma}$ fixed, yields

$$n^{1/2}(\hat{\mu}_r^* - \mu_r^0) = -\left[E\left\{\frac{\partial g(\mu_r^0)}{\partial \mu_r}\right\}\right]^{-1} n^{-1/2}\sum_{i=1}^{n}\left[h(\mu_r^0, \hat{\gamma}) + L_0(u_r^0, \xi^*)\right] + o_p(1). \tag{3.42}$$

whereas, the expansion of (3.28) yields

$$n^{1/2}(\hat{\mu}_r - \mu_r^0) = -\left[E\left\{\frac{\partial g(\mu_r^0)}{\partial \mu_r}\right\}\right]^{-1} n^{-1/2}\sum_{i=1}^{n}\left[h(\mu_r^0, \hat{\gamma}) + L_0(\mu_r^0, \hat{\xi})\right] + o_p(1), \tag{3.43}$$

where

$$\begin{aligned}
&\frac{\partial g(\mu_r^0)}{\partial \mu_r} = \left\{\left[\frac{I(U = t_r, \Delta = 1)}{\omega_r(\hat{\gamma})} + \frac{I(U \le t_r, \Delta = 0)}{\omega_\infty(\hat{\gamma})}\right] - L_{01}' - L_{02}'\right\},\\
&L_{01}' = \sum_{k=1}^{r-1}\left\{\frac{I(U = t_r, \Delta = 1) - \widetilde{\lambda}_r(\gamma^*)I(U \ge t_r)}{\widetilde{K}_r(\gamma^*)}\right\} p_r(\xi_{1r}^*)\\
&-\frac{I(U = t_r, \Delta = 1) - \widetilde{\lambda}_r(\gamma^*)I(U \ge t_r)}{\widetilde{\omega}_r(\gamma^*)},\\
&L_{02}' = \sum_{k=1}^{r-1}\left\{\frac{I(U = t_r, \Delta = 1) - \widetilde{\lambda}(\gamma^*)_r I(U \ge t_r)}{\widetilde{K}_r(\gamma^*)}\right\} P(C \le t_k | X, \xi_{1r}^*).
\end{aligned}$$

(1)Assume coarsening models are correctly specified, that is $\gamma^* = \gamma^0$. Thus,

$$-\left[E\left\{\frac{\partial g(\mu_r^0)}{\partial \mu_r}\right\}\right] = 1.$$

Equation (3.42) and (3.43) became(3.42) and (3.43) as follows,

$$n^{1/2}(\hat{\mu}_r^* - \mu_r) = n^{-1/2}\sum_{i=1}^{n}\left[h(\mu_r^0, \hat{\gamma}) + L_2(\mu_r^0, \xi^*)\right] + o_p(1), \tag{3.44}$$

$$n^{1/2}(\hat{\mu}_r^* - \mu_r) = n^{-1/2}\sum_{i=1}^{n}\left[h(\mu_r^0, \hat{\gamma}) + L_2(\mu_r^0, \hat{\xi})\right] + o_p(1). \tag{3.45}$$

Taking difference between (3.44) and (3.45), we obtain that

$$n^{1/2}(\hat{\mu}_r^* - \hat{\mu}_r) = \left[n^{-1/2}\sum_{i=1}^{n} L_2(\mu_r^0, \hat{\gamma}, \xi^*) - n^{-1/2}\sum_{i=1}^{n} L_2(\mu_r^0, \hat{\gamma}, \hat{\xi})\right] + o_p(1). \tag{3.46}$$

The proof is complete if we can show the term in the square brackets above converges in probability to zero.

Expand $n^{-1/2}\sum_{i=1}^{n} L_2(\mu_r^0, \hat{\gamma}, \hat{\xi})$ about ξ^*, we have

$$n^{-1/2}\sum_{i=1}^{n} L_2(\mu_r^0, \hat{\gamma}, \xi^*) + n^{-1}\sum_{i=1}^{n}\frac{\partial L_2(\mu_r^{in}, \hat{\gamma}, \xi^{in})}{\partial \xi} n^{1/2}(\hat{\xi} - \xi^*), \tag{3.47}$$

where ξ^{in} are intermediate values between ξ^* and $\hat{\xi}$. Plug (3.46) into (3.47), we obtain

$$n^{1/2}(\hat{\mu}_r^* - \hat{\mu}_r) = n^{-1}\sum_{i=1}^{n}\frac{\partial L_2(\mu_r^0, \hat{\gamma}, \xi^{in})}{\partial \xi} n^{1/2}(\hat{\xi} - \xi^*) + o_p(1). \tag{3.48}$$

Since $\hat{\gamma} \xrightarrow{P} \gamma^0$, and $\hat{\xi} \xrightarrow{P} \xi^*$, then under suitable regularity conditions, the sample average in equation (3.48) is

$$n^{-1}\sum_{i=1}^{n}\frac{\partial L_2(\mu_r^0, \hat{\gamma}, \xi^{in})}{\partial \xi} \to E\frac{\partial L_2(\mu_r^0, \gamma^0, \xi^0)}{\partial \xi} = 0. \tag{3.49}$$

It is not difficult to show (3.49) is 0 with correct specification of coarsening models. Since $n^{1/2}(\hat{\xi} - \xi_r^0)$ are bounded in probability, then a simple application of Slutsky's theorem can be applied to show that (3.48) converges in probability to zero. Thus,

we proved (3.40) is correct.

(2)Coarsening models are not correctly specified, that is $\gamma^* \neq \gamma^0$. At this point,

$$-\left[E\left\{\frac{\partial g(\mu_r^0)}{\partial \mu_r}\right\}\right] = a, a \neq 1.$$

Following the similar procedure to obtain equation (3.48), instead, we have,

$$(\hat{\mu}_r^* - \hat{\mu}_r) = \frac{1}{a} n^{-1} \sum_{i=1}^{n} \frac{\partial L_2(\mu_r^0, \hat{\gamma}, \xi^{in})}{\partial \xi} (\hat{\xi} - \xi^*) + o_p(1). \tag{3.50}$$

Hence, we still have

$$n^{-1} \sum_{i=1}^{n} \frac{\partial L_2(\mu_r^0, \hat{\gamma}, \xi^{in})}{\partial \xi} \to E \frac{\partial L_2(\mu_r^0, \gamma^0, \xi^0)}{\partial \xi}. \tag{3.51}$$

Although it is not equal to 0 any more, however, it is still bounded. Note that $\hat{\xi} \to \xi^*$, Slutsky's theorem can be applied to show (3.50) converges in probability to zero. □

Preprosition2 and Preposition 3 together guarantee the double robustness of the proposed estimator. The consistency of estimator proposed by Johnson and Tsiatis (2004) relied on correct specification of cause-specific hazard model only. However, our approach can provide a protection against misspecified models.

3.3.2 Efficiency

The procedure described in the previous sections is aimed to obtain the optimal estimator which has the smallest asymptotic variance among the class of all influence functions of RAL estimators for μ_r given by

$$\{\varphi(Z)\} = (Y - \mu_r) \left\{ \frac{I(U = t_r, \Delta = 1)}{\omega_k} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty} \right\} + \Lambda_2, \tag{3.52}$$

by assuming we have specified all the models in Λ_2. After identifying the influence function on observed coarsening data, we posited working models for propensity score and outcomes regressions, but these posited models may or may not be correct and the parameters are estimated by MLE method. Therefore, if the posited models are all correctly specified, then

$$P(C > c|X, \hat{\xi}_1), E(Y_r^*|X, \hat{\xi}_2) \text{ and } E(Y_C^*|X, \hat{\xi}_2), r = 1, \cdots, m,$$

will be consistent estimators of

$$P(C > c|X, \xi_{10}), E(Y_r^*|X, \xi_{20}) \text{ and } E(Y_C^*|X, \xi_{20}), r = 1, \cdots, m.$$

In this case, $L_0(U, \Delta, X, Y, \hat{\gamma}, \hat{\xi}_1^*, \hat{\xi}_2^*, \mu_r^0)$ will converge to

$$\prod \left[(Y - u_r) \left\{ \frac{I(U = t_r, \Delta = 1)}{\omega_r(\hat{\gamma})} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty(\hat{\gamma})} \right\} |\Lambda_2 \right],$$

the projection onto space Λ_2. Thus, the corresponding influence function is

$$\begin{aligned} &(Y - u_r) \left\{ \frac{I(U = t_r, \Delta = 1)}{\omega_r(\gamma)} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty(\gamma)} \right\} \\ &- \prod \left[(Y - u_r) \left\{ \frac{I(U = t_r, \Delta = 1)}{\omega_r(\gamma)} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty(\gamma)} \right\} |\Lambda_2 \right], \end{aligned}$$

which is the most efficient one, that is, having the smallest asymptotic variance, among the class of (3.52).

In practice, there is no guarantee for the correct specification of models desired and it is very likely that we can not feasibly construct the most efficient estimator. Nonetheless, the study of efficiency will aid us in constructing the more efficient estimators even if we are not able to derive the most efficient one. Generally, the attempt to estimate projection of h onto Λ_2 by positing working models leads to

more efficient estimators than the Inverse Probability Weighting (IPW) estimators (Tsiatis, 2006).

The influence function of IPW estimator proposed by Johnson and Tsiatis (2004) belongs to the class of all influence functions of RAL estimators for μ_r given by

$$\{\varphi(Z)\} = (Y - \mu_r)\left\{\frac{I(U = t_r, \Delta = 1)}{\omega_k} + \frac{I(U \leq t_r, \Delta = 0)}{\omega_\infty}\right\} + \Lambda_2,$$

which is a special case by choosing $L_2 = 0$. This implies that our estimator is more efficient than Johnson and Tsiatis's estimator. Therefore, we fulfill the task of improving efficiency talked at the beginning of this chapter.

3.4 Simulation Study

In this section, we conduct several simulation studies to compare the performance of the estimator we proposed with the estimator Johnson and Tsiatis proposed in 2004. We focus on two properties: efficiency and double robustness.

We duplicate the scenarios in Johnson and Tsiatis (2004), but here we assume that patients are assigned to one of a finite number of treatment duration policies at value $t = (t_1, \cdots, t_3)$. For simplicity, we only consider time-independent covariates.

In the first simulation, let $t = (15, 25, 30)$. We consider a single covariate, X, following a standard normal distribution, for each individual. We then generate a treatment-censoring random variable $\mathcal{C}$ as an exponential $\{p(X)\}$ random variable, where

$$p(X) = 0.01 \exp(\beta X),$$

$\beta = -2$. We firstly generate potential outcomes $Y_r^*, r = 1, \cdots, m$ for each patient if the patient had completed treatment duration $t_r, r = 1, 2, 3$, then potential outcomes

Y_C^*. We assume the following models hold:

$$Y_r^* \sim N(u_{1r}, 1), \text{ where } u_{1r} = \eta_{or}^1 + \eta_{1r}^1 X,$$

and

$$\text{For } t_r < C \leq t_{r+1}, Y_C^* \sim N(u_{2r}, 1), \text{ where } u_{2r} = \eta_{or}^2 + \eta_{1r}^2 X.$$

With full data generated above, it is convenient to estimate true value of $\mu_r, r = 1, 2, 3$, and we approximate values by simulation.

The treatment duration data are simulated according to the following algorithm, which is similar to the simulation scheme in Johnson and Tsiatis(2004): Start by letting $r = 1$.

1. if $C < t_r$, then define $U = C$ and $\Delta = 0$.

2. For $C \geq t_r$, generate a Bernoulli random variable Q_r, the indicator variable for stopping treatment at time t_r, with probability $\widetilde{\lambda}_r(X)$ where

$$\text{logit}(\widetilde{\lambda}_r(X)) = \gamma_{0r} + \gamma_1 X.$$

3. If $Q_r = 1$, then assign $U = t_r$ and $\Delta = 1$; if $Q_r = 0$ and $r < m$, then increment r to r+1 and goto step 1.

Results for the simulation studies are presented in the Table 3.1. When propensity score models, that is, cause-specific hazard function $\widetilde{\lambda}_r(X)$ are correctly specified, our AIPW estimator showed improved efficiency over IPW estimator, consistent with theoretical derivative, no matter whether the conditional mean models and conditional distribution are correctly specified. If conditional mean models and conditional distribution in augmentation part are correctly specified, more efficiency would be gained. If propensity score models are not correctly specified, AIPW estimator is necessarily more efficient than IPW estimator, although results in the Table 3.1 showed a slight

improvement on asymptotic variance. However, when propensity score models are not correctly specified, IPW estimators have a relatively large bias while AIPW estimator showed double robustness, having a slight bias.

3.5 Analysis of the ESPRIT Infusion Trial

The ESPRIT (Enhanced Suppression of the Platelet IIb/IIa receptor with Integrilin Therapy) trial, which motivated this article, targeted patients with coronary artery disease scheduled to undergo percutaneous coronary intervention (PCI) with stent implantation in a native coronary artery. The main objective of ESPRIT was to compare eptifibatide (Integrilin) therapy to placebo on the basis of the composite binary endpoint of death, myocardial infarction (MI), or urgent target vessel revascularization within 30 days. The study enrolled 2064 eligible patients who were randomized to either study drug (1040) or palcebo (1024) regimen. The experimental treatment regimen consisted of an eptifibatide bolus and a continuous eptifibatide infusion for 18-24 hours, with a similar regimen for the placebo group. This study protocol required that patients experiencing serious complications, such as abrupt closure, immediately discontinue the infusion process. We identified any complication as treatment-terminating events. The main study report suggested that drug regimen in the study is superior to placebo.

We apply the methods developed for improving efficiency and robustness described in the previous section to data from patients in the ESPRIT trial who receive eptifibatide. The observed infusion length are discretized by taking t_j to be the midpoint of five intervals I_j, namely $I_j = \{(t_{j-1} + t_j)/2, (t_j + t_{j+1})/2\}$ for $t = (t_1, t_2, t_3, t_4.t_5) = (16, 18, 20, 22, 24)$, and we redefine the random variable $U_i = t_j$ for any patient, where $U_i \in I_j$ and $\Delta = 1$. We also included the following potential confounders in our analysis: diabetes(0/1), percutaneous transluminal coronary

Table 3.1: Simulation Summary

γ_1	t	True Value	IPW (SE.)	MC SD.	Bias	AIPW(SE.)	MC SD	Bias	Ratio [1]	Effi. Gain [2](%)
Correct	1	-1.278	-1.278 (0.159)	0.179	0.001	-1.278 (0.105)	0.107	0.001	0.436	56.4
all models	2	-2.076	-2.076 (0.129)	0.138	0.002	-2.076 (0.093)	0.094	0.002	0.519	47.1
	3	-4.007	-4.006 (0.069)	0.069	0.000	-4.006 (0.062)	0.063	0.000	0.807	19.3
Correct PS	1	-1.279	-1.278 (0.159)	0.179	0.001	-1.305 (0.138)	0.157	0.006	0.753	24.7
	2	-2.077	-2.076 (0.129)	0.138	0.002	-2.063 (0.106)	0.113	0.010	0.675	31.5
	3	-4.006	-4.006 (0.069)	0.066	0.000	-4.006 (0.064)	0.065	0.001	0.860	14.0
Correct	1	-1.279	-1.495 (0.240)	0.264	0.216	-1.280 (0.223)	0.261	0.002	0.863	13.7
Conditional	2	-2.077	-2.294 (0.207)	0.219	0.218	-2.078 (0.195)	0.242	0.001	0.887	11.3
Expectation	3	-4.010	-4.050 (0.144)	0.144	0.039	-4.013 (0.133)	0.136	0.002	0.853	14.7

[1]Ratio=AVAR(AIPW)/AVAR(IPW).
[2]Effi. Gain=[AVAR(IPW)-AVAR(AIPW)]/AVAR(IPW).
n=1000, 1000 Monte Carlo datasets.
All models are correctly specified.
γ_1 is the coefficient in the propensity score.

angioplgsty (0/1), angina (0/1), heparin (0/1) and weight, in kilograms, which are identified in the earlier paper (Johnson and Tsiastis, 2004).

In Table 3.5, we present two estimators $\hat{\mu}$ and $\mu*$, obtained from the JT's method and the method proposed in this paper, respectively. Two estimators gave similar results both in terms of the estimates and the standard errors, however, our estimator gained more efficiency than the JT's estimator.

Table 3.2: Analysis of the ESPRIT trial data

t_r	IPW (SE)	New (SE)
16	0.040 (0.016)	0.042 (0.015)
18	0.066 (0.010)	0.067 (0.010)
20	0.078 (0.017)	0.077 (0.016)
22	0.071 (0.024)	0.074 (0.023)
24	0.121 (0.035)	0.131 (0.032)

3.6 Discussion

We have identified a class of augmented inverse probability weighted estimators and derived the locally efficient and doubly robust estimator for the mean outcome for a particular treatment policy, where treatment is censored by terminating events. Simulation study showed that the proposed estimator is more efficient than the IPW or JT estimator whenever propensity score is correctly specified and has smaller bias than the IPW or JT estimator due to the protection provided by double robustness.

When all the models are correctly specified and parameters in the models are estimated consistently, our estimator is the most efficient one among the class of augmented inverse probability weighted estimators. In spite of this fact, it is always true in practice. However, generally, as Tsiatis (2005) mentioned, the attempt to estimate augmentation part by positing a working model to estimate unknown nuisance parameters often leads to a more efficient estimators even if the model was incorrect.

An interesting findings were found when we conducted the simulation studies, which is more efficiency gained with more strong confounding in the propensity score models. We have not found out that there are any literature specifically reporting this phenomenon or discussing it in theory. Whether it is theoretically supported or just a coincidence corresponding to a special setting is not known yet. Therefore, theory and application behind this phenomenon is still open to discussion. This topic can potentially be part of future research.

An limitation of this method is that it can only handle the case that duration can take on only a finite number of values, $t_1, \cdots, t_m$. In truth, treatment duration in infusion studies is a continuous random variables. Hence, treating treatment duration as a continuous variable is more realistic. This is one of future work.

In addition, Tan(2006) and Cao, et al.(2009) have developed estimators which is more efficient than AIPW estimator when outcome regression are misspecified based on large sample theory. Therefore, extension of their work to our setting is another part of future work.

Chapter 4

Nonparametric Method Using Boosting Algorithm To Estimate Mean Potential Outcomes

We proposed an efficient doubly robust semiparametric estimator for the mean of potential outcome in the previous chapters and we also investigated their properties in theory and simulations studies. We found that semiparametric estimators perform poorly in small samples and working models of outcome regressions are not correctly specified. In addition, when confounding covariates are more than observations, we would fail to construct semiparametric estimators unless we limit the number of covariates in the model. However, modeling selection could exclude important confounding and decrease efficiency. Therefore, in this topic we propose a nonparametric method to estimate mean potential outcomes as an alternative analysis in those cases where semiparametric method did not perform very well.

4.1 Introduction

Because "estimating the causal effects of treatments in a non-randomized observational study may be viewed as a missing data problem" (Rubin, 1983, p.41), estimating mean outcome in the presence of missing values is the primary interest in this chapter. Without loss of generalization, we assume that the complete data are realizations of random variables

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

A random sample of incomplete data is

$$(X_1, Y_1, \delta_1), \cdots, (X_n, Y_n, \delta_n),$$

where all the X are observed and $\delta = 0$ if Y is missing, otherwise $\delta = 1$. We will estimate $E(Y)$ based on observational data by modeling $E(Y|X) = f(x)$, where $f(X)$ is left unspecified.

4.1.1 Nonparametric Regression

We firstly briefly introduce the form, objective of nonparametric regression and common methods to estimate functions in the nonparametric regression. The general nonparametric regression model fits the model

$$y_i = f(x_i) + \varepsilon_i,$$

where x_i is a vector of predictors for the ith of n observations; the error ε_i are assumed to be normally and independently distributed with mean 0 and constant variance σ^2. The function f is left unspecified. The object of nonparametric regression is to estimate function $f(\cdot)$ directly, rather than to estimate parameters. Most methods of nonparametric regression implicitly assume that $f(\cdot)$ is a smooth, continuous function.

The advantage of nonparametric models is that they make fewer assumptions. We typically made assumptions in parametric regression models, but not in nonparametric models, include normality, linearity, and homoscedasticity (constant variance). If these parametric assumptions are, in fact, true (at least approximately), then estimates from the parametric model are more precise that those from a nonparametric model. If the assumptions are not true, the nonparametric model is demonstrably better(Wright, 2010).

Another advantage of nonparametric regression is that it can handle the cases where there are many predictors in a more flexible way. Several more restrictive models have been developed. One such model is the additive regression model,

$$y_i = a + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}) + \varepsilon_i.$$

Additive regression models are an alternative to unconstrained nonparametric regression with several predictors. It provides more flexibility to models especially when there is high dimension curse.

There are three common estimating methods of nonparametric regression, which are kernel estimation, local-polynomial regression (which is a generalization of kernel estimation), and smoothing splines. There is a large literature on nonparametric regression analysis, both in scientific journals and in texts. (For more extensive introductions to this subject, see in particular, Bowman and Azzalini (1997), Fox (2000), Hastie, Tibshirani and Hastie(2001), Tibshirani and Friedman(1990), and Simonoff(1996)).

Recently, new algorithms have been proposed to improve the prediction ability of nonparametric regression methods. Bagging and boosting, two well-known method in the machine learning, have been considered to improve the performance of nonparametric regression. Bagging and boosting are both ensemble methods (a weighted

average of predictions of individual classifiers) for improving unstable estimation or classification schemes. In this paper, we adopt boosting algorithm nonparametric regression rather than bagging. There are two main reasons we prefer boosting over bagging. To begin with, the main disadvantage of bagging is hard to interpret. Moreover, empirical studies have shown boosting method has appreciably smaller misclassification rates than bagging (Borra and Ciaccio, 2002). Boosting algorithms have been reconsider in terms of gradient descendent algorithms on several potential functions and theoretical and empirical results about boosting for regression problem has received focus and investigation. In the following section, we give an overview of boosting algorithms.

4.1.2 Boosting Algorithms

Boosting, as ensemble methods, is one of the most successful and practical methods in machine learning. Over the past few years, it has been connected to statistical fields and proved to be a successful tool to improve prediction capability of classification and nonparametric regression methods. Much recent work has been on the "AdaBoost" boosting algorithm and its extensions. We briefly overview about origin of boosting, connection to statistics, general algorithm, basic elements in algorithm and its application to improve performance in the nonparametric regression.

Boosting began within the field of machine learning during the 1990s. It rooted from a theoretical framework called probably approximately correct (PAC) learning model(Freund and Schapire, 1999). Kearns and Valiant (1994) were the first to pose the question of whether a weak learning algorithm which performs just slightly better than random guessing in the PAC model can be boosted into an arbitrarily accurate strong learning algorithm.

Later, boosting has been applied to classification problem in statistics and its practical aspects have been tried on substantial datasets empirically proved to im-

pressively improve performance for statistical models. Early boosting algorithms have some difficulties in practice. The AdaBoost algorithm, introduced in 1995 by Freund and Schapire (1997), solved many difficulties of the earlier boosting algorithms. Various versions of AdaBoost have proven to be very competitive in terms of improving prediction ability in many application. Practically, AdaBoost has many advantages. First of all, it is fast, simple and easy to program. Second, it has no principle parameters to tune, except for the number of iteration. It does not requires prior knowledge about the weak learner (we explain this concept in the next section) and so can be flexibly combined with any method. Finally, there is a set of theoretical guarantees given sufficient data and a weak learner that can reliably provide only moderately accurate weak hypotheses. Thus, we can instead focus on finding weak learning algorithms that only need to be better than random, instead of trying to design a learning algorithm that is accurate over the entire space. In 1998 AdaBoost algorithm has been observed that it can be viewed as a functional gradient descent algorithm in function space by Breiman (1998). Moreover, Friedman et al. (2000) linked AdaBoost and other boosting algorithms to the framework of statistical estimation and additive basis expansion. Here, additive does not mean a model fit which is additive in covariates, but refer to the fact that boosting is an additive combination of simple function estimators. Their work built a foundation of application of boosting to statistics in a wider fields other than just classification.

Boosting methods have been originally recognized as ensemble methods. The essence of ensemble scheme is multiple prediction and aggregation. Specifically, ensemble schemes construct multiple function estimates or predictions from re-weighted data and use a linear combination for producing the final, aggregated estimator or prediction (Buhlmann and Hothorn, 2007). A general ensemble scheme is described below:

(1) We specify a base procedure (also called as base learner, or weak learner)

which constructs a function estimate, based on some data $(X_1, Y_1), \cdots, (X_n, Y_n)$:

$$(X_1, Y_1), \cdots, (X_n, Y_n) \overset{\text{base procedure}}{\longrightarrow} \hat{g}(\cdot).$$

(2) Generating an ensemble from the base procedures, i.e., an ensemble of function estimates or predictions,

$$\text{re-weighted data 1} \overset{\text{base procedure}}{\longrightarrow} \hat{g}^1(\cdot)$$

$$\text{re-weighted data 2} \overset{\text{base procedure}}{\longrightarrow} \hat{g}^2(\cdot)$$

$$\text{re-weighted data M} \overset{\text{base procedure}}{\longrightarrow} \hat{g}^M(\cdot)$$

$$\text{aggregation} : \hat{f}_A = \sum_{m=1}^{M} \alpha_m \hat{g}^m(\cdot)$$

where re-weighted data means that every of n sample points has been assigned individual data weights. Different choice of weights $\{\alpha_m\}_{m=1}^M$ results in different ensemble schemes. When data weights in iteration m depend on the results from the previous iteration $m-1$, ensemble schemes are characterized as sequential ensemble schemes. Most boosting methods are sequential ensemble schemes.

However, the scheme above is too general to be of any use. Every boosting algorithm requires the specification of a base procedure, which conducts a function estimate based on the data. Considering some structural properties of the boosting algorithm, estimate usually is more interesting as it allows for "better interpretation of the resulting model" (Buhlmann and Hothorn, 2007). Some important choices include componentwise linear least squares of linear models, componentwise smoothing spline for additive models and regression tree. The principle of choosing base procedure is low variance at the price of larger estimation bias (Buhlmann and Hothorn, 2007).

As mentioned before, rather than being viewed as an ensemble methods, boosting algorithms can also be seen as functional gradient descent(FGD) techniques. After Breiman(1998) showed that the AdaBoost algorithm can be represented as a steepest descent algorithm in function space, Friedman et al.(2000, 2001) then developed a more general statistical framework and directly interpreted boosting as a method for function estimation.

The goal of functional gradient descent method is to estimate a function by minimizing an expected loss

$$E[\rho(Y, f(X))],$$

where $\rho(\cdot, \cdot)$ is a loss function which is usually assumed to be differentiable and convex with respect to the second argument, based on data $(X_1, Y_1), \cdots, (X_n, Y_n)$. Friedman(2001) gave a generic FGD algorithm.

(1) Initialize $\hat{f}^0(\cdot)$ with an offset value. Common choices are

$$\hat{f}^0(\cdot) = \text{argmin} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, c).$$

Set $m = 0$.

(2) Increase m by 1. Compute the negative gradient $\frac{\partial \rho(Y,f)}{\partial f}$ and evaluate at $\hat{f}^{m-1}(X_i)$:

$$U_i = -\frac{\partial \rho(Y_i, f)}{\partial f}|_{f=\hat{f}^{m-1}(X_i)}, i = 1, \cdots, n.$$

(3) Fit the negative gradient vector $U_1, \cdots, U_n$ to $X_1, \cdots, X_n$ by the real-valued base procedure

$$(X_i, U_i)_{i=1}^n \overset{\text{base procedure}}{\longrightarrow} \hat{g}^m(\cdot).$$

(4) Update $\hat{f}^m(\cdot) = \hat{f}^{m-1}(\cdot) + \upsilon \hat{g}^m(\cdot)$, where υ is a steplength factor.

(5) Iterate steps 2 to 4 until $m = m_{stop}$ for some stopping iteration m_{stop}.

We need to determine two parameters in the algorithm above, m_{stop} and υ. The

stopping iteration can be decided via cross-validation or some information criterion, such as corrected AIC criterion. The choice of the step-length factor υ is of minor importance, as long as it is "sufficiently small"(e.g., $\upsilon = 0.1$).

Another key step to use the algorithm above is to define the form of loss function $\rho(y, f)$. Different loss functions result in different version of boosting algorithms. The most popular loss functions include the following choices: (1) For binary response $Y \in \{0, 1\}$, loss function $\rho(y, f)$ is usually chosen as

$$\exp(-(2y-1)f),$$

or

$$\log_2(1 + \exp(-(2y-1)f));$$

(2) For continuous response $Y \in R$, loss function $\rho(y, f)$ is usually chosen as

$$\frac{1}{2}|y-f|^2 \tag{4.1}$$

The choice of the last loss function defines the L_2Boosting, which is the simplest and perhaps most instructive boosting algorithm. L_2Boost algorithm is very useful for regression, in particular in presence of many covariates. We describe the specific L_2Boosting as below:

Step 1 (Initialization). Given data $\{(Y_i, X_i) : i = 1, \cdots, n\}$. fit a real-valued learner,

$$\hat{F}_0(x) = h(x; \hat{\theta}, x),$$

where $\theta = \text{argmin} \sum_{i=1}^{n} (Y_i - h(X_i; \theta)^2)$. Set m=0;

Step 2 (Projection of gradient to learner). Compute the negative gradient vector,

$$U_i = -\frac{\partial \rho(Y_i, F)}{\partial F} \mid_{F=\hat{F}_m(X_i)} = Y_i - \hat{F}_m(X_i), i = 1, \cdots, n,$$

and fit the real-valued learner to the gradient vector,

$$\hat{f}_{m+1}(x) = h(x; \hat{\theta}_U, X),$$

where $\hat{\theta}_U, X = \arg\min \sum_{i=1}^{n} (U_i - h(X_i; \theta)^2)$. Update

$$\hat{F}_{m+1}(\cdot) = \hat{F}_m(\cdot) + \hat{f}_{m+1}(\cdot).$$

Step 3 (iteration). Increase iteration index m by 1, and repeat step 2.

The algorithm above with different base learners results in different versions of L_2boosting algorithm. Using componentwise linear least squares results in GamBoost algorithm. GamBoost is especially useful for high-dimensional data. Tutz and Binder(2008) conducted extensive simulation to compare GamBoost algorithm with other methods and showed it was favorably for fitting generalized additive models when there were many predictors. Using regression tree as base learner results in another popular boosting algorithm: BlackBoost algorithm. BlackBoost was developed by Jerome Friedman of Stanford University. It has the advantage to be invariant under monotone transformations of variables and we do not need to search for good data transformations. Moreover, regression trees can handle continues or categorical covariates in a unified way.

In this paper, we adopt blackBoost algorithm rather than Gamboost algorithm due to the following reasons: (1) Although Gamboost is a popular algorithm to handle high dimensional covariates in nonparametric regression, its assumption of additive effects of covariates on outcomes may be not consistent of true model. Instead, black-

Boost algorithm brings more freedom to the form of model. (2)As the most popular boosting algorithm in the machine learning community, blackboost algorithm has the advantage to be invariant under monotone transformations of predictor variables, i.e., we do not need to search for good data transformations. We introduce decision tree in the next section.

4.1.3 Decision Tree

Decision tree is a common method used in machine learning community. The goal is to create a binary tree to predict the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; Each leaf represents a value of the predicted target variable given the values of the input variables represented by the path from the root to the leaf. If the target variable is continuous, then a regression tree is generated. If the target variable is categorical, then a classification tree is generated. The following Figure 4.1.3 will help understand what a tree looks like.

How do we develop a decision tree from any given dataset? Recursive partitioning (RP) was developed to address the problem of decision tree construction. There are many different versions of recursive partitioning available and each has its own unique details, such as unbiased RP and model-based RP, etc. However, the overall methodology is consistently the same regardless of the exact implementation. Figure 4.1.3 illustrates the general methodology involved in Recursive Partitioning. There are several basic elements related to recursive partitioning in the Figure 4.1.3, which are partitioning the training set, deciding which question to ask(splitting criteria) and selecting stopping criteria.

The first concept is partitioning the training set recursively, i.e., splitting the dataset. Starting from the original entire sample space, we consider a question that we ask in order to direct the user down the appropriate path. For simplicity, lets consider

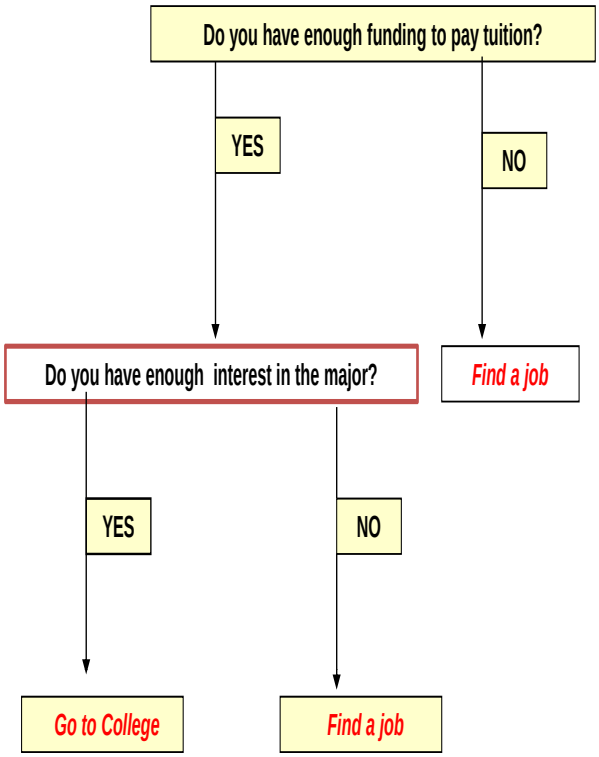


Figure 4.1: A simple decision tree for making decision on going to college or finding a job

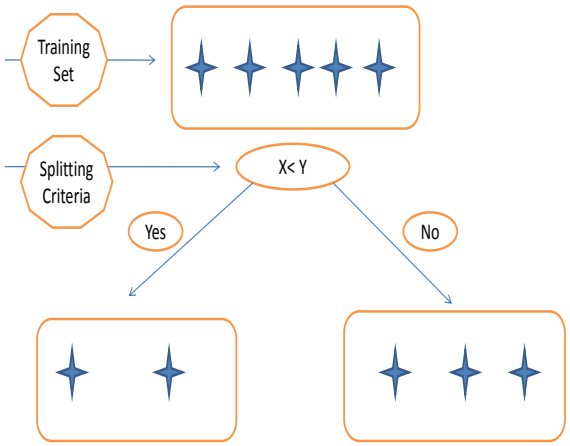


Figure 4.2: Recursive Partitioning

that each potential question can have a true or false answer, thus, any particular node will have at most two paths leading from it to the next nodes in the path. Every possible value of every possible feature within the training set represents a potential split that could be done. For example, for the data collected in the training set as $\{Y_i, i = 1, \cdots, n\}$ where all the Y range from 0 to 10, we can split the data according to whether $Y = 5$. The result is that we will be able to go down the right path or the left path based upon the data and we will effectively split the data at each node into two independent groups − this is partitioning. Once we have two new nodes linked to a previous node, we can repeat the process for each node independently using only the observations present in that node − this is the recursive step.

The second concept is related to how we choose a question to ask, i.e., splitting criteria. There are many criteria to decide how to split dataset. We could ask any question of the above form for every possible value of every possible feature within our

training set. In the previous example, the question whether $Y = 5$ can be replaced by any other forms such as $Y = 2$, $Y = 3$, or $Y < 5$, etc. For the purpose of choosing the most appropriate question, we have to develop a measure that we can use to decide which split is the best possible split from our choices. For example, one of the criterions for the splitting can be the minimum of within terminal node sum of squared errors (RSS_i) (Chu, Singfat (2001)). Breiman et al. (1984) and Shih (1999) have established criteria for binary splitting, and multiple splitting is available by utilizing the work of OBrien (2004). Hothorn et al. (2006) proposed a linear statistic which induced a two-sample statistic measuring the discrepancy between the samples $\{Y_i|X_i \in A; i = 1, \cdots, n\}$ and $\{Y_i|X_i \overline{\in} A; i = 1, \cdots, n\}$, where A are all possible subsets of the sample space for unbiased recursive partitioning.

Finally, we must have a stopping criteria. If we were to allow the splitting process to continue until each leaf only had 1 observation, we would have a perfect tree. However, such resulting decision tree is no much meaning for the purpose of prediction, because this sort of tree is over fit for the training data and will not perform well on new data. To avoid this situation, we need to determine a stopping criteria to halt the recursive partitioning process. stopping criteria can have many different forms, including: (1)A maximum number of nodes in the tree. Once this maximum is reached, the process is halted. (2) A minimum number of observations in a particular node. Once the number of observations in a node is less than or equal to a minimum value, we will not continue partitioning of that node, and that node becomes a leaf. (3)A threshold of a fit statistics, such as the sum of squared errors of all terminal node $RSS_i, i = 1, \cdots, m$ where m is the number of total nodes or predictive deviance.

An original tree may be too big to use, so pruning a tree is necessary in order to find a good tree. Cross-validation method is a more successful and widely-used approach. We randomly divide our data into a training set and a testing set, for example, 50 % training and 50 % testing. We first apply the basic tree-growing algorithm described

previously to the training data only. We then use cross-validation to prune the tree. At each pair of leaf nodes with a common parent, we evaluate the error on the testing data, and see whether the testing sum of squares would shrink if we removed those two nodes and made their parent a leaf. If so, we prune. We can prune the tree by minimizing the sum of (1) the output variable variance in the validation data, taken a terminal node at a time, or (2) the product of the cost complexity factor and the number of terminal nodes. Larger values of the cost complexity factor result in smaller trees, or other meaningful self-defined measurement. This is repeated until pruning no longer improves the error on the testing data.

There are lots of other cross-validation tricks for trees. One trick that has been commonly is to alternate growing and pruning. We divide the data into two parts, as before, and first grow and then prune the tree. We then exchange the role of the training and testing sets, and try to grow our pruned tree to find the second half. We then prune again, on the first half. We keep alternating in this manner until the size of the tree doesn't change.

Tree-based method has its advantages in the application of non-parametric regression. Firstly, both regression trees and classification trees are two approaches in a class of nonparametric predictive model. That means no assumption made on data, compared to generalized parametric model assuming dependent variable follows a certain distribution such as normal distribution or poisson distribution. Another reason we would like to grow a tree model to predict dependent outcomes is that tree model can easily deal with the cases where effect of covariates on dependent variable has a complicated features, such as nonlinear and interactions, compared to traditional linear regression which usually has a nice and neat form.

However, using decision trees has several major drawbacks, especially in large and complex trees. First of all, it is relatively easy to understand when there are few decisions and outcomes included in the tree. Large trees usually include dozens of

decision leaves and may be too complex to read. Second, while one of the decision tree advantages is its listing comprehensive information and all possible solutions to an issue. However, such comprehensiveness may be unnecessary sometimes. It could bring excessive information which distracts the decision makers's primary interest and slows down decision-making capacity. Moreover, large tree needs more advanced computing capability to determine the best split of each node, which could be time consuming.

A decision tree with only two terminal nodes (i.e., a tree with only one split) is called a tree stump. A tree stump is very simple but work surprisingly well in boosting (Schonlau, 2005).

4.1.4 Nonparametric Regression Analysis with Missing data

Statistical inference with missing data has a long historical background. The regression analysis of missing data has been developed since Yates (1933) formulates the idea of substituting least square estimates for the missing values. While parametric regression analysis with missing data has been developed for years, non-parametric literature in the case of the response variable having missing observations gained little attention. Titterington and Mill (1983) has discussed issues of density estimation. Chu and Cheng (1995) investigated the local behavior of the nonparametric regression estimation. Cheng and Wei (1986) and Cheng(1994) studied the estimation of the mean of the response variable, which built a fundamental basis for our topic. Gonzalez-Manteiga and Perez-Gonzalez (2004) studied the effect of missing observations on the response variable in the estimation of a multivariate regression function.

Observing the shared nature of these literature, imputation has become essential when missing value is present. The imputation of values where data are missing is an area of statistics which has developed much since the 1980s. Simple regression

method imputation may be used for data by using non-missing data to predict the values of missing data. Note that this may "over-correct", introducing unrealistically low levels of noise in the data (Penne, 2009). The regression method has the problem that all cases with the same values on the independent variables will be imputed with the same value on the missing variable, causing a portion of the same problems as mean substitution, which creates a spiked distribution at the mean in frequency distributions and causes attenuation in correlation of the item with others, and underestimates variance (Penne, 2009). The simple imputation method also assumes that the same model explains the data for the non-missing cases as for the missing cases, which is not necessarily true. Another possible method to estimate the regression function with missing observation is multiple imputation. Multiple imputation is a simulation-based approach to the statistical analysis of incomplete data. Rubin (1987) is a pioneer in this technique. Over the last decades multiple imputation has been widely used and recently it has come up in a few nonparametric studies. The procedure consists of replacing each missing observation by various observations from a probability distribution, giving rise to various complete data sets which are analyzed through statistical procedures, finally combining those results for the final result. However, the choice of imputation method is not the focus of this topic. We demonstrate our method by using the simple imputation and we welcome any application of our method to multiple imputation in the future.

4.2 Method

In this section, we extend Cheng's estimator (1994) for a mean of potential outcome with incomplete data to the two-stage design. We apply boosting algorithm to estimate nonparametric regression function.

4.2.1 Point Estimate

Without loss of generalization, we only present the case where potential outcome has two levels, Y_1^* and Y_0^*, corresponding to the consequence that if a patents received one of two different second-line treatments ($R = 1$ or $R = 0$) after he failed on the initial treatment, respectively. Observed data $\{Y, \Delta, R, X\}$ have been described in the previous chapters, where Y is observed outcome and Δ is the indicator of failing on the initial treatment. Our method to estimate the mean of potential outcome EY_1^* for a two-stage HIV data (the same procedure to estimate EY_0^*) is outlined as follows: (1) For those patients who did not fail on the initial treatment ($\Delta = 0$), we assume they have the same potential outcomes as their observed outcomes, i.e., $I(\Delta = 0)Y_1^* = I(\Delta = 0)Y_0^* = I(\Delta = 0)Y$; (2) For those patient who have failed on the initial treatment ($\Delta = 1$), we assume that potential outcome Y_1^* fits nonparametric regression model which is $Y_1^* = f_1(X)+\varepsilon$, and function f_1 is left unspecified. Complete cases of $\{\Delta = 1, R = 1\}$ are used to estimate function f_1 using boosting algorithm which specifies regression tree as base learner and minimizes L_2 Loss defined as (4.1) ; (3) Potential outcome Y_1^* on the other treatment group (R=0) will be predicted and imputed based on the nonparametric regression established in step(2); (4) Mean of potential outcome Y_1^* will be sample average of all patients' predicted outcome, which is $n^{-1}\sum\{(1-\Delta_i)Y_i+\Delta_i R_i Y_i+\Delta_i(1-R_i)\hat{f}(X_i)\}$; (5) Standard error of estimated mean potential outcome Y_1^* will be estimated using bootstrap method. In the following paragraphs, we provide a detailed description of our method.

Assume we have full data $\{Y_1^*, Y_0^*, X, \Delta\}$, where Y_1^* is the potential outcome if he/she had received the treatment, and Y_0^* is potential outcome if he/she had received placebo. Δ is failure indicator and $X = \{X_1, \cdots, X_p\}$ are covariates. Potential outcome Y_j^* follows the following model:

$$E(Y_j) = f_j(X_1, \cdots, X_p), j = 1, 0,$$

where f_i are unspecified and unknown functions and p is the number of covariates. Furthermore, we assume that effect of the covariates on potential outcomes is different between patients who failed on the initial treatment and those who did not. Full data implies that no matter whether failure happened to the patient, he/she would have two potential outcomes corresponding to two treatments. With full data, once a suitable method has been developed well, it is easy to make statistical inference based on full data, including estimating the parameter of interest which is mean of potential outcomes.

$$\mu_j = E(Y_j^*), j = 1, 0.$$

However, we are not able to observe full data. We can only observe $\{Y, \Delta, \Delta R, X\}$, where R is a two-level variable (1 if treatment group, 0 if control group). Only those patients who failed on the initial treatments have chance to receive one of the second-line treatments. Because each individual receives only one treatment, either Y_1^* or Y_0^* is missing. More specifically, since the objective is to estimate mean of Y_1^* and only those patients who actually received treatment can be observed to have outcome Y_1^*. For those patients who received placebo, their potential outcome Y_1^* is missing. Therefore, treatment R plays the role of indicating missing of potential outcome Y_1, which implied the fact that

$$\Delta Y = \Delta R Y_1^* + \Delta(1 - R)Y_0^*. \tag{4.2}$$

We have to consider those patients who did not fail on the initial treatment ($\Delta = 0$). In the previous two chapters, we have established an assumption about these patients' potential outcomes and observed outcomes, and LTD justified this assumption in their paper in 2002, which is

$$(1 - \Delta)Y = (1 - \Delta)Y_1^* = (1 - \Delta)Y_0^*. \tag{4.3}$$

Combine (4.2) and (4.3), we identify the connection between full data and observed data:

$$Y = (1-\Delta)Y_1^* + \Delta R Y_1^* + \Delta(1-R)Y_0^*,$$

or

$$Y = (1-\Delta)Y_0^* + \Delta R Y_1^* + \Delta(1-R)Y_0^*.$$

Thus, when the specific objective of estimating $E(Y_1^*)$, we actually need to focus on how to deal with those patient with missing potential outcome. I describe the nonparametric estimation scheme for the parameter of interest $\mu_1 = E(Y_1^*)$ below. The same procedure can be applied to estimate $\mu_0 = E(Y_0^*)$. Let $f_1(x) = E(Y_1^*|X = x)$. As a consequence, $\mu_1 = E(Y_1^*) = E(E(Y_1^*|X)) = Ef_1(X)$. A natural estimator of μ_1 is

$$\hat{\mu}_1 = 1/n \sum_{i=1}^{n} [(1-\Delta)Y_i + \Delta(RY_i + (1-R)\hat{f}(X_i))] \tag{4.4}$$

The estimator above implies that each missing potential outcome Y_1^* for patients in the control group are imputed by $\hat{f}(X)$. This estimator is an extension of Cheng's estimator (1994) to a two-stage HIV design.

Therefore, in order to obtain $\hat{\mu}_1$ the essential procedure is to accurately estimate $f_i(X), i = 1, 0.$ with complete case and efficiently predict missing potential outcomes for incomplete case. As we have discussed before, we adopt boosting algorithm using regression tree as base procedures to estimate $f_i(X), i = 1, 0.$ The stopping iteration is decided via 5-fold cross-validation by minimizing the L_2 loss function.

4.2.2 Variance Estimate

We use bootstrap method to estimate variance. A possible application of bootstrap method to estimate variance of $\hat{\mu}$ discussed by Gonzalez-manteiga and Perez-Gonzalez(2004). In the case of no missing data, the ordinary bootstrap method can be applied as described by Efron and Tibshirani(1986). When there are imputed

missing data, naive bootstrap estimators are obtained by treating imputed data as original data. However, Shao and Sitter (1996) argued that naive bootstrap method lead to serious underestimation of the variance, because it ignores the imputation process. Instead, " the bootstrap data set should also be imputed in the same way as the original data set was imputed". In addition, Shao and Sitter (1996) proved that this is the only method that works without any restriction on the sampling design, the imputation method, or the type of statistics. Therefore, we proceed the bootstrap process to estimate variance as follows:

(1) Draw a simple random sample with replacement with size n from the original data set.

(2) Partition the drawn random sample into two parts: $A = \{i, R_i = 1\}$ and $B = \{j, R_j = 0\}$, where A and B denotes the set of no missing and with missing in the bootstrap sample. Estimate the nonparametric regression using the set A with the same procedure used in the original data set, then impute missing respondent in B using the model obtained from A, which results in the same imputation procedure used in constructing the imputation of missing values in the original data set.

(3) Obtain the bootstrap estimator $\hat{\mu}_b$.

(4) Repeat procedure (1)-(3) B times.

(5) Apply Monte Carlo approximations $\widehat{\mathrm{Var}}(\mu_1) = 1/(B-1)\sum_{b=1}^{B}(\hat{\mu}_b - \bar{\hat{\mu}})^2$ to obtain bootstrap variance, where $\hat{\mu}_b$ is the Bootstrap estimate and $\bar{\hat{\mu}}$ is the average of B Bootstrap estimates.

4.3 Simulation

In this section, we conduct simulation studies to compare the performance of nonparametric regression estimators using boosting algorithm with semiparametric estimator proposed in the first topic. For the convenience of presenting methods,

we assume all the patients have failed on the initial treatment. To construct true propensity score models and outcome regression models, we follow simulation procedures similar to those in the first topic, which are also from Cao (2009). For each i, $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ was generated as standard multivariate normal, and the elements of $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ were defined as $X_{i1} = \exp(Z_{i1}/2), X_{i2} = Z_{i2}/(1 + \exp(Z_{i1})) + 10, X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$ and $X_{i4} = (Z_{i1} + Z_{i2})^2$, so that Z_i may be expressed in terms of X_i. The true propensity score model is $\pi_0 = \text{expit}(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$. We examine the performances of these estimators in the scenarios listed in the following Table 4.1. These scenarios focuses on the issues either misidentification of PS or OR models or high dimension of covariates, in small sample size. we conducte simulation for each scenario, where 200 Monte Carlo datasets were generated. The similar procedure to generate data has been described in the previous chapter. Estimators using IPW, AIPW, Tan, CTD methods and non-parametric method are presented with their bias and Monte Carlo standard deviation.

Specifically we consider four scenarios. In the first scenario, the propensity score model is specified correctly and potential outcome Y_1^* (and Y_0^*) is generated from normal distribution with mean being a linear combination of covariates $Z_1 - Z_4$. For the working models, PS model and OR model are correctly specified as the true models. In the second scenarios, while working propensity score models have been specified correctly as true models, working outcome regression models are not. Outcomes are generated from a highly skewed distribution, however, working outcome regression models neglect the high skewness by positing a normal distribution on the outcomes. Scenario 3 misspecified both PS model and OR model by replacing $Z_1 - Z_4$ by $X_1 - X_4$. Scenario 4 considers the case where there are many covariates and number of covariates is larger than the number of observations. This is a case which semiparametric method could not handle unless dropping important covariates

Table 4.1: Simulation Scenarios List

scenario	True PS	True OR	High Skewed Outcome	Num. of variables	Correct Working PS	Correct Working OR
1	$Z_1 - Z_4$	$Z_1 - Z_4$	No	4	Yes	Yes
2	$Z_1 - Z_4$	$Z_1 - Z_4$	Yes	4	Yes	No
3	$Z_1 - Z_4$	$Z_1 - Z_4$	Yes	4	No	No
4	$Z_1 - Z_{50}$	$Z_1 - Z_{50}$	No	50	Yes	Yes

from the models.

The purposes to conduct simulations for the scenarios above are: (1)In the previous topics, we found that semiparametric estimators we proposed did perform very well, in efficiency and double robustness, when sample size is large. However, performance of semiparametric estimators highly depends on whether models are correctly specified in small sample size. Therefore, we mainly intend to examine the small sample size performances here. (2) When all the models are correctly specified, efficient semiparametric estimators we proposed have similar estimation, we do not expect nonparametric methods give a very different estimators and draw opposite conclusion in the first scenario. (3)Some proposed semiparametric estimators gave a considerably large variance estimation, and we believe this is because of influence of both misspecification of models and small sample size. Therefore, we hope that nonparametric method which does not rely on specification of model and can work in small sample size has a better performance than semiparmetric estimator in scenario 2 and scenario 3. (4) when number of covariates is no smaller than number of observations, semiparametric methods fails to construct a valid estimator because in this case, because estimates in logistic regression and linear regression may be questionable. Hence, nonparametric estimator is an alternative method for this case as presented in scenario 4.

Results for all the scenarios are presented in the Table 4.2 below. For the first scenario where both models are correct specified, we can see that AIPW, Tan and CTD estimates and nonparametric estimator with regression tree as base learner perform similarly, and they all showed improved efficiencies than IPW estimators. For the second and third scenario where outcomes regression models are incorrectly specified badly, proposed semiparametric estimators did not show much improvement in efficiency compare to IPW estimator, especially CTD estimator which has the largest Monte Carlo standard deviation compared to other semiparametric estimators. How-

ever, nonparametric estimators perform better than any semiparametric estimators in efficiency. The flip side of nonparametric estimator is that it has larger bias. "Badly" incorrectly specifying outcome regression means a highly skewed distributed outcome is assumed to follow the normal distribution in the working models. In the last scenario, when we force 50 important covariates in the model, it is difficult to construct semiparametric estimators. However, containing 50 or even more covariates in the model is not an problem for non-parametric estimator.

4.4 Application to ACTG A5095 Data

We apply nonparametric methods to ACTG A5095 Data. Background, treatment and outcomes definition have been described in detail in the Chapter 2. Different from the method used in the chapter 2, we do not need to build propensity score model and outcome regression model; instead, we estimate the parameters of interest, $E(Y_1^*)$ and $E(Y_0^*)$, by directly modeling relationship between potential outcomes Y_1^* and Y_0^* and auxiliary covariates.

We include the following potential confounders in our analysis: age, height, weight, baseline CD4 cell counts, baseline CD8 and time to first failure, which is consistent with analysis in the Chapter 2.

We present naive estimators, semiparametric estimators and nonparametric estimator in the Table 4.3 below. The findings are summarized as follows: semiparametric estimators proposed in the Chapter 2 such as Tan estimator suggest there are significant differences in cumulative HIV RNA, proportion of time with suppressed HIV RNA and cumulative CD4 cell counts between patients switching early versus late to second-line ARV regimens on the the combined efavirenz-containing arm. However, CTD estimators did not detect any difference for any endpoint. On the other hand, nonparametric method showed that patients switching early had higher proportion of

time with suppressed HIV RNA and cumulative CD4 cell counts than patients who are late to second-line ARV regimens on the the combined efavirenz-containing arm (p-values < 0.05).

4.5 Discussion

In this chapter, we attempted to find an alternative estimating method for semiparametric method we proposed in the second chapter when the semiparametric estimator has large bias and excessively large standard error in those cases where outcome regression models have been "badly" incorrectly specified in small sample size or semiparametric estimators are failed to be constructed due to high dimension of covariates. We extended Cheng's estimator (1994) to two-stage HIV data by applying boosting algorithm to estimate nonparametric regression function. The nonparametric method does not require assuming any form of working models for the data, and it is not difficult to implement. In addition, our simulation results showed that in those cases where semiparametric estimators did not perform very well, nonparametric method displayed its advantages. We also found that in those cases where semiparametric estimators have good performance, non-parametric would not draw an opposite conclusion. In the application of non-parametric estimator to ACTG5095 data, nonparametric method detected patients switching early had higher proportion of time with suppressed HIV RNA and cumulative CD4 cell counts than patients who are late to second-line ARV regimens on the the combined efavirenz-containing arm.

The nonparametric method has its shortcomings. First of all, the key component of our nonparametric estimator is the application of boosting algorithm. Ability of variance reduction of boosting algorithm is based on the trade off large bias. Another disadvantage existing in the nonparametric method we proposed is that we chose simple implementation to predict missing values, which might underestimate

the standard error. Moreover, it has difficulty in calculating standard error, although bootstrap method can be used to estimate standard error. Last, although our method is easy to implement, its involvement with cross-validation to estimate the tuning parameter took considerable long time to implement this method. These issues form the basis of future study.

We proposing non-parametric method to resolve our scientific problem does not mean we are not in favor of semiparametric estimators. When we have confidence in the degree of correct form of working models, we are still in favor of semiparametric method, because semiparametric method is not only producing smaller bias but also efficiency. In those special cases we have identified before, non-parametric method has better performance than semiparametric method in improving efficiency; however, the price is to introduce large bias. Therefore, in practice applying semiparametric method or nonparametric method depends on the specific data and scientific purpose.

Table 4.2: Simulation results based on 200 Monte Carlo replications. $\mu = E(Y_1^*)$. True value=210.

Method	Bias	MCSD	RMSE	COV	Bias	MCSD	RMSE	COV
	PS correct OR correct				PS correct OR incorrect			
IPW	1.04	20.39	20.39	0.88	1.68	46.81	46.77	0.87
AIPW	0.04	3.47	3.48	0.97	1.08	43.71	43.74	0.88
Tan	0.04	3.48	3.47	0.97	0.04	46.77	46.84	0.88
CTD	0.05	3.48	3.48	0.97	0.45	69.04	69.60	0.89
Boosting	1.03	4.09	4.22	0.95	5.98	31.97	32.59	0.91
	PS incorrect OR incorrect				50 Variables			
IPW	12.93	81.96	82.83	0.85				
AIPW	1.45	51.29	51.23	0.89				
Tan	0.58	52.12	52.10	0.88				
CTD	4.37	122.12	122.07	0.90				
Boosting	6.83	31.72	32.12	0.90	1.685	12.601	12.781	0.92

Bias, Monte Carlo bias; MCSD, Monte Carlo standard deviation;
RMSE, Root Mean Square Error; COV, Monte Carlo coverage of 95% Wald confidence interval
OR, Outcome Regression; PS, Propensity Score

Table 4.3: Estimates of mean outcomes, 744 patients, full model

		RNA [1]		Suppression [2]		CD4 [3]	
Method	Switch	Est. (SE)	T	Est. (SE)	T	Est. (SE)	T
Naive	Early	2.600 (0.181)	0.513	0.592 (0.054)	0.800	2.436 (0.055)	0.458
	Late	2.685 (0.068)		0.546 (0.023)		2.466 (0.026)	
IPW	Early	1.835 (0.041)	4.970	0.837 (0.030)	2.720	2.621 (0.093)	0.369
	Late	1.914 (0.032)		0.787 (0.011)		2.564 (0.015)	
AIPW	Early	1.848 (0.048)	2.325	0.829 (0.033)	1.614	2.593 (0.035)	0.764
	Late	1.915 (0.033)		0.787 (0.011)		2.563 (0.015)	
RRZ	Early	1.833 (0.043)	4.218	0.828 (0.01)	19.860	2.600 (0.015)	9.800
	Late	1.914 (0.033)		0.787 (0.011)		2.561 (0.015)	
Tan	Early	1.835 (0.040)	4.948	0.830 (0.011)	21.235	2.599 (0.014)	18.326
	Late	1.914 (0.033)		0.788 (0.011)		2.563 (0.015)	
Non	Early	1.849 (0.048)	1.192	0.808 (0.012)	7.087	2.593 (0.017)	5.364
	Late	1.899 (0.030)		0.788 (0.010)		2.567 (0.015)	

NOTE: The estimated endpoint is reported for combination of initial ARV treatment regimen (A=the combined efavirenz-containing) and switching status (S);

We report the Wald test statistic for a test of the null hypothesis that the average causal effect (ACE) is zero;

[1] HIV RNA level: Length-adjusted AUC of Virus Load, logarithm scale;

[2] Virologic Suppression: Rate of Time Suppression of HIV RNA;

[3] CD4 cell counts: Length-adjusted AUC of CD4 cell counts, logarithm scale

Chapter 5

Summary and Future Work

This dissertation aims to solve two problems. One is to evaluate the effect of different treatment switching strategies in HIV studies and the second is to evaluate the effect of different treatment durations in infusion studies. There are several common features shared by these two problems. First of all, they are both more interested in one treatment policy than a single treatment assignment. Second, direct comparison by randomization experiments is not available. Inference is based on observational data. However, the problem in the infusion studies is more complex than the first problem in HIV studies, because in infusion study we have to consider the terminating event which censored the treatment duration. Through the dissertation, we made causal inference on the effect of treatment policy by applying semiparametric theory to missing data problem, and we proposed double robust and locally efficient estimators for the population mean response on the basis of censored observational data.

The first primary goal of this dissertation was to address a scientific question in HIV/AIDS research where there is an abundance of conjecture and speculation but limited evidence: *is it better to switch early or late from a failing ARV regimen?* Where an ordinary randomized trial could easily answer this question, the clinical literature suggests that such a randomized trial is difficult to enroll. Alternatively,

one can use data from other studies where assignment to switch ARV regimen early or late depends on patient-specific characteristics. Although standard sample averages and two-sample tests cannot be used to analyze such data, methods of causal inference may be utilized and have become a staple of modern statistical inference. To answer the above scientific question using data from ACTG A5095, we adopt methods based on potential outcomes (Rubin, 1974), an extension of two-stage designs (Lunceford et al., 2002) via Murphy et al. (2001), and Tan's (2006, 2007) adaptive doubly-robust estimator. Using this combination of techniques, we found that patients who started a standard combination antiretroviral regimen of nucleoside analogues and efavirenz, then made regimen changes within eight weeks of confirmed virologic failure on initial ARV regimen were associated with lower cumulative viral load level, higher cumulative CD4 cell counts, and spent a larger proportion of the follow-up period with suppressed viral load levels, on average. Although other authors have recently reported similar results for switching off of HAART, the endpoints used here are very different and can be computed even when a mortality outcome is not available. The ACTG A5095 study is an example of a clinical trial that whose primary objective was to test the efficacy of initial regimens but we used it in a secondary analysis of regimen change. When it is difficult to design and enroll a completely randomized study of regimen change, data like that from ACTG A5095 and the framework employed here will be germane for evaluating the effect of early regimen change.

The second primary goal of this dissertation was to address a scientific question in infusion research when treatment infusion is informatively right-censored. This scientific question looks similar to the first one, because they both compare the effect of timing of an event (treatment) and treatment assignment is not randomized, instead observational. However, the second question is more complex than the treatment switch problem in the first topic. Firstly, treatment duration is not binary, at least it is ordinary, having some sort of ordering. Secondly, treatment-termination

events exists and censor the treatment duration. We must consider the terminating event, because it is specified in the protocol and it is part of treatment plan. In addition, different from survival analysis where treatment duration only is of interest and treatment duration is censored by terminating event, we regard terminating event as a possible consequence related to treatment duration, instead of outcome. In this case, we identified a class of augmented inverse probability weighted estimators and derived the locally efficient and doubly robust estimator for the mean outcome for the particular treatment policy. Simulation study showed that the proposed estimator is more efficient than the IPW estimator whenever propensity score is correctly specified and has smaller bias than the IPW estimator due to the protection provided by double robustness. In addition, our efforts in identifying augmentation part in double robust estimator built a fundamental ground work for future efficiency improvement using techniques in Tan(2006) or Cao et al(2009).

An limitation of semiparametric method to deal with the case where treatment is informatively censored is that it can only handle the situation that duration can take on a finite number of values, $t_1, \cdots, t_m$. In truth, treatment duration in infusion studies is a continuous random variables. Hence, treating treatment duration as a continuous variable is more realistic. This is one of future work. In addition, Tan(2006) and Cao, et al.(2009) have developed estimators which is more efficient than AIPW estimator when outcome regression are misspecified based on large sample theory. Therefore, extension of their work to our setting is another part of future work. Nonparametric method we proposed in the first topic needs more investigation on the reason why it has relatively large bias.

Bibliography

Bather, J. "an Introduction to Dynamic Programming and Sequential Decisions." *Chichester: Wiley.* (2000).

Bickel, P., Klassen, C., Ritov, Y., and Wellner, J. "Efficient and Adaptive estimation for Semiparametric Models." *Baltimore: Johns Hopkins University Press* (1993).

Binder, H. and Tutz, G. "A comparison of methods for the fitting of generalized additive models." *Stat Comput*, 99:18–87 (2008).

Borra, S. and Ciaccio, A. "Improving nonparametric regression methods by bagging and boosting." *Computational Statistics and Data Analysis*, 38:407–420 (2002).

Bowman, A. and Azzalini, A. "Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations." Oxford University Press, Oxford (1997).

Breiman, L. "Arcing classifiers." *The Annals of Statistics*, 26:801–849 (1998).

Breiman, L., Friedman, J., Olshen, R., and Stone, C. "Classification and Regression Trees." *Journal of Computational and Graphical Statistics*, Wadsworth:651–674 (1984).

Buhlmann, P. and Yu, B. "Boosting with the L2 Loss: Regression and Classification." *J. Amer. Statist. Assoc.*, 98:324–340 (2003).

Cao, W., Tsiatis, A. A., and Davidian, M. "Improving efficiency and robustness of the doubly robust." *Biometrika*, 96:723–734 (2009).

Cassel, C. M., Sarndal, C., and Wretman, J. H. "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations." *Biometrika*, 63:615–620 (1976).

Cheng and Chu. "Nonparametric regression estimation with missing data." *Journal of Statistical Planning and Inference*, 48:85–99 (1995).

Cheng, P. "Nonparametric Estimation of Mean Functionals with Data Missing at Random." *Journal of the American Statistical Asso*, 89:81–87 (1994).

Chu, S. "Pricing the C's of Diamond Stones." *Journal of Statistics Education*, 9:2 (2001).

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. "Probabilistic Networks and Expert Systems." *New York: Springer.* (1999).

Deeks, S. G. "Treatment of antiretroviral-drug-resistant HIV-1 infection." *Lancet*, 362:2002–2011 (2003).

Fox, J. "Nonparametric Simple Regression: Smoothing Scatterplots." *Sage, Thousand Oaks CA* (2000a).

—. "Multiple and Generalized Nonparametric Regression." *Sage, Thousand Oaks CA* (2000b).

Freund, Y. and Schapire, R. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Japanese Society for Artificial Intelligence*, 55:119–139 (1997).

—. "A Short Introduction to Boosting." *Journal of Japanese Society for Artificial Intelligence*, 14:771–780 (1999).

Friedman, J. "Greedy function approximation: a gradient boosting machine." *The Annals of Statistics*, 29:1189–1232 (2001).

Friedman, J., Hastie, T., and Tibshirani, R. "Additive logistic regression: a statistical view of boosting (with discussion)." *The Annals of Statistics*, 28:337–407 (2000).

Gulick, R., Ribaudo, H., Shikuma, C., Lalama, C., Schackman, B., Meyer, W. I., Acosta, E., Schouten, J., Squires, K., Pilcher, C., Murphy, R., Koletar, S., Carlson, M., Reichman, R., Bastow, B., Klingman, K., Kuritzkes, D., and Team, A. A. S. "Three- vs four-drug antiretroviral regimens for the initial treatment of HIV-1 infection: a randomized controlled trial." *JAMA*, 296:769–781 (2006).

Gulick, R. M., Ribaudo, H. J., Lustgarten, S., Squires, K. E., Meyer, W. A., Acosta, E. P., Schackman, B. R., Pilcher, C. D., Murphy, R. L., Maher, W. L., Witt, M. D., Reichman, R. C., Snyder, S., Klingman, K. L., and Kuritzkes, D. R. "Triple-Nucleoside Regimens versus Efavirenz-Containing Regimens for the Initial Treatment of HIV-1 Infection." *The New England Journal of Medicine*, 350:1850–1861 (2004).

Hammersley, J. M. and Handscomb, D. C. *Monte Carlo Methods*. London: Methuen (1964).

Hastie, T. R., Tibshirani, R. J., and Freedman, J. "The Elements of Statistical Learning: Data Mining, Inference and Prediction." *Springer, New York* (2001).

Heckerman, D. "A tutorial on learning with Bayesian networks. In Learning in Graphical Models (ed. M. I. Jordan)." *Dordrecht: Kluwer.*, 301–354 (1998).

Holland, P. W. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81:945–960 (1986).

Horvitz, D. G. and Thompson, D. J. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association*, 47:663–685 (1952).

Hothorn, T. "Families of Splitting Criteria for Classification Trees." *Statistics and Computing*, 9:309–315 (1999).

—. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, 15:651–674 (2006).

Johnson, B. A. and Tsiatis, A. A. "Estimating Mean Response as a Function of Treatment Duration in an Observational Study, Where Duration may be Informatively Censored." *Biometrics*, 60:315–23 (2004).

—. "Semiparametric Inference In Observational Duration-response Studies, With Duration Possibly Right-censored." *Biometrika*, 92:605–618 (2005).

Kearns, M. and Vazirani, U. "An Introduction to Computational Learning Theory." *MIT Press* (1994).

Lavori, P. W. and Dawson, R. "A design for testing clinical strategies: biased adaptive within-subject ran- domization." *J R. Statist. Soc. A*, 163:29–38 (2000).

Levin, J. "Cumulative, Current Viral Replication and Low CD4 Increases NHL Risk: Immune deficiency, uncontrolled HIV replication and non-Hodgkin's lymphoma, ANRS C03 Auitaine Cohort, 1998-2006." *Abstract, IAS, Capetown*, July:531–41 (2009).

Long, Q., Little, R. J., and Lin, X. "Causal Inference in Hybrid Intervention Trials Involving Treatment Choice." *Journal of the American Statistical Association*, 103:474 (2008).

Lunceford, J. K., Davidian, M., and Tsitatis, A. A. "Estimation of Survival Distributions of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials." *Biometrics*, 58:48–57 (2002).

M., S. "Boosted Regression (Boosting): An introductory tutorial and a Stata plugin." *The Stata Journal*, 3:330–354 (2005).

Murphy, S. "Optimal structural nested models for optimal sequential decisions." *Journal of the Royal Statistical Society. B*, 65:331–366 (2003).

Murphy, S. A., van der Laan, M. J., and Robins, J. M. "Marginal Mean Models for Dynamic Regimes." *Journal of the American Statistical Association*, 96:1410–23 (2001).

Newey, W. K. "Semiparametric Efficiency Bounds." *Journal of Applied Econometrics*, 5:99–135 (1990).

Neyman, J. "On the Application of Probability Theory To Agricultural Experiments." *Statist. Sci.*, 5:465–480 (1923).

OBrien, S. "Cutpoint Selection for Categorizing a Continuous Predictor." *Biometrics*, 60:504–509 (2004).

Penne, M. "Attrition and Missing Data: Understanding,Detection and Solutions for Care Grantees." *RTI International Presented at OAPP Webinar* (2009).

Petersen, M., Deeks, S., Martin, J., and van der Laan, M. "History-adjusted Marginal Structural Models for Estimating Time-varying Effect Modification." *American Journal of Epidemiology*, 166:985–993 (2007).

Petersen, M. L., Van der Laan, M. J., Sonia, N., Eron, J. J., Moore, and Deeks., S. G. "Long-term Consequences of the Delay Between Virologic Failure of Highly Active Antiretroviral Therapy and Regimen Modification." *AIDS*, 22:2097–2106 (2008).

Riddler, S., Jiang, H., Tenorio, H., A.and Huang, Kuritzkes, D., Acosta, E., Landay, A., Bastow, B., Haas, D., Tashima, K., Jain, M., Deeks, S., and Bartlett, J. "A Randomized Study of Antiviral Medication Switch at Lower- Versus Higher-switch Thresholds: AIDS Clinical Trials Group Study A5115." *Antiviral therapy*, 12:531–41 (2007).

Robins, J. "Optimal structural nested models for optimal sequential decisions." *Springer*, New York:189–326 (2004).

Robins, J. M. "A New Approach To Causal Inference In Mortality Studies With Sustained Exposure Periods — Application To Control of the Healthy Worker Survivor Effect." *Mathematical Modelling*, 7:1393–1512 (1986).

—. "Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect"." *Comput. Math. Applic.*, 14:923–945. (1987).

—. "The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Health Service Research Methodology: a Focus on AIDS (eds L. Sechrest, H. Freeman and A. Mulley)." *US Public Health Service.*, 113–159 (1989).

Robins, J. M., Rotnizky, A., and Zhao, L. P. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association*, 89:846–866 (1994).

Robins, J. M. and Wasserman, L. "Estimation of effects of sequential treatments by reparameterizing Discussion on the Paper by Murphy 355 directed acyclic graphs." *In Proc. 13th Conf Uncertainty in Artificial Intelligence (eds D. Geiger and P. Shenoy)*, 409–442 (1997).

Rosenbaum, P. R. and Rubin, D. B. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society. Series B (Methodological)*, 45:212–218 (1983a).

—. "The Central Role of the Propensity Score in Observational Studies For Causal Effects." *Biometrika*, 70:41–55 (1983b).

—. "Reducing Bias in Observational Studies Using Subclassification On the Propensity Score." *Journal of the American Statistical Association*, 79:516–524 (1984).

Rotnitzky, A., Scharfstein, D., Su, T., and Robins, J. "Methods for Conducting Sensitivity Analysis of Trials with Potentially Nonignorable Competing Causes of Censoring." *Biometrics*, 57:103–113 (2001).

Rubin, D. B. "Estimating Causal Effects of Treatments In Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66:688–701 (1974).

Scharfstein, D. O., Rotnizky, A., and Robins, J. M. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association*, 94:1096–1120 (1999).

Shafer, R., Smeaton, L., Robbins, G., V., D. G., Snyder, S., D'Aquila, R., Johnson, V., Morse, G., Nokta, M., Martinez, A., Gripshover, B., Kaul, P., Haubrich, R., Swingle, M., McCarty, S., Vella, S., Hirsch, M., Merigan, T., and AIDS Clinical Trials Group 384 Team. "Comparison of four-drug regimens and pairs of sequential three-drug regimens as initial therapy for HIV-1 infection." *The New England Journa ofl Medicine*, 349(24):2304–15 (2003).

Simonoff, J. "Smoothing Methods in Statistics." *Springer, New York* (1996).

Spritzler, J., DeGruttola, V. G., and Pei, L. "Two-Sample Tests of Area-Under-the-

Curve in the Presence of Missing Data." *The International Journal of Biostatistics*, 4:1–20 (2008).

Stone, R. M., Berg, D. T., George, S. L., Dodge, R. K., Paciucci, P. A., Schulman, P., Lee, E. J., Moore, J. O., Powell, B. L., and Schiffer, C. A. "Granulocyte-macrophage Colony-stimulating Factor After Initial Chemotherapy for Elderly Patients With Primary Acute Myelogenous Leukemia." *The New England Journal of Medicine*, 322:1671–1677 (1995).

Study, U. C. H. C. "Treatment Switches After Viral Rebound in HIV-infected Adults Starting Antiretroviral Therapy: Multicentre Cohort Study." *AIDS*, 22:1943–1950 (2008).

Tan, Z. "A Distributional Approach for Causal Inference Using propensity scores." *Journal of the American Statistical Association*, 101:1619–1637 (2006).

—. "Understanding OR,PS and DR." *Statist. Sci.*, 22:560–568 (2007).

Tsiatis, A. A. *Semiparametric Theory and Missing Data*. Springer (2006).

Tutz, G. and Reithinger, F. "A boosting approach to flexible semiparametric mixed models." *Statistics in Medicine.*, 26:28722900 (2007).

Van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press (2000).

Wahed, A. S. and Tsiatis, A. A. "Optimal Estimator for the Survival Distribution and Related Quantities for Treatment Policies in Two-Stage Randomization Designs in Clinical Trials." *Biometrics*, 60:124–133 (2004).

Wright, S. "An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education." *SAS White Paper* (2010).

Yeh, K. C. and Kwan, K. C. "A Comparison of Numerical Integrating Algorithms by Trapezoidal, Lagrange and Spline Approximation." *Journal of Pharmacokinetics and Pharmacodynamics*, 6:79–87 (1978).