Causal Inference in Health Science Research

by

Qixing Liang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Min Zhang, Chair
Research Assistant Professor Zhi He
Professor Alexander Tsodikov
Associate Professor Haojie Zhu

Qixing Liang

liangqx@umich.edu

ORCID iD: 0000-0003-2479-4733

To my parents

ACKNOWLEDGEMENTS

I would like to express my great appreciation to my advisor, Dr. Min Zhang for her invaluable guidance, mentorship, and patience throughout my PhD study. She has made me a better student and a better researcher and taught me how to think as a statistician. This dissertation could never have been finished without her instruction, motivation and encouragement.

I would like to gratefully thank my committee members Dr. Zhi He, Dr. Haojie Zhu, and Dr. Alexander Tsodikov, for their insightful suggestions and constructive input for my dissertation.

I would like to thank my GSRA advisors, Dr. Francis Pagani, Dr. Donald Likosky, Dr. Keith Aaronson and Dr. Michael Thompson. Thank you all for providing me the working opportunity, teaching me the clinical knowledge and providing data for the application of the dissertation. In particular, I would like to sincerely thank Dr. Pagani for providing valuable suggestions from a clinical perspective for the application of Chapter III and IV; and sincerely thank Dr. Likosky for spending much time facilitating and supporting my GSRA work in the last four and a half years. I also thank Xiaoting Wu for preparing data for the application part of Chapter II. Finally, I would like to thank all other members in the collaborative research group, especially Raymond Strobel, Sarah Ward and Joshua Bourque for working together on projects.

Thanks to all the faculty in the Department of Biostatistics for their wonderful lectures and guidance. Thanks to Dr. Thomas Braun and Dr. Hyun Min Kang for

their concerns and encouragements during my PhD study.

I would like to thank my parents Zuoqin Liang and Shanyun Hu for their love, encouragement and support during my Master's and PhD studies at the University of Michigan. This dissertation is dedicated to you.

Thanks to all my friends in Ann Arbor.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Causal inference methods including propensity score (PS) matching and weighting have been widely used for comparative effectiveness research based on observational clinical databases. There are new challenges of causal inference in medical studies, including how to handle clustered data structure, how to improve efficiency of the traditional methods and etc. To overcome these challenges, we develop novel causal inference method for estimating treatment effects. The proposed methods are motivated by and applied to various studies of cardiovascular diseases and cardiac surgeries.

In Chapter II, we aim to estimate causal treatment effect in clustered observational data and in our application clustered data structure arises from patients being nested within hospitals. We propose a strategy to combine PS matching and outcome regression model for estimating treatment effect while accounting for the hierarchical nature of the data. We show that this method enjoys the double robustness property, i.e. when either the PS or outcome model is correctly specified, the bias is negligible. The proposed method has better performance than the usual PS method and the existing doubly robust PS weighted method, and is more robust than the outcome regression method.

Chapter III is motivated by comparing different types of ventricular assist devices (VAD) for end-stage heart failure patients where patients are likely to receive a heart transplant after receiving a VAD. We propose to treat heart transplants as dependent censoring and propose an augmented inverse probability weighted method to

estimate the treatment-specific difference in potential restricted mean lifetimes, had no patients received heart transplant. Specifically, we first derive an estimator that combines inverse probability of treatment weighting and inverse probability of censoring weighting to account for the imbalance in baseline characteristics and censoring that may depend on time-dependent confounders, respectively. Then we propose an augmentation method to improve the efficiency of estimation. Large-sample properties of the proposed methods are studied and simulation studies are conducted to assess the finite-sample performance.

In Chapter IV, we further extend and refine the work in Chapter III and develop methods for estimating more meaningful causal treatment effects as opposed to the average treatment effect. The goal is to overcome two potential problems related to estimating the average treatment effect. Namely, depending on the specific application the average treatment effect may not be the most clinical meaningful and relevant quantity. Also it is known that estimators for average treatment effect often have large variance even with the use of more sophisticated methods for improving efficiency (e.g., augmentation). We propose augmented methods of matching weights to estimate the treatment-specific difference in potential restricted mean lifetimes for the matched population, had no patients received heart transplant. Simulation studies show that the proposed methods considerably improve the efficiency compared to the existing methods.

CHAPTER I

Introduction

Observational data are widely used in health science research. In observational data, the covariates are often unbalanced between treatment groups and need to be adjusted when estimating causal treatment effect. Causal inference methods including propensity score (PS) matching and weighting have been widely used for comparative effectiveness research based on observational clinical databases. However, there are new challenges of causal inference in health science research that cannot be well handled by current methods, including how to handle clustered data structure, how to improve efficiency of the traditional methods, how to define the clinically meaningful estimand and etc. In this dissertation, to overcome these challenges, we develop novel causal inference method for estimating treatment effects. The proposed methods are motivated by and applied to studies of cardiovascular diseases and cardiac surgeries.

In Chapter II, we aim to estimate causal treatment effect in clustered observational data. In our application clustered data structure arises from patients being nested within hospitals. Compared to cross-sectional data, the complexity of data structure and treatment assignment mechanisms brings in additional challenges for estimating causal effects. Regression modeling is a typical method to handle covariate imbalance between treatment groups. By appropriately adjusting the confounding covariates, treatment effects can be consistently estimated from regression model. But

the accuracy of estimation of causal effects highly relies on the correct specification of outcome model, which is often difficult to achieve because the relationships between outcome and covariates are complicated. Alternatively, PS matching is a popular one in medical researches as it is intuitive and easy to implement. Instead of modeling outcomes or specifying the heterogeneity of treatment effect, it aims to equalize the covariate distributions between treatment groups. However, although PS matching has a mature application procedure in cross-sectional data, it has not been well studied in clustered data. We propose a strategy to combine PS matching and outcome regression model for estimating treatment effect while accounting for the hierarchical nature of the data. We show that this method enjoys the double robustness property, i.e. when either PS or outcome model is correctly specified, the bias is negligible. The proposed method has better performance than the usual PS matching method and the existing doubly robust PS weighted method, and is more robust than the outcome regression method. We also study various types of matching strategies for clustered data and compared their performances.

Chapter III is motivated by comparing different types of ventricular assist devices (VAD) for end-stage heart failure patients where patients are likely to receive a heart transplant after receiving a VAD. In this chapter, our aim is to estimate the causal treatment effects but there are two challenges. First, as this is an observational study, the distributions of patient baseline covariates are different between groups. Second, patients may receive heart transplants after receiving VAD implantation. So their survival outcomes we observed are due to not only VAD implantation but also heart transplant. Also, we know that the patients with worse post-implant situation were more likely to need a heart transplant. To overcome these two challenges, we treat heart transplant as dependent censoring and propose an augmented inverse probability weighted method to estimate the treatment-specific difference in potential restricted mean lifetimes had no patients received heart transplant. Specifically,

we derive an estimator that combines inverse probability of treatment weighting and inverse probability of censoring weighting to account for imbalance in baseline characteristics and censoring that may depend on time-dependent confounders, respectively. Then we propose an augmentation method to improve the efficiency of estimation. Large-sample properties of the proposed methods are studied and simulation studies are conducted to assess the finite-sample performance.

In Chapter IV, we further refine the work in Chapter III and develop methods for estimating more meaningful causal treatment effects as opposed to the average treatment effect. The goal is to overcome two potential problems related to estimating the average treatment effect. Namely, depending on the specific application the average treatment effect may not be the most clinical meaningful and relevant quantity. Also it is known that estimators for the average treatment effect (ATE) often have large variance even with the use of more sophisticated methods for improving efficiency (e.g., augmentation). In this chapter we focus on estimating the average treatment effect on the matched population (ATM). In the application of Chapter III, a small fraction (3.1%) of patients received BiVAD and the majority (96.9%) of patients received LVAD. It is possible that the physicians only provide the option of BiVAD implantation to the patients with more severe clinical conditions. In this case, we think it may be more reasonable to estimate the treatment effect for the patients who can receive either treatment. Hence, in this chapter we are interested in estimating the treatment-specific difference in potential restricted mean lifetimes on the matched population, had no patients received heart transplant. PS matching is a common method for estimating the ATM, but it has drawbacks in practice, including difficulty in estimating variance of PS matching estimator and developing methods to improve efficiency. *Li and Greene* (2013) proposed a matching weight (MW) method which is an analogue to one-to-one PS caliper matching method without replacement and showed it is more efficient and has better variance estimation than the PS match-

ing. However, how to apply the MW method for causal inference for survival outcome and further improve the efficiency remains unclear. In this chapter, motivated by the same application with Chapter III, we adopt the matching weight method to estimate the treatment-specific difference in potential restricted mean lifetimes had no patients received heart transplant, had no patients received heart transplant. Then we develop augmented methods to improve the efficiency.

CHAPTER II

Doubly Robust Propensity Score Matching Methods for Clustered Data

2.1 Introduction

In medical research, randomized design is gold standard of assessing causal treatment effects. However, randomized trials are not always feasible due to high cost and potential ethical issues. Alternatively, observational data are often used in medical research. In observational studies, the covariates (e.g., age, gender, race, comorbidity) are often unbalanced between treatment groups and need to be adjusted when estimating causal treatment effects. Regression modeling is a typical method to handle the covariate imbalance between treatment groups. By appropriately adjusting the confounding covariates, treatment effects can be consistently estimated from regression model. However, if differences in the characteristics across groups are large, the causal effect estimated from regression models may be quite problematic because valid estimations rely on model extrapolations which may be sensitive to model misspecification (*Rubin*, 1979). In other words, the accuracy of estimation of causal effects highly relies on the correct specification of outcome models, which is often difficult to achieve because the relationships between outcome and covariates are often complicated.

Rosenbaum and Rubin (1983) proposed the propensity score (PS) methods to reduce bias in estimating causal effects. Instead of modeling outcomes or specifying the heterogeneity of treatment effect, PS methods aim to equalize the covariate distributions between treatment groups. There are four types of PS methods (*Austin*, 2011), including matching (*Rosenbaum and Rubin*, 1985; *Austin*, 2008; *Stuart*, 2010), inverse probability weighting (IPW) (*Lunceford and Davidian*, 2004), stratification and covariate adjustment using propensity score (*Rosenbaum and Rubin*, 1984). Among these methods, PS matching is popular in medical research as it is intuitive and easy to implement. In last decades PS matching in cross-sectional data has been well studied and improved in many aspects, including PS model fitting and matching strategy (*Rosenbaum*, 1989; *Gu and Rosenbaum*, 1993; *Stuart*, 2010).

However, although PS matching has a mature application procedure in cross-sectional data, it has not been well studied in clustered data, which is a common type of medical data. In medical data patients are often nested in hospitals, and the patient characteristics and treatment assignment mechanism may vary greatly among hospitals. The complexity of data structure and treatment assignment mechanisms introduces additional challenges for estimating causal effects using PS matching. In the last decade, a few publications have explored PS matching in clustered data. The performances of within-cluster and across-cluster matching have been compared in the cases of different cluster sizes (*Steiner et al.*, 2013; *Kim and Steiner*, 2015). Single-level and two-level PS models together with new two-level matching strategies have been studied (*Thoemmes and West*, 2011; *Rickles and Seltzer*, 2014; *Arpino and Cannas*, 2016). Furthermore, to overcome the common problem of unobserved cluster-level confounders when estimating causal treatment effects, *Arpino and Mealli* (2011) and *Oelrich* (2014) developed PS matching methods, and *Li et al.* (2013) and *Yang* (2016) developed inverse probability weighting methods, respectively. To summarize, the previous researches have mainly focused on taking into account the

hierarchy of data structure by using multi-level PS models or improving matching strategies, or solving the problem of unobserved cluster-level confounders.

Medical researchers often perform PS matching and estimate treatment effects by calculating the difference of the average of observed outcomes between treatment groups. Unpaired or paired two-sample tests are used to assess the significance of the estimated treatment effects. However, generally PS matching cannot remove all the imbalances between groups. Directly comparing the average outcome between groups after matching precludes further removing residual imbalances and improving the efficiency. It has been suggested that post-matching outcome regression may further remove the residual imbalances in matched data and thus improve the performance of PS matching (*Rubin*, 1973, 1979; *Rubin and Thomas*, 2000; *Ho et al.*, 2007; *Stuart*, 2010). However, to our knowledge, no study has evaluated the performance of combining PS matching and outcome regression using simulation study, especially for clustered data.

This chapter is motivated by a multi-center observational cardiac study. The patient data were collected from multiple hospitals with widely varying sizes, and two treatment groups were unbalanced in patient-, surgeon- and hospital-level characteristics. The aim of the study is to estimate the causal effect of cardiac treatment. In this chapter, we propose a strategy to combine PS matching and outcome regression, while accounting for the hierarchical nature of the data. We also study different matching methods in clustered data and compare the performances. The remainder of the chapter is organized as follows: we propose the methods in Section 2.2; Section 2.3 is the simulation study; Section 2.4 is the application; Finally Section 2.5 is the discussion.

2.2 Methods

2.2.1 Assumptions for Causal Inference in Two-level Data

Without loss of generality, we consider a two-level data structure, including hospital and patient levels. Considering a two-level observational data structure where n patients are nested into m hospitals, i.e. $n = \sum_{i=1}^{m} n_i$, $i = 1, ..., m$, where n_i denotes number of patients in hospital i. Let Z_{ij} indicate whether patient j in hospital i is assigned to treatment ($Z_{ij} = 1$) or control ($Z_{ij} = 0$). Let X denote the vector of patient-level covariates, V denote the vector of hospital-level covariates, and Y denote the continuous outcome. The propensity score e_{ij} is defined as the probability of being assigned to treatment for patient j in hospital i, conditional on patient- and/or hospital-level covariates, i.e. $e_{ij} = Pr(Z_{ij} = 1|X_{ij}, V_i)$.

To estimate the causal treatment effect, we adopt the potential outcome framework. Unlike cross-sectional data, in two-level data analysis, the framework need to be adjusted to account for the hierarchical nature of data and the existence of hospital-level covariates. We assume patient j in hospital i has two potential outcomes Y_{ij}^1 and Y_{ij}^0, which denote the hypothetical outcomes if the patient has taken treatment ($Z_{ij} = 1$) or control ($Z_{ij} = 0$) in his/her observed hospital, respectively. There are two common estimands including the average treatment effect (ATE) and the average treatment effect on the treated (ATT). In this chapter we are interested in estimating the ATE, i.e. $E(Y_{ij}^1 - Y_{ij}^0)$. Valid estimation of the ATE depends on three assumptions. We first make the stable unit treatment value assumption (SUTVA), stating that the potential outcomes for one patient are not affected by the treatment assignment of other patients in either the same or different hospitals, i.e. $Y_{ij} = Z_{ij}Y_{ij}^1 + (1 - Z_{ij})Y_{ij}^0$. The second assumption is unconfoundedness, claiming that

$$(Y_{ij}^1, Y_{ij}^0) \perp Z_{ij}|X_{ij}, V_i \quad \text{or} \quad (Y_{ij}^1, Y_{ij}^0) \perp Z_{ij}|(X_{ij}, \text{hospital } i). \tag{2.1}$$

The symbol "$\perp$" denotes independence. *Rosenbaum and Rubin* (1983) proved that for cross-sectional data, if the unconfoundedness assumption i.e. $(Y^1, Y^0) \perp Z|X$ holds, the potential outcomes are independent of treatment assignment conditional on propensity score i.e. $(Y^1, Y^0) \perp Z|e(X)$, which is the theory basis for the PS matching methods. Here, we modify the unconfoundedness assumption for the two-level data setting. Specifically, we assume that the potential outcomes for one patient are independent of treatment assignment conditional on both patient- and hospital-level covariates, or conditional on patient-level covariates and hospital membership. The second part of Equation (2.1) applies to matching within hospital. Matching within hospital does not require observing hospital-level covariates, because the hospital-level covariates are automatically balanced as the matched pairs are in the same hospital. In this case, the hospital membership is sufficient for making valid estimation of the ATE, which has also been claimed by *Thoemmes and West* (2011) and *Steiner et al.* (2013). Based on the second part of Equation (2.1), matching within hospital is a method to make valid inference in the presence of unobserved cluster-level covariates. The third assumption is the overlap assumption, which states the patients in each hospital have positive probabilities of being assigned to either the treatment or the control, i.e.

$$0 < e_{ij} < 1. \tag{2.2}$$

Rosenbaum and Rubin (1983) considered it as "strong ignorable" if both unconfoundedness and overlap assumptions are satisfied.

2.2.2 Propensity Score Models

The propensity scores are often not observed and need to be estimated from models (e.g. logistic model). Depending on how to model the hierarchical effect of hospitals, we consider three types of PS models. The first one is single-level logistic model, for which we ignore the hierarchical data structure and model the probability of being

assigned to treatment using patient-level covariates, i.e.

$$\text{Single-level PS model: } \text{logit}\{P(Z_{ij} = 1|X_{ij})\} = X_{ij}\beta.$$

As two-level PS models have been shown to reduce bias of causal treatment effects in clustered data (*Arpino and Mealli*, 2011; *Li et al.*, 2013), we consider two types of two-level PS models: the random-effect logistic model and the fixed-effect logistic model, i.e.

$$\text{Random-effect PS model: } \text{logit}\{P(Z_{ij} = 1|X_{ij}, V_i)\} = b_i + X_{ij}\beta + V_i\gamma,$$

where b_i is a random intercept for hospital i.

$$\text{Fixed-effect PS model: } \text{logit}\{P(Z_{ij} = 1|X_{ij}, V_i)\} = H_i + X_{ij}\beta,$$

where H_i is the indicator for hospital i.

It is important to include all covariates related to treatment assignment into the PS model. If it is not clear that which characteristics are associated with treatment assignment mechanisms, it has been recommended to include as many covariates as possible to avoid potential confounding (*Stuart*, 2010), although the trade-off is a slight inflation of variance of the estimator (*Brookhart et al.*, 2006).

2.2.3 Matching

After estimating PS from models, the next step is to match patients between groups using estimated PS. In cross-sectional data analysis, various matching strategies have been proposed and studied, including nearest neighbor matching, optimal matching and nearest neighbor matching within a caliper (caliper matching). The

caliper matching is a popular method in medical research and has been shown to be less biased than nearest neighbor matching and optimal matching (*Austin*, 2014). *Gu and Rosenbaum* (1993) have shown that optimal matching does not improve the performance of balancing between groups compared to other matching algorithms, although it performs better at minimizing the within pair difference. In this chapter, the primary interest is to estimate the ATE instead of achieving best individual matching pair, so we use the caliper matching with caliper equal to 0.10 standard deviation of the estimated propensity score. Let us take the caliper matching within hospital as an example; let M_{ij} be the matched set for patient j in hospital i who received treatment, S_{i0} be the set of all control patients in hospital i, j' be the patient id in S_{i0}. Then M_{ij} can be formally expressed as

$$M_{ij} = \left\{ j' \in S_{i0} : \min_{j' \in S_{i0}} |\widehat{e}_{ij} - \widehat{e}_{ij'}| \leq 0.10 \widehat{\sigma}_e \right\}.$$

With regards to matching with or without replacement, we consider three methods, including matching without replacement, matching with replacement, and modified matching with replacement. Matching without replacement is frequently used in medical research. It means that one patient can only be matched once and after s/he is matched, s/he would not be considered in future matching. Matching with replacement means that the patients are allowed to match with multiple patients. It has been argued that matching with replacement may decrease bias but cause more complicated inference due to the violation of independence assumption (*Stuart*, 2010). For matching with replacement, *Dehejia and Wahba* (1999) and *Hill et al.* (2004) proposed to incorporate weights into outcome analysis. For example, if 1 control is matched to 5 treated patients, then each of these 5 patients receive a weight of 1/5. Beside these two traditional methods, we propose a third way that we call "modified matching with replacement". Specifically, we first match the patients

without replacement, then for the unmatched patients, we attempt to match them with the patients that have already been matched. When an unmatched patient can be matched to multiple patients, we match her/him to the patient with fewest times being matched.

A new challenge for PS matching in clustered data is how to handle the hierarchical data structure during matching. Generally there are two common matching methods, including within-hospital matching and across-hospital matching. Within-hospital matching only matches the patients in the same hospital. It achieves perfect hospital-level balance between groups, but the matching rates may be low, especially when hospital sizes are small. Across-hospital matching ignores the hierarchical structure and matches patients regardless of whether they are in the same hospital, which has been shown to perform better for reducing bias. In this chapter, we consider both within-hospital and across-hospital matching, and also propose a matching method called "modified across-hospital matching", that is, we first match patients within the same hospital, and for the patients who cannot be matched within hospital, we match them across hospitals. This is similar to the preferential matching proposed by *Arpino and Cannas* (2016).

2.2.4 Post-matching Analysis

Traditionally people perform PS matching and then estimate the ATE or ATT by calculating the difference in average outcomes between two matched groups. Unpaired or paired two-sample tests are used to assess the significance. As stated in Section 2.1, it has been suggested that post-matching covariate adjustment would potentially improve the estimation. Hence, instead of directly comparing the averages between two matched groups, we propose a method of modeling outcome using covariates after matching. Similar to PS models, we consider three types of outcome models

depending on how to account for hierarchical effects of hospitals:

$$\text{Single-level outcome model: } Y_{ij} \sim Z_{ij}\beta + X_{ij}\gamma \tag{2.3}$$

$$\text{Random-effect outcome model: } Y_{ij} \sim \alpha_i + Z_{ij}\beta + X_{ij}\gamma + V_i\theta \tag{2.4}$$

$$\text{Fixed-effect outcome model: } Y_{ij} \sim H_i + Z_{ij}\beta + X_{ij}\gamma, \tag{2.5}$$

where α_i is a random intercept for hospital i, and H_i is the indicator for hospital i.

There are still debates on whether to account for matching pairs in the post-matching analysis (*Rubin*, 1973; *Hill and Reiter*, 2006; *Schafer and Kang*, 2008; *Stuart*, 2008); we have chosen not to account for matching pairs in this chapter. In the situation that we do not observe all the hospital-level confounders, the random-effect PS model will be biased due to the endogeneity problem (*Mundlak*, 1978). In this case, the fixed-effect model is still unbiased because the effect of the missing hospital-level covariates would be absorbed into hospital indicator H_i in Model (2.5).

2.3 Simulation Study

We perform a simulation study to evaluate the performance of the proposed methods using 1000 Monte Carlo datasets. The number of hospitals is 30. We consider three scenarios by varying hospital sizes. The hospital sizes in three scenarios are 30, 100 and a random number generated from the uniform distribution (30,170), respectively.

Let X_{ij} denote the patient-level covariate vector and $X_{ij} = (X_{1,ij}, X_{2,ij}, X_{3,ij})^T$. $X_{1,ij}$ and $X_{3,ij}$ are generated from standard normal distribution $\mathrm{N}(0,1)$. $X_{2,ij}$ is generated from Bernoulli(0.6). The hospital-level covariate V_i is simulated from $\mathrm{N}(0,1)$. The treatment indicator Z_{ij} is generated from the Bernoulli distribution with parameter $\text{expit}(X_{1,ij} + X_{2,ij} + 0.5V_i + \eta_i)$, where $\eta_i \sim \mathrm{N}(0,1)$. The outcome Y_{ij} is generated

from a two-level random-effect model:

$$Y_{ij} = \alpha_i + 1.5X_{1,ij} + X_{1,ij}^2 + 2.5X_{2,ij} + 1.5X_{3,ij} + 3Z_{ij} + V_i + \epsilon_{ij},$$

where $\alpha_i \sim N(0, 1.2)$, $\epsilon_{i,j} \sim N(0, 2)$.

To estimate propensity scores, we consider five PS models: (1) the benchmark model, (2) the single-level model with linear forms of all patient-level covariates, (3) the single-level model with linear forms of all patient-level covariates except $X_{1,ij}$, (4) the random-effect model with linear forms of all patient-level covariates, and (5) the fixed-effect model with linear forms of all patient-level covariates. For the PS Models (2)-(5), we omit hospital-level covariate in order to mimic the realistic situation that we do not observe hospital-level confounders.

PS model 1 (benchmark): $\text{logit}(e_{ij}) = b_i + X_{1,ij} + X_{2,ij} + V_i$

PS model 2: $\text{logit}(e_{ij}) = X_{1,ij} + X_{2,ij} + X_{3,ij}$

PS model 3: $\text{logit}(e_{ij}) = X_{2,ij} + X_{3,ij}$

PS model 4: $\text{logit}(e_{ij}) = b_i + X_{1,ij} + X_{2,ij} + X_{3,ij}$

PS model 5: $\text{logit}(e_{ij}) = H_i + X_{1,ij} + X_{2,ij} + X_{3,ij}$.

After estimating propensity scores, we consider three ways to match patients as described in Section 2.2: within-hospital, across-hospital and modified across-hospital matching. Regarding matching with or without replacement, we also consider three ways: with replacement, without replacement and modified with replacement matching. After matching, we further perform outcome regression modeling using five models, including (1) benchmark model, (2) single-level model with Z_i as the only covariate, (3) mixed-effect model with Z_i as the only covariate, (4) random-effect model with Z_i and linear forms of all patient-level covariates, and (5) fixed-effect model with Z_i and linear forms of all patient-level covariates. For outcome model (2)-(5), we omit hospital-level covariate V_i.

Outcome model 1 (benchmark): $Y_{ij} \sim b_i + Z_{ij} + X_{1,ij} + X_{1,ij}^2 + X_{2,ij} + X_{3,ij} + V_i$

Outcome model 2: $Y_{ij} \sim Z_{ij}$

Outcome model 3: $Y_{ij} \sim b_i + Z_{ij}$

Outcome model 4: $Y_{ij} \sim b_i + Z_{ij} + X_{1,ij} + X_{2,ij} + X_{3,ij}$

Outcome model 5: $Y_{ij} \sim H_i + Z_{ij} + X_{1,ij} + X_{2,ij} + X_{3,ij}$.

In each scenario, in addition to the proposed methods, we also evaluate the outcome regression method, PS matching method and Li's doubly robust weighting methods, in terms of bias and Monte Carlo standard deviation. We estimate the standard error via the bootstrap approach. In each scenario, we mainly focus on matching without replacement; the detailed simulation results of the matching with replacement and the modified matching with replacement can be found in Appendix A.1 and A.2.

Tables 2.1-2.6 show the simulation results for matching without replacement in three scenarios. We first describe the simulation results for Scenario 1 (Table 2.1 and 2.2) from the following six aspects.

First, we show that the proposed method has the double robustness property. Double robustness means that when either the PS model or the outcome model is correctly specified, the bias of the estimation is negligible. In Table 2.1, when the PS model is wrong (PS model 3), if we directly compare the two groups without covariate adjustment after matching (outcome models 2 and 3), the biases are large. In contrast, if we perform outcome regression using patient-level covariates after matching (outcome models 1, 4 and 5), the biases are greatly reduced. Although outcome models 4 and 5 are not exactly correct (because they omit the hospital-level confounder and polynomial terms of the patient-level confounders), it still works well for reducing bias. Similarly, when the outcome models are wrong (models 2 and 3), if we use correct or almost correct PS models (models 1,4 and 5), the biases are greatly reduced. Thus, the simulation results show that when either the PS or the outcome model is almost correct, the biases of the proposed method are low.

Second, the proposed method is more efficient than PS matching without subsequent outcome modeling. As we described in Section 2.2, the PS matching method is equivalent to PS matching followed by outcome model 2 or 3. Table 2.1 shows that the proposed methods using outcome model with covariate adjustment (outcome models 1, 4 and 5) always have smaller variance compared to PS matching without subsequent covariate adjustment (outcome model 2 and 3).

Third, we show that our method is more robust than outcome regression. The biases for outcome regression are large when the outcome models are greatly misspecified (model 2 and 3), and not negligible when outcome models are almost correct (model 4 and 5). The proposed method combines outcome regression with PS matching, and if the PS models are correct or almost correct (model 1, 4 and 5), the biases are greatly reduced compared to outcome regression. Even when the PS models are very wrong (model 3), the proposed method is still less biased than outcome regression.

Fourth, we show that the proposed method has comparable or better performance than Li's doubly robust IPW method in our simulation settings. Most of the biases and variances for the proposed methods are smaller than those for Li's method conditional on same PS models and outcome models. Especially when either the PS model or outcome model is very wrong, the biases from Li's method are much larger than the proposed method.

Fifth, in Scenario 1, the biases for matching within hospital are smaller compared to across-hospital and modified across-hospital matchings. In particular, when either the PS model or outcome are very wrong (PS model 3, outcome model 2 and 3), the across-hospital and modified across-hospital matchings are much more biased compared to within-hospital matching. One possible reason is that matching within hospital achieves perfect balance on the hospital level, hence avoiding the potential hospital-level confoundings. Its variances are slightly larger compared to matching

across hospital and modified matching across hospital. A possible reason may be that many patients cannot be successfully matched within the same hospital, so we lose a large percentage of patients during matching. The low matching rate is a particular common problem when hospital sizes are small. When the hospital size increases, the variance of matching within hospital would be smaller (Scenario 2 and 3).

Sixth, the coverage probabilities for the proposed methods appear to be around 95% when using PS models 1,4 and 5 and outcome models 1,4 and 5.

Tables 2.3-2.6 show simulation results for Scenario 2 and 3 in which the hospital sizes increase to 100 and random numbers between 30 and 170, respectively. As the hospital sizes increase, the variances of the proposed methods are smaller than that in Scenario 1, making the proposed methods much more efficient than Li's method. Additionally, the variances of within-hospital matching methods are still larger than but closer to the modified across-hospital matching. The conclusion from the simulation results for Scenario 2 and 3 are generally consistent with the above seven conclusions for Scenario 1. It suggests that the proposed methods perform well in either small or large hospital sizes.

Table A.1-A.12 show simulation results for matching with replacement and modified matching with replacement in three scenarios. The conclusions from Table A.1-A.12 are generally consistent with matching without replacement. One remarkable difference is that when the sample size is 30, matching with replacement and modified matching with replacement are more efficient than matching without replacement. This makes sense because when the sample size is small, matching without replacement greatly limits the number of matchings, but matching with replacement greatly increases the chance of being matched. In contrast, when the sample size increases to 100 or to a random number between 30 and 170, matching without replacement are more efficient than matching with replacement or modified matching with replacement. This seems counter-intuitive, but actually the efficiency depends

on the number of unique patients in matched data, instead of the total number of the matched patients. In the medium or large hospitals, matching with replacement is likely to create more matching pairs but the number of unique matched patients could be smaller than matching without replacement because one patient can be matched with multiple patients.

Table 2.1: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
Outcome Only	Regression		0.004	0.168	1.764	0.327	1.679	0.253	−0.051	0.205	−0.089	0.206
Without replacement	Within hospital matching	PS Model 1	0.000	0.232	0.008	0.300	0.008	0.300	0.002	0.242	0.002	0.242
		PS Model 2	0.000	0.245	0.006	0.289	0.006	0.289	0.003	0.257	0.003	0.257
		PS Model 3	−0.004	0.206	1.189	0.261	1.189	0.261	0.002	0.237	−0.009	0.238
		PS Model 4	0.000	0.234	0.008	0.275	0.008	0.275	0.003	0.244	0.003	0.244
		PS Model 5	0.000	0.233	0.010	0.272	0.010	0.272	0.003	0.244	0.003	0.244
	Across hospital matching	PS Model 1	0.009	0.191	−0.142	0.279	−0.190	0.273	0.005	0.223	0.002	0.224
		PS Model 2	0.003	0.186	0.414	0.343	0.179	0.232	0.052	0.201	0.000	0.204
		PS Model 3	0.004	0.183	1.366	0.327	1.224	0.244	0.050	0.216	0.002	0.218
		PS Model 4	0.005	0.196	−0.085	0.253	−0.190	0.240	0.009	0.224	0.015	0.226
		PS Model 5	0.006	0.195	0.030	0.257	−0.097	0.241	0.005	0.225	0.021	0.227
	Modified across hospital matching	PS Model 1	0.005	0.191	−0.118	0.269	−0.110	0.257	0.014	0.208	0.018	0.208
		PS Model 2	0.002	0.181	0.346	0.307	0.250	0.221	0.045	0.195	0.004	0.198
		PS Model 3	0.003	0.176	1.338	0.287	1.311	0.231	0.036	0.207	−0.002	0.209
		PS Model 4	0.006	0.192	−0.078	0.235	−0.104	0.225	0.023	0.209	0.021	0.209
		PS Model 5	0.006	0.195	−0.017	0.242	−0.056	0.228	0.010	0.214	0.013	0.213
Li et al method		PS Model 1	0.008	0.192	0.166	0.428	0.201	0.405	0.003	0.282	−0.003	0.282
		PS Model 2	0.007	0.173	0.441	0.383	0.377	0.286	0.046	0.237	0.007	0.238
		PS Model 3	0.004	0.168	1.361	0.308	1.283	0.219	−0.055	0.206	−0.094	0.207
		PS Model 4	0.008	0.191	0.235	0.401	0.215	0.380	0.007	0.280	−0.003	0.280
		PS Model 5	0.007	0.222	0.006	0.530	0.008	0.494	0.004	0.334	0.004	0.333

Table 2.2: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
Outcome Only	Regression		0.089	0.945	0.137	0.000	0.136	0.000	0.108	0.883	0.109	0.862
Without replacement	Within hospital matching	PS Model 1	0.241	0.962	0.406	0.994	0.362	0.987	0.264	0.968	0.265	0.968
		PS Model 2	0.251	0.960	0.422	0.993	0.375	0.989	0.274	0.968	0.276	0.970
		PS Model 3	0.211	0.957	0.343	0.029	0.311	0.018	0.261	0.973	0.262	0.973
		PS Model 4	0.241	0.960	0.406	0.998	0.362	0.992	0.264	0.961	0.265	0.963
		PS Model 5	0.238	0.959	0.402	0.997	0.358	0.993	0.261	0.972	0.262	0.971
	Across hospital matching	PS Model 1	0.202	0.957	0.330	0.965	0.305	0.930	0.228	0.951	0.230	0.955
		PS Model 2	0.193	0.954	0.286	0.657	0.291	0.955	0.217	0.956	0.221	0.961
		PS Model 3	0.182	0.953	0.274	0.005	0.277	0.005	0.223	0.955	0.227	0.961
		PS Model 4	0.202	0.957	0.331	0.986	0.305	0.959	0.229	0.949	0.230	0.949
		PS Model 5	0.202	0.955	0.334	0.989	0.308	0.983	0.230	0.953	0.232	0.953
	Modified across hospital matching	PS Model 1	0.197	0.951	0.329	0.976	0.298	0.966	0.221	0.957	0.221	0.960
		PS Model 2	0.184	0.950	0.284	0.745	0.277	0.901	0.206	0.952	0.209	0.961
		PS Model 3	0.175	0.954	0.274	0.002	0.265	0.000	0.215	0.950	0.217	0.958
		PS Model 4	0.197	0.957	0.330	0.989	0.298	0.982	0.221	0.957	0.221	0.956
		PS Model 5	0.197	0.945	0.331	0.991	0.299	0.985	0.221	0.956	0.222	0.960
Li et al method		PS Model 1	0.172	0.918	0.429	0.888	0.392	0.864	0.235	0.923	0.235	0.919
		PS Model 2	0.151	0.912	0.340	0.674	0.315	0.726	0.204	0.911	0.204	0.916
		PS Model 3	0.136	0.886	0.251	0.002	0.227	0.000	0.168	0.882	0.167	0.853
		PS Model 4	0.171	0.917	0.421	0.861	0.389	0.864	0.233	0.926	0.233	0.926
		PS Model 5	0.207	0.935	0.521	0.954	0.479	0.952	0.281	0.939	0.280	0.936

Table 2.3: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 100.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
Outcome Only	Regression		0.000	0.088	1.751	0.252	1.657	0.138	−0.083	0.108	−0.096	0.108
Without replacement	Within hospital matching	PS Model 1	0.001	0.104	0.016	0.134	0.016	0.134	0.004	0.108	0.004	0.108
		PS Model 2	−0.004	0.106	0.005	0.126	0.005	0.126	−0.001	0.110	−0.001	0.110
		PS Model 3	0.000	0.099	1.190	0.134	1.190	0.134	0.001	0.113	−0.002	0.112
		PS Model 4	−0.003	0.103	0.008	0.117	0.008	0.117	−0.001	0.108	−0.001	0.108
		PS Model 5	−0.003	0.103	0.010	0.117	0.010	0.117	0.000	0.108	0.000	0.108
	Across hospital matching	PS Model 1	0.001	0.098	−0.034	0.173	−0.143	0.151	0.010	0.120	0.015	0.121
		PS Model 2	0.000	0.097	0.409	0.288	0.124	0.130	0.010	0.104	−0.006	0.104
		PS Model 3	0.000	0.093	1.366	0.261	1.192	0.139	0.010	0.108	−0.005	0.108
		PS Model 4	−0.001	0.100	−0.012	0.162	−0.142	0.134	0.010	0.121	0.014	0.121
		PS Model 5	−0.001	0.099	0.024	0.161	−0.113	0.133	0.009	0.120	0.014	0.120
	Modified across hospital matching	PS Model 1	0.001	0.098	−0.022	0.133	−0.015	0.129	0.017	0.106	0.017	0.106
		PS Model 2	0.000	0.094	0.292	0.218	0.266	0.121	0.014	0.100	0.003	0.100
		PS Model 3	0.000	0.092	1.318	0.206	1.311	0.132	0.000	0.105	−0.012	0.105
		PS Model 4	−0.001	0.099	−0.018	0.117	−0.019	0.111	0.014	0.106	0.014	0.106
		PS Model 5	−0.002	0.098	−0.005	0.116	−0.009	0.111	0.010	0.105	0.010	0.105
Li et al method		PS Model 1	0.001	0.114	0.043	0.406	0.055	0.377	−0.006	0.224	−0.007	0.223
		PS Model 2	0.002	0.090	0.424	0.298	0.325	0.165	0.011	0.123	−0.002	0.123
		PS Model 3	0.000	0.088	1.351	0.247	1.257	0.125	−0.088	0.110	−0.100	0.109
		PS Model 4	0.001	0.113	0.069	0.392	0.058	0.366	−0.005	0.219	−0.007	0.218
		PS Model 5	0.000	0.129	−0.025	0.528	−0.023	0.487	−0.006	0.279	−0.006	0.277

Table 2.4: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 100.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
Outcome Only	Regression		0.089	0.945	0.137	0.000	0.136	0.000	0.108	0.883	0.109	0.862
Without replacement	Within hospital matching	PS Model 1	0.111	0.972	0.189	0.992	0.170	0.988	0.124	0.977	0.124	0.977
		PS Model 2	0.112	0.966	0.191	0.996	0.172	0.992	0.125	0.977	0.125	0.977
		PS Model 3	0.105	0.963	0.173	0.000	0.158	0.000	0.130	0.973	0.130	0.973
		PS Model 4	0.111	0.971	0.189	0.998	0.171	0.996	0.124	0.978	0.124	0.978
		PS Model 5	0.111	0.966	0.189	0.998	0.170	0.996	0.124	0.976	0.124	0.976
	Across hospital matching	PS Model 1	0.108	0.965	0.178	0.959	0.165	0.880	0.122	0.952	0.123	0.950
		PS Model 2	0.104	0.959	0.155	0.371	0.159	0.920	0.117	0.962	0.118	0.967
		PS Model 3	0.099	0.959	0.150	0.000	0.152	0.000	0.121	0.974	0.122	0.971
		PS Model 4	0.108	0.965	0.178	0.970	0.165	0.912	0.122	0.951	0.123	0.950
		PS Model 5	0.108	0.971	0.179	0.967	0.165	0.945	0.123	0.957	0.123	0.952
	Modified across hospital matching	PS Model 1	0.104	0.970	0.177	0.993	0.160	0.987	0.118	0.972	0.118	0.972
		PS Model 2	0.097	0.960	0.154	0.529	0.147	0.579	0.109	0.971	0.109	0.970
		PS Model 3	0.094	0.957	0.150	0.000	0.144	0.000	0.115	0.966	0.116	0.967
		PS Model 4	0.104	0.964	0.177	0.996	0.160	0.998	0.118	0.973	0.118	0.972
		PS Model 5	0.104	0.960	0.177	0.996	0.160	0.997	0.118	0.972	0.118	0.971
Li et al method		PS Model 1	0.106	0.935	0.284	0.950	0.264	0.932	0.155	0.940	0.155	0.940
		PS Model 2	0.084	0.925	0.188	0.422	0.176	0.517	0.114	0.928	0.114	0.926
		PS Model 3	0.075	0.899	0.137	0.000	0.125	0.000	0.093	0.807	0.093	0.774
		PS Model 4	0.106	0.935	0.282	0.937	0.263	0.932	0.155	0.938	0.155	0.938
		PS Model 5	0.116	0.942	0.315	0.978	0.293	0.966	0.171	0.943	0.171	0.943

Table 2.5: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD
Outcome Only	Regression		0.000	0.090	1.736	0.267	1.660	0.139	−0.088	0.112	−0.099	0.112
Without replacement	Within hospital matching	PS Model 1	−0.001	0.107	0.011	0.140	0.011	0.140	0.002	0.113	0.002	0.113
		PS Model 2	−0.001	0.107	0.007	0.128	0.007	0.128	0.000	0.113	0.000	0.113
		PS Model 3	−0.001	0.100	1.188	0.131	1.188	0.131	−0.001	0.113	−0.004	0.113
		PS Model 4	−0.001	0.104	0.012	0.120	0.012	0.120	0.002	0.111	0.002	0.111
		PS Model 5	−0.001	0.105	0.014	0.121	0.014	0.121	0.002	0.111	0.002	0.111
	Across hospital matching	PS Model 1	0.002	0.102	−0.032	0.184	−0.142	0.160	0.010	0.124	0.015	0.125
		PS Model 2	0.002	0.099	0.374	0.309	0.118	0.136	0.011	0.105	−0.003	0.105
		PS Model 3	0.001	0.097	1.344	0.279	1.194	0.135	0.012	0.113	−0.001	0.114
		PS Model 4	0.001	0.103	−0.008	0.170	−0.139	0.141	0.012	0.124	0.015	0.125
		PS Model 5	0.002	0.101	0.029	0.172	−0.109	0.142	0.011	0.124	0.016	0.125
	Modified across hospital matching	PS Model 1	−0.001	0.101	−0.026	0.143	−0.020	0.137	0.012	0.109	0.012	0.108
		PS Model 2	0.000	0.095	0.265	0.232	0.256	0.129	0.012	0.102	0.002	0.102
		PS Model 3	−0.001	0.094	1.299	0.214	1.306	0.132	−0.002	0.109	−0.012	0.109
		PS Model 4	−0.001	0.098	−0.017	0.119	−0.019	0.114	0.012	0.107	0.012	0.106
		PS Model 5	−0.001	0.099	−0.003	0.120	−0.008	0.115	0.009	0.107	0.009	0.107
Li et al method		PS Model 1	0.001	0.109	0.060	0.247	0.075	0.235	0.000	0.162	−0.001	0.161
		PS Model 2	−0.002	0.094	0.391	0.327	0.314	0.167	0.005	0.130	−0.006	0.130
		PS Model 3	0.000	0.090	1.326	0.263	1.251	0.124	−0.093	0.113	−0.104	0.113
		PS Model 4	0.001	0.109	0.085	0.239	0.078	0.226	0.000	0.161	−0.001	0.161
		PS Model 5	0.001	0.116	0.005	0.265	0.008	0.250	0.003	0.173	0.003	0.172

Table 2.6: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching without replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
Outcome Only	Regression		0.088	0.952	0.137	0.000	0.136	0.000	0.108	0.860	0.108	0.840
Without replacement	Within hospital matching	PS Model 1	0.111	0.955	0.189	0.993	0.170	0.987	0.124	0.964	0.124	0.964
		PS Model 2	0.112	0.959	0.190	0.998	0.171	0.996	0.125	0.965	0.125	0.965
		PS Model 3	0.104	0.958	0.173	0.000	0.158	0.000	0.129	0.976	0.129	0.973
		PS Model 4	0.111	0.968	0.189	1.000	0.170	0.997	0.124	0.970	0.124	0.971
		PS Model 5	0.111	0.965	0.188	0.999	0.170	0.995	0.124	0.967	0.124	0.968
	Across hospital matching	PS Model 1	0.108	0.963	0.178	0.939	0.165	0.870	0.122	0.944	0.123	0.942
		PS Model 2	0.104	0.962	0.155	0.422	0.158	0.932	0.117	0.963	0.117	0.964
		PS Model 3	0.099	0.951	0.150	0.000	0.152	0.000	0.121	0.957	0.122	0.962
		PS Model 4	0.108	0.959	0.178	0.961	0.165	0.889	0.122	0.942	0.123	0.942
		PS Model 5	0.108	0.970	0.179	0.965	0.165	0.930	0.123	0.940	0.123	0.940
	Modified across hospital matching	PS Model 1	0.104	0.955	0.177	0.984	0.160	0.977	0.118	0.960	0.118	0.960
		PS Model 2	0.097	0.955	0.154	0.575	0.147	0.602	0.109	0.966	0.109	0.966
		PS Model 3	0.094	0.951	0.150	0.000	0.144	0.000	0.115	0.958	0.116	0.956
		PS Model 4	0.104	0.968	0.177	0.997	0.160	0.993	0.118	0.969	0.118	0.970
		PS Model 5	0.104	0.961	0.177	0.996	0.160	0.994	0.118	0.968	0.118	0.968
Li et al method		PS Model 1	0.104	0.936	0.264	0.930	0.246	0.904	0.146	0.933	0.145	0.932
		PS Model 2	0.084	0.928	0.189	0.465	0.178	0.548	0.115	0.915	0.115	0.920
		PS Model 3	0.075	0.909	0.137	0.000	0.125	0.000	0.093	0.782	0.093	0.756
		PS Model 4	0.104	0.934	0.263	0.931	0.246	0.925	0.145	0.932	0.145	0.931
		PS Model 5	0.112	0.935	0.286	0.954	0.266	0.949	0.156	0.940	0.156	0.940

2.4 Application

Cardioplegia is a blood or crystalloid based solution that induces myocardial electrical silence and provides buffers and basic cellular nutrients to myocardium during cardiac surgery. Del Nido cardioplegia (DC) and blood-based cardioplegia (BC) are two types of cardioplegia used in cardiac surgery. We apply our methods to compare the post-operative length of stay (LOS) (days) between DC and BC using multi-center clinical data. The data contain 11 hospitals and 14,339 patients in total. The hospital sizes vary between 40 and 4,211 and median hospital size is 1,077. The patients received either DC ($n = 5,005$) or BC ($n = 9,394$). The possible patient characteristics associated with the post-operative LOS and treatment assignment are age, gender, hypertension, diabetes, previous cardiovascular disease and etc. Imbalances of the patient characteristics and post-operative LOS were observed between hospitals (Table 2.7).

Table 2.7: Distribution of post-operative length of stay (days), age, gender, hypertension, diabetes and previous cardiovascular disease by hospital.

Hospital	Post-operative LOS (mean)	Age (mean)	Female (%)	Hypertension (%)	Diabetes (%)	Previous cardiovascular disease (%)
1	7.4	62.6	34.4	71.9	27.0	14.0
2	6.3	65.2	30.0	90.0	40.0	7.5
3	10.7	63.0	31.1	84.9	42.5	5.3
4	7.6	65.6	28.9	85.2	35.9	5.9
5	7.5	69.1	31.2	85.1	41.6	3.5
6	7.3	66.7	32.1	92.1	43.3	5.8
7	7.6	66.7	30.8	87.4	42.9	3.6
8	7.2	67.5	31.2	81.9	38.4	6.0
9	8.3	64.7	33.4	85.6	40.9	4.4
10	12.2	67.3	30.3	89.4	39.4	7.6
11	8.2	63.1	37.1	85.4	46.7	7.3

We perform the data analysis using the proposed method. According to the simulation results, we choose matching within hospitals without replacement as it generally has the best performance when hospital sizes are larger than 30. We use the fixed-effect PS model and outcome models because we assume there were unmeasured hospital-level confounders. We include the linear forms of all the possible confounders mentioned above in both the PS and outcome models. Table 2.8 shows the results of the analyses. The proposed method shows that the post-operative LOS for DC is slightly longer than that for BC although the difference is not significant.

The standard error of the proposed method is smaller compared to other methods. δ: difference in the post-operative LOS (DC minus BC). SE: standard error.

Table 2.8: Estimation of the difference in the post-operative LOS (days) between DC and BC.

Method	$\widehat{\delta}$	$\widehat{SE}(\widehat{\delta})$	P-value
Outcome regression	0.001	0.117	0.503
PS matching	−0.058	0.118	0.311
Li et al method	0.110	0.127	0.193
Proposed method	0.068	0.114	0.275

2.5 Discussion

In this chapter, we propose doubly robust PS matching methods for estimating the ATE in clustered observational data. The proposed methods combine PS matching and outcome regression while accounting for the hierarchical nature of the data. We further study and compare various matching methods for clustered data, including matching within hospital, across hospital and modified across hospital. The simulation results show that the proposed methods perform well for either small or large cluster sizes.

We show that as long as we fit either PS model or outcome model correct or relatively correct, the bias of proposed method is negligible. We recommend to use hierarchical models instead of single-level models for causal inference in clustered data. It is important to include as many important covariates as possible when fitting PS or outcome model. Regarding the matching approaches, matching within hospital has best performance on reducing bias. Also if people suspect there are important hospital-level confounders unobserved which likely happen in reality, we recommend matching within hospital. As we stated in Section 2.4, matching with replacement is less efficient when hospital sizes are medium or large. Our results are consistent with

the previous research by *Austin* (2014) which claimed that matching with replacement does not reduce or even have larger standard errors of the ATE estimates compared to matching without replacement. Yet when hospital sizes are very small, matching with replacement should be a better choice for reducing variance. In this chapter, we did not study how to decide the caliper to achieve best baseline balance between groups in this chapter.

We use bootstrap approach to calculate standard errors. *Abadie and Imbens* (2008) claimed that bootstrap standard errors for matching estimator are not valid and proposed an alternative way for standard error for nearest neighbor matching estimator. However, we use caliper matching and consider matching in clustered data, which make the estimation of standard error much more challenging. When we use matching with replacement, the coverage probabilities seem not close to the nominal level, which suggests that the bootstrap standard errors are more problematic when patients have different weights in the matched data. Although we did not propose a method to solve this in this chapter, it may be an interesting research topic in future.

CHAPTER III

Augmented Double Inverse-Weighted Method for Causal Inference Based on Restricted Mean Lifetime

3.1 Introduction

In medical studies, it is often of interest to compare the survival outcomes between treatment groups using observational data. In observational studies, the distributions of patient baseline covariates are often different between treatment groups. To estimate the treatment effect, it is therefore necessary to adjust for covariate imbalances. Since being proposed in 1972 (*Cox*, 1992), the Cox model has been a dominant method of covariate adjustment in survival analysis and is often used to compare hazards between groups. However, if the proportional hazard assumption is violated, the hazard ratio estimated from the Cox model is difficult to interpret. In such cases, it is more reasonable to compare survival times instead of hazards between groups. As the durations of medical studies are often finite, the time to the event of interest may be administratively censored. Therefore, the restricted mean lifetime instead of mean lifetime has been widely used in medical research.

This Chapter is motivated by a study of cardiac surgery, in which we were interest-

ed in comparing the restricted mean lifetime between two types of cardiac surgeries for patients with advanced heart failure (AHF). Heart transplants used to be the dominant treatment for AHF. However, during the last decade, the implantation of ventricular assist devices (VAD) has grown dramatically and become the most popular treatment. Depending on patients' clinical conditions, VAD implantations can be performed on either the left side only (LVAD) or both sides (BiVAD) of the heart. The cardiac surgeons were interested in comparing the LVAD and BiVAD in terms of restricted mean lifetime. However, the comparison was challenging due to two reasons. The first one was that two types of VAD implantations were not randomized, which meant that the baseline patient covariates were not balanced between groups; patients receiving BiVAD had more severe conditions than those receiving LVAD. For example, patients were more likely to receive BiVAD implantation if s/he was experiencing the critical cardiogenic shock or her/his central venous pressure was greater than 15 mmHg before the surgery. The second reason was that it is not uncommon that patients receive heart transplants after the VAD implantations, which may not be balanced between the two groups. We knew that the patients with a worse post-implant situation were more likely to receive heart transplants. For example, it has been shown that the patients with high post-operative creatinine and albumin were more likely to need a subsequent heart transplant compared to other patients. Hence, the time of receiving heart transplant was not independent of survival time given treatment assignment and patient baseline covariates. Various approaches have been proposed to handle the events like transplant (*Zhang and Wang*, 2012, 2013). In this chapter, we treat the heart transplant as dependent censoring and then estimate the treatment-specific difference in restricted mean lifetimes, had no patients received heart transplants. To reach this goal, we need to handle both group-level baseline imbalance and dependent censoring.

Generally, two strategies can be used to overcome these two challenges. The first

strategy is directly modeling the relationship between survival time and covariates using group-specific models, then average the fitted mean lifetimes over the entire sample (*Chen and Tsiatis*, 2001; *Zhang and Schaubel*, 2011). *Zhang and Schaubel* (2011) further improved this strategy by incorporating weighting methods to accommodate the dependent censoring. The second strategy is using inverse probability weighting to equalize group-specific covariate distributions and handle dependent censoring. Instead of modeling survival outcome, this strategy only requires modeling treatment assignment and censoring. In our motivating study, the relationship between survival time and covariates was complicated and hence difficult to model. However, the treatment assignment and transplant decision were much better understood. The surgeons typically have standard protocols to decide which treatment to use according to the patient's conditions. Therefore, we prefer the second strategy.

Inverse probability of treatment weighting (IPTW) was proposed by *Robins et al.* (1994) to adjust baseline covariate imbalances between treatment groups. To solve the dependent censoring problem, inverse probability of censoring weighting (IPCW) was proposed by *Robins and Rotnitzky* (1992). To addressing both baseline imbalances and dependent censoring, several researchers have proposed combining IPTW and IPCW. *Anstrom and Tsiatis* (2001) proposed a double weighted method for time-lagged data. *Schaubel and Wei* (2011) proposed a double weighted estimator to estimate the cumulative hazard and restricted mean lifetime. *Zhang and Schaubel* (2012b) developed a double robust estimator using both IPTW and IPCW to estimate the restricted mean lifetime. However, one disadvantage of inverse probability weighting (IPW) methods is inefficiency, for which various strategies have been proposed. For example, *Zhang et al.* (2008) proposed a broadly applicable augmentation approach to improve efficiency using baseline auxiliary covariates. More recently, *Zhang* (2015) developed a robust method to use patient covariates to improve efficiency in randomized clinical trials which made no assumptions of the relationship

between survival outcome and covariates.

In this chapter, we propose a method to estimate the treatment-specific difference in potential restricted mean lifetimes, had no patients received heart transplant, where the heart transplant is treated as dependent censoring for the potential lifetime. Specifically, we first derive an estimator that combines IPTW and IPCW to account for the imbalance in baseline characteristics and receipt of heart transplant, respectively. Then we propose augmentation method to improve the efficiency of the estimation.

The remainder of this chapter is organized as follows. In Section 3.2, we describe the notation and data structure. In Section 3.3, we propose our method and state its connection to existing methods. We evaluate the performance of our method using a simulation study in Section 3.4. In Section 3.5, we apply the method to cardiac surgery data obtained from the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS). Finally, we conclude and discuss in Section 3.6.

3.2 Notation and Data Structure

Let A indicate two non-randomized treatment groups ($A = 1$ means treatment, $A = 0$ means control). We denote the survival time without heart transplant by T. Let C denote dependent censoring (heart transplant). In practice, one observes the minimum of the survival time and time to censoring. We let $U = T \wedge C$ denote the observation time and $\Delta = I(T \leq C)$ denote the indicator for observing the death time. We let Z be the baseline covariate vector, and $Z(t)$ be the time-dependent covariate at time t. Note that $Z(0)$ would be the elements of Z. We let $\tilde{Z}(t) = \{Z(u); u \in [0, t)\}$ be the history of the baseline and time-dependent covariates up to time t. The observed data can be summarized as $O_i = \{A_i, U_i, \Delta_i, Z_i, \tilde{Z}_i(u_i)\}$, where the O_i is assumed to be identically and independent distributed (i.i.d.) across $i = 1, ..., n$. Note that

the O_i is redundant because $\tilde{Z}(t)$ includes all the baseline covariate Z_i, but it is more convenient for presentation. We denote the observed event process and at-risk process by $N_i(t) = I(U_i \leq t, \Delta = 1)$ and $Y_i(t) = I(U_i \geq t)$, respectively.

We define the parameter of interest using the potential outcome framework proposed by *Rubin* (1974, 1978). Let T^k (k = 0,1) denote the potential lifetime for a randomly selected patient from the population if, possibly contrary to fact, s/he were assigned to group k, and not censored by a heart transplant. We make the assumption that there are no unmeasured baseline confounders. We also assume that C_i is conditionally independent of T_i given $\{A_i, Z_i, \tilde{Z}_i(t)\}$, formally expressed as

$$\begin{aligned}
&\lim_{\xi \to 0} \xi^{-1} P\{t \leq U_i < t + \xi, \Delta_i = 0 | U_i \geq t, A_i, \tilde{Z}_i(t), T_i\} \\
&\qquad = \lim_{\xi \to 0} \xi^{-1} P\{t \leq U_i < t + \xi, \Delta_i = 0 | U_i \geq t, A_i, \tilde{Z}_i(t)\}.
\end{aligned}$$

Our estimand δ is the difference in restricted mean lifetime up to time L between two treatment groups, had no patients received heart transplants. Let μ_k denote the restricted mean lifetime for group k and $S_k(t) = P(T^k > t)$, then

$$\begin{aligned}
\delta &= \mu_1 - \mu_0 \\
&= E\{\min(T^1, L)\} - E\{\min(T^0, L)\} \\
&= \int_0^L S_1(t)dt - \int_0^L S_0(t)dt.
\end{aligned}$$

$S_k(t)$ can be estimated as by $\exp\{-\Lambda_k(t)\}$, where $\Lambda_k(t)$ is the cumulative hazard for T^k.

3.3 Methods

3.3.1 Existing Methods

We first introduce existing methods and their connections with the proposed method. Let $\Lambda_k(t)$ denote the marginal cumulative hazard function of T^k. We estimate the group-specific restricted mean lifetime μ_k by $\int_0^t \exp\{-\widehat{\Lambda}_k(u)\}du$. If we assume that, possibly contrary to fact, all patients were assigned to group k, and survival time is independent of censoring given their treatment assignment, i.e. $T \perp\!\!\!\perp C|A$. Based on these assumptions, we can use the Nelson-Aalen estimator to estimate for the marginal cumulative hazard for group k Λ_k

$$\widehat{\Lambda}_k^{NA}(u) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u)}.$$

Actually, the Nelson-Aalen estimator can be viewed as a solution of the following estimating equation

$$\sum_{i=0}^{n} \{dN_{ik}(t) - Y_{ik}(t)d\Lambda_k(t)\} = 0.$$

In observational studies, not everyone is assigned to group k and there are baseline imbalances between groups. If we assume C is independent of T given A and there are no unmeasured baseline confounder, the unbiased IPTW estimating equation for $d\Lambda_k(t)$ can be generated as follows:

$$\sum_{i=1}^{n} \frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})} = 0, \tag{3.1}$$

where $A_{ik} = I(A_i = k)$ and $p_{ik}(\widehat{\theta})$ estimates $Pr(A_{ik} = 1|Z_i)$ through a model with parameter θ. By solving the Equation (3.1), the IPTW estimator for $\Lambda_k(t)$ is

$$\widehat{\Lambda}_k^{IPTW}(t) = \int_0^t \frac{\sum_{i=1}^n w_{ik}(\widehat{\theta}) dN_i(u)}{\sum_{i=1}^n w_{ik}(\widehat{\theta}) Y_i(u)},$$

where $w_{ik}(\widehat{\theta}) = I(A_{ik} = 1)/p_{ik}(\widehat{\theta})$.

In observational studies, the assumption that C is independent of T given A is too restrictive. A more realistic assumption would be that C is independent of T given $\{A, Z, \tilde{Z}(t)\}$ as assumed in this chapter. Under this assumption, the IPTW estimator is biased as it does not account for the dependent censoring. As IPCW can be used to handle dependent censoring, if we can incorporate IPTW and IPCW in a proper way, we can obtain an unbiased estimator.

Before explaining the proposed method, we will first introduce the coarsening concept first. We say that the full data one would like to observe are coarsened because of treatment assignment and censoring (*Tsiatis*, 2007). For example, if $A_{ik} = 0$, T_i^k is completely missing, we say it is most coarsened; if $A_{ik} = 1$, $C_i = t < T_i^k$, T_i^k is partially observed, we say it is less coarsened; if $A_{ik} = 1$, $C_i \geq T_i^k$, T_i^k is completely observed, it is not coarsened. By *Tsiatis* (2007), one can inverse weight an unbiased estimating function based on full data by the probability of observing the complete case (not being coarsened), i.e. the probability of being assigned to treatment k and not being censored. Based on this principle, we can combine IPTW and IPCW into a single inverse probability weighting framework and derive the unbiased double weighted estimating equation for $d\Lambda_k(t)$, that is:

$$\sum_{i=1}^n \frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta}) p_i^c(\widehat{\gamma}, t)} = 0,$$

where $p_i^c(\widehat{\gamma}, t)$ estimates $Pr\{C_i \geq t|A_i, Z_i, \tilde{Z}_i(t)\}$ through a model like the Cox model

with parameter γ. By solving the estimating equation, the double inverse probability weighted (DIPW) estimator for $\Lambda_k(t)$ can be obtained:

$$\widehat{\Lambda}_k^{DIPW}(t) = \int_0^t \frac{\sum_{i=1}^n w_{ik}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) dN_i(u)}{\sum_{i=1}^n w_{ik}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) Y_i(u)}.$$

Here $w_i^c(\widehat{\gamma}, t) = I(C_i \geq t)/p_i^c(\widehat{\gamma}, t)$. Please note that, $\widehat{\Lambda}_k^{DIPW}(t)$ are consistent only if the models for $P(A_{ik} = 1|Z_i)$ and $P\{C_i \geq t|A_i, Z_i, \tilde{Z}_i(t)\}$ are correctly specified. As stated in Section 3.1, we assume that we are able to correctly specify both models in our motivating study, and therefore we can consistently estimate $\Lambda_k(t)$ using $\widehat{\Lambda}_k^{DIPW}(t)$.

3.3.2 Proposed Augmentation Method

So far we have introduced IPW approaches that use baseline and time-varying covariates to reweight patients in order to equalize the covariate distributions between groups. As we stated in Section 3.1, one drawback of the simple inverse probability weighting method is inefficiency. In order to improve the efficiency of the DIPW estimator, we propose an augmentation method. Motivated by *Tsiatis* (2007) and *Zhang et al.* (2008), we construct unbiased augmented double weighted estimating equation by adding an augmentation term with expectation equal to zero to the double weighted estimating equation:

$$\sum_{i=1}^n \left[\frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta}) p_i^c(\widehat{\gamma}, t)} - \frac{A_{ik} - p_{ik}(\theta)}{p_{ik}(\widehat{\theta})} h_k(t, Z_i) dt \right] = 0. \tag{3.2}$$

In this equation, $h_k(t, Z_i)$ is an arbitrary function of baseline covariate Z_i at time t. The expectation of the second part of this equation is 0, because $E\big[\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})} h_k(t, Z_i)dt\big] = E\big[E\big\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})} h_k(t, Z_i)dt\big|Z_i\big\}\big] = E\big[h_k(t, Z_i)dtE\big\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}\big|Z_i\big\}\big] = 0$. Therefore Equation (3.2) is still an unbiased estimating equation. The rationale for adding the sec-

ond part of the left part of Equation (3.2), is that if we can find an 'optimal' $h_k(t, Z_i)$, we may reduce the standard deviation of the resulting estimator. It is straightforward to show that the expectation of $\left[\frac{A_{ik}\{dN_i(t)-Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)} - \frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}h_k(t, Z_i)dt\right]$ is 0. Hence the estimators derived from Equation (3.2) are M-estimators. According to the theory of M-estimators, we show that the asymptotic variance of the estimator derived from this equation is

$$\frac{E\left[\frac{A_{ik}\{dN_i(t)-Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)} - \frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}h_k(t, Z_i)dt\right]^2}{E^2\left[\frac{A_{ik}Y_i(t)}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\right]}.$$

The optimal $h_k(t, Z_i)$ corresponds to the one that minimizes $E\left[\frac{A_{ik}\{dN_i(t)-Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)} - \frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}h_k(t, Z_i)dt\right]^2$, which is the variance of the left part of Equation (3.2).

By straightforward algebra, we can show that $h_k(t, Z_i)$ that minimizes the above asymptotic variance is $E(dM_{ik}|Z_i, A_{ik} = 1)$. This solution involves a conditional expectation, which needs to be estimated from a model that we inevitably misspecify in reality. If we misspecify the model, the efficiency is not guaranteed to be improved, and may be even worsen. Thus, this may not be a good strategy for improving the efficiency. Instead, we apply the augmentation strategy proposed by *Zhang* (2015). Specifically, we consider a subclass where $h_k(t, Z_i)$ is of the form $\beta_k^T(t)g_k(Z_i)$, and where $g_k(Z_i)$ are basis functions in Z_i, including intercept, linear term, possibly polynomial terms and interaction terms in Z_i. Then we derive the optimal $h_k(t, Z_i)$ in this subclass, which is equivalent to finding the optimal $\beta_k(t)$. Therefore we only need to identify $\beta_k(t)$ to minimize the variance of the estimating equation. We show that the optimal $\beta_k(t)$ that minimizes the variance of the estimating equation can be estimated by ordinary least square method, where we treat $\left[\frac{A_{ik}\{dN_i(t)-Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\right]$ as the outcome, and $\left\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\right\}$ as covariates. We show that the optimal $\beta_k(t)dt$

is

$$\beta_{k,opt}(t)dt = E^{-1}\Bigg[\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}^T\Bigg|A_{ik}=1\Bigg]$$
$$E\Bigg[\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}\frac{A_{ik}\{dN_i(t)-Y_i(t)d\Lambda_k(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\Bigg|A_{ik}=1\Bigg],$$

and can be consistently estimated by

$$\widehat{\beta}_{k,opt}(t)dt = \Bigg[\sum_{i=1}^{n}\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}^T\Bigg]^{-1}$$
$$\Bigg[\sum_{i=1}^{n}\Bigg\{\frac{A_{ik}-p_{ik}(\widehat{\theta})}{p_{ik}(\widehat{\theta})}g_k(Z_i)\Bigg\}\frac{A_{ik}\{dN_i(t)-Y_i(t)d\widehat{\Lambda}_k^{DIPW}(t)\}}{p_{ik}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\Bigg].$$

if we correctly model treatment assignment and dependent censoring. Note that since we do not know $d\Lambda_k(t)$, we substitute it by its unbiased estimator $d\widehat{\Lambda}_k^{DIPW}(t)$. The estimator developed using this strategy is guaranteed to improve efficiency compared to the double inverse weighted estimator. The reason is that when $\beta_k(t) = 0$, the estimating equation reduces to the DIPW estimating equation. In other words, this subclass includes the DIPW estimator as a special case. Since we minimize variance within this subclass, the optimal estimator must be equivalent to or better than the DIPW estimator, in terms of efficiency.

3.3.3 Proposed Augmented Double Inverse Weighted Estimators

We fit models for treatment assignment and censoring conditional on A_i and Z_i. First, we model treatment assignment by logistic regression:

$$\text{logit}\{P(A_{ik} = 1|Z_i)\} = \theta^T X_i,$$

where $X_i = (1, Z_i)$. We assume that we correctly model the treatment assignment, thus the maximum likelihood estimator for θ, $\widehat{\theta}$ consistently estimate θ. Secondly, we assume the hazard model of censoring is a Cox model:

$$\lambda(t|V_i) = \lambda_0(t)\exp(\gamma^T V_i),$$

where $V_i = \{A_i, Z_i, Z_i(t)\}$. We assume it is correctly specified, thus $\Lambda_0(t)$ and γ can be consistently estimated by the Breslow estimator and the maximum partial likelihood estimator, respectively. Therefore, the proposed augmented double inverse probability weighted (ADIPW) estimator for $\Lambda_k(t)$ is

$$\widehat{\Lambda}_k^{ADIPW}(t) = \int_0^t \frac{\sum_{i=1}^n w_{ik}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) dN_i(u) + \{1 - w_{ik}(\widehat{\theta})\}\widehat{\beta}_{k,opt}(u) g(Z_i, u) du}{\sum_{i=1}^n w_{ik}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) Y_i(u)}. \tag{3.3}$$

$S_k(t)$ can be estimated by $\widehat{S}_k(t) = e^{-\widehat{\Lambda}_k(t)}$, $\mu_k(t)$ can be estimated by $\widehat{\mu}_k = \int_0^L \widehat{S}_k(u)du$, and hence $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_0$. The asymptotic properties of the proposed estimator are derived in Appendix B.1.

3.4 Simulation Study

We perform a simulation study to evaluate the performance of our proposed methods. The reported results are based on 1000 Monte Carlo data sets. The sample size n is 500 for each data set.

Each Monte Carlo data set is generated as follows: baseline covariate vector $Z = (Z_1, Z_2, Z_3)^T$, with Z_1, Z_2 and Z_3 generated from multivariate normal distribution with mean 0, unit variance and all pairwise correlations 0. All the normal covariates are truncated at ± 3 to satisfy the regularity conditions. The treatment indicator A is generated from the Bernoulli distribution with parameter $\text{expit}(1.5Z_1 - Z_2)$. The censoring time C is generated from the exponential distribution with parameter $0.15\exp(1.5Z_3 + 0.3Z(t))$. To generate the time-varying covariate $Z(t)$, we first generate a latent variable V from the standard normal distribution, then generate $Z(t)$ as a binary variable depending on V and A, $Z(t) = \text{I}(0.5 + (5 + V - A)t \geq 5)$. $t^* = 4.5/(t + V - A)$ is the time that $Z(t)$ jumps from 0 to 1. The survival time T is generated from the three scenarios below.

The details of generating survival time T are in Appendix B.2. Briefly, in the first scenario, the survival time T is generated from a Cox model with an exponential baseline survival time distribution. In the second scenario, T is generated from an accelerated failure time (AFT) model with lognormal baseline survival time distribution. In the third scenario, T is generated from a Cox model with lognormal baseline survival time distribution.

In our simulation, we use the correct treatment assignment model and censoring model for the estimation of the restricted mean lifetime. Specifically, we use the treatment assignment model

$$\text{logit}\{P(A_i = 1|Z_i)\} = \beta_1 Z_1 + \beta_2 Z_2,$$

and the Cox model for censoring

$$\lambda^c(t|Z_i, Z_i(t), A_i) = \lambda_0^c(t)\exp\{\gamma_1 Z_{3,i} + \gamma_2 Z_i(t)\},$$

where $\lambda^c(t|Z_i, Z_i(t), A_i)$ is the conditional hazard for censoring. For the proposed method, when we do augmentation, we let $g_k(Z, u)$ be a basis function including intercept and linear forms of all baseline covariates, i.e. $g_k(Z, u) = (1, Z_1, Z_2, Z_3)^T$. We estimate the standard error via bootstrap approach. In our simulation study, we are interested in the restricted mean lifetime up to $L = 1.5$. We include a detailed practical implementation procedure in Appendix B.3.

Table 3.1: Estimation of the difference in restricted mean lifetimes between treatment and control. The results are based on 1000 Monte Carlo datasets. Sample size $n = 500$.

	$\widehat{\delta}$					$\widehat{\mu}_1$				$\widehat{\mu}_0$			
Method	Bias	ESD	ASE	CP	RE	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP
Scenario 1													
N-A	−0.457	0.053	0.053	0.00		−0.175	0.038	0.038	0.00	0.282	0.038	0.036	0.00
IPTW	−0.056	0.074	0.084	0.90		0.016	0.046	0.049	0.96	0.072	0.069	0.067	0.75
DIPW	−0.026	0.084	0.090	0.94	1	0.000	0.050	0.052	0.96	0.027	0.077	0.073	0.91
Proposed	0.008	0.077	0.083	0.96	1.29	0.015	0.048	0.055	0.95	0.007	0.070	0.066	0.92
Scenario 2													
N-A	−0.319	0.047	0.046	0.00		−0.115	0.035	0.034	0.07	0.204	0.031	0.032	0.00
IPTW	−0.050	0.077	0.072	0.85		0.016	0.043	0.040	0.93	0.065	0.070	0.059	0.70
DIPW	−0.024	0.095	0.086	0.91	1	−0.001	0.047	0.043	0.94	0.023	0.087	0.070	0.87
Proposed	0.001	0.090	0.087	0.93	1.19	0.010	0.044	0.046	0.95	0.009	0.082	0.075	0.92
Scenario 3													
N-A	−0.430	0.048	0.047	0.00		−0.157	0.034	0.033	0.00	0.273	0.034	0.034	0.00
IPTW	−0.066	0.070	0.075	0.85		0.021	0.047	0.047	0.95	0.086	0.061	0.058	0.62
DIPW	−0.025	0.090	0.086	0.93	1	−0.002	0.050	0.050	0.95	0.022	0.078	0.069	0.90
Proposed	−0.009	0.084	0.084	0.96	1.22	0.014	0.047	0.055	0.96	0.005	0.073	0.066	0.93

Bias:Monte Carlo bias. ESD: Monte Carlo standard deviation. ASE: Monte Carlo average of estimated standard errors. CP: coverage probability of nominal 95% Wald confidence intervals. RE: Monte Carlo mean squared error for the DIPW estimator divided by that for the indicated estimator

We use the proposed method to estimate the restricted mean lifetime and compare its performance with Nelson-Aalen, IPTW and DIPW methods in three scenarios. Under all scenarios, the Nelson-Aalen estimator greatly underestimates the difference

of restricted mean lifetime between groups, which leads to large bias, as it does not account for both baseline imbalances and dependent censoring. The bias of the IPTW method is much smaller compared to the Nelson-Aalen method because it accounts for the baseline imbalances between groups. However, there are still some bias as the IPTW method does not account for the dependent censoring. Both DIPW and the proposed method are approximately unbiased and the coverage probabilities approximately achieve the nominal levels.

We compare the efficiency of the unbiased estimators which are DIPW and the proposed estimators. Compared to the DIPW estimator, the variances of the proposed estimators are smaller under all scenarios. We use relative efficiency to quantify the difference in efficiency between DIPW and the proposed method. The relative efficiency is defined as the ratio of the mean squared errors of DIPW and the proposed estimators. Under the three scenarios, the relative efficiencies are between 1.19 and 1.29 for $\widehat{\delta}$, which suggests that the proposed method improves efficiency in different situations of survival outcome distributions and generating models.

3.5 Application

We apply the proposed method to a cardiac surgical data set obtained from INTERMACS. It contains 5,856 AHF patients who received VAD implantations from 2008 to 2014. Among these patients, 5,672 (96.9%) received LVAD and the other 184 (3.1%) received BiVAD. We are interested in comparing the restricted mean lifetime between LVAD and BiVAD up to 365 days. Approximately 58.4% of the survival time was administratively censored, and 1.4% received heart transplants after VAD implantations. As stated in previous sections, we treat the heart transplant as dependent censoring. The data contains a rich set of covariates that are expected to be related to survival outcome, and decisions regarding treatment and transplant. Based

on cardiac surgeons' recommendation, we identified variables for the estimation of the restricted mean lifetime by the proposed method. Specifically, 7 variables were related to treatment assignment, including 5 variables measured within 48 hours before the implantation surgery–extracorporeal membrane oxygenation (ECMO), mechanical ventilation, intra-aortic balloon pump, dialysis or ultrafiltration and feeding tube, and also whether central venous pressure was larger than 15 mmHg and INTERMACS profile. Eight variables were related to the transplant decision, including age, gender, race, creatinine, albumin, number of infections, device strategy, and INTERMACS profile, where the number of infections is a time-varying covariate. Up to 13 variables were thought to be related to survival outcome, including age, gender, race, the Elixhauser comorbidity index, hemoglobin and diastolic blood pressure.

Preliminary analyses (Table 3.2) showed, as expected, that there were imbalances in baseline covariates between BiVAD and LVAD. Generally, the patients in BiVAD group needed more clinical interventions prior to surgery, which implied that they were more ill compared to the patients in LVAD. In particular, there were large imbalances in the interventions within 48 hours before surgery, including ECMO, ventilation, IABP, dialysis and feeding tube. The percentage of central venous pressure higher than 15 mmHg is larger for BiVAD patients.

We fit a logistic model for treatment assignment and a Cox model for the time to receive heart transplants after VAD implantations. Then we applied the proposed method to estimate the difference of the restricted mean lifetime up to 365 days.

Table 3.3 shows the estimates of the restricted mean lifetime for the two treat-

Table 3.2: Characteristics of the study cohort stratified by treatment group

Patient Characteristic	Category	Overall	LVAD	BiVAD	P-value
Age (mean ± sd)		62.3 ± 11.2	62.4 ± 11.2	59.8 ± 12.4	0.002
Female (n, %)		1152, 19.7	1115, 19.7	37, 20.1	0.954
Race (n, %)	White	4182, 71.4	4059, 71.6	123, 66.8	0.106
	Black	1260, 21.5	1219, 21.5	41, 22.3	
	Other	414, 7.1	394, 6.9	20, 10.9	
Extracorporeal membrane oxygenation (ECMO) (n, %)		85, 1.5	63, 1.1	22, 12.0	<0.001
Mechanical ventilation (n, %)		269, 4.6	236, 4.2	33, 17.9	<0.001
Intra-aortic balloon pump (IABP) (n, %)		1385, 23.7	1318, 23.2	67, 36.4	<0.001
Dialysis or continuous veno-venous ultrafiltration (n, %)		126, 2.2	112, 2.0	14, 7.6	<0.001
Feeding tube (n, %)		81, 1.4	71, 1.3	10, 5.4	<0.001
Central venous pressure higher than 15 mmHg (n, %)		1343, 22.9	1282, 22.6	61, 33.2	0.001
Creatinine (mg/dL) (mean ± sd)		1.5 ± 0.8	1.5 ± 0.7	1.6 ± 1.1	0.004
Albumin (g/dL) (mean ± sd)		3.4 ± 0.6	3.4 ± 0.6	3.2 ± 0.7	<0.001
Device strategy (n, %)	Group 1	1289, 22.0	1246, 22.0	43, 23.4	0.008
	Group 2	3019, 51.6	2943, 51.9	76, 41.3	
	Group 3	1548, 26.4	1483, 26.1	65, 35.3	
INTERMACS profile (n, %)	1	698, 11.9	617, 10.9	81, 44.0	<0.001
	2	2191, 37.4	2123, 37.4	68, 37.0	
	3	1729, 29.5	1707, 30.1	22, 12.0	
	4-7	1238, 21.1	1225, 21.6	13, 7.1	
Elixhauser comorbidity index (mean ± sd)		3.2 ± 1.5	3.2 ± 1.5	2.6 ± 1.4	<0.001
Previous cardiac operation (n, %)		2071, 35.4	2000, 35.3	71, 38.6	0.395
Total bilirubin (mg/dL) (mean ± sd)		1.3 ± 1.5	1.3 ± 1.4	2.4 ± 3.0	<0.001
Diastolic blood pressure (mean ± sd)		63.9 ± 11.5	63.9 ± 11.5	62.0 ± 11.7	0.02
Hemoglobin (g/dL) (mean ± sd)		11.3 ± 2.0	11.4 ± 2.0	10.7 ± 2.1	<0.001

Table 3.3: Estimation of the difference in restricted mean lifetime (days) up to 365 days between BiVAD and LVAD.

	δ		BiVAD		LVAD		
Method	$\widehat{\delta}$	$\widehat{SE}(\widehat{\delta})$	$\widehat{\mu}_1$	$\widehat{SE}(\widehat{\mu}_1)$	$\widehat{\mu}_0$	$\widehat{SE}(\widehat{\mu}_0)$	P-value
Nelson-Aalen	−135.7	17.2	176.7	17.3	312.4	1.5	<0.001
IPTW	−125.3	21.9	186.6	22.0	311.8	1.5	<0.001
DIPW	−125.0	21.9	186.9	22.0	311.9	1.5	<0.001
Proposed	−87.1	25.6	222.8	25.5	309.9	1.9	<0.001

ment groups and their differences. The proposed method yields a larger estimate for

BiVAD and smaller estimate for LVAD, and hence smaller absolute difference between the two groups, compared to the other methods. This seems reasonable because that the patients in BiVAD are more ill than those in LVAD; thus, failing to adjust for the imbalances will overestimate the LVAD and underestimate the BiVAD. Regarding the standard errors, the proposed method does not show improvement compared to the DIPW method. One possible reason is that we did not estimate the standard error accurately, which is also a concern in our simulation study (Table 3.1). Recent work by *Li and Ding* (2019) introduced the concepts "S-optimal" and "C-optimal", which indicated the optimality based on the uncertainty of the sampling distribution and estimated distribution, respectively. According to the definitions, our proposed estimators are S-optimal and may differ with the C-optimal estimators.

3.6 Discussion

In this chapter, we propose an augmented double inverse weighted method to estimate the difference in restricted mean lifetime between groups had no patients received heart transplant. We incorporate two types of inverse weighting methods into one single weighting framework to build an estimator and further improve its efficiency by the augmentation strategy. The asymptotic properties of the proposed estimator are heuristically proved, and the performance of the proposed methods in finite samples is shown by simulation. The proposed method is applied to the INTERMACS cardiac surgery data.

The proposed method allows the clinical analysts to compare restricted mean lifetime in observational study without modeling the complicated relationship between the outcome and patient covariates. In many cases, the clinicians know much better about treatment decision than the relationships between outcomes and patient covariates. The consistency of the proposed estimator requires correct modeling of

treatment assignment and dependent censoring. If the relationship between patient covariates and the treatment assignment or dependent censoring is not clear, the proposed method may not be a good choice. In that case, it may be a better choice to model the outcome directly if they have good knowledge about the outcome.

We treat the heart transplant as censoring and ignore the information after the initiation of the heart transplant. There are other strategies which can be used to deal with such problems. For instance, the heart transplant can be treated as a secondary treatment and then handled using a marginal structural Cox model that was proposed by *Zhang and Wang* (2012). This method would let us gain more efficiency. However, we did not consider this strategy because the effect of heart transplant is not specified; moreover, the further efficiency gain may not be sufficiently large because the percentage of patients receiving heart transplants is small. Another method is to consider the heart transplant as a time-varying covariate and use an outcome model to estimate the treatment effect. This method is not applicable in our motivating study because the percentage of patients receiving heart transplants is small.

CHAPTER IV

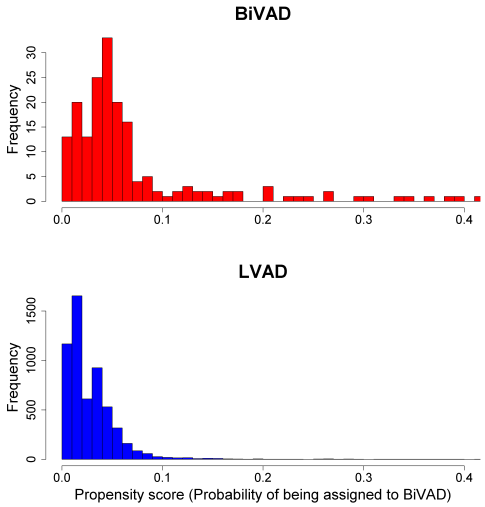
Methods for Estimating More Meaningful Causal Treatment Effects as Opposed to The Average Treatment Effect

4.1 Introduction

In Chapter III, we estimated the average treatment effect (ATE), where the average was calculated over the whole population. Although the ATE is a popular estimand in medical research, it is not always the most clinically meaningful quantity of interest. In Section 3.5, we showed that the patient characteristics vary between two groups (Table 3.2). Figure 4.1 shows that the discrepancy of the distributions of the probability of receiving BiVAD between groups. Compared to BiVAD group, a higher percentage of LVAD patients had a probability of receiving BiVAD close to 0, and a much smaller percentage of LVAD patients had probability of receiving BiVAD larger than 0.15. In practice, not all patients can receive either BiVAD or LVAD. For example, the very sick patients may only receive the BiVAD, while healthier patients are more likely to receive the LVAD. When doctors compare BiVAD and LVAD, they may want to exclude the patients who are unlikely to receive either treatment or control, and mainly focus on the patients who may receive either. Therefore,

in this chapter, we are interested in estimating the average treatment effect on the matched population (ATM), which is the treatment-specific difference in potential restricted mean lifetimes on the matched population, had no patients received heart transplants. Here the matched population means the population that can be matched between treatment and control using propensity scores (PS).

Figure 4.1: Distribution of PS by treatment groups



According to the definition of ATM, we know that the PS matching is a natural method for estimating the ATM. However, PS matching has some drawbacks in practice. First, it is challenging to estimate the variances of the PS matching estimators. To our best knowledge, the asymptotic properties of the PS matching methods have not yet been well studied. As an alternative, the bootstrap method has been used for estimating the standard errors of PS matching estimators. However, *Abadie and Imbens* (2008) showed that bootstrap standard errors are generally not valid for the matching estimators. *Abadie and Imbens* (2006) proposed a method for estimating standard errors for nearest neighbor matching with replacement, but this matching approach is infrequently used in medical research. How to calculate standard errors for the popular matching methods, including one-to-one PS matching without replacement and optimal matching, remains unclear. Secondly, we always need to

subjectively choose a matching algorithm or tuning parameters like the caliper size for the caliper matching, in order to achieve the best balance between groups. However, the criteria for whether achieving the best balances is often vague. Thirdly, it is difficult to develop methods to improve efficiency of PS matching. In Chapter II, we proposed methods to improve the efficiency of PS matching, but they lacked rigorous theoretical derivation.

Unlike PS matching, the inverse probability weighting (IPW) methods have a rigorous theoretical framework, which can be used to derive the asymptotic properties and improve the efficiency of estimation using strategies like augmentation. However, the IPW is generally used to estimate the ATE instead of the ATM. *Li and Greene* (2013) proposed a PS weighting method called the matching weight (MW) method, which is an analogue to the one-to-one PS caliper matching method without replacement. They showed that compared to PS matching, the MW method has an asymptotically identical estimand but better variance calculation and efficiency of estimation. The non-stabilized IPW weights are the inverse of the probability of being assigned to the treatment or control, while the matching weight is defined as a product of the usual non-stabilized IPW weight and a term $\min(e_i, 1-e_i)$, i.e. the matching weight $w_i^{mw} = \frac{\min(e_i,1-e_i)}{A_i e_i+(1-A_i)(1-e_i)}$ where e_i is the conditional probability of being assigned to treatment and A_i is the indicator of being assigned to treatment for subject i. Intuitively, suppose there are k patients whose propensity scores are close to e^*, then approximately ke^* patients would be assigned to the treatment, and $k(1-e^*)$ would be assigned to the control. If $e^* < 0.5$, fewer patients would be assigned to the treatment and more patients would be assigned to the control, all treated patients can be matched while only a subset of control patients can be matched. Therefore, the probability of being matched for treated patients is about 1, and the probability of being matched for control is about $e^*/(1-e^*)$. As we can see, these two probabilities are exactly the w_i^{mw} conditional on A_i and e_i. In other

words, intuitively the w_i^{mw} can be viewed as the probability of being matched. Please note that the one-to-one PS caliper matching discards the patients that cannot be matched, while the MW method keeps all the patients and only down-weights some patients whose PSs are close to 0 or 1.

Li and Greene (2013) proposed the MW method for continuous outcomes and discussed the possibility of extending to binary outcomes. However, how to apply the MW method for causal inference for survival outcomes and further improve the efficiency of estimation has not been studied. In this chapter, motivated by the same application as in Chapter III, we adopt the MW method for estimating the treatment-specific difference in potential restricted mean lifetimes for the matched population, had no patients received heart transplants. Then we propose augmentation methods to improve the efficiency of estimation.

The remainder of the chapter is organized as follows: in Section 4.2, we describe the notation. In Section 4.3, we propose the augmented MW methods. Then we evaluate the performance of the proposed methods using a simulation study in Section 4.4. In Section 4.5, we apply our methods to analyze the cardiac surgery data obtained from the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS). Finally, we conclude and discuss in Section 4.6.

4.2 Notations

The notations are the same as in Chapter III. We let A denote treatment groups that are not randomized ($A = 0$ control, $A = 1$ treatment). We let T denote the survival time without heart transplant and C denote the dependent censoring (heart transplant). $U = T \wedge C$ denotes the observation time, and $\Delta = I(T \leq C)$ is the indicator for observing the death time. We let Z denote baseline covariate vector, and $Z(t)$ represent the time-dependent covariate at time t. We let $\tilde{Z}(t) = \{Z(u); u \in$

$[0, t)\}$ denote the history of the baseline and time-dependent covariates up to time t. We denote the observed event process and at-risk processes by $N_i(t) = I(U_i \leq t, \Delta = 1)$ and $Y_i(t) = I(U_i \geq t)$, $i = 1, ..., n$, respectively.

We define the parameter of interest by using the potential outcome framework which was studied by *Rubin* (1974, 1978, 1980). Let T^k (k = 0,1) denote the potential (or counterfactual) lifetime for a randomly selected subject from the population if, possibly contrary to fact, s/he was assigned to group k, and not censored by heart transplant. We assume that there are no unmeasured baseline confounders. We also assume that C_i is conditionally independent of T_i given $\{A_i, Z_i, \tilde{Z}_i(t)\}$, formally expressed as:

$$\begin{aligned}
&\lim_{\xi \to 0} \xi^{-1} P\{t \leq U_i < t + \xi, \Delta_i = 0 | U_i \geq t, A_i, \tilde{Z}_i(t), T_i\} \\
&\qquad = \lim_{\xi \to 0} \xi^{-1} P\{t \leq U_i < t + \xi, \Delta_i = 0 | U_i \geq t, A_i, \tilde{Z}_i(t)\}.
\end{aligned}$$

The estimand of interest δ is the difference in restricted mean lifetime up to time L between two treatment groups on the matched population, had no patients received heart transplants. Let μ_k denote the restricted mean lifetime on the matched population in group k and $S_k(t) = P\{T^k > t | A = 1\}$, then

$$\begin{aligned}
\delta &= \mu_1 - \mu_0 \\
&= E\{\min(T^1, L) | A = 1\} - E\{\min(T^0, L) | A = 1\} \\
&= \int_0^L S_1(t) - \int_0^L S_0(t) dt.
\end{aligned}$$

$S_k(t)$ can be estimated as by $\exp\{-\Lambda_k(t)\}$, where $\Lambda_k(t)$ is the cumulative hazard for T^k.

4.3 Methods

4.3.1 Matching Weight Method

Li and Greene (2013) defined the matching weight for patient i as

$$w_i^{mw} = \frac{\min(e_i, 1-e_i)}{A_i e_i + (1-A_i)(1-e_i)}, \tag{4.1}$$

where $e_i = Pr(A_i = 1|Z_i)$. Using the same double inverse weighting method as in Chapter III, we combine the IPTW and IPCW to develop the double inverse weighted MW estimating equation for $\Lambda_k(t)$ $(k = 0, 1)$, that is

$$\sum_{i=1}^{n} \frac{\min(\widehat{e}_i, 1-\widehat{e}_i)}{\Psi} \left[\frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{\{A_i\widehat{e}_i + (1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma}, t)} \right] = 0, \tag{4.2}$$

where $A_{ik} = I(A_i = k)$, and $\Psi = E\{\min(e_i, 1-e_i)\}$. $\widehat{e}_i \equiv e_i(\widehat{\theta})$ estimates e_i through a model with parameter θ, and $p_i^c(\widehat{\gamma}, t)$ estimates informally the probability of not being censored conditional on $\{A_i,\ Z_i,\ \tilde{Z}_i(t)\}$ through a model with parameter γ. By solving Equation (4.2), the double inverse weighted MW estimator for $\Lambda_k(t)$ is

$$\widehat{\Lambda}_k^{mw}(t) = \int_0^t \frac{\sum_{i=1}^n A_{ik} w_i^{mw}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) dN_i(u)}{\sum_{i=1}^n A_{ik} w_i^{mw}(\widehat{\theta}) w_i^c(\widehat{\gamma}, u) Y_i(u)},$$

where $w_i^{mw}(\widehat{\theta}) = \frac{\min(e_i(\widehat{\theta}), 1-e_i(\widehat{\theta}))}{A_i e_i(\widehat{\theta}) + (1-A_i)(1-e_i(\widehat{\theta}))}$ and $w_i^c(\widehat{\gamma}, u) = I(C_i > t)/p_i^c(\widehat{\gamma}, t)$.

4.3.2 Proposed Augmented MW Methods

Starting from the MW estimator for $\Lambda_k(t)$ that we derived, we propose augmentation methods to improve its efficiency. Similar with Equation (3.2), we propose the

augmented double inverse weighted MW estimating equation for $\Lambda_k(t)$, that is

$$\sum_{i=1}^{n} \frac{\min(\widehat{e}_i, 1-\widehat{e}_i)}{\Psi} \left[\frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{\{A_i\widehat{e}_i + (1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma}, t)} - \frac{A_{ik} - \widehat{e}_{ik}}{\widehat{e}_{ik}} h_k(t, Z_i)dt \right] = 0, \tag{4.3}$$

where $\widehat{e}_{ik}$ estimates $P(A_{ik} = 1|Z_i)$ and $h_k(t, Z_i)$ is an arbitrary function of baseline covariate Z_i at time t. We consider two augmentation strategies. The first strategy is the augmentation method proposed in Chapter III; that is, we consider a subclass of Equation (4.3) where $h_k(t, Z_i) = \beta_k(t)g(Z_i)$. Then the Equation (4.3) can be rewritten as

$$\sum_{i=1}^{n} \frac{\min(\widehat{e}_i, 1-\widehat{e}_i)}{\Psi} \left[\frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{\{A_i\widehat{e}_i + (1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma}, t)} - \frac{A_{ik} - \widehat{e}_{ik}}{\widehat{e}_{ik}} \beta_k(t)g(Z_i)dt \right] = 0. \tag{4.4}$$

Then we use the ordinary least square (OLS) method to obtain $\widehat{\beta}_k^{opt}(t)$. Specifically, when we apply the OLS method to estimate $\beta_k^{opt}(t)$, we let the outcome be

$$O = \sum_{i=1}^{n} \frac{\min(\widehat{e}_i, 1-\widehat{e}_i)}{\Psi} \left[\frac{A_{ik}\{dN_i(t) - Y_i(t)d\Lambda_k(t)\}}{\{A_i\widehat{e}_i + (1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma}, t)} \right].$$

As we don't know $d\Lambda_k(t)$, we substitute it with $d\widehat{\Lambda}_k^{mw}(t)$. We let the covariate be

$$C = \sum_{i=1}^{n} \frac{\min(\widehat{e}_i, 1-\widehat{e}_i)}{\Psi} \left[\frac{A_{ik} - \widehat{e}_{ik}}{\widehat{e}_{ik}} \beta_k(t)g(Z_i)dt \right].$$

Then the optimal $\beta_k(t)$ can be estimated by $\widehat{\beta}_k^{opt}(t)dt = (C^TC)^{-1}C^TO$. Plugging $\widehat{\beta}_k^{opt}(t)dt$ into Equation (4.4) and solving the equation, the augmented double inverse

weighted MW estimator for $\Lambda_k(t)$ using OLS method is

$$\widehat{\Lambda}_k^{mw,OLS}(t) = \int_0^t \frac{\sum_{i=1}^n \min(\widehat{e}_i, 1-\widehat{e}_i)\big[\frac{A_{ik}dN_i(u)}{\{A_i\widehat{e}_i+(1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma},u)} + \frac{A_{ik}-\widehat{e}_{ik}}{\widehat{e}_{ik}}\widehat{\beta}_k^{opt}(u)g(Z_i)du\big]}{\sum_{i=1}^n \frac{\min(\widehat{e}_i,1-\widehat{e}_i)A_{ik}Y_i(u)}{\{A_i\widehat{e}_i+(1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma},u)}}.$$

Alternatively, motivated by *Zhang and Schaubel* (2012a) and *Zhang and Schaubel* (2012b), we consider using the Cox model for augmentation. The augmented double inverse weighted MW estimator for $d\Lambda_k(t)$ using Cox model is

$$\widehat{\Lambda}_k^{mw,Cox}(t) = \int_0^t \frac{\sum_{i=1}^n \min(\widehat{e}_i, 1-\widehat{e}_i)\big[\frac{A_{ik}dN_i(u)}{\{A_i\widehat{e}_i+(1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma},u)} - \frac{A_{ik}-\widehat{e}_{ik}}{\widehat{e}_{ik}}e^{-\widehat{\Lambda}_{ik}(u)}d\widehat{\Lambda}_{ik}(u)\big]}{\sum_{i=1}^n \frac{\min(\widehat{e}_i,1-\widehat{e}_i)A_{ik}Y_i(u)}{\{A_i\widehat{e}_i+(1-A_i)(1-\widehat{e}_i)\}p_i^c(\widehat{\gamma},u)}},$$

where $\widehat{\Lambda}_{ik}(u)$ and $d\widehat{\Lambda}_{ik}(u)$ were estimated through a Cox model.

4.4 Simulation Study

We perform a simulation study to evaluate performance of the proposed methods. The reported results are based on 5000 Monte Carlo data sets. In each data set, the sample size n is 500. We use the same simulation settings and scenarios as in Chapter III; please refer to Section 3.4 for the detailed information of the simulation design.

In our simulation, we use correct treatment assignment model and censoring model for the estimation of restricted mean lifetime. Specifically, we use treatment assignment model

$$\text{logit}\{P(A_i = 1|Z_i)\} = \beta_1 Z_1 + \beta_2 Z_2,$$

and Cox model for censoring

$$\lambda^c(t|Z_i, Z_i(t), A_i) = \lambda_0^c(t)\exp\{\gamma_1 Z_{3,i} + \gamma_2 Z_i(t)\},$$

where $\lambda^c(t|Z_i, Z_i(t), A_i)$ is the conditional hazard for censoring. For the proposed methods, when we use the OLS augmentation method, we let $g_k(Z, u)$ be a basis function including intercept and linear forms of all baseline covariates, i.e. $g_k(Z, u) = (1, Z_1, Z_2, Z_3)^T$. When we use the Cox augmentation method, we use the following model to calculate $\widehat{\Lambda}_{ik}(t)$ and $d\widehat{\Lambda}_{ik}(t)$

$$\lambda(t|Z_i, Z_i(t), A_i) = \lambda_0(t)\exp\{\eta_1 Z_{1,i} + \eta_2 Z_{2,i} + \eta_3 Z_{3,i}\}.$$

We estimate standard errors of estimators via the bootstrap approach.

We evaluate the performance of the proposed methods, together with the methods described in Chapter III, including the Nelson-Aalen, IPTW and DIPW methods. The simulation results are summarized in Table 4.1.

Table 4.1: Estimation of the difference in restricted mean lifetime between treatment and control. The results are based on 5000 Monte Carlo datasets. Sample size $n = 500$.

	$\widehat{\delta}$					$\widehat{\mu}_1$				$\widehat{\mu}_0$			
Method	Bias	ESD	ASE	CP	RE	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP
Scenario 1													
NA	−0.456	0.053	0.053	0.000		−0.175	0.039	0.038	0.004	0.281	0.036	0.036	0.000
IPTW	−0.054	0.077	0.084	0.900		0.015	0.047	0.049	0.954	0.069	0.070	0.067	0.754
DIPW	−0.022	0.087	0.091	0.942		−0.001	0.051	0.052	0.961	0.021	0.078	0.073	0.908
MW	0.001	0.080	0.078	0.941	1	0.040	0.048	0.045	0.818	0.039	0.069	0.062	0.854
Proposed (OLS)	0.002	0.076	0.073	0.947	1.11	0.041	0.046	0.045	0.810	0.039	0.067	0.061	0.858
Proposed (Cox)	0.002	0.074	0.080	0.965	1.17	0.039	0.045	0.048	0.852	0.038	0.067	0.062	0.852
Scenario 2													
NA	−0.318	0.046	0.046	0.000		−0.114	0.034	0.033	0.071	0.205	0.032	0.032	0.000
IPTW	−0.051	0.074	0.072	0.847		0.016	0.039	0.040	0.942	0.066	0.067	0.059	0.682
DIPW	−0.025	0.091	0.083	0.905		−0.002	0.044	0.044	0.954	0.023	0.083	0.070	0.864
MW	−0.010	0.079	0.073	0.934	1	0.022	0.042	0.040	0.880	0.032	0.069	0.060	0.864
Proposed (OLS)	−0.009	0.078	0.072	0.936	1.03	0.023	0.041	0.040	0.881	0.031	0.068	0.061	0.873
Proposed (Cox)	−0.010	0.077	0.079	0.960	1.05	0.022	0.041	0.044	0.920	0.032	0.068	0.062	0.884
Scenario 3													
NA	−0.430	0.048	0.047	0.000		−0.156	0.033	0.033	0.004	0.274	0.034	0.034	0.000
IPTW	−0.064	0.069	0.076	0.853		0.021	0.045	0.047	0.946	0.085	0.060	0.058	0.610
DIPW	−0.026	0.085	0.085	0.930		−0.001	0.050	0.050	0.956	0.025	0.075	0.068	0.883
MW	−0.003	0.076	0.072	0.943	1	0.025	0.045	0.042	0.877	0.028	0.066	0.058	0.880
Proposed (OLS)	−0.002	0.072	0.068	0.939	1.12	0.026	0.043	0.041	0.878	0.027	0.065	0.057	0.882
Proposed (Cox)	−0.003	0.071	0.077	0.968	1.15	0.025	0.042	0.045	0.915	0.028	0.065	0.059	0.900

Bias:Monte Carlo bias. ESD: Monte Carlo standard deviation. ASE: Monte Carlo average of estimated standard errors. CP: coverage probability of nominal 95% Wald confidence intervals. RE: Monte Carlo mean squared error for the MW estimator divided by that for the indicated estimator

In all scenarios, the biases of $\widehat{\delta}$ for the MW and proposed methods are negligible compared to the other methods. The 95% coverage probabilities appear to achieve the nominal level. There are biases for $\widehat{\mu}_1$ and $\widehat{\mu}_0$ from the MW and proposed methods. This is because the true μ_1 and μ_0 values that we used to calculate biases are the expectation of potential outcomes for the whole population. However, here the $\widehat{\mu}_1$ and $\widehat{\mu}_0$ estimated by the MW and proposed methods are for a subpopulation, that is the patients who can be matched between two groups. Hence, it is actually unfair to compare the biases of group-level estimates between MW methods and usual IPW methods (IPTW and DIPW), because they estimate different estimands.

Next, we compare the efficiency of the methods. As we show in Chapter III, the simple IPW estimators including IPTW and DIPW reduce bias but lose efficiency. The MW method is also a simple double inverse weighted estimator, but compared to DIPW method, the variances are much smaller. As we stated in Section 4.1, the MW method reduces variances because it down-weights some of the patients. The resulting weighted population is more homogeneous compared to the weighted population by the usual IPW methods.

In Table 4.1, we also see that the proposed augmentation methods reduce the variance of the MW estimators. We use relative efficiency to quantify the difference in efficiency between the MW and the proposed methods. The relative efficiency is defined as the ratio of mean squared errors for the MW method and the proposed methods. In three scenarios, for proposed OLS methods, the relative efficiencies are between 1.03 and 1.12 for $\widehat{\delta}$. For the proposed Cox methods, the relative efficiencies are even larger, which are between 1.05 and 1.17 for $\widehat{\delta}$. This results suggest that the proposed methods successfully improve efficiency of MW estimators in different situations of outcome distributions and generating models.

Furthermore, we see that the proposed Cox method has larger efficiency gains in Scenario 1 and 3 compared to Scenario 2. This may be because that we generate

survival time from Cox models in Scenario 1 and 3, but from an accelerated failure time (AFT) model in Scenario 2. Although the analysis outcome model we use in Scenario 2 is incorrect, the proposed Cox method still shows some efficiency gain.

4.5 Application

We apply the proposed method to the same application data as in Chapter III. The 5,856 advanced heart failure patients received either LVAD or BiVAD implantations. Of these patients, 5,672 (96.9%) patients received LVAD and 184 (3.1%) patients received BiVAD. We showed the baseline imbalances between the two groups in Table 3.2. As in Chapter III, we consider BiVAD as the treatment and LVAD as the control. We are interested in estimating treatment-specific difference in potential restricted mean lifetimes on the matched population, had no patients received heart transplants. The proposed methods show smaller estimates of the differences between BiVAD and

Table 4.2: Estimation of the difference in restricted mean lifetime (days) up to 365 days between BiVAD and LVAD.

	δ		BiVAD		LVAD		
Method	$\widehat{\delta}$	$\widehat{SE}(\widehat{\delta})$	$\widehat{\mu}_1$	$\widehat{SE}(\widehat{\mu}_1)$	$\widehat{\mu}_0$	$\widehat{SE}(\widehat{\mu}_0)$	P-value
Nelson-Aalen	−135.7	17.2	176.7	17.3	312.4	1.5	<0.001
IPTW	−125.3	21.9	186.6	22.0	311.8	1.5	<0.001
DIPW	−125.0	21.9	186.9	22.0	311.9	1.5	<0.001
MW	−117.8	17.1	177.1	17.3	294.9	4.0	<0.001
Proposed (OLS)	−119.2	17.9	184.7	17.1	303.9	5.5	<0.001
Proposed (Cox)	−92.0	16.4	202.5	16.5	294.5	3.7	<0.001

LVAD compared to the DIPW method. The MW shows smaller standard errors than the DIPW method. The proposed Cox augmentation method has smaller standard errors than the MW method.

4.6 Discussion

In this chapter, we discuss the clinically meaningful estimand and then propose methods to estimate that estimand. Specifically, we propose an augmented double inverse MW method to estimate the difference in the restricted mean lifetime on the matched population between groups. We first develop an MW estimator and then propose augmentation strategies to improve the efficiency. We show the performance of the proposed methods in finite sample using simulation. The proposed methods are then applied to INTERMACS cardiac surgery data.

The MW method is an analog to the one-to-one caliper matching without replacement, but it has better properties of variance estimation and efficiency improvement. *Li and Greene* (2013) showed the MW method was more efficient than the one-to-one caliper matching without replacement in their continuous outcome setting. Beside one-to-one caliper matching, there are many other types of PS matching methods, such as matching with replacement, nearest neighbor matching, optimal matching, and many-to-one matching. It would be interesting to compare the performance of the MW method with these types of matching. We can also apply the doubly robust PS matching method proposed in Chapter II for survival outcomes and compare the performance with the proposed OLS and Cox methods.

Our simulation results show that the MW method is more efficient than DIPW method. As we know, the DIPW method use the usual IPW weight which is the inverse of probability of being assigned to one group. The usual IPW weights can be very large when the propensity scores are small, which greatly increase the variability of the estimation and make the estimation unstable. But the MW method down-weights the patients with extreme propensity scores and control the variance inflation. This is why we see the MW method is generally more efficient than DIPW method.

We use the bootstrap approach to calculate standard errors for the MW and proposed augmented estimators. We can also study the asymptotic properties and

develop the closed form of standard errors, because the MW method provides a good framework for these theoretical derivation.

APPENDICES

APPENDIX A

Appendix for Chapter II

A.1 Simulation Results for Matching with Replacement

Table A.1: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
With replacement	Within hospital matching	PS Model 1	0.001	0.120	0.031	0.158	0.031	0.158	0.010	0.126	0.010	0.126
		PS Model 2	−0.002	0.111	0.005	0.132	0.005	0.132	0.000	0.116	0.000	0.116
		PS Model 3	0.000	0.114	1.158	0.148	1.158	0.148	−0.006	0.130	−0.008	0.130
		PS Model 4	0.001	0.119	0.032	0.143	0.032	0.143	0.011	0.125	0.011	0.125
		PS Model 5	0.001	0.120	0.040	0.146	0.040	0.146	0.013	0.126	0.013	0.126
	Across hospital matching	PS Model 1	0.003	0.128	−0.060	0.224	0.039	0.193	0.014	0.149	0.014	0.149
		PS Model 2	−0.003	0.111	0.369	0.315	0.328	0.153	0.007	0.120	−0.007	0.120
		PS Model 3	−0.001	0.106	1.325	0.273	1.247	0.140	0.003	0.123	−0.011	0.124
		PS Model 4	0.000	0.129	−0.038	0.217	0.037	0.181	0.011	0.148	0.010	0.148
		PS Model 5	0.003	0.129	0.022	0.215	0.091	0.183	0.013	0.152	0.013	0.152
	Modified across hospital matching	PS Model 1	0.002	0.125	0.082	0.196	0.139	0.177	0.034	0.140	0.034	0.140
		PS Model 2	−0.001	0.108	0.173	0.185	0.235	0.139	0.010	0.116	0.004	0.116
		PS Model 3	−0.001	0.113	1.168	0.147	1.164	0.147	−0.005	0.129	−0.008	0.129
		PS Model 4	0.000	0.124	0.091	0.184	0.138	0.163	0.033	0.139	0.033	0.139
		PS Model 5	0.002	0.124	0.131	0.182	0.171	0.164	0.038	0.139	0.038	0.139

Table A.2: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
With replacement	Within hospital matching	PS Model 1	0.094	0.877	0.163	0.952	0.147	0.921	0.105	0.897	0.105	0.898
		PS Model 2	0.099	0.928	0.170	0.987	0.153	0.976	0.110	0.939	0.110	0.939
		PS Model 3	0.095	0.902	0.157	0.000	0.143	0.000	0.118	0.924	0.118	0.916
		PS Model 4	0.094	0.882	0.163	0.970	0.147	0.941	0.105	0.902	0.105	0.902
		PS Model 5	0.094	0.881	0.163	0.958	0.146	0.934	0.105	0.893	0.105	0.893
	Across hospital matching	PS Model 1	0.088	0.819	0.161	0.823	0.146	0.847	0.103	0.834	0.103	0.831
		PS Model 2	0.096	0.906	0.155	0.427	0.157	0.438	0.111	0.931	0.112	0.926
		PS Model 3	0.099	0.939	0.152	0.000	0.151	0.000	0.121	0.944	0.121	0.943
		PS Model 4	0.088	0.812	0.161	0.851	0.146	0.877	0.103	0.822	0.103	0.821
		PS Model 5	0.088	0.802	0.160	0.858	0.146	0.836	0.103	0.808	0.103	0.811
	Modified across hospital matching	PS Model 1	0.088	0.833	0.158	0.849	0.143	0.783	0.101	0.837	0.101	0.839
		PS Model 2	0.088	0.897	0.155	0.767	0.145	0.636	0.103	0.922	0.103	0.919
		PS Model 3	0.094	0.901	0.156	0.000	0.142	0.000	0.116	0.921	0.116	0.916
		PS Model 4	0.087	0.839	0.158	0.867	0.144	0.809	0.101	0.836	0.101	0.836
		PS Model 5	0.087	0.834	0.157	0.828	0.143	0.743	0.101	0.825	0.101	0.824

Table A.3: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 100.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
With replacement	Within hospital matching	PS Model 1	0.002	0.115	0.031	0.147	0.031	0.147	0.013	0.121	0.013	0.121
		PS Model 2	0.000	0.108	0.007	0.129	0.007	0.129	0.003	0.112	0.003	0.112
		PS Model 3	0.000	0.110	1.159	0.148	1.159	0.148	−0.005	0.126	−0.008	0.126
		PS Model 4	0.001	0.116	0.029	0.138	0.029	0.138	0.011	0.121	0.011	0.121
		PS Model 5	0.000	0.118	0.038	0.143	0.038	0.143	0.013	0.124	0.013	0.124
	Across hospital matching	PS Model 1	0.004	0.124	−0.051	0.220	0.044	0.187	0.017	0.146	0.017	0.146
		PS Model 2	0.000	0.109	0.405	0.292	0.340	0.154	0.013	0.119	−0.002	0.119
		PS Model 3	−0.001	0.104	1.349	0.258	1.249	0.143	0.004	0.121	−0.012	0.121
		PS Model 4	0.004	0.122	−0.027	0.205	0.045	0.174	0.019	0.145	0.019	0.145
		PS Model 5	0.004	0.122	0.025	0.208	0.095	0.176	0.019	0.147	0.019	0.146
	Modified across hospital matching	PS Model 1	0.003	0.121	0.090	0.189	0.149	0.167	0.038	0.137	0.039	0.137
		PS Model 2	0.001	0.104	0.202	0.181	0.254	0.136	0.015	0.113	0.008	0.113
		PS Model 3	0.000	0.110	1.171	0.149	1.166	0.148	−0.004	0.126	−0.008	0.126
		PS Model 4	0.003	0.121	0.099	0.183	0.148	0.160	0.039	0.138	0.039	0.138
		PS Model 5	0.003	0.121	0.137	0.182	0.180	0.160	0.041	0.137	0.041	0.137

Table A.4: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
With replacement	Within hospital matching	PS Model 1	0.094	0.897	0.163	0.970	0.146	0.942	0.105	0.920	0.105	0.920
		PS Model 2	0.100	0.930	0.171	0.992	0.153	0.982	0.111	0.954	0.111	0.954
		PS Model 3	0.095	0.902	0.157	0.000	0.143	0.000	0.118	0.935	0.118	0.936
		PS Model 4	0.094	0.892	0.163	0.984	0.146	0.967	0.105	0.917	0.105	0.917
		PS Model 5	0.094	0.877	0.163	0.972	0.146	0.949	0.104	0.906	0.104	0.906
	Across hospital matching	PS Model 1	0.088	0.848	0.160	0.847	0.146	0.870	0.103	0.830	0.103	0.830
		PS Model 2	0.096	0.904	0.154	0.366	0.157	0.421	0.111	0.929	0.112	0.936
		PS Model 3	0.099	0.935	0.152	0.000	0.152	0.000	0.121	0.949	0.122	0.945
		PS Model 4	0.088	0.847	0.161	0.875	0.146	0.890	0.103	0.828	0.103	0.829
		PS Model 5	0.088	0.847	0.160	0.869	0.146	0.837	0.103	0.835	0.103	0.833
	Modified across hospital matching	PS Model 1	0.087	0.848	0.158	0.877	0.143	0.774	0.101	0.835	0.101	0.835
		PS Model 2	0.088	0.897	0.155	0.719	0.145	0.594	0.103	0.929	0.103	0.925
		PS Model 3	0.094	0.909	0.156	0.000	0.143	0.000	0.117	0.932	0.117	0.933
		PS Model 4	0.087	0.843	0.158	0.860	0.143	0.796	0.101	0.835	0.101	0.838
		PS Model 5	0.087	0.838	0.157	0.816	0.143	0.725	0.101	0.839	0.101	0.838

Table A.5: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD
With replacement	Within hospital matching	PS Model 1	0.001	0.120	0.031	0.158	0.031	0.158	0.010	0.126	0.010	0.126
		PS Model 2	−0.002	0.111	0.005	0.132	0.005	0.132	0.000	0.116	0.000	0.116
		PS Model 3	0.000	0.114	1.158	0.148	1.158	0.148	−0.006	0.130	−0.008	0.130
		PS Model 4	0.001	0.119	0.032	0.143	0.032	0.143	0.011	0.125	0.011	0.125
		PS Model 5	0.001	0.120	0.040	0.146	0.040	0.146	0.013	0.126	0.013	0.126
	Across hospital matching	PS Model 1	0.003	0.128	−0.060	0.224	0.039	0.193	0.014	0.149	0.014	0.149
		PS Model 2	−0.003	0.111	0.369	0.315	0.328	0.153	0.007	0.120	−0.007	0.120
		PS Model 3	−0.001	0.106	1.325	0.273	1.247	0.140	0.003	0.123	−0.011	0.124
		PS Model 4	0.000	0.129	−0.038	0.217	0.037	0.181	0.011	0.148	0.010	0.148
		PS Model 5	0.003	0.129	0.022	0.215	0.091	0.183	0.013	0.152	0.013	0.152
	Modified across hospital matching	PS Model 1	0.002	0.125	0.082	0.196	0.139	0.177	0.034	0.140	0.034	0.140
		PS Model 2	−0.001	0.108	0.173	0.185	0.235	0.139	0.010	0.116	0.004	0.116
		PS Model 3	−0.001	0.113	1.168	0.147	1.164	0.147	−0.005	0.129	−0.008	0.129
		PS Model 4	0.000	0.124	0.091	0.184	0.138	0.163	0.033	0.139	0.033	0.139
		PS Model 5	0.002	0.124	0.131	0.182	0.171	0.164	0.038	0.139	0.038	0.139

Table A.6: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
With replacement	Within hospital matching	PS Model 1	0.094	0.877	0.163	0.952	0.147	0.921	0.105	0.897	0.105	0.898
		PS Model 2	0.099	0.928	0.170	0.987	0.153	0.976	0.110	0.939	0.110	0.939
		PS Model 3	0.095	0.902	0.157	0.000	0.143	0.000	0.118	0.924	0.118	0.916
		PS Model 4	0.094	0.882	0.163	0.970	0.147	0.941	0.105	0.902	0.105	0.902
		PS Model 5	0.094	0.881	0.163	0.958	0.146	0.934	0.105	0.893	0.105	0.893
	Across hospital matching	PS Model 1	0.088	0.819	0.161	0.823	0.146	0.847	0.103	0.834	0.103	0.831
		PS Model 2	0.096	0.906	0.155	0.427	0.157	0.438	0.111	0.931	0.112	0.926
		PS Model 3	0.099	0.939	0.152	0.000	0.151	0.000	0.121	0.944	0.121	0.943
		PS Model 4	0.088	0.812	0.161	0.851	0.146	0.877	0.103	0.822	0.103	0.821
		PS Model 5	0.088	0.802	0.160	0.858	0.146	0.836	0.103	0.808	0.103	0.811
	Modified across hospital matching	PS Model 1	0.088	0.833	0.158	0.849	0.143	0.783	0.101	0.837	0.101	0.839
		PS Model 2	0.088	0.897	0.155	0.767	0.145	0.636	0.103	0.922	0.103	0.919
		PS Model 3	0.094	0.901	0.156	0.000	0.142	0.000	0.116	0.921	0.116	0.916
		PS Model 4	0.087	0.839	0.158	0.867	0.144	0.809	0.101	0.836	0.101	0.836
		PS Model 5	0.087	0.834	0.157	0.828	0.143	0.743	0.101	0.825	0.101	0.824

A.2 Simulation Results for Modified Matching with Replacement

Table A.7: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use modified matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
Outcome Only	Regression		0.004	0.168	1.764	0.327	1.679	0.253	−0.051	0.205	−0.089	0.206
With modified replacement	Within hospital matching	PS Model 1	−0.002	0.224	0.010	0.286	0.010	0.286	0.003	0.231	0.003	0.231
		PS Model 2	−0.001	0.238	0.005	0.278	0.005	0.278	0.003	0.248	0.003	0.248
		PS Model 3	−0.001	0.198	1.186	0.246	1.186	0.246	0.001	0.227	−0.009	0.228
		PS Model 4	−0.001	0.224	0.012	0.260	0.012	0.260	0.005	0.232	0.005	0.232
		PS Model 5	0.001	0.223	0.019	0.257	0.019	0.257	0.007	0.232	0.007	0.232
	Across hospital matching	PS Model 1	0.006	0.184	−0.148	0.259	−0.142	0.256	0.007	0.212	0.016	0.212
		PS Model 2	0.004	0.178	0.417	0.333	0.253	0.219	0.043	0.193	−0.002	0.195
		PS Model 3	0.004	0.175	1.359	0.314	1.266	0.230	0.041	0.206	−0.002	0.207
		PS Model 4	0.005	0.184	−0.088	0.231	−0.140	0.222	0.014	0.208	0.013	0.208
		PS Model 5	0.006	0.184	0.030	0.235	−0.046	0.224	0.011	0.210	0.019	0.211
	Modified across hospital matching	PS Model 1	0.005	0.180	−0.065	0.249	−0.046	0.244	0.017	0.199	0.017	0.200
		PS Model 2	0.003	0.177	0.388	0.310	0.303	0.213	0.044	0.191	0.004	0.193
		PS Model 3	0.003	0.174	1.354	0.287	1.311	0.224	0.023	0.202	−0.016	0.204
		PS Model 4	0.003	0.180	−0.025	0.216	−0.048	0.210	0.022	0.197	0.017	0.197
		PS Model 5	0.005	0.180	0.069	0.217	0.030	0.210	0.015	0.200	0.015	0.199

Table A.8: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
	Within hospital matching	PS Model 1	−0.002	0.224	0.010	0.286	0.010	0.286	0.003	0.231	0.003	0.231
		PS Model 2	−0.001	0.238	0.005	0.278	0.005	0.278	0.003	0.248	0.003	0.248
		PS Model 3	−0.001	0.198	1.186	0.246	1.186	0.246	0.001	0.227	−0.009	0.228
		PS Model 4	−0.001	0.224	0.012	0.260	0.012	0.260	0.005	0.232	0.005	0.232
		PS Model 5	0.001	0.223	0.019	0.257	0.019	0.257	0.007	0.232	0.007	0.232
With modified replacement	Across hospital matching	PS Model 1	0.006	0.184	−0.148	0.259	−0.142	0.256	0.007	0.212	0.016	0.212
		PS Model 2	0.004	0.178	0.417	0.333	0.253	0.219	0.043	0.193	−0.002	0.195
		PS Model 3	0.004	0.175	1.359	0.314	1.266	0.230	0.041	0.206	−0.002	0.207
		PS Model 4	0.005	0.184	−0.088	0.231	−0.140	0.222	0.014	0.208	0.013	0.208
		PS Model 5	0.006	0.184	0.030	0.235	−0.046	0.224	0.011	0.210	0.019	0.211
	Modified across hospital matching	PS Model 1	0.005	0.180	−0.065	0.249	−0.046	0.244	0.017	0.199	0.017	0.200
		PS Model 2	0.003	0.177	0.388	0.310	0.303	0.213	0.044	0.191	0.004	0.193
		PS Model 3	0.003	0.174	1.354	0.287	1.311	0.224	0.023	0.202	−0.016	0.204
		PS Model 4	0.003	0.180	−0.025	0.216	−0.048	0.210	0.022	0.197	0.017	0.197
		PS Model 5	0.005	0.180	0.069	0.217	0.030	0.210	0.015	0.200	0.015	0.199

Table A.9: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use modified matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 100.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD	Bias	MCSD
With modified replacement	Within hospital matching	PS Model 1	0.000	0.100	0.017	0.127	0.017	0.127	0.004	0.104	0.004	0.104
		PS Model 2	−0.002	0.101	0.007	0.118	0.007	0.118	0.001	0.104	0.001	0.104
		PS Model 3	0.000	0.095	1.181	0.128	1.181	0.128	0.000	0.106	−0.003	0.106
		PS Model 4	−0.002	0.099	0.013	0.109	0.013	0.109	0.002	0.102	0.002	0.102
		PS Model 5	−0.002	0.099	0.015	0.110	0.015	0.110	0.002	0.103	0.002	0.103
	Across hospital matching	PS Model 1	0.002	0.093	−0.035	0.154	−0.083	0.140	0.008	0.110	0.010	0.110
		PS Model 2	0.001	0.093	0.409	0.277	0.210	0.125	0.006	0.099	−0.008	0.099
		PS Model 3	0.000	0.090	1.351	0.249	1.240	0.130	0.005	0.104	−0.009	0.104
		PS Model 4	0.001	0.095	−0.013	0.141	−0.083	0.123	0.008	0.111	0.009	0.111
		PS Model 5	0.001	0.094	0.025	0.140	−0.052	0.122	0.007	0.111	0.010	0.111
	Modified across hospital matching	PS Model 1	0.002	0.092	0.092	0.132	0.096	0.124	0.016	0.101	0.014	0.101
		PS Model 2	0.000	0.092	0.373	0.232	0.338	0.115	0.018	0.098	0.007	0.098
		PS Model 3	0.000	0.090	1.350	0.211	1.311	0.134	−0.014	0.103	−0.026	0.103
		PS Model 4	0.000	0.093	0.100	0.117	0.092	0.107	0.013	0.102	0.011	0.102
		PS Model 5	−0.001	0.092	0.124	0.116	0.112	0.107	0.011	0.101	0.009	0.101

Table A.10: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use modified matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size equals to 30.

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
With modified replacement	Within hospital matching	PS Model 1	0.102	0.953	0.174	0.995	0.157	0.982	0.114	0.970	0.114	0.970
		PS Model 2	0.104	0.961	0.176	0.999	0.159	0.995	0.115	0.977	0.115	0.977
		PS Model 3	0.097	0.961	0.160	0.000	0.146	0.000	0.120	0.970	0.120	0.971
		PS Model 4	0.102	0.960	0.174	1.000	0.157	0.997	0.114	0.971	0.114	0.971
		PS Model 5	0.102	0.959	0.174	0.999	0.157	0.997	0.114	0.969	0.114	0.969
	Across hospital matching	PS Model 1	0.097	0.956	0.165	0.958	0.150	0.928	0.111	0.954	0.111	0.953
		PS Model 2	0.095	0.946	0.145	0.339	0.147	0.727	0.107	0.962	0.108	0.964
		PS Model 3	0.092	0.948	0.141	0.000	0.141	0.000	0.112	0.969	0.113	0.964
		PS Model 4	0.097	0.951	0.165	0.976	0.150	0.957	0.111	0.952	0.111	0.948
		PS Model 5	0.097	0.965	0.165	0.971	0.151	0.978	0.111	0.953	0.111	0.952
	Modified across hospital matching	PS Model 1	0.095	0.952	0.162	0.958	0.147	0.939	0.108	0.960	0.108	0.959
		PS Model 2	0.092	0.951	0.146	0.362	0.142	0.310	0.104	0.963	0.105	0.969
		PS Model 3	0.091	0.949	0.144	0.000	0.140	0.000	0.112	0.960	0.112	0.957
		PS Model 4	0.095	0.953	0.163	0.969	0.147	0.972	0.108	0.955	0.108	0.957
		PS Model 5	0.095	0.951	0.163	0.960	0.147	0.951	0.108	0.959	0.108	0.959

Table A.11: Monte Carlo bias and standard deviation using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use modified matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD	bias	MCSD
With modified replacement	Within hospital matching	PS Model 1	−0.001	0.102	0.016	0.133	0.016	0.133	0.003	0.108	0.003	0.108
		PS Model 2	−0.002	0.102	0.007	0.120	0.007	0.120	0.000	0.108	0.000	0.108
		PS Model 3	0.000	0.096	1.179	0.126	1.179	0.126	−0.002	0.109	−0.005	0.109
		PS Model 4	−0.002	0.101	0.015	0.113	0.015	0.113	0.001	0.106	0.001	0.106
		PS Model 5	−0.002	0.100	0.018	0.113	0.018	0.113	0.002	0.106	0.002	0.106
	Across hospital matching	PS Model 1	0.001	0.097	−0.036	0.158	−0.084	0.145	0.008	0.113	0.010	0.113
		PS Model 2	0.002	0.095	0.377	0.301	0.205	0.128	0.006	0.101	−0.006	0.101
		PS Model 3	0.001	0.092	1.330	0.268	1.238	0.130	0.006	0.108	−0.007	0.108
		PS Model 4	0.001	0.097	−0.011	0.142	−0.081	0.125	0.009	0.112	0.010	0.113
		PS Model 5	0.002	0.096	0.026	0.144	−0.050	0.126	0.009	0.113	0.011	0.113
	Modified across hospital matching	PS Model 1	0.000	0.094	0.093	0.137	0.097	0.130	0.012	0.104	0.010	0.104
		PS Model 2	0.001	0.093	0.349	0.252	0.336	0.123	0.017	0.101	0.007	0.101
		PS Model 3	0.000	0.092	1.331	0.220	1.308	0.134	−0.016	0.107	−0.026	0.107
		PS Model 4	0.000	0.094	0.105	0.117	0.098	0.109	0.012	0.102	0.010	0.102
		PS Model 5	0.000	0.094	0.129	0.117	0.118	0.109	0.009	0.103	0.008	0.103

Table A.12: Monte Carlo average of estimated standard errors and coverage probability using the outcome regression method, proposed methods and Li et al method. For each method, we use 5 PS models (not for the outcome regression method) and 5 outcome models described in Section 2.3. We use modified matching with replacement. The results are based on 1000 Monte Carlo data sets with number of hospitals equals to 30 and hospital size follows uniform distribution U(30,170).

		PS Model	**Outcome Model 1**		**Outcome Model 2**		**Outcome Model 3**		**Outcome Model 4**		**Outcome Model 5**	
			ASE	CP	ASE	CP	ASE	CP	ASE	CP	ASE	CP
With modified replacement	Within hospital matching	PS Model 1	0.102	0.946	0.174	0.988	0.156	0.975	0.114	0.958	0.114	0.958
		PS Model 2	0.103	0.955	0.176	0.997	0.158	0.990	0.115	0.964	0.115	0.964
		PS Model 3	0.097	0.951	0.160	0.000	0.146	0.000	0.120	0.968	0.120	0.967
		PS Model 4	0.102	0.957	0.174	0.998	0.156	0.994	0.114	0.956	0.114	0.956
		PS Model 5	0.102	0.957	0.174	0.997	0.156	0.994	0.114	0.958	0.114	0.958
	Across hospital matching	PS Model 1	0.097	0.949	0.165	0.949	0.150	0.918	0.111	0.940	0.111	0.938
		PS Model 2	0.095	0.956	0.145	0.390	0.146	0.743	0.107	0.961	0.108	0.958
		PS Model 3	0.092	0.953	0.141	0.000	0.141	0.000	0.112	0.958	0.113	0.958
		PS Model 4	0.097	0.947	0.165	0.981	0.150	0.955	0.111	0.944	0.111	0.942
		PS Model 5	0.097	0.961	0.165	0.975	0.150	0.977	0.111	0.943	0.111	0.940
	Modified across hospital matching	PS Model 1	0.094	0.948	0.162	0.948	0.147	0.936	0.107	0.954	0.107	0.956
		PS Model 2	0.092	0.949	0.146	0.405	0.141	0.314	0.104	0.955	0.104	0.961
		PS Model 3	0.091	0.954	0.144	0.000	0.140	0.000	0.112	0.959	0.112	0.953
		PS Model 4	0.095	0.956	0.162	0.963	0.147	0.959	0.107	0.962	0.107	0.962
		PS Model 5	0.094	0.952	0.162	0.947	0.147	0.946	0.107	0.964	0.107	0.962

APPENDIX B

Appendix for Chapter III

B.1 Asymptotic Properties

We heuristically prove that the estimator that solves Equation (3.2) is consistent and asymptotically normal.

The Equation (3.2) can be written into the form:

$$\sum_{i=1}^{n}\left[\frac{A_{ik}I(C_i \geq t)\{dN^*_{ik}(t) - Y^*_{ik}(t)d\Lambda_k(t)\}}{p_{ik}(\theta)p^c_i(\gamma,t)} - \frac{A_{ik}-p_{ik}(\theta)}{p_{ik}(\theta)}h_k(t,Z_i)dt\right] = 0. \quad \text{(B.1)}$$

We define $m(t,\theta,\gamma)$ as

$$m(t,\theta,\gamma) = \frac{A_{ik}I(C_i \geq t)\{dN^*_{ik}(t) - Y^*_{ik}(t)d\Lambda_k(t)\}}{p_{ik}(\theta)p^c_i(\gamma,t)} - \frac{A_{ik}-p_{ik}(\theta)}{p_{ik}(\theta)}h_k(t,Z_i)dt.$$

Then the left part of Equation (B.1) is a summation of independent and identically distributed quantities $m(t, \theta, \gamma)$. The expectation of the first part of $m(t, \theta, \gamma)$ is:

$$\begin{aligned}
&E\left[\frac{A_{ik}I(C_i \geq t)\{dN^*_{ik}(t) - Y^*_{ik}(t)d\Lambda_k(t)\}}{p_{ik}(\theta)p^c_i(\gamma, t)}\right] \\
&= E\left[E\left\{\frac{A_{ik}I(C_i \geq t)\{dN^*_{ik}(t) - Y^*_{ik}(t)d\Lambda_k(t)\}}{p_{ik}(\theta)p^c_i(\gamma, t)}\middle| Z_i\right\}\right] \\
&= E\{dN^*_{ik}(t) - Y^*_{ik}(t)d\Lambda_k(t)\}E\left\{\frac{A_{ik}I(C_i \geq t)}{p_{ik}(\theta)p^c_i(\gamma, t)}\middle| Z_i\right\} \\
&= E\{dM^*_{ik}(t)\}E\left\{\frac{A_{ik}I(C_i \geq t)}{p_{ik}(\theta)p^c_i(\gamma, t)}\middle| Z_i\right\} \\
&= 0 * E\left\{\frac{A_{ik}I(C_i \geq t)}{p_{ik}(\theta)p^c_i(\gamma, t)}\middle| Z_i\right\}\Bigg] \\
&= 0,
\end{aligned}$$

and the expectation of the second part of $m(t, \theta, \gamma)$ is:

$$\begin{aligned}
E\left\{\frac{A_{ik} - p_{ik}(\theta)}{p_{ik}(\theta)}h_k(t, Z_i)dt\right\} &= E\left[E\left\{\frac{A_{ik} - p_{ik}(\theta)}{p_{ik}(\theta)}h_k(t, Z_i)dt\middle| Z_i\right\}\right] \\
&= E\left[h_k(t, Z_i)dtE\left\{\frac{A_{ik} - p_{ik}(\theta)}{p_{ik}(\theta)}\middle| Z_i\right\}\right] \\
&= E\Big[h_k(t, Z_i)dt * 0\Big] \\
&= 0.
\end{aligned}$$

Thus, the expectation of $m(t, \theta, \gamma)$ is equal to 0. As $m(t, \theta, \gamma)$ has expectation 0, $\widehat{\Lambda}_j(t)$ is an M-estimator and therefore is consistent and asymptotically normal.

B.2 Generating Censoring and Survival Times

For ease of presentation, we denote $X = (1, Z_1, Z_2, Z_3, A)^T$.

B.2.1 Censoring Time

The censoring time C is generated from a Cox model and $\lambda^c(t|Z(t), Z) = \lambda_0^c\{\exp(1.5Z_3 + 0.3Z(t))\}$. We assume C is exponentially distributed and $\lambda_0^c(t) = \lambda^c = 1.5$. To generate C, we follow the same procedure with generating the survival time T in Scenario 1.

B.2.2 Scenario 1

In the first scenario, the survival time T is generated from a Cox model and $\lambda(t|Z(t), X) = \lambda_0(t)\exp\{\eta X + \beta Z(t)\}$. We assume T is exponentially distributed and $\lambda_0(t) = \lambda$, then then cumulative hazard function

$$\begin{aligned}\Lambda(t|Z(t), X) &= \lambda \int_0^t \exp\{\eta X + \beta Z(u)\}du \\ &= \lambda \exp(\eta X) \int_0^t \exp\{\beta Z(u)\}du.\end{aligned}$$

If $t < t^*$,

$$\Lambda(t|Z(t), X) = \lambda \exp(\eta X) \int_0^t \exp\{0\}du = \lambda \exp(\eta X)t;$$

If $t \geq t^*$,

$$\begin{aligned}\Lambda(t|Z(t), X) &= \lambda \exp(\eta X) \left\{ \int_0^{t^*} \exp(0)du + \int_{t^*}^t \exp(\beta)du \right\} \\ &= \lambda \exp(\eta X)\{1 - \exp(\beta)\}t^* + \lambda \exp(\eta X + \beta)t.\end{aligned}$$

As $S(t|Z(t), X) = \exp\{-\Lambda(t|Z(t), X)\}$ and $S(t|Z(t), X)$ follows uniform distribution $U(0, 1)$,

$$T = \begin{cases} \dfrac{-\log(u)}{\lambda \exp(\eta X)} & \text{, if } -\log(u) \leq \lambda \exp(\eta X) t^* \\ \dfrac{-\log(u) - \lambda \exp(\eta X)\{1 - \exp(\beta)\} t^*}{\lambda \exp(\eta X + \beta)} & \text{, if } -\log(u) > \lambda \exp(\eta X) t^*, \end{cases}$$

where $u \sim U(0, 1)$.

In this scenario, we let $\lambda = 0.6$, $\eta = (-0.5,\ 1,\ -1,\ 1,\ -0.5)$ and $\beta = 0.3$.

B.2.3 Scenario 2

In the second scenario, the survival time T is generated from an accelerated failure time (AFT) model. For the time dependent covariate $Z(t)$, *Cox* (2018) proposed an extension of the AFT model:

$$U = \int_0^T \exp\{\beta Z(s)\} ds,$$

where U is the latent baseline survival time.

Based on this idea, we consider both baseline covariates and time-varying covariate:

$$U = \int_0^T \exp\{\eta X + \beta Z(s)\} ds.$$

If $T \leq t^*$,

$$U = \int_0^T \exp\{\eta X + 0\} ds = \exp(\eta X) t;$$

If $T > t^*$,

$$\begin{aligned} U &= \int_0^{t^*} \exp(\eta X) ds + \int_{t^*}^{T} \exp(\eta X + \beta) ds \\ &= \exp(\eta X) t^* + \exp(\eta X + \beta)(T - t^*). \end{aligned}$$

We let $U = e^{\epsilon}$, where $\epsilon \sim$N(0,1), then

$$T = \begin{cases} \dfrac{U}{\exp(\eta X)} & \text{, if } \exp(\eta \mathrm{X})\mathrm{U} \leq t^* \\ \dfrac{U - \exp(\eta X)\{1 - \exp(\beta)\}t^*}{\exp(\eta X + \beta)} & \text{, if } \exp(\eta \mathrm{X})\mathrm{U} > t^*. \end{cases}$$

In this scenario, we let $\eta = (0.2, -0.5, 0.5, -0.5, 0.3)$ and $\beta = -0.3$.

B.2.4 Scenario 3

In the third scenario, the survival time T is generated from a Cox model and we assume T follows lognormal distribution. The cumulative hazard function

$$\Lambda(t|Z(t), X) = \int_0^t \lambda_0(u) \exp\{\eta X + \beta Z(u)\} du.$$

If t $< t^*$,

$$\begin{aligned} \Lambda(t|Z(t), X) &= \int_0^t \lambda_0(u) \exp\{\eta X + 0\} du \\ &= \exp(\eta X) \Lambda_0(t). \end{aligned}$$

Hence

$$\Lambda_0(t) = \frac{\Lambda(t|Z(t), X)}{\exp(\eta X)}. \tag{B.2}$$

If $t \geq t^*$,

$$\begin{aligned}\Lambda(t|Z(t),X) &= \int_0^{t^*} \lambda_0(u)\exp(\eta X)du + \int_{t^*}^{t} \lambda_0(u)\exp(\eta X+\beta)du \\ &= \exp(\eta X)\Lambda_0(t^*) + \exp(\eta X+\beta)\{\Lambda_0(t)-\Lambda_0(t^*)\} \\ &= \exp(\eta X)[1-\exp(\beta)]\Lambda_0(t^*) + \exp(\eta X+\beta)\Lambda_0(t).\end{aligned}$$

Hence

$$\Lambda_0(t) = \frac{\Lambda(t|Z(t),X) - \exp(\eta X)[1-\exp(\beta)]\Lambda_0(t^*)}{\exp(\eta X+\beta)}. \tag{B.3}$$

The cumulative baseline hazard function for the lognormal distributed T is:

$$\Lambda_{LN,0}(t) = -\log\Big\{1-\Phi\Big(\frac{\log(t)}{\sigma}\Big)\Big\}. \tag{B.4}$$

As $S(t|Z(t),X) = \exp\{-\Lambda(t|Z(t),X)\}$ and $S(t|Z(t),X)$ follows uniform distribution $U(0,1)$, we generate T by inversing the cumulative baseline hazard function (B.4), plugging in the right sides of Equation (B.2) and (B.3) respectively, and replacing $\Lambda(t|Z(t),X)$ with $\{-\log(u)\}$, where u is from the uniform distribution $U(0,1)$, i.e.

$$T = \begin{cases} \Lambda_{LN,0}^{-1}\Big(-\dfrac{\log(u)}{\exp(\eta X)}\Big) & \text{, if } -\dfrac{\log(u)}{\exp(\eta X)} \leq t^* \\ \Lambda_{LN,0}^{-1}\Big(-\dfrac{\log(u)+\exp(\eta X)[1-\exp(\beta)]\Lambda_0(t^*)}{\exp(\eta X+\beta)}\Big) & \text{, if } -\dfrac{\log(u)}{\exp(\eta X)} > t^*. \end{cases}$$

In this scenario, we let $\eta = (0.2,\ 1,\ -1,\ 1,\ -0.5)$ and $\beta = 0.3$.

B.3 Practical Implementations

Take the treatment ($A = 1$) as an example, we show how to implement our method.

Step 1. Fit treatment assignment model

We fit a logistic model with parameter θ for treatment assignment using Z as the covariates, i.e. $\text{logit}\{P(A_i = 1|Z_i)\} = \theta^T Z_i$. θ is estimated through maximum likelihood estimator $\widehat{\theta}$, which solves the estimating equation

$$\sum_{i=1}^{n} Z_i\{A_i - \text{expit}(\theta^T Z_i)\} = 0.$$

Step 2. Calculate IPTW weight

We define $p_{i1}(\theta) = P(A_i = 1|Z_i, \theta)$. The IPTW weight $w_{i1}(\widehat{\theta}) = I(A_i = 1)/p_{i1}(\widehat{\theta})$.

Step 3. Fit censoring model

We fit a Cox model for censoring where the time to censoring is outcome. Let $X(t) = \{A, Z_3, Z(t)\}^T$, then the Cox model can be fit as $\lambda_{1,i}\{t|X_i(t)\} = \lambda_{01}(t)\exp\{\gamma_1^T X_i(t)\}$. Estimators for γ and $\Lambda_{01}(t)$ can be obtained by the maximum partial likelihood (PL) estimator, $\widehat{\gamma}$ and the Breslow estimator, $\widehat{\Lambda}_{01}(t)$, respectively. The Breslow estimator for Λ_{01} is defined as

$$\widehat{\Lambda}_{01}(t) = \int_0^t \frac{\sum_{i=1}^n dN_{i1}(t)}{\sum_{i=1}^n \exp(\widehat{\gamma}^T X_i(t))Y_{i1}(t)}.$$

Step 4. Calculate IPCW weight

Let $p_i^c(\gamma, t)$ denote the probability that C_i is greater than t given $\{X_i(t), \gamma\}$. For each patient at each time point, we calculate the IPCW weight $w_i^c(\widehat{\gamma}, t) = I(C_i \geq$

$t)/p_i^c(\widehat{\gamma}, t)$.

Step 5. Find optimal β_1

As described in Section 3.3, we used the OLS approach to minimize the variance of estimators within the subclass. To obtain the OLS estimator, we let $A_{i1} = I(A_i = 1)$, and we treat $\left[\frac{A_{i1}\{dN_i(t) - Y_i(t)d\widehat{\Lambda}_1^{DIPW}(t)\}}{p_{i1}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\right]$ as the outcome, and $\left[\frac{A_{i1}-p_{i1}(\widehat{\theta})}{p_{i1}(\widehat{\theta})}g_1(Z_i)\right]$ as covariates, where $g_1(Z_i)$ is a vector of basis function of Z_i, e.g. $g_1(Z_i) = (1, Z_{1,i}, Z_{2,i}, Z_{3,i})$. Then the optimal $\beta_1(t)$ was estimated by

$$\widehat{\beta}_{1,opt}(t)dt = \left[\sum_{i=1}^{n}\left\{\frac{A_{i1} - p_{i1}(\widehat{\theta})}{p_{i1}(\widehat{\theta})}g_1(Z_i)\right\}\left\{\frac{A_{i1} - p_{i1}(\widehat{\theta})}{p_{i1}(\widehat{\theta})}g_1(Z_i)\right\}^T\right]^{-1}$$
$$\left[\sum_{i=1}^{n}\left\{\frac{A_{i1} - p_{i1}(\widehat{\theta})}{p_{i1}(\widehat{\theta})}g_1(Z_i)\right\}\frac{A_{i1}\{dN_i(t) - Y_i(t)d\widehat{\Lambda}_1^{DIPW}(t)\}}{p_{i1}(\widehat{\theta})p_i^c(\widehat{\gamma},t)}\right].$$

Step 6. Augmented estimator

The cumulative harzard $\Lambda_1(t)$ is estimated by $\widehat{\Lambda}_1(t)$,

$$\widehat{\Lambda}_1(t) = \int_0^t \frac{\sum_{i=1}^n w_{i1}(\widehat{\theta})w_i^c(\widehat{\gamma}, u)dN_i(u) + \{1 - w_{i1}(\widehat{\theta})\}\widehat{\beta}_{1,opt}(u)g_1(Z_i)du}{\sum_{i=1}^n w_{i1}(\widehat{\theta})w_i^c(\widehat{\gamma}, u)Y_i(u)}.$$

Step 7. Estimate survival probability

$S_1(t)$ is estimated by $\widehat{S}_1(t) = e^{-\widehat{\Lambda}_1(t)}$.

Step 8. Estimate restricted mean lifetime

$\mu_1(t)$ is estimated by $\widehat{\mu}_1 = \int_0^L \widehat{S}_1(u)du$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A., and G. W. Imbens (2006), Large sample properties of matching estimators for average treatment effects, *econometrica*, *74*(1), 235–267.

Abadie, A., and G. W. Imbens (2008), On the failure of the bootstrap for matching estimators, *Econometrica*, *76*(6), 1537–1557.

Anstrom, K. J., and A. A. Tsiatis (2001), Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data, *Biometrics*, *57*(4), 1207–1218.

Arpino, B., and M. Cannas (2016), Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score, *Statistics in medicine*.

Arpino, B., and F. Mealli (2011), The specification of the propensity score in multilevel observational studies, *Computational Statistics & Data Analysis*, *55*(4), 1770–1780.

Austin, P. C. (2008), A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003, *Statistics in medicine*, *27*(12), 2037–2049.

Austin, P. C. (2011), An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate behavioral research*, *46*(3), 399–424.

Austin, P. C. (2014), A comparison of 12 algorithms for matching on the propensity score, *Statistics in medicine*, *33*(6), 1057–1069.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer (2006), Variable selection for propensity score models, *American journal of epidemiology*, *163*(12), 1149–1156.

Chen, P.-Y., and A. A. Tsiatis (2001), Causal inference on the difference of the restricted mean lifetime between two groups, *Biometrics*, *57*(4), 1030–1038.

Cox, D. R. (1992), Regression models and life-tables, in *Breakthroughs in statistics*, pp. 527–541, Springer.

Cox, D. R. (2018), *Analysis of survival data*, Routledge.

Dehejia, R. H., and S. Wahba (1999), Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American statistical Association*, *94*(448), 1053–1062.

Gu, X. S., and P. R. Rosenbaum (1993), Comparison of multivariate matching methods: Structures, distances, and algorithms, *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

Hill, J., and J. P. Reiter (2006), Interval estimation for treatment effects using propensity score matching, *Statistics in medicine*, *25*(13), 2230–2256.

Hill, J. L., J. P. Reiter, and E. L. Zanutto (2004), A comparison of experimental and observational data analyses, *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family*, pp. 49–60.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007), Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Political analysis*, *15*(3), 199–236.

Kim, J.-S., and P. M. Steiner (2015), Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy, in *Quantitative Psychology Research*, pp. 293–306, Springer.

Li, F., A. M. Zaslavsky, and M. B. Landrum (2013), Propensity score weighting with multilevel data, *Statistics in medicine*, *32*(19), 3373–3387.

Li, L., and T. Greene (2013), A weighting analogue to pair matching in propensity score analysis, *The international journal of biostatistics*, *9*(2), 215–234.

Li, X., and P. Ding (2019), Rerandomization and regression adjustment, *arXiv preprint arXiv:1906.11291*.

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

Mundlak, Y. (1978), On the pooling of time series and cross section data, *Econometrica: journal of the Econometric Society*, pp. 69–85.

Oelrich, O. (2014), Causal inference using propensity score matching in clustered data.

Rickles, J. H., and M. Seltzer (2014), A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study, *Journal of Educational and Behavioral Statistics*, *39*(6), 612–636.

Robins, J. M., and A. Rotnitzky (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS epidemiology*, pp. 297–331, Springer.

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

Rosenbaum, P. R. (1989), Optimal matching for observational studies, *Journal of the American Statistical Association*, *84*(408), 1024–1032.

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

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

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

Rubin, D. B. (1973), The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics*, pp. 185–203.

Rubin, D. B. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies., *Journal of educational Psychology*, *66*(5), 688.

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

Rubin, D. B. (1979), Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association*, *74*(366a), 318–328.

Rubin, D. B., and N. Thomas (2000), Combining propensity score matching with additional adjustments for prognostic covariates, *Journal of the American Statistical Association*, *95*(450), 573–585.

Schafer, J. L., and J. Kang (2008), Average causal effects from nonrandomized studies: a practical guide and simulated example., *Psychological methods*, *13*(4), 279.

Schaubel, D. E., and G. Wei (2011), Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring, *Biometrics*, *67*(1), 29–38.

Steiner, P., J.-S. Kim, and F. Thoemmes (2013), Matching strategies for observational multilevel data, in *JSM proceedings*, pp. 5020–5032.

Stuart, E. A. (2008), Developing practical recommendations for the use of propensity scores: Discussion of'a critical appraisal of propensity score matching in the medical literature between 1996 and 2003'by peter, *Statistics in medicine*, *27*(12), 2062–2065.

Stuart, E. A. (2010), Matching methods for causal inference: A review and a look forward, *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Thoemmes, F. J., and S. G. West (2011), The use of propensity scores for nonrandomized designs with clustered data, *Multivariate Behavioral Research*, *46*(3), 514–543.

Tsiatis, A. (2007), *Semiparametric theory and missing data*, Springer Science & Business Media.

Yang, S. (2016), Propensity score weighting for causal inference with multi-stage clustered data, *arXiv preprint arXiv:1607.07521*.

Zhang, M. (2015), Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials, *Lifetime data analysis*, *21*(1), 119–137.

Zhang, M., and D. E. Schaubel (2011), Estimating differences in restricted mean lifetime using observational data subject to dependent censoring, *Biometrics*, *67*(3), 740–749.

Zhang, M., and D. E. Schaubel (2012a), Contrasting treatment-specific survival using double-robust estimators, *Statistics in medicine*, *31*(30), 4255–4268.

Zhang, M., and D. E. Schaubel (2012b), Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies, *Biometrics*, *68*(4), 999–1009.

Zhang, M., and Y. Wang (2012), Estimating treatment effects from a randomized clinical trial in the presence of a secondary treatment, *Biostatistics*, *13*(4), 625–636.

Zhang, M., and Y. Wang (2013), Adjusting for observational secondary treatments in estimating the effects of randomized treatments, *Biostatistics*, *14*(3), 491–501.

Zhang, M., A. A. Tsiatis, and M. Davidian (2008), Improving efficiency of inferences in randomized clinical trials using auxiliary covariates, *Biometrics*, *64*(3), 707–715.