UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ
DEGLI STUDI
DI PERUGIA

iNδAM
Istituto Nazionale
di Alta Matematica

Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXIV

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Matteo Focardi

Weighting Methods For Causal Inference With Survival Outcomes

Settore Scientifico Disciplinare SECS-S/01

Tutore
Prof. Fabrizia Mealli
Mealli

Dottorando:
Anahita Nodehi
Anahita Nodehi

Co-tutore
Prof. Alessandra Mattei
Alessandra Mattei

Coordinatore
Prof. Matteo Focardi
Matteo Focardi

Anni 2018/2022

Weighting methods for causal inference with survival outcomes

Anahita Nodehi

Supervisor: Fabrizia Mealli
Co-Supervisor: Alessandra Mattei

Contents

Abstract

This dissertation aims to propose a technique for estimating the causal effects of exposure on survival outcomes using the Rubin Causal Model (RCM), a framework for defining causal estimands, discussing assumptions, and developing methods for drawing inferences on causal effects. From a substantive perspective, the research was motivated by the evaluation of the effect of two different treatments (*Interferon* versus *Azathioprine*) on time to the first worsening of Multiple Sclerosis (MS) disease. The study uses data from an observational study on patients with MS collected between 1981 and 2019 in Tuscany, Italy. The causal analysis of this study raises several challenges due to the unknown treatment assignment mechanism, and the survival outcome is subject to two different covariate-dependent censoring mechanisms: administrative censoring and treatment switching. Then, using Marginal Structural Cox models, we propose a new weighting method to adjust for observed confounders and correct selection bias due to different types of censoring.

Acknowledgements

I would like to thank everyone who contributed to the work described in this thesis. First and foremost, I thank my supervisor, Professor Fabrizia Mealli, for accepting me into her group. She has been a mentor to me and, because of her deep and broad knowledge of statistics and her ‘humanistic’ approach towards it, also an inexhaustible source of inspiration.

My deepest gratitude also goes to my co-supervisor, Professor Alessandra Mattei. I want to thank her for all the hours she spent talking with me and her valuable suggestions. Her advice about this thesis and, more generally, about the different aspects of academia has been constructive for my training as a researcher. Additionally, I would like to thank my committee members, Professor Pacini (University of Pisa), Professor Catelan (University of Padova) and Professor Geraci (Sapienza University of Rome) for their valuable comments on my work.

I wish to thank Professor Stijn Vansteelandt, who gave me the opportunity to join his research group at Ghent University in Belgium. He has been extremely kind to me since the first day, and his extensive knowledge of causal inference significantly impacted my education. Although my stay in Ghent was short, it was one of the best periods of my Ph.D.

Many thanks also to Professor Luca Massacesi and Dr. Alice Mariottini from the Multiple Sclerosis Regional Referral Centre, Careggi Hospital and Department of Neurosciences, Drug and Child Health, University of Florence, Florence, Italy, for providing data on the Multiple Sclerosis study, which is the motivating study of this thesis, and for their helpful feedback and suggestions.

Finally, I would like to acknowledge my family, who supported me during my life. First and foremost, I would like to thank Mom, Dad, Shahab, and Mehrab for their constant love and support. In addition, all my gratitude goes to Majid, an unexpected gift along the way for whom no thanks would ever be enough. He has given me the strength to hold on even in the most challenging times and has constantly encouraged me to believe in myself. We shared every moment of this Ph.D. journey from the beginning, and now that it is coming to an end, my greatest hope is that it provided us with the awareness and the tools to face a much more challenging journey called life. Hand by hand, always together.

Chapter 1

Introduction

In medical research, making inferences about the effects of treatments and interventions is challenging. Estimating the effects of treatments is known as causal inference. Observational medical studies are used in many research investigations to estimate the effects of treatments, exposures, and interventions on health outcomes.

The potential outcomes approach of causal inference is a framework that allows the definition of a treatment's causal effect as a comparison of potential outcomes, the discussion of assumptions that allow for the identification of such causal effects from available data, and the development of methods for estimating causal effects under these assumptions [Rubin, 1975, 1978, Imbens and Rubin, 2015]. We refer to this framework as the Rubin Causal Model after Holland [1986a] (RCM). The RCM defines a *treatment's causal effect* as a contrast of potential outcomes, one of which will be observed while the others will be missing and become counterfactual once the treatment is assigned. In that respect, the assignment mechanism is explicitly defined as a probability model for how units receive the different treatment levels. A causal inference problem is thus understood from this perspective as a problem of missing data, where the assignment mechanism is explicitly modeled as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects. In experimental study, the assignment mechanism is known and controlled by the researcher. In contrast, in observational studies, treatment conditions and assignment timing are observed after data collection. Frequently, the data are collected for purposes other than the study. As a result, the researcher does not control the treatment assignment mechanism. In addition, the lack of randomization makes it challenging to ensure that covariates are balanced between treatment groups, resulting in systematic differences between the treatment group and the control group. Practitioners do not traditionally distinguish between the design and analysis phases when analyzing observational data. To make objective causal inference from observational studies, we must address these challenges [Dominici et al., 2020].

On the other hand, survival analysis is concerned with statistical procedures in which the *outcome* variable of interest is *time-to-event*, such as time to death, time to recovery, or time to infection. Nevertheless, the precise event times are unknown in many studies. A most common complexity of observed survival data is the presence of *censoring* on the survival time. On that note, some patients may not experience the event of interest and are censored due to the end of the follow-up (i.e., administrative censoring). A subject may be lost to follow-up before the event occurs (censoring due to loss of follow-up), or he or she may switch treatments due to unwanted side effects that prevent him or her from continuing to take the treatment (censoring due to switching the treatment). Survival analysis has been expanded to deal with situations where the precise event times are unknown. Some statistical methods designed to account for censored observations imply that patients' withdrawal from a study is independent of the event of interest (completely ignorable censoring). However, some covariates might be associated with lifetime and censoring mechanisms in real-world situations, resulting in ignorable censoring conditional on those covariates. Applying the classical survival techniques assumes independent censoring (i.e., completely ignorable censoring) may under-or over-estimate the survival time in the case of covariate-dependent censoring (i.e., ignorable censoring conditional on covariates). Statistical methods in survival analysis were mainly developed to address the presence of censoring (under completely ignorable or covariate-dependent censoring assumption) and the non-symmetric shape of the distribution of survival time.

This research aims to provide an understanding of the methodologies that can be adopted to estimate causal effects of exposure on survival outcomes in medical observational studies: the problem of how to adjust for observed confounders and how to deal with covariate-dependent censoring mechanisms. Notably, we must deal with the unknown assignment mechanism and simultaneously adjust for two different covariate-dependent censoring (administrative and switching the treatment). The objective of this research is to provide both methodological and empirical contributions. In literature, Robins and colleagues [Robins et al., 2000, Robins, 2000a, Hernán et al., 2000] have proposed a new class of causal models called marginal structural models. According to [Hernán et al., 2000], the analysis based on weighted samples gives an asymptotically unbiased estimate of the causal parameter of interest. In particular, we propose a novel WEIGHTING method in the Marginal Structural Model to overcome all challenges simultaneously. Furthermore, apply the proposed method to a medical observational study aimed to assess the relative effectiveness of two treatments Interferon (INF) and Azathioprine (AZA), for patients with Multiple Sclerosis (MS) on Progression-Free

Survival (PFS) in Italy. Prespecification of the functional form of covariates for model building is problematic, thus prompting the use of relatively fast algorithms.

The remainder of this thesis is organized as follows: Chapter 2 introduces our motivating study. Chapter 3 presents notation for causal inference in observational studies and a common theoretical background. Furthermore, Chapter 4 discusses survival analysis, explains the difficulties that arise based on different censoring mechanisms (i.e., administrative censoring and treatment switching), and discusses the Cox models' estimation methods and their diagnostic. The definition of causal estimands in survival settings is presented in Chapter 5. Moreover, in Chapter 5, we demonstrate how to overcome all of the concerns raised in the MS data set by using the WEIGHTING methods in the Marginal Structural Cox Model. It is the thesis's novelty. Next, we investigate our proposals in comprehensive simulation analysis. Notably, simulation studies' results help analysts choose which methods are most appropriate for answering research questions using their data. After that, we apply the proposed method to estimate the causal effects of two treatments on worsening among patients with a high risk of Multiple Sclerosis (MS) localized the Careggi hospital data set in Section 6. In Chapter 7, we address additional concerns related to decisions regarding the cost-effectiveness of two treatments over a shorter period (such as five years). Finally, we end with a discussion and suggest some further research in Section 8.

Chapter 2

Motivating Study

Multiple Sclerosis (MS) involves an immune-mediated process in which an abnormal response of the body's immune system is directed against the central nervous system. Symptoms of MS are unpredictable and vary in type and severity from one person to another and the same person over time. Worldwide, more than 2.3 million people have a diagnosis of MS. In the United States, a recently completed prevalence study funded by the National MS Society has estimated that nearly 1 million people over 18 live with a diagnosis of MS. There is no cure for multiple sclerosis. Treatment typically focuses on speeding recovery from attacks, slowing the progression of the disease, and managing MS symptoms. Several consensus working groups in MS [Lublin et al., 2014, Scolding et al., 2015, MAGNIMS, 2015] have highlighted the need for further research to establish optimum treatment and monitoring strategies in MS.

This thesis is motivated by an *observational study* aimed to assess the relative effectiveness of two treatments - Interferon (INF) and Azathioprine (AZA) - on progression-free survival for MS patients. Henceforth, we refer to his study as the MS study. The study involves 594 patients enrolled between 1981 and 2019 at Tuscany region, Italy, and exposed to either Azathioprine or Interferon. Interferon is the treatment category that slows the disease's progression. These are a group of proteins that normally produce cells in response to viral infection and other incentives. For many years, several studies have shown interferon is effective using outcome variables defined in terms of relapses, disability, or MRI at various doses and by different routes of administration [IFNB et al., 1993, Jacobs et al., 1996, Arbizu et al., 2000, Paolillo et al., 2002, Milanese et al., 2003, Trojano et al., 2007, Brown et al., 2007, Coles et al., 2008, Melendez-Torres et al., 2018]. For instance, Shirani et al. [2012] investigated the association between Interferon exposure and disability progression (based on time to `EDSS`) in patients with relapsing-remitting MS. The 10-point Kurtzke Expanded Disability Status Scale (`EDSS`) is the most widely accepted clinical disability scale. The `EDSS` is considered the standard for monitoring patients with multiple sclerosis (MS). The paper by Karim et al. [2014] was the first to use the Marginal structural Cox model to estimate the causal associations between INF and time to reach a sustained EDSS score to 6 in the presence of time-dependent confounding (relapse rates) and selection bias. To remove the possible confounding effects of both time-varying and baseline confounders, they used weights [Hernán et al., 2000]. Their analysis found no association between INF and time to development of a sustained EDSS score of 6 over the follow-up, which is consistent with findings of other studies [Ebers et al., 2010, Shirani et al., 2012].

Although there is rich literature about Interferon, the validity of some findings has been brought to question [Brown et al., 2007, Gout, 2008, Koch et al., 2008] because of major methodological issues including selection bias [Trojano et al., 2007, Koch et al., 2008, Dimick and Livingston, 2010], small sample sizes [Paolillo et al., 2002, Pozzilli et al., 2005, Coppola et al., 2006], and insufficient follow-up [Arbizu et al., 2000, Milanese et al., 2003, Coppola et al., 2006]. Moreover, in some patients, Interferon shows no or little efficacy or is not well-tolerated Filippini et al. [2003]. This is because autoimmune pathogenetic mechanisms against central nervous system white matter underlying the development of MS lesions. Immunosuppressive medications have also been successfully used in the treatment of this disease. Azathioprine is the most widely used immunosuppressive treatment for MS. It is also an alternative to Interferon for treating MS because it is less expensive. Massacesi et al. [2005] evaluated the efficacy of Azathioprine therapy on new brain lesion suppression in MS. Massacesi et al. [2005] indicated for the first time that Azathioprine was effective in reducing MS new brain inflammatory lesions. A few studies have been undertaken over many years to assess the effectiveness of Interferon *versus/combined to* Azathioprine to prevent long-term disability accumulation. For instance, Havrdova et al. [2009] tried to assess the efficacy of combining Interferon with classical immunosuppressive agent groups such as Azathioprine (AZA) or corticosteroids on annualized relapse rate (ARR) at two years. They showed that combination treatment did not show superiority over Interferon monotherapy. Etemadifar et al. [2007] compared the relative efficacy of Interferon (INF) and Azathioprine (AZA) in the treatment of relapsing-remitting multiple sclerosis. Statistical analysis was based on an intention-to-treat principle. Comparison between groups receiving INF and AZA was made using an independent t-test and analysis of variance with repeated measures over time. The results demonstrated that both the INF formulations and AZA treatment groups decreased EDSS and relapse rate 12 months after the start of treatment [Etemadifar et al., 2007]. However, in next years, some studies were conducted by Benedetti et al. [2012], Massacesi et al. [2013], Massacesi et al. [2014] and Massacesi et al. [2016] to compare Azathioprine efficacy versus Interferon in relapsing-remitting

Multiple Sclerosis. These studies' results indicated that Interferon's efficacy is not superior to that of Azathioprine for patients with relapsing-remitting multiple sclerosis. The outcomes of these studies were Annualized Relapse Rate (ARR) and the number of new brain MRI lesions. All analyses were performed using the intention-to-treat (ITT) principle and a per-protocol analysis. The efficacy between the two treatments was judged by Poisson regression for ARR. The number of new brain MRI lesions was analyzed using the χ^2 test with one degree of freedom for rate comparison (based on Poisson regression); χ^2 test with two degrees of freedom for the number of relapsed patients; Kaplan-Meier curves, log-rank test and Cox proportional-hazards model for time to first relapse and Fisher's exact test for patients with no confirmed disability progression.

To the best of our knowledge, there is no study to compare Azathioprine efficacy versus Interferon over long-term follow-up. Consequently, the goal of this study is **to assess the relative effectiveness of** these two treatments on PFS, which is the time from treatment initiation until disease progression or worsening. A typical complexity of observed survival data is the presence of right censoring on the survival time. In particular, in our study different individuals will have different administrative censoring times due to the staggered entry of the study. This indicates that censoring time for all patients is not fixed (i.e. the follow-up is not fix-ended) and it is dependent on patient characteristics in our case. For example, young adults are at higher risk of suffering from MS and need treatment than older adults. This implies that the event is observed more often in younger than in old patients. On that note, in our example, some patients do not experience the event of interest and are censored due to the end of the follow-up (i.e., administrative censoring). Moreover, some patients in this study may experience treatment *switches* from one disease-modifying treatment to another or treatment discontinuation primarily due to ethical considerations, lack of efficacy, side effects, risk of the long-term adverse event, and pregnancy. Treatment switching often has a crucial impact on estimates of the effectiveness and cost-effectiveness of new treatments. It may be a clinically relevant question to estimate the efficacy that would have been observed if no patients had switched, for example, to estimate "real-life" clinical effectiveness for a health technology assessment. Several commonly used statistical methods are available that try to adjust time-to-event data to account for treatment switching, ranging from naive exclusion and censoring approaches to more complex inverse probability of censoring weighting and rank-preserving structural failure time models [Watkins et al., 2013]. We consider treatment switching as an additional censoring. The switching censoring mechanism may depend on the covariates that have led to this change of treatment. Thus, in our data set, both censoring mechanisms are assumed to be covariate-dependent censoring mechanisms, possibly conditional on different sets of covariates. We aim to adjust for observed confounders under the assumption that the assignment mechanism is strongly ignorable [Rosembaum and Rubin, 1983] and for selection bias due to dependent-censoring. Of note, we consider time-fixed treatment and first record for all patients, regardless of whether they have switched treatments.

Table 1 shows the number of patients in each group. In general, 85 patients switched the treatment during follow-up (it shows in Table 1 as $74 + 9 + 2 = 85$). Some covariates are recorded for each unit, and their descriptions are presented in Table 2. Based on expert-knowledge and a low missing rate (less than 5%), we consider complete cases in our data set. As a result, there are 562 patients, 211 exposed to AZA and 351 exposed to INF. Two important covariates are pre-treatment annualized relapse rate (`ARR_pre`) and progression index (`PI_pre`). `ARR_pre` measures an MS patient's average number of attacks per year. This measure is used to help detect and quantify levels of sustained disability. Formally,

$$\texttt{ARR\_pre} = \frac{\texttt{Relaps\_pre}}{\left(\frac{\texttt{Disease\_durat}}{12}\right)}.$$

`PI_pre` is a measure of the rapidity of the accumulation of disability, defined as the `Baseline.EDSS` divided by disease duration in years:

$$\texttt{PI\_pre} = \frac{\texttt{Baseline.EDSS}}{\left(\frac{\texttt{Disease\_durat}}{12}\right)}.$$

Table 1: Number of each groups

	Number	N_{AZA}	N_{INF}
Number of records	696	254	442
Patients	594	215	379
Patients (complete cases)	562	211	351
Single_obs	477	154	323
Patients who switch 1 time (with 2 records)	74	47	27
Patients who switch 2 times (with 3 records)	9	8	1
Patients who switch 3 times (with 4 records)	2	2	0

Table 2: The covariates which are recorded for all units (N = 562)

Baseline Covariates	*Description*
Age	Age
ARR_pre	Annualized Relapse Rate (ARR): average number of attacks per year
Disease_durat	Duration of disease (months)
Dummy_EDSS	0 if Baseline. EDSS < 4 and 1 otherwise
Baseline.EDSS	Expanded Disability Status Scale Baseline.EDSS $\in (0, 10)$
Gender	0 for Male and 1 for Female
PI_pre	The progression index
Relaps_pre	Number of relapse before therapy
Relapse_Dummy	1 if Relapse_pre is missing and 0 otherwise
Year	The year of treatment started

Chapter 3

Notation & theoretical background

In this chapter, causal inference primitives are introduced, along with various notions and assumptions that are required in observational studies.

1 Causal inference primitives and assignment mechanisms

1.1 Potential outcomes and SUTVA

We have information on a sample of n units, indexed by i ($i = 1, 2, \cdots, n$). Each unit i can be subject or exposed to a specific treatment or to alternative treatments, which could be different active treatments or no treatment at all. This thesis considers studies where the treatment variable is binary: $z = 1$ for the active treatment and $z = 0$ for the control. In our motivating study, a unit is a patients with MS disease who can be potentially exposed to either AZA ($z = 1$) or INF ($z = 0$).

The goal is to assess outcome Y at a certain time after each unit has been treated. For that purpose, each treatment-unit pair is linked to a potential outcome. As a result, two potential outcomes are following treatment for each unit i. They would be the $Y_i(1)$ value of the outcome variable Y if the unit were exposed to the active treatment and the $Y_i(0)$ value of Y at the same future point in time if the unit was exposed to the control treatment. Under the "Stable Unit Treatment Value Assumption [SUTVA, Rubin, 1980]" these definitions are valid.

Assumption 1 (Stable Unit Treatment Value Assumption)**.** *The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for* *each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

The depiction in Table 1 already requires assumptions for it to be adequate. In particular, it requires SUTVA to hold, which comprises two sub-assumptions. The SUTVA points out that notation such as $Y_i(z)$ effectively presupposes (I) that if individual i is given treatment z then individual i's outcome under treatment z does not depend on which treatment individual $i' \neq i$ received and (II) that there do not exist multiple versions of treatment a which might give rise to different outcomes depending on which version is administered. The first of these assumptions is sometimes referred to as "no-interference" which Rubin [1980] attributes to Cox [1958]; the second assumption is a "no-versions-of-treatment assumption" which Rubin attributes to Neyman and Iwaszkiewicz [1935].

Let $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$ be $n-$dimensional vectors with i^{th} elements equal to $Y_i(0)$ and $Y_i(1)$, respectively, and let $\boldsymbol{X}$ be a $n \times p$ matrix with i^{th} row equal to $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})$ collecting values of p covariates for unit i. Covariates are variables that are unaffected by the treatment. The $n \times (p+2)$ matrix of the covariates and the potential outcomes, $[\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)]$, is commonly referred to as *the Science* in the context of causal inference [Mattei et al., 2022].

Causal effects are defined at the unit level: a comparison of treatment and control possible outcomes, $Y_i(1)$ vs $Y_i(0)$, determines the treatment's *unit-level causal effect.* The difference, $Y_i(1) - Y_i(0)$, log-difference, $\log\{Y_i(1)\} - \log\{Y_i(0)\}$, or ratio $Y_i(1)/Y_i(0)$ between treatment and control potential outcomes are examples of unit-level causal effects. Individual causal effects are often summarized, with causal effects defined at the level of collections of units. *Summary causal effects* are comparisons between $Y_i(1)$ and $Y_i(0)$ for a common set of units, that is, comparisons of the ordered sets $\{Y_i(1), i \in G\}$ and $\{Y_i(0), i \in G\}$, where G is a collection of units. For example, one may be interested in summary causal effects for all units or sub-groups of units specified by, say, covariate values.

Table 1 summarizes the basic concepts of the potential outcome framework: the Science (covariates and potential outcomes), the collection of individual unit-level causal effects, and summary causal effects.

Some summary causal effects are typical unit-level causal effects in that they summarize unit-level causal effects for a group of units and hence correspond to characteristics of the joint distribution of potential outcomes. Other summary causal effects are marginal in the sense that they compare features of the marginal distributions of $Y_i(1)$ and $Y_i(0)$ for a set of units. For example, typical unit-level causal effects are the median and quantiles of a collection of units, while marginal causal effects are the difference between the medians or quantiles of $Y_i(1)$ and $Y_i(0)$ for a collection of units. *Average treatment effects*, which are both typical unit-level and marginal causal effects, are commonly defined causal estimands.

Table 1: The Science and causal estimands

				Potential Outcomes			
	Covariates			Treatment	Control	Unit-level	Summary
Units	X_1	$\dots$	X_p	$Y(1)$	$Y(0)$	Causal Effects	Causal Effects
1	X_{11}	$\dots$	X_{1p}	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ vs $Y_1(0)$	Comparison of $Y_i(1)$ vs $Y_i(0)$ for a common set of units
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
i	X_{i1}	$\dots$	X_{ip}	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ vs $Y_i(0)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	X_n	$\dots$	X_{np}	$Y_n(1)$	$Y_n(0)$	$Y_n(1)$ vs $Y_n(0)$	

1.2 Defining causal estimands

A causal study can focus on causal estimands either for the finite set of n units participating in the study (*finite-population perspective*) or for a large super-population from which the n units are considered as a random sample (*super-population perspective*). From a finite-population perspective, where potential outcomes are viewed as fixed quantities, the *Finite Population Average Treatment Effect* is

$$\tau_{FP}^{ATE} = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)] = \frac{1}{n}\sum_{i=1}^{n}Y_i(1) - \frac{1}{n}\sum_{i=1}^{n}Y_i(0) = \overline{Y}(1) - \overline{Y}(0)$$

From a super-population perspective, where potential outcomes are viewed as random variables because sampling from the super-population induces a distribution of the two potential outcomes for each unit, the *Super-Population Average Treatment Effect* is the expectation of the unit-level causal effect under the distribution induced by sampling from the super-population:

$$\tau_{SP}^{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

A different estimand is the population average treatment effect on the treated, defined by averaging over the subpopulation of treated units

$$\tau_{SP}^{ATT} = \mathbb{E}\left[Y_i(1) - Y_i(0) | Z_i = 1\right].$$

In some observational studies τ_{SP}^{ATT} is a more interesting estimand than τ_{SP}^{ATE}.

Sometimes average treatment effects are defined as averages over subpopulations defined in terms of covariates. *Conditional average treatment effects*, that is, averages of unit-level causal effects over sub-populations described in terms of covariates, are sometimes of interest:

$$\tau_{SP}(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right].$$

1.3 Assignment mechanism

The fundamental problem of causal inference [Rubin, 1978, Holland, 1986b] is that we cannot observe both potential outcomes $Y_i(0)$ and $Y_i(1)$ for any unit. In this sense, the problem of causal inference is, as pointed out in Rubin [1976], Mealli and Rubin [2015] and Mattei et al. [2022], a missing data problem: given any treatment assigned to an individual unit, the potential outcome associated with any alternative treatment is missing. Formally, for each unit i, let Z_i denote the treatment assigned. The observed and missing outcomes are $Y_i^{\text{obs}} = Z_i Y_i(1) - (1 - Z_i)Y_i(0)$ and $Y_i^{\text{mis}} = Z_i Y_i(0) - (1 - Z_i)Y_i(1)$, respectively. Drawing inference on causal effects for any particular unit, which are functions of both the missing and observed potential outcomes, will generally require predicting or imputing the missing potential outcome. Then, to learn about the causal effects of interest it is crucial to posit a treatment assignment mechanism, i.e., the process that determines which units receive which treatments, which potential outcomes are realized, and which are missing.

Definition 1 (Assignment Mechanism)**.** *The assignment mechanism is a row-exchangeable function of all covariates and of all potential outcomes, giving the probability of any vector of treatment assignments given the Science:*

$$\Pr\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right)$$

with $\sum_{\boldsymbol{Z}\in\{0,1\}^n} \Pr\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) = 1$ *for all* $\boldsymbol{X}$*,* $\boldsymbol{Y}(0)$*, and* $\boldsymbol{Y}(1)$.

Definition 2 (Unconfounded assignment mechanism)**.** *An unconfounded assignment mechanism [Rubin, 1990] is free of dependence on either* $\boldsymbol{Y}(0)$ *or* $\boldsymbol{Y}(1)$*:*

$$\Pr\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) = \Pr(\boldsymbol{Z} \mid \boldsymbol{X})$$

Definition 3 (Probabilistic)**.** *If each unit has a positive probability of receiving either treatment, the assignment mechanism is probabilistic:*

$$0 < \Pr\left(Z_i = 1 \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) < 1$$

Definition 4 (Strongly ignorable)**.** *If the assignment mechanism is unconfounded and probabilistic, it is called strongly ignorable.*

Definition 5 (Confounded assignment mechanism)**.** *A confounded assignment mechanism is one that depends on the potential outcomes:*

$$\Pr\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) \neq \Pr(\boldsymbol{Z} \mid \boldsymbol{X})$$

Definition 6 (Ignorable assignment mechanisms)**.** *The ignorable assignment mechanisms [Rubin, 1978] is a special class of possibly confounded assignment mechanisms. Ignorable assignment mechanisms are defined as being free of dependence on any missing potential outcomes*

$$\Pr\left(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) = \Pr(\boldsymbol{Z} \mid \boldsymbol{X}, Y^{\text{obs}})$$

Randomized experiments versus observational studies

After discussing some possible constraints on the assignment mechanism, let us now use them to distinguish between randomized and non-randomized experiments. A `randomized experiment` is a probabilistic assignment mechanism controlled by the researcher and is a known function of its arguments. A specific class of randomized experiments is the class of classical randomized experiments, where the assignment mechanism is also (i) individualistic and (ii) unconfounded. An individualistic assignment mechanism posits constraints on the dependence of the treatment assignment of a unit on the outcomes and assignments for other units. If the treatment assignment mechanism is individualistic, the probability that a unit is assigned to the active treatment does not depend on the covariates or potential outcomes of the other units.

In contrast, an `observational study` (known as non-randomized experiments) is an assignment mechanism if the functional form of the assignment mechanism is unknown. Cochran [1965] defined an observational study to be an empirical investigation in which the "objective is to elucidate cause-and-effect relationships [in settings in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures" [Cochran, 1965]. An observational study, by this definition, has the same goal as a randomized experiment: to estimate causal effects. However, a key difference between them is in one design issue: the use of randomization to assign units to treatment and control groups.

The objective of this thesis is observational research, and we must make assumptions to draw inference about causal effects.

2 Causal Inference in observational Studies

2.1 Designing observational studies: the role of the propensity score

Although the validity of causal results from randomized experiments is generally understood, inference from observational studies frequently encounters issues that bring the validity of the derived causal conclusions into question. To deal with such issues, one should try to create observational studies that as closely as possible resemble a randomized experiment. According to Rubin [2008], *design* means "all contemplation, collection, organization, and analysis of data that occurs before viewing any outcome data." It is important to note that the design phase does not include outcome data. This is a benefit of the approach. The design phase allows us to avoid any conflict of interest, which may be a relevant problem with traditional approaches to causal inference involving fitting regression models where the estimated causal effect is given by the estimated coefficient of an indicator variable for exposure to an intervention and the estimated answers are constantly being seen and modified as models are fitted and refitted. Depending on what the company expects to see, this process may result in a variety of responses from which the analyst can pick [Rubin and Waterman, 2006].

In practice, under unconfoundedness, it is suggested to determine the degree of balance in covariate distributions by comparing covariate distributions in the treated and control subsamples. Scaled differences in average covariate values by treatment status can be employed; alternatively, the propensity score distribution can be examined.

The key part of analyzing observational studies under unconfoundedness is using the *propensity score* (noted as $e(X_i)$).

Definition 7 (Propensity Scores)**.** *Suppose that the assignment mechanism is unconfounded. The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables*

$$e(X_i) = \Pr(Z_i = 1 \mid X_i).$$

Assumption 2 (Overlap (also known as probabilistic or positivity))**.** *Under unconfounded assumption, we consider Overlap assumption as*

$$0 < \Pr\left(Z_i = 1 \mid \boldsymbol{X}\right) < 1$$

Rubin argues that the use of propensity score techniques has a benefit in that it allows observational research to be constructed similarly to randomized experiment [Rubin, 2001]. This is because the propensity score is a *balancing score*[Rosembaum and Rubin, 1983].

Definition 8 (Balancing Scores)**.** *A balancing score $b(\cdot)$ is a function of the covariates such that*

$$Z_i \perp X_i \mid b(X_i).$$

Balancing scores are not unique. By definition, the vector of covariates X_i itself is a balancing score, and any one-to-one function of a balancing score is also a balancing score.

Lemma 1 (Balancing Property of the Propensity Score)**.** *The propensity score is a balancing score which means*

$$Z_i \perp X_i \mid e(X_i)$$

so that the covariate distribution is the same in treatment and control units with the same propensity score.

Proof. We show that $Z_i \perp X_i \mid e(X_i)$ or equivalently,

$$\Pr(Z_i = 1 \mid X_i, e(X_i)) = \Pr(Z_i = 1 \mid e(X_i))$$

implying that Z_i is independent of X_i given the propensity score. First, consider the left hand side:

$$\Pr(Z_i = 1 \mid X_i, e(X_i)) = \Pr(Z_i = 1 \mid X_i) = e(X_i)$$

where the first equality follows because the propensity score is a function of X_i and the second is by the definition of the propensity score. Second, consider the right hand side. By the definition of probability and iterated expectations,

$$\Pr(Z_i = 1|e(X_i)) = \mathbb{E}[Z_i|e(X_i)] = \mathbb{E}[\mathbb{E}[Z_i|X_i, e(X_i)]|e(X_i)] = \mathbb{E}[e(X_i)|e(X_i)] = e(X_i)$$

■

Balancing scores have an important property: if assignment to treatment is unconfounded given the full set of covariates, then assignment is also unconfounded conditioning only on a balancing score:

Lemma 2 (Unconfoundedness given a Balancing Score)**.** *Suppose assignment to treatment is unconfounded. Then the assignment is unconfounded given any balancing score:*

$$Z_i \perp Y_i(0), Y_i(1) \mid b(X_i).$$

Proof. We show that

$$\Pr(Z_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = \Pr(Z_i = 1 \mid b(X_i))$$

which is equivalent to the statement in the lemma. By iterated expectations we can write

$$\begin{aligned}\Pr(Z_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) &= \mathbb{E}[Z_i \mid Y_i(0), Y_i(1), b(X_i)] \\ &= \mathbb{E}[\mathbb{E}[Z_i \mid Y_i(0), Y_i(1), X_i, b(X_i)] \mid Y_i(0), Y_i(1), b(X_i)].\end{aligned}$$

By unconfoundedness, the inner expectation is equal to $\mathbb{E}[Z_i \mid X_i, b(X_i)]$ and by the definition of balancing scores, this is equal to $\mathbb{E}[Z_i \mid b(X_i)]$. Hence the last expression is equal to

$$\mathbb{E}[\mathbb{E}[Z_i \mid b(X_i)] \mid Y_i(0), Y_i(1), b(X_i)] = \mathbb{E}[Z_i \mid b(X_i)] = \Pr(Z_i = 1 \mid b(X_i))$$

which is equal to the right hand side. ■

The first implication of Lemma 2 is that given a vector of covariates that ensure unconfoundedness, adjustment for treatment-control differences in balancing scores suffices for removing all biases associated with differences in the covariates [Imbens and Rubin, 2015]. As a result, even if a covariate is associated with the potential outcomes, differences in covariates between treated and control units do not lead to bias because they cancel out by averaging over all units with the same value for the balancing score. The situation is similar to a completely randomized experiment in which both treatment arms have the same covariates distribution.

Because the propensity score is a balancing score, Lemma 2 implies that conditional on the propensity score, treatment assignment is unconfounded. But within the class of balancing scores, the propensity score has a special place, formally described in the following Lemma:

Lemma 3 (Coarseness of Balancing Scores)**.** *The propensity score is the coarsest balancing score. That is, the propensity score is a function of every balancing score.*

Proof. Let $b(x)$ be a balancing score. Suppose that we can not write the propensity score as a function of the balancing score. Then it must be the case that for two values x and x' we have $b(x) = b(x')$, and at the same time $e(x) = e(x')$. Then, $\Pr(Z_i = 1|X_i = x) = e(x) = e(x') = \Pr(Z_i = 1|X_i = x')$, and so Z_i and X_i are not independent given $b(X_i) = b(x)$, which violates the definition of a balancing score. ■

Since the propensity score is the coarsest balancing score, it presents an enormous decrease in the number of variables we need to adjust for. However, there is a challenge in that in observational studies; we do not generally know the true value of the propensity score for all units; therefore, we cannot directly exploit this result. However, it can be estimated using the study data.

The propensity score is most often estimated using a logistic regression model $e(\boldsymbol{X}, \boldsymbol{\beta}) = \{1 + e(-\boldsymbol{X}^\top \boldsymbol{\beta})\}^{-1}$. Interaction and higher-order terms may also be included. Here, $\boldsymbol{\beta}$ may be estimated by the maximum likelihood (ML) estimator $\hat{\boldsymbol{\beta}}$ solving

$$\sum_{i=1}^{n} \frac{Z_i - e(X_i, \boldsymbol{\beta})}{e(X_i, \boldsymbol{\beta})\{(1 - e(X_i, \boldsymbol{\beta}))\}} \partial/\partial\boldsymbol{\beta}\{e(X_i, \boldsymbol{\beta})\} = \mathbf{0}. \tag{1}$$

We assume that the analyst is proficient at modelling $e(\boldsymbol{X}, \boldsymbol{\beta})$ so that it is correctly specified. The estimated propensity score is the predicted probability of treatment derived from the fitted regression model. Although logistic regression appears to be the most commonly used method for estimating the propensity score, the use of bagging or boosting [Lee et al., 2011, McCaffrey et al., 2004], recursive partitioning, or tree-based methods [Lee et al., 2011, Setoguchi et al., 2008], random forests [Lee et al., 2011], and neural networks [Setoguchi et al., 2008] for estimating the propensity score have been examined.

The goal is to obtain an estimated propensity score that balances the covariates between treated and control subpopulations rather than one that estimates the true propensity score as accurately as possible. Subjects with the same propensity score have the same distribution of the observed potential confounders, whether they are treated or not. In the design phase, the estimated propensity score is an effective tool for obtaining overlap and constructing a comparison group through matching, stratifying, or weighting observations [Rubin, 2008].

we follow Rubin and colleagues' recommendation of explicitly distinguishing "design" and "analysis" phases Rubin [2001, 2008], Stuart and Rubin [2008]. In particular, what follows is a review of all PS techniques that achieve balance on measured covariates in the design phase:

2.1.1 Stratification on the Propensity Score

In *stratification*, based on the value of the propensity score, the sample is split into subclasses, and the data within the subclasses are evaluated as if they came from a completely randomized experiment [Rosembaum and Rubin, 1984, Lunceford and Davidian, 2004]. A common approach is to divide subjects into five equal-size groups using the quintiles of the estimated propensity score. Cochran [1968] demonstrated that stratifying on the quintiles of a continuous confounding variable eliminated approximately 90% of the bias due to that variable. Rosembaum and Rubin [1984] extended this result to stratification (or subclassification) on the propensity score, stating that stratifying on the quintiles of the propensity score eliminates approximately 90% of the bias due to measured confounders when estimating a linear treatment effect.

2.1.2 Propensity Score Matching

Forming matched sets of treated and untreated individuals who share a similar value of the propensity score is known as propensity score *matching* [Rosembaum and Rubin, 1983]. The most common implementation of propensity score matching is one-to-one or pair matching, in which pairs of treated and untreated subjects are formed such that matched subjects have similar values of the propensity score. There are *three* steps to follow in “design” phase as

(i) Defining “closeness”: the distance measure used to determine whether an individual is a good match for another

(ii) Implementing a matching method, given that measure of closeness,

(iii) Assessing the quality of the resulting matched samples and perhaps iterating with Steps (i) and (ii) until well-matched samples result.

Abadie and Imbens [2002] used the diagonal matrix obtained using the diagonal elements of the inverse of the covariance-variance matrix of the covariates:

$$d_{AI}(x, z) = (x - z)^\top \operatorname{diag}(\Sigma_X^{-1})(x - z)$$

where Σ_X^{-1} is the covariance matrix of the covariates. The most common choice is the Mahalanobis metric (e.g., Rosenbaum and Rubin [1985]) which uses the inverse of the covariance matrix of the pre-treatment variables:

$$d_M(x, z) = (x - z)^\top \Sigma_X^{-1}(x - z)$$

This metric has the attractive property that it reduces differences in covariates within matched pairs in all directions. See for more formal discussions Rubin and Thomas [1992a].

To form matched pairs of treated and untreated subjects when matching on the propensity score, there are several methods. Focus on the case where only treated units are matched, then we can conduct

1. *Nearest-neighbor matching*: One of the most common, and easiest methods is nearest-neighbor matching[Rubin, 1973a]. Nearest-neighbor matching chooses the untreated subject whose propensity score is closest to that of the treated subject for matching to a specific treated subject. One concern is that, without any restrictions, matching can lead to some poor matches, if for example, there are no control individuals with propensity scores similar to a given treated individual. One strategy to avoid poor matches is to impose a caliper and only select a match if it is within the caliper [Stuart, 2010].

2. *Nearest-neighbor matching within a specified calliper*: Nearest neighbour matching within a specified calliper distance is similar to nearest neighbour matching; however, the absolute difference in the propensity scores of matched patients must be less than a predetermined threshold (the calliper distance) [Austin, 2011].

3. *Optimal matching*: An alternative to simple nearest neighbour matching is optimal matching, in which matches are formed to minimize the total within-pair difference of the propensity score [Gu and Rosenbaum, 1993].

4. *Matching without replacement*: Once an untreated subject has been chosen to be matched to a specific treated subject using matching without replacement, that untreated subject is no longer available for consideration as a potential match for subsequently treated subjects. As a result, each untreated subject is included in at most one matched set Rosenbaum [2002].

5. *Matching with replacement*: The same untreated subject can be included in several matched sets using replacement matching.

2.1.3 Inverse of treatment probability weighting

Weighting is a well-known non-parametric balancing technique in which weights are applied to the sample of units in each treatment group to balance the covariate distribution of a target population. In that respect, weighting removes confounding by creating a pseudo-population in which the exposure is independent of the observed confounders [Robins et al., 2000, Hirano et al., 2003, Hernán and Robins, 2006, Morgan, 2014, Li et al., 2018]. Inverse Probability Weight (IPW) originating from survey research is a popular approach used to adjust for confounding due to differences between two groups that arise in observational data [Austin and Stuart, 2015, Jones et al., 2018]. IPW's purpose is to construct a pseudo-population in which there is no association between variables X and treatment Z. Assuming that all important confounders are measured, this approach is appealing for its simplicity.

As long as all advantages of IPW offer better adjustment and separate the design phase from the analysis phase, a major limitation of IPW is that it may be inefficient in the presence of extreme propensity scores. To address these problems, trimming methods [Crump et al., 2009] and Stabilized IPW [Robins et al., 2000] have been proposed.

As mentioned earlier, Inverse probability weighting (IPW) is a popular approach used to adjust for confounding due to differences between comparator groups that arise in observational data. In this setting, to assess balance, the standardized differences allow researchers to quantitatively compare balance in measured baseline covariates between treated and control subjects in the weighted sample Austin and Stuart [2015]. Moreover, the Love plot (a graphical diagnostic) is a summary plot of covariate balance before and after weighting proposed by Dr Thomas E. Love [Ahmed et al., 2007]. In a visually appealing and clear way, balance can be presented to demonstrate to readers that balance has been met within a threshold and that balance has improved after weighting.

2.2 Estimation

The estimated propensity score is used in many existing practices for estimating and assessing causal effects [Stuart, 2010, Harder et al., 2010]. This referred to as *analysis* phase. Methods for applying the PS include matching [Rosenbaum and Rubin, 1985, Dehejia and Wahba, 1999, 2002, Ho et al., 2007, Stuart, 2010, Stuart et al., 2011] , stratification (or subclassification) [Rosembaum and Rubin, 1984, Lunceford and Davidian, 2004], weighting [Robins et al., 2000, Hirano et al., 2003, Morgan, 2014, Li et al., 2018] and use of the PS for covariate adjustment[Austin, 2008b]. In other words, several literatures tried to find a clear guidance to make a sensible choice among these various PS methods for any given database[Austin, 2008b, 2009b,c, 2011, Stuart, 2008, 2010, Harder et al., 2010, Elze et al., 2017, Zhou et al., 2020].

We sketch how different procedures work.

2.2.1 Stratification

We shall focus on the population Average Treatment Effect (ATE):

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

The popular approach using stratification on estimated propensity scores to estimate τ involves the following steps
i. Calculate estimated propensity scores for i (denoted as $\hat{e}_i$)
ii. Form K strata according to the sample quantiles of the $\hat{e}_i$, where the $j-$th sample quantile $\hat{q}_j$, $j = 1, \cdots, K$, is such that the proportion of unit with $\hat{e}_i \leq \hat{q}_j$ is roughly $\dfrac{j}{K}$, $\hat{q}_0 = 0$, and $\hat{q}_K = 1$.
iii. Within each stratum, calculate the difference of sample means of the Y_i for each treatment
iv. Estimate τ by a weighted sum of the differences of sample means across strata, where weighting is the proportion of observations falling in each stratum.

Defining $\hat{Q}_j = (\hat{q}_{j-1}, \hat{q}_j]$ and $n_j = \sum_{i=1}^{n} \mathrm{I}(\hat{e}_i \in \hat{Q}_j)$ as the number of individuals in stratum j; and $n_{1j} = \sum_{i=1}^{n} Z_i \, \mathrm{I}(\hat{e}_i \in \hat{Q}_j)$ is the number of these who are treated, the estimator using a weighted sum is

$$\hat{\tau}_{strat} = \sum_{j=1}^{K}(\frac{n_j}{n})\left\{\frac{\sum_{i=1}^{n} Z_i Y_i \, \mathrm{I}(\hat{e}_i \in \hat{Q}_j)}{n_{1j}} - \frac{\sum_{i=1}^{n}(1 - Z_i) Y_i \, \mathrm{I}(\hat{e}_i \in \hat{Q}_j)}{n_j - n_{1j}}\right\} \tag{2}$$

As the weights $\frac{n_j}{n} \approx \frac{1}{K}$, they may be replaced by $\frac{1}{K}$ to yield an average across strata. Since treatment exposure is essentially random for individuals with the same propensity value, we expect mean comparisons within this group to be unbiased. We expected mean comparisons within this group to be unbiased since treatment exposure is essentially at random for individuals with the same propensity value. In practice, identifying individuals with the same propensity score may be unfeasible; therefore, stratification aims to create groups in which this holds approximately. Consequently, $\hat{\tau}_{strat}$ may be biased as some residual confounding within strata may remain. Rosembaum and Rubin [1983, 1984] advocate using quantiles ($K = 5$) which require that the propensity model be correctly specified. Thus, it is recommended [Rosembaum and Rubin, 1984, Perkins et al., 2000] that the balance within each stratum is examined using standard statistical tests.

D'Agostino [1998] proposed a variation of $\hat{\tau}^{strat}$. Here, steps (*iii*) and (*iv*) are modified as follows:

iii. within each stratum $j = 1, \cdots, K$, fit a regression model of the form $m^{(j)}(\boldsymbol{Z}, \boldsymbol{X}, \alpha^{(j)})$ representing the postulated regression relationship $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X})$ within stratum j and, based on the resulting estimate $\alpha^{(j)}$, estimate treatment effect in stratum j by averaging over X_i in j as

$$\hat{\tau}_{strat}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n} \mathrm{I}(\hat{e}_i \in \hat{Q}_j)\Big\{m^{(j)}(Z_i = 1, X_i, \hat{\alpha}^{(j)}) - m^{(j)}(Z_i = 0, X_i, \hat{\alpha}^{(j)})\Big\} \tag{3}$$

iv. estimate τ by the average or weighted sum of the $\hat{\tau}_{strat}^{(j)}$, e.g. using the average

$$\hat{\tau}_{SR} = \frac{1}{K} \sum_{j=1}^{K} \hat{\tau}_{strat}^{(j)} \tag{4}$$

Within-stratum regression modelling is intended to eliminate any remaining imbalances within strata [Lunceford and Davidian, 2004].

A pooled estimate of the variance of the estimated treatment effect can be obtained by pooling the variances of the stratum-specific treatment effects. For a greater discussion of variance estimation, the reader is referred to Rosembaum and Rubin [1984] and Lunceford and Davidian [2004].

2.2.1.1 Theoretical Properties:

In this section, we summarize the properties of the estimators based on subclassification of the propensity score and highlight the practical insights that can be deduced from them. As it is common to take a constant number of strata (i.e. K) regardless of sample size ($K = 5$ is most common), we consider K to be fixed (and hence independent of n), and we assume propensity score is correctly specified. We may rewrite equation (2) in an asymptotically equivalent form by replacing $\frac{n_j}{n}$ with its limit $\frac{1}{K}$ and writing $\hat{p}_j = \frac{n_{1j}}{n}$ as

$$\hat{\tau}_{strat} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{K}\Big\{\sum_{j=1}^{K}\frac{\mathrm{I}(\hat{e}_i \in \hat{Q}_j)}{\hat{p}_j}\Big\} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{K}\Big\{\sum_{j=1}^{K}\frac{\mathrm{I}(\hat{e}_i \in \hat{Q}_j)}{\frac{1}{K}-\hat{p}_j}\Big\}. \quad (5)$$

Thus, considering the asymptotically equivalent form (5), we may replace $\hat{e}_i$, $\hat{Q}_j$ and $\hat{p}_j$ by their true values and apply the law of large numbers directly to see that $\hat{\tau}_{strat}$ converges in probability to $\tau^*_{strat} = \mu^*_1 - \mu^*_0$, where

$$\hat{\mu}^*_1 = K^{-1}\sum_{j=1}^{K}\frac{\mathbb{E}[Y(1)e\,\mathrm{I}(e \in Q_j)]}{\mathbb{E}[e\,\mathrm{I}(e \in Q_j)]}$$

and

$$\hat{\mu}^*_0 = K^{-1}\sum_{j=1}^{K}\frac{\mathbb{E}[Y(0)(1-e)\,\mathrm{I}(e \in Q_j)]}{(K^{-1} - \mathbb{E}[e\,\mathrm{I}(e \in Q_j)])}.$$

However, in general, $\tau^*_{strat} \neq \tau$, so that $\hat{\tau}_{strat}$ is not consistent.

Lunceford and Davidian [2004] proved a similar argument for $\hat{\tau}_{SR}$. Now, substituting $\frac{n_j}{n} \approx K^{-1}$ in equation (3), we may rewrite

$$\hat{\tau}_{SR} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\mathrm{I}(\hat{e}_i \in \hat{Q}_j)\Big\{m^{(j)}(Z_i = 1, X_i, \hat{\alpha}^{(j)}) - m^{(j)}(Z_i = 0, X_i, \hat{\alpha}^{(j)})\Big\}$$

Then, applying the law of large numbers, $\hat{\tau}_{SR}$ converges in probability to

$$\tau^{**}_{strat} = \sum_{j=1}^{K}\mathbb{E}[\mathrm{I}(e \in Q_j)\{m^{(j)}(\boldsymbol{Z} = 1, \boldsymbol{X}, \alpha^{(j)}_*) - m^{(j)}(\boldsymbol{Z} = 0, \boldsymbol{X}, \alpha^{(j)}_*)\}]$$

where $\alpha^{(j)}_*$ depend on the functions $m^{(j)}$ used. Thus, $\tau^{**}_{strat} = \tau$. This demonstrates that $\hat{\tau}_{SR}$ is a consistent estimator for τ as long as the $m^{(j)}$ has the same form as the true regression relationship.

Theoretical results demonstrate that the frequent version of stratification using estimated propensity scores based on within-stratum sample mean differences and a fixed number of strata can lead to biased inference due to residual confounding, with the effect of this bias becoming more serious as sample size increases. Although using more strata can increase the sample size at which the trade-off of bias and variability involved in efficiency takes place, stratifying on quintiles seems to be the most popular approach in practice, even for large sample sizes. As a result, because the "trade-off" point for each given case will be unknown, this approach should be utilized with caution [Lunceford and Davidian, 2004].

2.2.2 Matching

Matching methods have been in use since the first half of the 20^{th} Century (e.g., Greenwood [1945] and Chapin [1947]), however a theoretical basis for these methods was not developed until the 1970s. This development began with papers by Cochran and Rubin [1973] and Rubin [1973a,b] for situations with one covariate and an implicit focus on estimating the ATT. In a series of papers in the 1990s, Rubin and Thomas [1992a,b, 1996] provided a theoretical basis for multivariate settings with affinely invariant matching methods and ellipsoidally symmetric covariate distributions (such as the normal or t-distribution). In "analysis" phase, one should follow all the following steps as

(i) Defining "closeness": the distance measure used to determine whether an individual is a good match for another

(ii) Implementing a matching method, given that measure of closeness,

(iii) Assessing the quality of the resulting matched samples, and perhaps iterating with Steps (i) and (ii) until well-matched samples result, and

(iv) Analysis of the outcome and estimation of the treatment effect, given the matching done in Step (iii).

It is worth noting that the first three steps are the same as the "design" phase and the *fourth* step represents the "analysis" phase.

The treatment effect can be assessed by directly comparing outcomes between treated and untreated subjects in the matched sample once a matched sample has been created. Matching estimators have been widely studied in practice and theory [Rubin, 1973a, Rosenbaum, 1989, Gu and Rosenbaum, 1993, Heckman et al., 1998, Dehejia and Wahba, 1999, 2002, Abadie and Imbens, 2002, Ho et al., 2007, Stuart, 2010, Stuart et al., 2011].

If the outcome is continuous, the effect of the treatment can be estimated as the difference between the mean outcome for treated subjects and the mean outcome for untreated subjects in the matched sample [Rosembaum and Rubin, 1983]. If the outcome is dichotomous, the effect of the treatment can be estimated as the difference between the proportion of subjects experiencing the event in each of the two groups (treated vs control) in the matched sample. With binary outcomes, the effect of the treatment can also be described using the relative risk [Rosembaum and Rubin, 1983, Austin, 2008a, 2010].

Again, the population average treatment effect is an estimand of interest

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Formally, given a sample, (Y_i, X_i, Z_i) $(i = 1, \cdots, n)$, let $\ell_m(i)$ be the index l that satisfies $Z_l \neq Z_i$ and

$$\sum_{j|Z_l \neq Z_i} 1\left\{||X_j - X_i|| \leq ||X_j - X_i||\right\}$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit i in terms of the distance measure based on the norm $||\cdot||$. In particular, $\ell_1(i)$ is the nearest match for unit i. Let $\mathscr{L}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathscr{L}_M(i) = \{\ell_1(i), \cdots, \ell_M(i)\}$. Define the imputed potential outcomes as:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } Z_i = 0 \\ \dfrac{1}{M}\sum_{j \in \mathscr{L}_M(i)} Y_j & \text{if } Z_i = 1 \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \dfrac{1}{M}\sum_{j \in \mathscr{L}_M(i)} Y_j & \text{if } Z_i = 0 \\ Y_i & \text{if } Z_i = 1 \end{cases}$$

The simple matching estimator discussed in Abadie and Imbens [2002] is then

$$\hat{\tau}_M^{sm} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right)$$

According to Abadie and Imbens [2002], the bias of this estimator is of order $O(n^{\frac{-1}{p}})$, where p is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by $\sqrt{n}$ (as can be justified by the fact that the variance of the estimator is of order $O(\frac{1}{n})$), the bias does not disappear if the dimension of the covariates is equal to two, and will dominate the large sample variance if p is at least three. Moreover, Abadie and Imbens [2002] proved that the simple matching estimator is consistent for the average treatment effect and that, without the bias term, is $\sqrt{n}$-consistent and asymptotically normal. We state the formal results of consistency and asymptotic normality based on Abadie and Imbens [2002] as

Theorem 1 (Consistency of the Simple Matching Estimator)**.** *Suppose the following assumptions hold:*

> ***assumption (1):*** *Let $\boldsymbol{X}$ be a random vector of continuous covariates distributed on $\mathbb{R}^p$ with compact and convex support $\mathbb{X}$, with the density bounded, and bounded away from zero on its support.*
>
> ***assumption (2):*** *For almost every $x \in \mathbb{X}$, the assignment mechanism is unconfounded and probabilistic (i.e. strongly ignorable).*
>
> ***assumption (3):*** *(Y_i, X_i, Z_i), $i = 1, \cdots, n$ are independent draws from the distribution of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$.*

If in addition $\mu_1(x) = \mathbb{E}[\boldsymbol{Y}(1) \mid \boldsymbol{X} = x]$ and $\mu_0(x) = \mathbb{E}[\boldsymbol{Y}(0) \mid \boldsymbol{X} = x]$ are continuous, then

$$\hat{\tau}^{sm} - \tau \xrightarrow{P} 0$$

Theorem 2 (Asymptotic Normality for the Simple Matching Estimator)**.** *Suppose assumptions (1) to (3) as well as*

> ***assumption (4):*** *if (i) $\mu_z(x)$ and $\sigma_z^2(x)$ are continuous in x for all z, and (ii) the fourth moments of the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{Z} = z$ and $\boldsymbol{X} = x$ exist and are uniformly bounded.*

hold and that $\mu_1(x)$ and $\mu_0(x)$ have bounded third derivatives. Then

$$\sqrt{n}(\hat{\tau}^{sm} - \mathfrak{B}^{sm} - \tau) \xrightarrow{d} \mathscr{N}(0, V^E + V^{\tau(x)})$$

where $\mathfrak{B}^{sm}$ is the bias term, V^E is the conditional variance of the simple matching estimator $\hat{\tau}^{sm}$ and $V^{\tau(x)}$ is the variance of conditional average treatment effects as $V^{\tau(x)} = \mathbb{E}[(\tau(x) - \tau)^2]$ where $\tau(x) = \mathbb{E}[\boldsymbol{Y}(1) - \boldsymbol{Y}(0) \mid \boldsymbol{X} = x]$.

Three points about the Abadie-Imbens result [Abadie and Imbens, 2002] should be highlighted. First, in this dimension, p, only continuous variables should be counted. Since matching with discrete variables is exact in large samples, such covariates do not contribute to the bias order. Second, suppose only the treated are matched, and the number of possible controls is substantially larger than the number of treated units. In that case, one can justify ignoring the bias by appealing to an asymptotic sequence where the number of potential controls increases faster than the number of treated units. Third, even though the order of the bias may be high, the actual bias may be minimal if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offset. The leading term in the bias, for example, is dependent on the regression function being nonlinear and the covariate density having a nonzero slope. The resulting bias may be reasonably minimal if one of these two conditions is at least close to being met. Abadie and Imbens [2002] proposed combining the matching procedure with a regression adjustment to eliminate bias. Another point made by Abadie and Imbens [2002] is that matching estimators are generally not efficient. Even when the bias is of low enough order to be dominated by the variance, the estimators are not efficient given a fixed number of matches. To reach efficiency, one would need to increase the number of matches with the sample size [Abadie and Imbens, 2002, Imbens, 2004, Austin, 2011].

2.2.3 Weighting

Another broad class of estimation methods is *weighting.* The weighting approach assigns a sampling weight to each member of the population, and the probability of being selected is proportional to this weight. Weighting procedures are not new and have a long history of being used in survey sampling. The inverse-probability weights (IPW) have dominated the weighting literature, e.g. Robins and Rotnitzky [1995], Hahn [1998], Robins et al. [2000], Hirano and Imbens [2001]; Hirano et al. [2003], Imbens [2004] and Crump et al. [2009]. The Horvitz-Thompson (HT) weight [Horvitz and Thompson, 1952], which is the inverse probability of that unit being allocated to the observed group, is a special case of IPW.

The Horvitz-Thompson estimator Horvitz and Thompson [1952] exploits the following two equalities

$$\mathbb{E}\Big[\frac{Z_i Y_i^{obs}}{e(X_i)}\Big] = \mathbb{E}[Y_i(1)], \qquad \mathbb{E}\Big[\frac{(1-Z_i)Y_i^{obs}}{1-e(X_i)}\Big] = \mathbb{E}[Y_i(0)] \tag{6}$$

These inequalities can be derived as follows. Because Y_i^{obs} is $Y_i(1)$ when $Z_i = 1$, it follows that

$$\mathbb{E}\Big[\frac{Z_i Y_i^{obs}}{e(X_i)}\Big] = \mathbb{E}\Big[\frac{Z_i Y_i(1)}{e(X_i)}\Big].$$

By iterated expectations, we can write this as

$$\mathbb{E}\Big[\frac{Z_i Y_i(1)}{e(X_i)}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\frac{Z_i Y_i(1)}{e(X_i)} \mid X_i\Big]\Big].$$

Under unconfoundedness assumption, Z_i is independent of $Y_i(1)$ conditional on X_i, so that the expectation of the product $Z_i Y_i(1)$ given X_i is the product of the conditional expectations:

$$\mathbb{E}\Big[\frac{Z_i Y_i(1)}{e(X_i)} \mid X_i\Big] = \frac{\mathbb{E}[Z_i \mid X_i]\ \mathbb{E}[Y_i(1) \mid X_i]}{e(X_i)} = \frac{e(X_i)\ \mathbb{E}[Y_i(1) \mid X_i]}{e(X_i)} = \mathbb{E}[Y_i(1) \mid X_i]$$

and thus

$$\mathbb{E}\Big[\frac{Z_i Y_i(1)}{e(X_i)}\Big] = \mathbb{E}\Big[\mathbb{E}[Y_i(1) \mid X_i]\Big].$$

The same argument leads to the second equality in (6) for the average control potential outcome. The two equalities in (6) suggest estimating $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$ as

$$\widehat{\mathbb{E}[Y_i(1)]} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i^{obs}}{e(X_i)} \qquad \text{and} \qquad \widehat{\mathbb{E}[Y_i(0)]} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i^{obs}}{1-e(X_i)}$$

and thus estimating the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ as a Horvitz-Thompson estimator

$$\hat{\tau}_{ipw1} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i^{obs}}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i^{obs}}{1-e(X_i)}$$

In observational studies, we don't know the true propensity score, so we use the estimated propensity score $\hat{e}(X_i)$ as

$$\hat{\tau}_{ipw1} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i^{obs}}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i^{obs}}{1-\hat{e}(X_i)}. \tag{7}$$

Since $\mathbb{E}[\frac{Z}{e(\boldsymbol{X})}] = \mathbb{E}[\frac{\mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}]}{e(\boldsymbol{X})}] = 1$ and $\mathbb{E}[\frac{(1-\boldsymbol{Z})}{1-e(\boldsymbol{X})}] = 1$, normalizing the weights to add up to one improve the mean-squared-error properties of the estimator. Thus,

$$\hat{\tau}_{ipw2} = \Big(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}(X_i)}\Big)^{-1}\Big(\sum_{i=1}^{n} \frac{Z_i Y_i^{obs}}{\hat{e}(X_i)}\Big) - \Big(\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}(X_i)}\Big)^{-1}\Big(\sum_{i=1}^{n} \frac{(1-Z_i)Y_i^{obs}}{1-\hat{e}(X_i)}\Big). \tag{8}$$

As $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$ involve weighting the observations in each group by the inverse of the probability of being in that group, "IPW" denotes "inverse probability weighting" and $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$ are popular approaches based on such weighting. They are; however, special cases of a larger class of estimators that may be deduced by viewing the situation as a "missing data" problem, as discussed in a landmark study by Robins et al. [1994].

Robins et al. [1994] Proposed the estimator within the class having the smallest (large-sample) variance, the (locally) semiparametric efficient estimator

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{Z_i Y_i^{obs} - (Z_i - \hat{e}(X_i))m_1(X_i, \hat{\alpha}_1)}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i^{obs} + (Z_i - \hat{e}(X_i))m_0(X_i, \hat{\alpha}_0)}{1-\hat{e}(X_i)}\Big). \tag{9}$$

Here $m_z(\boldsymbol{X}, \alpha_z) = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{Z} = z, \boldsymbol{X}]$ is the regression of the response on $\boldsymbol{X}$ in group z, $z = \{0, 1\}$, depending on parameters α_z, and $\hat{\alpha}_z$ is an estimator for α_z based on the data from subjects with $Z = z$. Each term in $\hat{\tau}_{DR}$ has the form of those in $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$ but 'augmented' [Robins, 2000b] by an expression involving the regression. This augmentation makes $\hat{\tau}_{DR}$ to be the efficient estimator in the class, and in large samples, it has a smaller variance than $\hat{\tau}_{ipw1}$ or $\hat{\tau}_{ipw2}$ [Lunceford and Davidian, 2004]. Moreover, [Scharfstein et al., 1999, Section 3.2.3] note that $\hat{\tau}_{DR}$ has a so-called "double-robustness" property that the estimator remains consistent if either the propensity score model or the regression models are correctly specified, not necessarily both. Double-robust estimators are desirable because they give analysts two chances to "get it right" and guard against model misspecification. Moreover, $\hat{\tau}_{DR}$ reaches the semiparametric efficiency bound of τ if both models are correctly specified [Hahn, 1998, Chernozhukov et al., 2018].

2.2.3.1 Theoretical Properties:

Here, we summarize the properties of the weighting estimators and highlight the practical insights that can be deduced from these. The large-sample properties for weighted estimators follow the general framework of Robins et al. [1994] and may also be obtained directly from the standard theory of M-estimation [Stefanski and Boos, 2002].

Properties of $\hat{\tau}_{ipw1}$, $\hat{\tau}_{ipw2}$ and $\hat{\tau}_{DR}$ when propensity score is correctly specified may be deduced by viewing them as solutions to a set of estimating equations. To obtain $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$, one need to solve equation (1) and then follow (7) and (8). Similarly, $\hat{\tau}_{DR}$ in equation (9) depends on α_0 and α_1, which are then estimated by solving equations. To do so, applying the theory of M-estimation [Stefanski and Boos, 2002] is essential. In the presence of the true values of parameter $\boldsymbol{\beta}$, $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$ are consistent for τ_0, the true value of τ (This may be seen equivalently by substituting the true values of $\boldsymbol{\beta}$ in (7) and (8) and applying the law of large numbers directly) [Lunceford and Davidian, 2004]. A similar argument shows that $\hat{\tau}_{DR}$ converges in probability to τ_0, even if the models $m_z(\cdot)$ are not correctly specified. The theory presented by Stefanski and Boos [2002], then implies that each estimator is such that $n^{\frac{1}{2}}(\hat{\tau} - \tau_0)$ converges in distribution to Normal.

First consider the (unlikely) case where $\boldsymbol{\beta}$ is known. One of the interesting property of estimating $\boldsymbol{\beta}$ is that even if its true value is known, $\hat{\boldsymbol{\beta}}$ leads to smaller (large-sample) variance for these estimators than using the true value. Thus, under these conditions, Lunceford and Davidian [2004] in an empirical study have shown that the large-sample variances of $\hat{\tau}_{ipw1}$ and $\hat{\tau}_{ipw2}$ (noted as Σ_{ipw1} and Σ_{ipw2}, respectively) are in general $\Sigma_{ipw1} \geq \Sigma_{ipw2}$. For $\hat{\tau}_{DR}$, the theory of Robins et al. [1994] guarantees that $\Sigma_{DR} \leq \Sigma_{ipw1}, \Sigma_{ipw2}$. In practice, the $\Sigma_{ipw1}, \Sigma_{ipw2}$ and Σ_{DR} may be estimated from the observed data, yielding approximate sampling variances for $\hat{\tau}_{ipw1}$, $\hat{\tau}_{ipw2}$ and $\hat{\tau}_{DR}$. Alternatively, variance estimates may be obtained via the empirical sandwich method [Stefanski and Boos, 2002].

In the presence of extreme values of estimated propensity scores, IPW has some drawbacks that include large weights for some patients and, therefore, bias in the estimated treatment effect [Stuart, 2010, Hirano and Imbens, 2001]. It has been argued that such a method may perform poorly even when the propensity score model appears to be correctly specified [Kang and Schafer, 2007]. To address these problems, trimming methods [Crump et al., 2009] and Stabilized IPW [Robins et al., 2000] have been proposed.

2.2.4 Covariate Adjustment Using the Propensity Score

Under covariate adjustment using the propensity score, the outcome is regressed on an indicator variable denoting treatment status and the estimated propensity score [Rosembaum and Rubin, 1983, Vansteelandt and Daniel, 2014]. It was proposed by Rosembaum and Rubin [1983] in their original article on the propensity score. The type of regression model to use would be determined by the outcome. For continuous outcomes, a linear model would be chosen; for dichotomous outcomes, a logistic regression model may be selected. The effect of treatment is determined using the estimated regression coefficient from the fitted regression model. For a linear model, the treatment effect is an adjusted difference in means, whereas for a logistic model it is an adjusted odds ratio. Even though it performs well in some cases, it may produce a result very similar (not even superior) to traditional covariate adjustment [Elze et al., 2017].

Formally, let's define

$$\nu_z(e) = \mathbb{E}[Y(z) \mid e(\boldsymbol{X}) = e].$$

By unconfoundedness this is equal to $\mathbb{E}[Y \mid Z = z, e(\boldsymbol{X}) = e]$. Given an estimator $\hat{\nu_z}(e)$, one can estimate the average treatment effect as

$$\hat{\tau}_{regprop} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\nu_1}(e(X_i)) - \hat{\nu_0}(e(X_i))\right].$$

Heckman et al. [1998] consider a local linear version of this for estimating the average treatment effect for the treated. Hahn [1998] considers a series version and shows that it is not as efficient as the regression estimator based on adjustment for all covariates.

2.2.5 Comparison of the Different Propensity Score Methods

For each of the *four* propensity score methods, adequate diagnostics exist to determine if the propensity score model has been appropriately specified. Once the specification of the propensity score model is satisfied, one might directly estimate the effect of treatment on outcomes in the matched, stratified, or weighted sample using propensity score matching, stratification, and weighting. Another distinction between the four propensity score methods is that weighting and propensity score covariate adjustment may be more sensitive to whether the propensity score model has been correctly specified Rubin [2004]. Some studies have indicated that propensity score matching eliminates a greater proportion of the systematic differences in baseline covariates between treated and control groups than does stratification on the propensity score or covariate adjustment using the propensity score [Austin and Mamdani, 2006, Austin et al., 2007, Austin, 2009a].

Across all common PS applications: matching, stratification, weighting, and use of PS as a covariate, we focus on *weighting* in this thesis. This is because it is easy to implement to deal with unknown assignment mechanisms even in a more complex situation like time-vary confounders. It uses the propensity score directly in estimating the effect of treatment. Furthermore, there is no need to know the number of the subclasses *apriori* or use ad hoc methods or exclude some individuals due to unmatched which leads to information excluded from the analysis.

2.3 Balancing Weights

The objective of comparative effectiveness studies is to evaluate the causal effect of a treatment or intervention that is unconfounded by differences in the characteristics of those assigned to the treatment and control conditions under current practice. Originating in the context of survey sampling and observational studies Horvitz and Thompson [1952], Lunceford and Davidian [2004], IPW assigns weights to the sample of units in each treatment group to match the covariate distribution of a target population, and the comparison is made between the weighted outcomes. In this section, we point out that IPW is a special case of the general class of propensity score weights, called the `balancing weights` [Li et al., 2018], many members of which could be used for covariate adjustment in observational studies.

Assume the marginal density of the covariates $\boldsymbol{X}$, $f(x)$, exists, for a base measure μ (a product of counting measure with respect to categorical variables and Lebesgue measure for continuous variables). We then consider the target population density by $f(x)h(x)$, where $h(\cdot)$ is pre-specified function of x, and average treatment effect (ATE) conditional on x is

$$\tau(x) = \mathbb{E}[\boldsymbol{Y}(1) - \boldsymbol{Y}(0) \mid \boldsymbol{X} = x].$$

A general class of estimands by the expectation of the conditional ATE $\tau(x)$

over the target population is defined as

$$\tau_h \equiv \frac{\int \tau(dx) f(x) h(x) \mu(dx)}{\int f(x) h(x) \mu(dx)}$$

where τ_h is the weighted average treatment effect (WATE) [Hirano et al., 2003]. Let $f_z(x) = \Pr(\boldsymbol{X} = x \mid \boldsymbol{Z} = z)$ be the density of X in the Z = z group, then

$$f_1(x) \propto f(x)e(x), \quad \text{and} \quad f_0(x) \propto f(x)(1 - e(x))$$

For a given $h(x)$, to estimate τ_h, we can weight $f_z(x)$ to the target population using the following weights (proportional up to a normalizing constant):

$$\begin{cases} w_1(x) \approx \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)} & \text{for } Z = 1 \\ w_0(x) \approx \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)} & \text{for } Z = 0 \end{cases} \tag{10}$$

The `balancing weights` is the name of the class of weights defined in (10) because they balance the weighted distributions of the covariates between comparison groups:

$$f_1(x) w_1(x) = f_0(x) w_0(x) = f(x) h(x).$$

According to different function forms of h, all weights that balance the covariate distributions between groups can be specified within this class. The selection of h specifies the target population, estimands, and weights. In general, Table 2 shows an overview of all balancing weights. When $h(x) = 1$, the corresponding target population $f(x)$ is the combined (treated and control) population, the weights (w_1, w_0) are the IPW weights and the estimand is the ATE for the combined population. When $h(x) = e(x)$, the target population is the treated subpopulation, and the estimand is the average treatment effect for the treated (ATT), $\tau^{ATT} = \mathbb{E}[Y(1) - Y(0) \mid Z = 1]$. When $h(x) = 1 - e(x)$, the target population is the control subpopulation, and the estimand is the average treatment effect for the control (ATC), $\tau^{ATC} = \mathbb{E}[Y(1) - Y(0) \mid Z = 0]$. In the presence of extreme value of propensity scores, Crump et al. [2009] recommended use of $1(\alpha < e(x) < 1 - \alpha)$ with a pre-specified $\alpha \in (0, \frac{1}{2})$ that defines a subpopulation with sufficient overlap of covariates between two groups. Li and Greene [2013] defined a weighting analogue to pair matching; a similar notion was discussed earlier in Dehejia and Wahba [1999] as $h(x) = \min(e(x), 1 - e(x))$.

2.3.1 Large-sample Properties of Nonparametric Estimators

To establish properties of the sample estimator of WATE, consider

$$\hat{\tau}_h = \frac{\sum_i w_1(x_i) Z_i Y_i}{\sum_i w_1(x_i) Z_i} - \frac{\sum_i w_0(x_i)(1 - Z_i) Y_i}{\sum_i w_1(x_i)(1 - Z_i)} \tag{11}$$

Table 2: An overview of all Balancing Weights

target population	estimand	h(x)	Weight $w_1(x)$	Weight $w_0(x)$
`combined`	ATE	1	$\frac{1}{e(x)}$	$\frac{1}{(1-e(x))}$
`treated`	ATT	$e(x)$	1	$\frac{e(x)}{(1-e(x))}$
`control`	ATC	$1-e(x)$	$\frac{1-e(x)}{e(x)}$	1
`overlap`	ATO	$e(x)(1-e(x))$	$1-e(x)$	$e(x)$
`truncated combined`		$1(\alpha < e(x) < 1-\alpha)$	$\frac{1(\alpha<e(x)<1-\alpha)}{e(x)}$	$\frac{1(\alpha<e(x)<1-\alpha)}{1-e(x)}$
`matching`		$\min(e(x), 1-e(x))$	$\frac{\min(e(x),1-e(x))}{e(x)}$	$\frac{\min(e(x),1-e(x))}{1-e(x)}$

where the sum is over a sample drawn from density $f(x)$. One of the main property of τ_h is presented in the following Theorem.

Theorem 3. *$\hat{\tau}_h$ is a consistent estimator of τ_h.*

Proof. See Appendix SM.1 for the details. ■

The next result concerns the component of variation due to residual (model) variation in $\hat{\tau}_h$ conditional on the sampled covariate design points $\boldsymbol{X}$, the first term of the decomposition $\mathbb{V}[\hat{\tau}_h] = \mathbb{E}\big[\mathbb{V}[\hat{\tau}_h \mid \boldsymbol{X}]\big] + \mathbb{V}\big[\mathbb{E}[\hat{\tau}_h \mid \boldsymbol{X}]\big]$ showing that it can be characterized after making only limited assumptions about residual variances, as in the Corollary below.

Theorem 4. *As $n \to \infty$, the expectation (over possible samples of covariate values) of the conditional variance of the estimator $\hat{\tau}_h$ given the sample $\boldsymbol{X} = (x_1, \cdots, x_n)$ converges:*

$$n \cdot \mathbb{E}[\mathbb{V}[\hat{\tau}_h \mid \boldsymbol{X}]] \to \int f(x)h(x)^2 \Big[\frac{v_1(x)}{e(x)} + \frac{v_0(x)}{1-e(x)}\Big] \frac{\mu(dx)}{C_h^2}$$

where $v_z(x) = \mathbb{V}[Y(z) \mid \boldsymbol{X}]$ and $C_h = \int h(x)f(x)d\mu(x)$ is a normalizing constant.

Consequently, if the residual variance is assumed to be homoscedastic across both groups, $v_1(x) = v_0(x) = v$, then the asymptotic variance of $\hat{\tau}_h$ simplifies to

$$n \cdot \mathbb{E}[\mathbb{V}[\hat{\tau}_h \mid \boldsymbol{X}]] \to \frac{v}{C_h^2} \int \frac{f(x)h(x)^2 \mu(dx)}{e(x)(1-e(x))}$$

Corollary 1. *The function $h(x) \propto e(x)(1-e(x))$ gives the smallest asymptotic variance for the weighted estimator $\hat{\tau}_h$ among all h's under homoscedasticity, and as $\mu \to \infty$*

$$n \cdot \min_h \big\{\mathbb{E}[\mathbb{V}[\hat{\tau}_h \mid \boldsymbol{X}]]\big\} \to \frac{v}{C_h^2} \int f(x)e(x)(1-e(x))\mu(dx).$$

In practice, the true propensity score is unknown and is replaced by the estimated propensity score. It is shown in Rosenbaum [1987] and Hirano et al. [2003], a consistent estimate of the propensity score leads to a more efficient estimation than the true propensity score.

2.3.2 Overlap Weighting

Li et al. [2018] proposed the Overlap Weights (OW), which weight each unit according to its probability of being assigned to the opposing group. In the Overlap Weights, let $h(x) = e(x)(1 - e(x))$, implying balancing weights

$$\begin{cases} w_1(x) \approx 1 - e(x) & \text{for } Z = 1 \\ w_0(x) \approx e(x) & \text{for } Z = 0 \end{cases} \tag{12}$$

Following Corollary 1, the corresponding nonparametric estimator $\hat{\tau}_h$ has the minimum asymptotic variance among all balancing weights. The target population of OW emphasizes subjects with the most overlap in their observed characteristic, and its corresponding estimand is the average treatment effect in the overlap population [Li et al., 2018]. The OW has properties that are likely to be beneficial in the presence of extreme values. By definition, the overlap weights are bounded between 0 and 1 and thus automatically overcome the large uncertainty issue caused by extreme propensity scores when using IPW. Overlap weights estimated from a logistic model also have a useful small-sample property: they yield *exact balance* between groups in the means of each covariate included in the model [Li et al., 2018]. The following Theorem shows this attractive small-sample property.

Theorem 5. *When the propensity scores are estimated by maximum likelihood under a logistic regression model,* $\operatorname{logit} e(X_i) = \beta_0 + X_i\boldsymbol{\beta}^\top$*, the overlap weights lead to an exact balance in the means of any included covariate between treatment and control groups. That is*

$$\frac{\sum_i x_{ik} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i x_{ik} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}, \qquad \textit{for} \quad k = 1, \ldots, p$$

where p is the dimension of the covariates, $\hat{e}_i = \left\{1 + \exp\left[-(\hat{\beta}_0 + X_i\hat{\boldsymbol{\beta}}^\top)\right]\right\}^{-1}$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is the MLE for the regression coefficients.

Proof. See Appendix SM.1 for the details. ∎

While the main effects model guarantees exact equality between groups for the mean of each included covariate, it is advisable to improve balance by including interactions and higher-order terms.

In this thesis, IPW and Overlap weighting are utilized to address observed confounders. Particularly, Overlap weights emphasize the target population with the greatest overlap in observed characteristics between treatments by continuously deweighting the units in the tails of the distribution of propensity scores. However, the target population of IPW is the entire study cohort. IPW creates a weighted pseudo-population in which both treatment groups resemble the total sample combined across treatment groups. Assuming that all important confounders are measured, IPW is appealing for its simplicity. Although the challenge of extreme propensities has been identified as a downside of IPW, OW bounded and smoothly reduced the influence of patients at the tails of the PS distribution without making any exclusions. Consequently, despite good study design and thoughtful inclusion/exclusion criteria, OW removes the arbitrary decisions involved in trimming and improves the characteristics of bias and precision [Li et al., 2018].

Chapter 4

Survival Analysis

In this chapter, some details about survival data are addressed. More specifically, in Section 1, some basic concepts are presented briefly. Furthermore, some conventional statistical methods for survival analysis focusing on the Cox model are introduced in Section 2. Finally, we discuss the estimation methods of the Cox models and their diagnostic.

1 Basic concepts in survival analysis

In general terms, survival analysis deals with statistical procedures for which the **outcome** variable of interest is **time-to-event**. This time variable gives the elapsed time between the starting point (e.g. beginning of the relevant observation due to diagnosis, treatment start, etc.) until the occurrence of the event of interest. This *time* could be measured in years, months, weeks, or days. The event (also known as a failure if negative) could be death, relapse from remission, recovery, or any designated experience of interest that may happen to an individual. Sometimes there is an event (typically, death) that prevents the event of interest (e.g., stroke) from happening to define a competing event: an individual who dies from other causes (say, cancer) cannot ever develop stroke. An alternative to handling competing events is to create a composite event that includes both the competing event and the event of interest (e.g., death and stroke) and conduct a survival analysis for the composite event. This chapter illustrates how survival analysis is used to estimate the time to a certain event of interest, and the basic concepts are presented.

A time origin denoted as t_0 is defined to measure the time to an event. This is when one starts counting the time to the event of interest. This could be a birthday, a diagnosis moment, etc. Let t_e denote the moment the event of interest occurs. The survival time is defined as $Y = t_e - t_0$, i.e. Y is the random variable representing the time to the event of interest.

1.1 The Survival Function

All functions of the event time distribution are defined over the interval $[0, \infty)$. The probability density function (p.d.f.) is denoted by f. The distribution of a random variable is completely and uniquely determined by its probability density function. Other useful functions exist that can be obtained from the probability density function. The most important one is the cumulative distribution function (c.d.f.) of Y as

$$F(t) = \Pr(Y \leq t) = \int_0^t f(s)ds$$

where $\Pr(\cdot)$ denotes the probability that event of interest occurs.

The *survival function*, $\mathbb{S}(t)$, expresses the probability that an individual survives to time t, and is defined as

$$\mathbb{S}(t) = 1 - F(t) = \Pr(Y > t) = \int_t^\infty f(s)ds.$$

The survival curve is monotone and non-increasing, the probability of survival is 1 at t_0 (i.e. $\mathbb{S}(t_0) = 1$) and goes to 0 when time goes to infinity (i.e. $\lim_{t \to \infty} \mathbb{S}(t) = 0$). A steep (fast decreasing) survival curve means a high probability of an event.

1.2 The Hazard Function

The hazard function is a key concept in survival analysis. This function is also known as mortality rate, incidence rate, mortality curve, failure rate, or force of mortality (depending on the field of use). The hazard function, denoted as $h(t)$, gives the rate at which an individual, who has survived to time t, will experience the event in the next instant of time. The hazard function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq Y \leq t + \Delta t | Y \geq t)}{\Delta t} = \frac{f(t)}{1 - F(t)}. \tag{1}$$

The hazard rate curve can take many different shapes. If the curve increases, the chances of experiencing the particular event are low near the starting point t_0 and increase with time. Decreasing hazard rates represent events with a higher frequency close to the starting point compared to the frequency at later time points. The hazard rate function is not necessarily monotone.

1.3 The Cumulative Hazard Function

Closely related to the hazard rate is the *cumulative hazard function*, $\mathbb{H}(t)$, which is defined as

$$\mathbb{H}(t) = \int_0^t h(s)ds \tag{2}$$

The cumulative hazard function does not represent a probability, but a measure of the risk of the occurrence of an event. It is easy to derive relations between the different notions; for example, equations 2 and 1 imply that

$$\mathbb{H}(t) = \int_0^t h(s)ds = \int_0^t (\frac{f(s)}{1 - F(s)})ds = -\ln[1 - F(t)]$$

and consequently

$$\mathbb{S}(t) = 1 - F(t) = \exp\left[-\mathbb{H}(t)\right] = \exp\Big[- \int_0^t h(s)ds\Big].$$

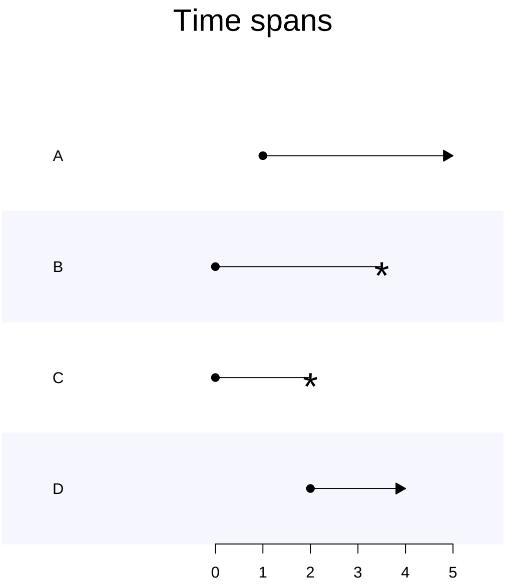
This equation is the main exponential formula of survival analysis. It presents a characterization of the distribution and survival function via the hazard function. Because of its relevant probabilistic interpretation and simplicity in probability expressions examined, the hazard function turns out to be more straightforward to work with than the density, distribution, or survival function.

1.4 Censoring

In most survival studies, some individuals in the study do not experience the prespecified event of interest during the observation period. The `censoring` is the name of this limitation on the observability of the event. There are different types of censoring: right, left and interval censoring.

Right Censoring
Figure 1 illustrated typical observed data in a study involving survival outcomes subject to censoring from time 0 (beginning of the study) to time 5 (end of the study), and the survival time is known exactly. The solid line represents the risk period for each patient. The line ending with an asterisk ($*$) indicates an occurrence of the event of interest. The line ending with an arrow indicates censoring, such as the end of follow-up or an occurrence of an event other than the event of interest, e.g. loss to follow-up, death due to causes other than the one under study or switching treatment. In Figure 1, for patients **B, C**, the time of occurrence of the events are known; hence, there is *no censoring* for these patients. Censoring may arise in the following ways:

Figure 1: The different types of right censoring mechanisms

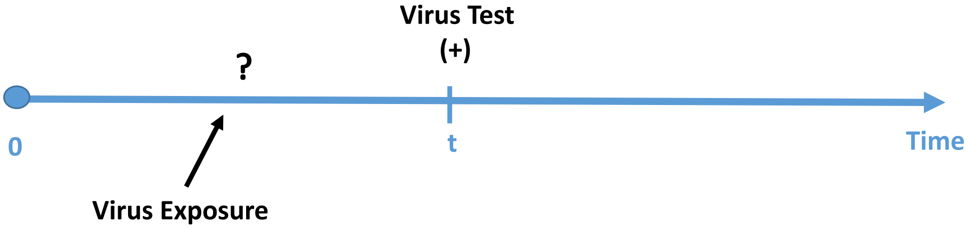


1. a patient has not (yet) experienced the event of interest, such as relapse or death, within the study time period (e.g. patient **A**);
2. a patient experiences a different event that makes further follow-up impossible such as switching the treatment (e.g. patient **D**).
This type of censoring, named *right censoring*, is often handled in survival analysis. This is because, for this example, the complete survival time interval, which we don't know, has been cut off (i.e., censored) at the right side of the observed survival time interval. The most common reasons for right censoring are:
I. **Administrative censoring:** The study ends before a person experiences the event.
II. **Not Administrative censoring:**
– `Drop out or lost to follow-up:` a person fails to return for a study visit
– `Competing event:` a competing event is an event (typically, death) that prevents the event of interest (e.g. stroke) from happening.

Left Censoring

Left censoring happens when the individual's true survival time is less than or equal to the observed survival time. An example of such a situation would be virus testing. For instance, if we have been following an individual and have recorded an event in which the person tests positive for a virus. As shown in Figure 2, we

Figure 2: Left censoring



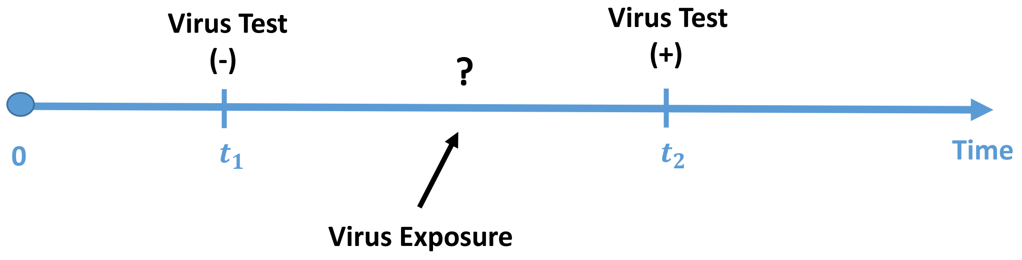
do not know when the individual was exposed to the disease. We only know that there was some exposure between 0 and the time they were tested.

Interval Censoring

For interval-censored data, the precise moment of the event is known to be within a time interval (t_1,t_2]. Thus, the event took place after time point t_1, but before or at t_2. This type of censoring may occur when periodic inspections are done, for example, if a patient visits a doctor every few months. If, at a specific visit, it is detected that a patient has experienced the event, e.g. recovery, then it is not known when the event happened exactly. The patient may have recovered only a day before the current visit, the day after the last visit, or sometime between. The exact event time is unknown, but it falls into the time interval between the last two visits. Using the virus testing example (Figure 3), if we have the situation whether we have performed testing on the individual at some timepoint t_1 and the test was negative. Then, the individual tested positive at a timepoint further on t_2. In this situation, we know the individual was exposed to the virus sometime between t_1 and t_2; however, we do not know the exact timing of the exposure.

Since exact event times are not known for some patients, analyzing survival data in the presence of censoring is more complicated. Hence, it is necessary to make assumptions about censoring when common statistical methods are used to analyze censored data. Some statistical methods designed to account for censored observations imply that patients' withdrawal from a study is independent of the event of interest. In practice, however, some covariates may be associated with lifetime and censoring mechanisms, resulting in ignorable censoring conditional on

Figure 3: Interval censoring



those covariates. Applying the classical survival techniques assumes independent censoring may under-or over-estimate the survival time in the case of covariate-dependent censoring. To do so, alternative methods are designed to account for dependent censoring.

2 Estimation of the Survival and Cumulative Hazard Functions

In the case of parametric inference, assumptions regarding the distribution of failure times must be made. This makes sense in some cases, especially when more information about the nature of the underlying process is known. On the other hand, nonparametric models are commonly used when we do not consider any parametric assumptions. It is critical to decide whether to employ a parametric or nonparametric model. Nonparametric models have the benefit of being flexible, allowing them to deal with any probability distribution. In contrast, depending on the model, parametric models often provide closed-form solutions to the hazard and survival function. Furthermore, even with small sample size, they usually make quality results. If the postulated model is valid, the estimate procedure is more efficient than nonparametric estimation.

2.1 Kaplan-Meier (KM)

Kaplan and Meier proposed the Product-Limit Estimator as an estimator for the survival function (often referred to as the Kaplan-Meier Estimator) [Kaplan and Meier, 1958]. The KM estimator assumes that at any time patients who are censored have the same survival probability as those who continue to be followed up. Furthermore, the KM also assumes that the survival probabilities are the same for subjects recruited early in the study versus late. It is a non-increasing step function with steps only at times of at least one failure.

Suppose that time points t_i denote the r distinct observation times, i.e. $t_0 < t_1 < \ldots < t_r$ $(r \leq n)$. The Kaplan–Meier estimator is

$$\hat{\mathbb{S}}(t) = \prod_{i \in R(t)} \left(1 - \frac{d_i}{\#R(t_i)}\right). \tag{3}$$

with d_i the number of events at time t_i. $R(t)$ denotes the set of indices of all individuals at risk at time t, meaning all individuals alive just before t. In addition, $\#R(t)$ denotes the number of individuals in the risk set at time t. The Product-Limit estimator can be used to estimate the cumulative hazard function, $\mathbb{H}(t) = -\ln \mathbb{S}(t)$, so

$$\hat{\mathbb{H}}(t) = -\ln \hat{\mathbb{S}}(t)$$

with $\hat{\mathbb{S}}(t)$ as defined in (3).

For the variance of the Kaplan–Meier estimate, the Greenwood formula [Greenwood, 1926] given by the expression

$$Var\left(\hat{\mathbb{S}}(t)\right) = \hat{\mathbb{S}}^2(t) \sum_{i \in R(t)} \left(\frac{d_i}{\#R(t_i) - d_i}\right)$$

is commonly used. Peterson [1977] shows that the Kaplan–Meier estimator is consistent, and [Breslow and Crowley, 1974] show its asymptotic normality.

2.2 The Cox proportional hazards model

So far, the models provided have dealt with the most basic situation of independent and identically distributed variables. This means that the population is homogenous. In most real applications, however, the population under study is not homogenous. Indeed, some variables are of particular interest as the impact of a treatment in a clinical study or confounders whose effect must be controlled for in the analysis. In both cases, we will use the notion covariate for these variables. For right-censored survival data, Cox [1972] proposed the Proportional Hazards Model to incorporate event time as a dependent variable in the survival model.

It is the most commonly used model in this area because it makes it simple to include information about known (observed) covariates in survival data models.

According to our predefined notation, let $\mathbf{X} = (X_1, \ldots, X_p)$ be the vector of p covariates. Assume that all covariates are time-independent, i.e. the values X_k, for $k = 1, \ldots, p$, do not change over time.

The Cox proportional hazards model is defined as

$$h(t|\mathbf{X}) = h_0(t)c(\beta, \mathbf{X})$$

where $h_0(t)$ is an arbitrary baseline hazard rate, and no structure is imposed on it, which gives the model great flexibility. The model assumes that all subjects in the study population share a baseline hazard (the risk of death or other events). This assumption can be relaxed, but it simplifies the presentation. The parameters of primary interest are contained in $c(\beta, \mathbf{X})$, often

$$c(\beta, \mathbf{X}) = \exp\{c(\beta^\top \mathbf{X})\}$$

yielding

$$h(t|\mathbf{X}) = h_0(t) \exp(\beta^\top \mathbf{X}) \tag{4}$$

with $\beta^\top = (\beta_1, \ldots, \beta_p)$ is the parameter vector. The covariates in this model work multiplicatively on the baseline hazard, adding additional risks based on the individuals' prognostic information. This provides for a simple and unambiguous understanding of the model. The essential thought is to separate the time effect in the baseline hazard function on one side and the influence of covariates in an exponential term on the other. In summary, this assumption (named as *proportionality assumption*) says that the hazards of two individuals at time t are related by a proportionality constant that does not depend on t.

The Cox model is called a semiparametric model because of the parametric nature of the covariate term and the nonparametric baseline hazard function. The Cox estimator is almost entirely inferred via asymptotic results [Andersen and Gill, 1982]. This semiparametric model is the most widely used in survival analysis. It is included in all statistical packages, is simple to use, and creates an output that is easy to understand. As a result, we will concentrate on this model in this thesis.

Estimating the coefficients

The maximum likelihood estimator may be used to estimate the coefficients in the Cox model; however, the maximum partial likelihood estimator is a commonly used alternative. The partial likelihood employs only a part of the full likelihood. In the Cox model, individual i has a hazard ratio at time t:

$$h(t|X_i) = h_0(t) \exp(\beta^\top X_i)$$

Let t_i, $i = 1, \cdots, n$ be the observed time points. We define the risk set $R_i = \{j \mid t_j \geq t_i\}$. We are assuming no tied t_i's. We may alternatively state that at time t_i subject $j \in R_i$ is still at risk, i.e. individual j has not failed or been censored by time t_i. People that are censored are also part of the risk set, but in that case, $\delta = 0$. Here, δ is an event indicator and obviously, for censoring patients is 0. As a result, the probabilities of censoring are not included in the partial likelihood. Given the risk set R_i, the conditional probability that subject i experience the event of interest at time point t_i is $\frac{h(t_i|X_i)}{\sum_{j\in R_i} h(t_j|X_j)}$. Then, the partial likelihood is:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{h(t_i|X_i)}{\sum_{j\in R_i} h(t_j|X_j)} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\beta^\top X_i}}{\sum_{j\in R_i} e^{\beta^\top X_j}} \right)^{\delta_i}$$

which this expression $h_0(t)$ drops out. That is why proportionality is such an attractive assumption to make. The partial log-likelihood is calculated as follows:

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \left(\beta^\top X_i - \log\,(\sum_{j\in R_i} e^{\beta^\top X_j}) \right)$$

The `coxph` function in the statistical package `R` computes the maximum of this function.

3 Diagnostics for the Cox Model

The proportional hazards regression model [Cox, 1972] estimates the effect of covariates on failure time. Because of its widespread applicability, before adopting the results of a fitted Cox model as valid, a few issues should be addressed: is the proportional hazards assumption satisfied? Are the variable's functional forms appropriate? Is there any evidence of outliers or influencing observations? Several methods have been proposed to address those issues, many of which rely on different forms of model *residuals*. On that note, residuals play a significant role in regression method diagnostics. Multiple types of residuals are defined for the Cox model, and they can often serve different purposes in model diagnostics. The common residuals for the Cox model include:

- *Schoenfeld residuals* to check the proportional hazards assumption
- *Martingale residuals* to assess nonlinearity
- *Deviance residuals* (symmetric transformation of the Martinguale residuals), to examine nonlinearity and influential observations.
- *Delta-beta residuals* to test Influential Observations.

3.1 Schoenfeld residuals

Schoenfeld [1980] proposed a chi-squared goodness-of-fit test for the proportional hazards regression model, which utilizes a residual of the form Expected - Observed. The formal definition and its properties were later discussed in Schoenfeld [1982].

Suppose there are n individuals indexed by $i = 1, \cdots, n$ and that each has a p-vector of covariates $X_i = (X_{i1}, \cdots, X_{ip})^\top$. The proportional hazards regression model specifies that the hazard function (Equation 4) of the i^{th} individual is

$$h(t|X_i) = h_0(t) \exp(\beta^\top X_i)$$

where β is a vector of p parameters and $h_0(t)$ is an arbitrary function. Let $\mathfrak{D}$ be the indices of the individuals who failed and let $\mathfrak{R}_i$ be the indices of those under observation when the i^{th} individual fails. Using partial [Cox, 1975] or marginal [Kalbfleisch and Prentice, 1980] likelihood, one can estimate the parameter. For $i \in \mathfrak{D}$, an index $m \in \mathfrak{R}_i$ is selected with probability

$$\frac{\exp\left(\beta^\top X_m\right)}{\sum_{k \in \mathfrak{R}_i} \exp\left(\beta^\top X_k\right)}.$$

In this model X_i, is a random variable with

$$\mathbb{E}(X_{ij} \mid \mathfrak{R}_i) = \frac{\sum_{k \in \mathfrak{R}_i} X_{kj} \exp\left(\beta^\top X_k\right)}{\sum_{k \in \mathfrak{R}_i} \exp\left(\beta^\top X_k\right)}, \quad j = 1, \cdots, p$$

and the maximum likelihood estimate of β is a solution to

$$\sum_{i \in \mathfrak{D}} \mathbb{E}\left(X_{ij} - \mathbb{E}(X_{ij} \mathfrak{R}_i)\right)$$

We substituted the solution (i.e. $\hat{\beta}$) into $\mathbb{E}(X_{ij} \mid \mathfrak{R}_i)$ and denoted by $\hat{\mathbb{E}}(X_{ij} \mid \mathfrak{R}_i)$. Let define the partial residual at t_i as the vector $\hat{\tau}_i = (\hat{\tau_{i1}}, \cdots, \hat{\tau_{ip}})^\top$, where

$$\hat{\tau}_{ik} = X_{ik} - \hat{\mathbb{E}}(X_{ik} \mid \mathfrak{R}_i).$$

Thus the residual is the difference between the observed value of X_i and its conditional expectation given $\mathfrak{R}_i$.

Remark 1 (Examining the proportional hazard assumption)**.** *If proportional hazards holds,* $\mathbb{E}(\hat{\tau}_i) \approx 0$ *and a plot of* $\hat{\tau}_{ik}$ *versus* t_i *will be centred about* 0.

In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of a violation of the PH assumption[Grambsch and Therneau, 1994]. Furthermore, the function *cox.zph()* in the `survival` package in R [Therneau, 2021] gives an easy way to verify the proportional hazards assumption for each covariate in a Cox regression model fit. To test for independence between residuals and time, the function *cox.zph()* correlates the matching set of scaled Schoenfeld residuals with time for each covariate. Furthermore, it does a global test on the entire model.

According to Keele [2010], while the Therneau and Grambsch test is commonly used because it is simple to conduct and interpret, its implementation requires considerable caution due to its sensitivity to various types of misspecification. Omitted predictors, interactions, and nonlinear covariate functional forms can all significantly impact the test result. The research also emphasized correcting the functional form for continuous covariates before testing for non-proportionality [Keele, 2010].

3.2 Martingale residuals

Martingale residuals are very useful and can be used for many of the usual purposes that we use residuals for in other models (identifying outliers, choosing a functional form for the covariate, etc). Martingale residual represents the discrepancy between the observed value of a subject's failure indicator and its expected value, integrated over the time for which that patient was at risk [Therneau and Grambsch, 2000]. Moreover, Martingale residuals may present any value in the range (-∞, +1):

- A value of martingale residuals near 1 represents individuals that "die too soon",
- Large negative values correspond to individuals that "live too long".

To describe the martingale residuals, we define the counting process, $N_i \equiv \{N_i(t), t \geq 0\}$ $(i = 1, \cdots, n)$ indicates the number of observed events experienced over the passage of time. These processes have the intensity $U_i(t)h_0(t)\exp(\beta^\top X_i(t))$ where $U_i(t)$ is a left continuous $0-1$ process indicating whether the i^{th} subject is in the risk set at time t, and $X_i(t)$ is a p-dimensional vector of left continuous covariate processes having right hand limits. The differences between the counting processes and their respective integrated intensity functions

$$M_i(t) = N_i(t) - \int_0^t U_i(v)\exp(\beta^\top X_i(v))h_0(v)dv.$$

are martingales [Therneau et al., 1990]. The martingale residuals are defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t U_i(v)\exp(\hat{\beta}^\top X_i(v))d\hat{\Lambda}_0(v). \tag{5}$$

where

$$\hat{\Lambda}_0(t) = \frac{\sum_{l=1}^n dN_l(v)}{\mathfrak{S}^{(0)}(\hat{\beta}, v)}$$

and $\mathfrak{S}^{(0)}(\beta, v) = \sum_{l=1}^{n} U_l(v) \exp\{\beta^\top X_l(v)\}$. Martingale residual $\hat{M}_i(t)$ can be interpreted as the difference at time t between the observed and expected numbers of events for the i^{th} subject.

Martingale residuals play an essential role in functional form diagnostics. Barlow and Prentice [1988] provided a more detailed discussion and illustrated that plots of such residuals might provide insight into the choice of the model form. Therneau et al. [1990] discussed the usage of martingale residuals in investigating the functional form of covariates.

Continuous covariates are frequently assumed to have a linear form. This assumption, however, needs to be tested. Plotting the Martingale residuals versus continuous covariates is a standard method for detecting nonlinearity or determining the functional form of a covariate. Patterns in the plot for a particular continuous covariate may indicate that the variable is not correctly fitted. Nonlinearity is not an issue for categorical variables, so we only examine plots of martingale residuals and partial residuals against a continuous variable. Furthermore, according to Therneau et al. [1990], it is common to plot the martingale and deviance residuals against time to check for possible outliers.

3.3 Deviance residuals

The primary drawback to the martingale residual ($\hat{M}_i(t)$) is its clear asymmetry (its upper bound is 1, but it has no lower bound). As a visual aid in certain plots, it may be helpful to transform the residual to achieve a more normal-shaped distribution. The Deviance residual is a normalized transformation of the martingale residual.

Inspired by the deviance residuals for GLM in McCullagh and Nelder [1983], Therneau et al. [1990] introduced the deviance residual for a Cox model as

$$D_i(t) = \text{sign}(\hat{M}_i(t))\sqrt{-2[\hat{M}_i(t) + \delta_i \ln(\delta_i - \hat{M}_i(t))]} \tag{6}$$

where δ_i is an event indicator and $\hat{M}_i(t)$ computed by 5. In Equation 6, the log function "expands" residuals close to one, while the square root contracts the large negative values. Note that the Deviance residual is zero if and only if $\hat{M}_i(t) = 0$

In the plot of Deviance residuals, positive values correspond to individuals that "died too soon" compared to expected survival times and negative values correspond to the individual that "lived too long". Besides, very large values are outliers, which are poorly predicted by the model. Both the martingale residual and the deviance residual are useful for assessing the functional form of a continuous variable in a Cox proportional hazards model and identifying outlying observations, but the deviance residual is less skewed and, therefore, more useful.

3.4 Delta-beta residuals

An influential measurement significantly affects model fit and may be measured in various ways. As methods of quantifying and analyzing influence, a variety of residuals (score residuals, Schoenfeld residuals, delta-beta residuals) have been proposed; we will focus on delta-beta residuals. The idea behind Delta-beta residuals is straightforward: let $\hat{\beta}_j^{(i)}$ denote the estimate of $\hat{\beta}_j$ obtained if we leave subject i out of the model. The Delta-beta residual for coefficient j and subject i is therefore defined as

$$\Delta_{ij} = \beta_j - \hat{\beta}_j^{(i)}$$

This may appear a computationally challenging task, but several computational tricks allow one to refit models reasonably quickly while leaving individual observations out. In the Delta-beta residuals plot, we determine the estimated changes in the regression coefficients upon deleting each observation.

Chapter 5

Causal survival analysis

This chapter goes through some specifics concerning causal survival analysis. More precisely, in Section 1, all necessary notations and assumptions dealing with the censoring mechanism, as well as causal estimands, are introduced. Furthermore, the Marginal Structural Cox Model, which is the core of this thesis, is presented in Section 2. In the current chapter, we mainly show how to overcome challenges by applying the new weighted method to the Marginal Structural Cox Model. It is the novelty of this thesis. Finally, we explore our proposal techniques in a comprehensive Simulation Study in Section 3.

1 Causal estimands

1.1 Notations

In causal survival analysis, let $Y_i(z) \geq 0$ and $C_i(z) \geq 0$ be the potential survival time and the potential censoring time for unit i under treatment assignment $z \in \{0, 1\}$. The follow-up starts and ends at specific calendar times, which determine a fixed duration of the study, c.

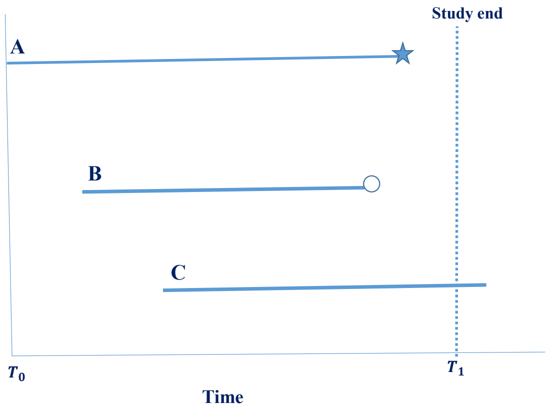
In some studies, patients may enter the study at different times and are subsequently assigned to different treatment arms and monitored until they either experience the event of interest or end the study. The censoring time is calculated by entering units, which are staggered over time. Thus, $C_i(z) \leq c$ represents the duration of entering the follow-up till the end of the study for unit i given treatment assignment z. This is so-called "administrative censoring".

Treatment switching occurs when patients in one group switch from the treatment arm to another in the trial. The situation arose mainly due to concern for the patient's health. In the presence of the switching behavior, let $S_i(z)$ be the potential switching time (the duration of entering the follow-up till switching the treatment) of unit i under treatment assignment z. The main point is that unit i can switch the treatment arms only before her/his survival time, implying a natural constraint $S_i(z) < Y_i(z)$. Let us define

- the survival time under the actual treatment assignment as
$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$
- the administrative censoring time under the actual treatment assignment as
$$C_i = Z_i C_i(1) + (1 - Z_i) C_i(0).$$
- the switching time under the actual treatment assignment as
$$S_i = Z_i S_i(1) + (1 - Z_i) S_i(0).$$

In general, the observed outcome is $\tilde{Y} = \min(Y_i, C_i, S_i)$. It is worth to noting that in our data set, we have the following cases (see Figure 1).
(a) Patients who experience the event during the follow-up and do not switch treatment. For these patients $\min(Y_i, C_i, S_i) = Y_i$ and $\tilde{Y} = Y_i$ (**e.g. Patient A**)
(b) Patients who switch treatment before experiencing the event. The natural constraint implies that $\min(Y_i, C_i, S_i) = S_i$ and $\tilde{Y} = S_i$ (**e.g. Patient B**).
(c) Patients who neither switch treatment nor experience the event. For these patients $\min(Y_i, C_i, S_i) = C_i$ and $\tilde{Y} = C_i$ (**e.g. Patient C**).

Figure 1: The observed outcome



1.2 Censoring Assumption

As in the missing data literature [Rubin, 1976, Tsiatis, 2006], three assumptions about the censoring mechanism have been proposed: censoring completely at random (CCAR), censoring at random (CAR), and censoring not at random (CNAR). In some cases, censoring due to administrative constraints, e.g., the planned start and end of the study, is unrelated to the study treatment or the underlying health condition. As a result, the event times will most likely be CCAR. In the presence of staggered entry (i.e., with a varying start of follow-up date) and variable end of follow-up for patients, the administrative censoring is unlikely to be CCAR. In our case, this is because it is dependent on patient characteristics. For instance, MS affects young adults rather than older individuals. This indicates that the event was observed more frequently in younger patients than in older patients. Furthermore, censored event times due to non-administrative reasons such as dropout or treatment switching are unlikely to be CCAR. For example, adverse events may cause individuals to drop out of the study or switch treatments. Common survival analysis methods assume CAR that patients censored at t and patients uncensored at t with the same history have the same distribution of the entire current and future variables. Even after accounting for their observed history, this assumption will be violated if sicker subjects are more likely to withdraw from the study, resulting in CNAR [Rubin, 1976].

Formally, some possible assumptions on the **censoring mechanism** are as

- No unmeasured confounding [Robins and Finkelstein, 2000, Hernán et al., 2001]:

$$(C_i(1), C_i(0), S_i(1), S_i(0)) \perp Z_i | X_i$$

It indicates that we have measured all variables that contribute to the censoring process and determines whether a subject is exposed or not. In that sense, it implies that exposure and censoring mechanism are conditional independent.

- Completely independent censoring (completely ignorable/ non-informative):

$$Y_i(0) \perp C_i(0),\ Y_i(1) \perp C_i(1)$$

and

$$Y_i(0) \perp S_i(0),\ Y_i(1) \perp S_i(1)$$

which implies that time to censoring is independent of the time to event.

- Covariate-dependent censoring:

$$Y_i(0) \perp C_i(0)|X_i,\ Y_i(1) \perp C_i(1)|X_i$$

and

$$Y_i(0) \perp S_i(0)|X_i,\ Y_i(1) \perp S_i(1)|X_i$$

which implies that time to censoring is conditional independent of the time to event given the covariates.

1.3 Definition of the estimands

The definition of the estimands when the outcome is survival times is challenging. This is because the survival outcome is non-negative, skewed, and subject to right censoring. As a result, the average causal effect (ATE), defined as the mean difference, may not be estimable or have the desired interpretation for many clinical studies. To compare treatment groups based on survival outcomes, five estimands in the literature have proper interpretation Royston and Parmar [2011, 2013], Uno et al. [2014], Andersen et al. [2017], Mao et al. [2018]. They are Average Survival Causal effect (ASCE), Restricted Average Survival Causal Effect (RACE), Survival Probability Causal Effect (SPCE), and Survival Quantile Effect (SQE). These quantities are defined as follows:

According to conterfactual survival function, let us define $\Pr(Y(1) > t) = \mathbb{S}_1(t)$ and $\Pr(Y(0) > t) = \mathbb{S}_0(t)$ as the survival functions under treatment and control.
1. `Average Survival Causal Effect` (ASCE) is

$$\Delta_{\text{ASCE}} = \int_0^{\infty} \mathbb{S}_1(t)dt - \int_0^{\infty} \mathbb{S}_0(t)dt$$

2. `Restricted Average Survival Causal Effect` (RACE) is

$$\Delta_{\text{RACE}} = \int_0^{t^*} \mathbb{S}_1(t)dt - \int_0^{t^*} \mathbb{S}_0(t)dt$$

where t^* is a pre-specified time point.
3. `Survival Probability Causal Effect` (SPCE) is

$$\Delta_{\text{SPCE}} = \mathbb{S}_1(t^*) - \mathbb{S}_0(t^*)$$

where t^* is the time point at which the survival probability is evaluated.
4. `Survival Quantile Effect` (SQE) is

$$\Delta_{\text{SQE}} = \mathbb{S}_1^{-1}(1-q) - \mathbb{S}_0^{-1}(1-q)$$

where q is a pre-specified number between 0 and 1. The median survival times are compared with $q = 0.5$.

Δ_{ASCE} is the mean difference in survival time when the entire population is placed under treatment and control. Δ_{RACE} is the expected survival time restricted by the upper bound t^*. It measures the between-group restricted averages and reduces to Δ_{ASCE} when t^* goes to infinity. Δ_{SPCE} is the difference between two survival probabilities at time t^*. Δ_{SQE} compares the q-quantile of the survival distribution between groups. According to Greenland et al. [1999], Δ_{SQE} has been recommended as an alternative for survival comparison especially when the proportional hazards assumption does not hold[Uno et al., 2014]. All these estimands (i.e. $\Delta_{\text{ASCE}}, \Delta_{\text{RACE}}, \Delta_{\text{SPCE}}$ and Δ_{SQE}) are causal estimands in the sense that they can be defined on Rubin's causal model framework. In this thesis, we focus on the first three causal estimands (i.e. $\Delta_{\text{ASCE}}, \Delta_{\text{RACE}}, \Delta_{\text{SPCE}}$).

2 Marginal Structural Cox Model

Robins and his colleagues [Robins, 2000a, Robins et al., 2000, Hernán et al., 2000] have proposed a new class of causal models called marginal structural models. In general, the Marginal Structural model is

$$\mathbb{E}(Y(z)) = G(z; \beta)$$

where the parameters β of this marginal structural model become the causal parameters of interest and the target of our estimation. One of the advantages of this approach is the flexibility and range of MSMs that can be fitted.

Marginal structural cox models (MSM) estimated using inverse probability of treatment weighting (IPW) for time-to-event outcomes were introduced by Hernán et al. [2000]. Other methods include estimation of MSMs using the g-formula (also called g-computation) [Robins, 1986, Daniel et al., 2011, Keil et al., 2014], structural nested accelerated failure time models [Robins, 1992, Hernán et al., 2005], structural nested failure time models [Robins et al., 1992, Vansteelandt and Joffe, 2014], structural nested cumulative failure time models [Picciotto et al., 2012], and structural nested cumulative survival time models [Seaman et al., 2020]. A recent review [Clare et al., 2019] found that the marginal structural Cox models are the most commonly used method in practice.

The Marginal structural Cox Model is defined as

$$h(t|Z_i) = h_0(t) \exp\left\{\gamma Z_i\right\} \tag{1}$$

where $h_0(t)$ is an unspecified baseline hazard function, γ encodes the unknown exposure effect. This model is a marginal structural model because it is a structural model for the marginal distribution of the counterfactual outcome. The Marginal Structural Cox model has been widely used in observational studies for the analysis of the effect of different therapies on the progress of various diseases, such as AIDS and hemodialysis [Cole et al., 2003, Sterne et al., 2005, Hernán and Robins, 2006].

Marginal structural models attempt to adjust for measured confounders to enhance group comparability and estimate causal effects in a similar way [Robins et al., 2000] To accomplish this, the partial likelihood function of the Cox model was modified such that the contribution of patient i to the risk set at time t was weighted to remove the possible confounding effects of baseline confounders [Hernán et al., 2000].

Let's exploit the profile partial score function of interest as

$$PL(\gamma) = -\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{c} \big[Z_i - \frac{\mathbb{E}_n\big\{ZR(t)\exp\big\{\gamma Z\big\}\big\}}{\mathbb{E}_n\big\{R(t)\exp\big\{\gamma Z\big\}\big\}}\big] dN_i(t)$$

where $\mathbb{E}_n(.)$ referring to the sample average and $R_i(t)$ is the at risk indicator (which is the product of the indicators $I(Y_i \geqslant t)$, $I(C_i \geqslant t)$ and $I(S_i \geqslant t)$), and $dN_i(t)$ the increment in the counting process with respect to the event time Y_i. It is shown that weights (say, $\hat{\omega}$) removes the possible confounding effects of baseline confounders [Hernán et al., 2000]. In the marginal structural Cox model, these weights ($\hat{\omega}$) are inserted in the partial likelihood function as follows:

$$PL(\gamma) = -\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{c} \hat{\omega} \times \big[Z_i - \frac{\mathbb{E}_n\big\{ZR(t)\hat{\omega}\exp\big\{\gamma Z\big\}\big\}}{\mathbb{E}_n\big\{R(t)\hat{\omega}\exp\big\{\gamma Z\big\}\big\}}\big] dN_i(t)$$

Under correct specification of the model, the maximum partial likelihood estimator $\hat{\gamma}$, defined as the solution to the score equation $PL(\hat{\gamma}) = 0$, is shown to converge in distribution to Normal distribution with mean zero and a covariance matrix which can consistently be estimated by $-\frac{\partial PL(\gamma)}{\partial \gamma}|_{\gamma=\hat{\gamma}} = 0$. Moreover, using stabilized weights also provides consistent estimates of γ [Hernán et al., 2000].

According to [Hernán et al., 2000], the analysis based on weighted samples gives an asymptotically unbiased estimate of the causal parameter of interest, which relies on the potential outcome framework. Under the four assumptions of SUTVA, unconfoundedness, positivity, and no misspecification of the model used to estimate the weights, weighting creates a pseudo-population in which the exposure is independent of the measured confounders [Hernán and Robins, 2006].

It is important to note that there is a distinction between structural assumptions and inference assumptions. All of the assumptions based on the valid causal inference are generally referred to as "*structure assumptions*", and they are as SUTVA, unconfoundedness, positivity, and the ignorability assumption related to censoring mechanisms. On the other hand, assumptions based on the correct functional forms of each equation in the (partial) likelihood (i.e. linearity), accurate measurements of all of the observed variables (i.e. the reliability of the available data), guarantees unbiased estimation of the model's parameters, and assess the proportionality of cox model in constructing censoring weights are referred to as "*inference assumptions*".

One of the difficulties in dealing with weights is how to construct them. More specifically, the construction of weights for marginal structural models requires a thoughtful procedure to determine which functional form of model optimizes bias reduction and precision. In the next section, we focus on the weighting framework for causal effect estimation and address how to construct it.

2.1 The weighting framework for causal effect estimation

Lack of balance is common in observational studies, and one of the initial choices is using standard parametric adjustment by regression. Nevertheless, it is often sensitive to model misspecification when groups differ significantly in observed characteristics [Rubin, 1979]. *Weighting* is a well-known non-parametric balancing strategy. In this regard, weights are applied to the sample of units in each treatment group to match the covariate distribution of a target population. The formal definition and different balancing weights are presented in Chapter 3, Section 2.3.

Regarding some interesting features shown in Table 1, in this thesis, we focus on IPW and Overlap weighting to deal with observed confounders. When propensity scores approach 0 or 1, IPW has limitations: large weights for individual patients, bias, and large variability in the estimated treatment effect [Stuart, 2010, Hirano and Imbens, 2001]. To address these problems, trimming methods have been proposed that exclude patients with very high predicted probabilities of being in the treatment group (or the control group). Despite the potential gains from trimming, the decision regarding how many patients to exclude is ad-hoc and can result in a substantial loss of sample size.

Robins et al. [2000] discuss a technique they refer to as *stabilization* that reduces the variability of the IPW weights and gives individuals with extreme weights less influence. The *stabilized* IPW is accomplished by multiplying the treatment and comparison weights (separately) by a constant, equal to the expected value of being in the treatment or comparison groups, respectively. Thus, the stabilized IPW weights are $\frac{\Pr(Z=1)}{\hat{e}(X)}$ for the treated and $\frac{\Pr(Z=0)}{1-\hat{e}(X)}$ for the control group. The stabilized weighting makes the narrower range of the weights for each individual. Furthermore, in many settings (e.g., time-varying or continuous treatments), it is recommended to use stabilized weights [Hernán and Robins, 2006, 2020]. According to Hernán and Robins [2006, 2020], one should always check that the estimated *stabilized* weights have mean 1. Deviations from 1 indicate *model misspecification* or possible violations, or near violations, of *positivity.*

2.2 Censor weights

In the presence of covariate-dependent censoring [Zheng and Klein, 1994, Huang and Wolfe, 2002, Braekers and Veraverbeke, 2005], many methods have been developed to analyze data. To named but a few, for administrative censoring, the `Inverse Probability of Censoring Weighting (IPCW)` [Robins and Rotnitzky, 1992, Robins, 1993, Robins and Finkelstein, 2000] is proposed. To dealing with selection bias due to switching the treatment, `standard intention-to-treat (ITT)` for analyzing RCTs[Moher et al., 2001], `Exclude switchers` [Watkins et al.,

Table 1: The overview of all weighting methods

Methods	Pros	Cons
Inverse Probability Weighting (IPW)	• It is the most famous balancing weight. • It can be used to adjust for measured confounding and selection bias • It is easy to implement.	• It is so sensitive to poor overlap and PS model misspecification. In practice, violations of the positivity assumption often manifest by the presence of limited overlap in the PS distributions between treatment groups.
Overlap Weighting (OW)	• The overlap weights are bounded between 0 and 1. • By definition, the overlap weights automatically overcome the large uncertainty issue caused by extreme propensity scores when using IPW. • The OW based on the PS estimated from a logistic model leads to exact balance between treatment groups for all covariates. • More robust to misspecification of the propensity score model and limited overlap than IPW.	• With the overlap weight, we estimate causal effects for a specific subpopulation.

2013], `Inverse Probability of Censoring Weighting (IPCW)` [Hernán et al., 2001], `two-stage adjustment`[Morden et al., 2011, Latimer et al., 2017, 2018, 2019] and `Rank preserving structural failure time models (RPSFTM)` [Robins and Tsiatis, 1991] have been proposed. Since the most famous method to deal with both administrative and switching treatment is the Inverse Probability of Censoring Weighted (IPCW), in this thesis, we focus on correcting selection bias due to covariate-dependent censoring by giving extra weight to subjects who are not censored at the certain time. In practice, a Cox proportional hazard model is assumed for the event time, while an inverse probability of censoring weight is applied to the Cox model score equation. The weight can be considered the inverse probability of remaining uncensored at the considered time, which can be estimated non- or semi-parametrically. With these weights, the survival function can be estimated in the absence of censoring.

The general form of the non-stabilized IPCW is

$$\frac{1}{\Pr(C_i > t|X_i)}$$

where the denominator represents the probability of an individual remaining uncensored conditional only on baseline confounders. In the IPCW, the censored individuals are replaced by copies of uncensored individuals with the same values of treatment and covariate.

To fit a model for censoring and controlling for all confounders, it is common to consider

1. **logistic regression:** This method is based on discrete-time, dividing follow-up into small intervals and using pooled logistic regression. It is worth noting that it only uses limited information as to whether an observation is censored or not [Hernán et al., 2001].
2. **Cox proportional hazards model:** This method is based on continuous-time, predicting the time to an event under the proportional hazards assumption. This method works when the censoring mechanism is independent or when the censoring mechanism may depend on the set of covariates [Jackson et al., 2014, Willems et al., 2018].

To assess the influence of covariates $\mathbf{X}_i$ on the probability of being censored for subject i, the Cox model for time to censoring is considered as

$$h_c(t|\mathbf{X}_i, Z_i) = h_0^c(t) \exp\left[\beta_c^\top \mathbf{X}_i + \gamma Z_i\right] \quad for\ all\ t > 0 \tag{2}$$

where $h_0^c(t)$ is the baseline hazard of censoring, β_c is the vector of model parameters. The hazard $h_c(t|\mathbf{X}_i, Z_i)$ indicates the estimated probability for patient i with covariates $\mathbf{X}_i$ being censored in the next instant of time, if this subject was not censored until time t. A Product-Limit estimator for time to censoring that includes covariates for subject i, is derived as

$$\hat{K}_i^x(t) = \prod_{\left\{j;t_j<t,\delta_j=1\right\}} \left[1 - \hat{h}_0^c(t_j)\exp\left[\hat{\beta}_c^\top \mathbf{X}_i + \hat{\gamma} Z_i\right]\right]$$

where the t_j's are the observation times, and $\hat{\beta}_c$ is the estimated vector of parameters equation (2). The δ_j indicates whether for subject j the event was observed or not, i.e. $\delta_j = 0$ for censored subjects and $\delta_j = 1$ if the event is observed. Therefore, the weights for each subject i are computed as $\omega_i^c = \frac{1}{\hat{K}_i^x}$. This is called "unstabilized weights". To avoid numerical problems, Robins [1993] proposed a modified version of the weights (called "stabilized weights"): $\omega_i^{sc} = \frac{\hat{K}_i^0}{\hat{K}_i^x}$ where $\hat{K}_i^0(t)$ denote the traditional product-limit estimator for the probability of being uncensored independent of the covariates $\mathbf{X}_i$. So, the stabilized IPCW (for both administrative and switching censoring) is computed seperately as

$$\omega_i^{sc} = \frac{\Pr(C_i > t)}{\Pr(C_i > t|X_i)} = \frac{\hat{K}_i^0}{\hat{K}_i^x} \tag{3}$$

$$\omega_i^{ss} = \frac{\Pr(S_i > t)}{\Pr(S_i > t|X_i)} = \frac{\hat{K}_i^0}{\hat{K}_i^x} \tag{4}$$

2.2.1 Functional form of censoring weights

Covariate adjustment poses difficulties in the presence of a covariate-dependent censoring mechanism. One of the difficulties in constructing censoring weights is predicting which variables will be adjusted for and which functional form (and how). The adjustment for variables raises additional concerns regarding model misspecification since any change in the adjustment set affects the censoring assumption and the treatment effect estimator. Regarding that, in this thesis, we consider all main effects and add some nonlinear terms (e.g., interaction and higher orders).

In practice, we usually do not know the proper form of the censoring weights. The typical work is to include all main effects and add interactions and polynomial terms, which produce a better model. With p covariates, one possible strategy would be to try all 2^p models based on including/excluding each covariate. One may then pick the optimal model according to a certain criterion, e.g., the model with the smallest AIC. This strategy is known as *best subset selection.* However, best subset selection is computationally very intensive. It requires the evaluation of 2^p models, which quickly becomes enormous. This makes it practically impossible to find the "best" model across all models that can contain up to all predictors.

Therefore, it is more common to use computationally faster forward, backward, and hybrid stepwise approaches. The *backward stepwise* regression works as follows. Let $\mathbb{M}_p$ be the full model which contains all predictors. For $k = p, \cdots, 1$, we then choose the "best" model among all k models that remove one predictor from the selected model $\mathbb{M}_k$, and call $\mathbb{M}_{k-1}$. We then choose the "best" model out of the selected models.

2.3 Our proposal

To calculate the survival probability $\hat{\mathbb{S}}(t|Z_i)$ specific to each treatment, we estimated the parameters of a marginal structural Cox proportional hazards model of the form $h(t|Z_i)$ in equation (1). This marginal structural model can be estimated by using weights ω_i under the assumptions of SUTVA [Hernán and Taubman, 2008, Cole and Frangakis, 2009], positivity [Hernán and Robins, 2006, Cole and Hernán, 2008], no unmeasured confounding or selection bias, and correct model specification as

$$h^{\omega_i}(t|Z_i) = h_0(t) \exp\big\{\theta Z_i\big\}.$$

where the superscript ω_i indicates that the hazard of MS disease for patient i at time t and treatment Z_i, that is $h(t|Z_i)$, is weighted by ω_i. Moreover, θ is the log hazard ratio, and $h^{\omega_i}(t|Z_i) \to h(t|Z_i)$ in distribution and therefore $\theta \to \gamma$ if the above stated assumptions are met.

To assign weights for each unit, we propose

$$\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \times \omega_i^{sc}$$

and

$$\hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss} \times \omega_i^{sc}$$

where ω_i^{sc} and ω_i^{ss} are the stabilized IPCW for administrative (equation 3) and switching censoring (equation 4), respectively. Furthermore, treatment weights (i.e. ω_i^{ipw} and ω_i^{ow}) are computed by fitting a logistic regression model for the propensity score. In general, to overcome the problem of dependent censoring due to treatment switching and administrative censoring and to adjust for observed confounders simultaneously, we propose to multiply all weights to obtain overall weights ($\hat{\omega}_i$) that made all confounding removed by creating "pseudo-population". Moreover, in order to reduce some concerns regarding misspecification assumptions in both censoring weights, we will consider all main effects as well as some nonlinear terms.

2.4 Alternative methods using PS weighting to estimate survival function

In the literature, when applied to an inverse probability weighted sample, the marginal structural cox model is a popular method for drawing causal inferences with survival outcomes in observational studies Cole and Hernán [2004] or the causal hazard ratio Austin and Schuster [2016]. Furthermore, Binder et al. [2014], Andersen et al. [2017], Zeng et al. [2021] proposed combining weighting methods with pseudo-observations to estimate the causal survival function. Each pseudo-observation is treated as an uncensored contribution to the target parameter, enabling the standard approach to continue as if outcomes are observed. Nevertheless, pseudo-observations are jackknife statistics requiring intensive computation for estimating survival functions with large sample sizes [Zeng et al., 2021]. A final approach is to combine IPW with inverse probability of censoring weight (IPCW), which is inherently connected to the Kaplan-Meier estimator [Robins and Finkelstein, 2000] to estimate counterfactual survival functions while accommodating covariate-dependent censoring. The challenge presented by extreme propensity scores in IPW also applies to survival outcomes [Austin and Schuster, 2016], and it makes sense that OW would help alleviate this challenge. However, investigation on the empirical performance of OW with survival outcomes has been limited with two exceptions: It was proved by Mao et al. [2018] that combining OW with a Cox outcome model results in efficiency gain from IPW for a variety of causal estimands. Zeng et al. [2021] combined OW with pseudo-observations and showed that OW leads to optimal efficiency. In the following, we briefly describe these two methods.

1. Mao's method:

Mao et al. [2018] proposed a unified analytical framework for propensity score weighting analysis with survival outcomes that includes a estimation framework for point and variance estimation. Suppose $r = \zeta(t)^\top \mathbf{a} + \lambda(\mathbf{t})^\top \mathbf{b}$ be a regression spline approximation for the log-hazard function for one treatment group of sample size n. The parameters are $\mathbf{a} = (\mathbf{a_0}, \cdots, \mathbf{a_L})^\top$ and $\mathbf{b} = (\mathbf{b_0}, \cdots, \mathbf{b_L})^\top$ and for simplicity, define $\alpha = (\mathbf{a}^\top, \mathbf{b}^\top)^\top$. The weighted log likelihood for subject i is

$$\ell_i = W_i \Big\{ \delta_i \mathbf{U}(Y_i)^\top \alpha - \int_0^{Y_i} \exp(\mathbf{U}(t)^\top \alpha) dt \Big\}$$

where W_i is the weight (IPW or OW) for subject i, $\mathbf{U}(\mathbf{t}) = (\zeta(\mathbf{t})^\top, \lambda(\mathbf{t})^\top)^\top$ is the known truncated power basis function of degree L with K knots. $\zeta(t) = (1, 2, \ldots, t^L)^\top$ and $\lambda(t) = (\lambda_1^L(t), \ldots, \lambda_K^L(t))^\top$. The penalized spline estimator of α can be calculated by maximizing the penalized log likelihood with respect to α using the **Newton-Raphson method**. Once the model parameters $\hat{\alpha}$ are obtained, the survival function in each treatment $\mathbb{S}_j(t)$ for $j = 0, 1$ is estimated as

$$\hat{\mathbb{S}}_j(t) = \exp\Big\{ - \int_0^t \mathbf{U}(\mathbf{x})^\top \hat{\alpha}_{\mathbf{j}} dx \Big\}.$$

2. Zeng's method:
Zeng et al. [2021] used the class of propensity score weighting estimators for survival outcomes based on the pseudo-observations. Furthermore, Zeng et al. [2021] defined a new closed-form variance estimator that takes into account the uncertainty due to both pseudo-observations calculation and propensity score estimation. The pseudo-observation is a leave-one-out jackknife approach to address right-censoring and provides a straightforward unbiased estimator of the function of uncensored data under the completely independent censoring assumption.

For a given time t, generally define $\theta^k(t) = \mathbb{E}\{\nu_k(Y_i; t)\}$ as a population parameter. The causal estimands of interest are based on two typical transformations of the potential survival times: (i) the at-risk function $\nu_1(Y_i; t) = 1\{Y_i \geq t\}$ and (ii) the truncation function $\nu_2(Y_i; t) = Y_i \wedge t$ where t is a given time point of interest. The pseudo-observation for each unit is written as

$$\hat{\theta}_i^k(t) = n\hat{\theta}^k(t) - (n-1)\hat{\theta}_{-i}^k(t)$$

where $\hat{\theta}^k(t)$ is the consistent estimator of $\theta^k(t)$, and $\hat{\theta}_{-i}^k(t)$ is the corresponding estimator with unit i left out. Regarding that, Zeng et al. [2021] proposed the following nonparametric Hajek-type estimator for the class of estimands:

$$\hat{\tau}_{j,j'}^{k,h}(t) = \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = j\}\hat{\theta}_i^k(t) w_j^h(\mathbf{X_i})}{\sum_{i=1}^n \mathbf{1}\{Z_i = j\} w_j^h(\mathbf{X_i})} - \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = j'\}\hat{\theta}_i^k(t) w_{j'}^h(\mathbf{X_i})}{\sum_{i=1}^n \mathbf{1}\{Z_i = j'\} w_{j'}^h(\mathbf{X_i})}, \quad j \neq j' \quad (5)$$

where the $w_j^h(X)$ for $j = 0, 1$ corresponds to the balancing weight $(\omega^{ipw}, \omega^{ow})$. For transformation ν_k with $k = 1, 2$, Zeng et al. [2021] consider the Kaplan-Meier estimator to construct $\theta^k(t)$. In detail, when the interest estimand lies in the survival functions (i.e. SPCE with $k = 1$),

$$\hat{\theta}_i^1(t) = n\hat{\mathbb{S}}(t) - (n-1)\hat{\mathbb{S}}_{-i}(t)$$

and when the interest estimand lies in the restricted mean survival times (i.e. RACE with $k = 2$),

$$\hat{\theta}_i^2(t) = n \int_0^t \hat{\mathbb{S}}(u) du - (n-1) \int_0^t \hat{\mathbb{S}}_{-i}(u) du$$

and then compute the estimand.

Estimator (5) can be extended to accommodate covariate-dependent censoring. In this case, one can consider inverse probability of censoring weighted pseudo-observation [Robins and Finkelstein, 2000, Binder et al., 2014]:

$$\hat{\theta}_i^k(t) = \frac{\nu_k(\tilde{Y}_i;t)\, 1\{C_i \geq \tilde{Y}_i \wedge t\}}{\hat{G}(\tilde{Y}_i \wedge t \mid X_i, Z_i)}$$

where $\hat{G}(u \mid X_i, Z_i)$ is a consistent estimator of the censoring survival function $\hat{G}(u \mid X_i, Z_i) = \Pr(C_i \geq u \mid X_i, Z_i)$, for example, given by the Cox proportional hazards regression.

In the presence of two different covariate-dependent censoring mechanisms, let's define

$$\hat{\theta}_i^k(t) = \frac{\nu_k(\tilde{Y}_i;t)\mathbf{1}\{\min(C_i, S_i) \geqslant \tilde{Y}_i \wedge t\}}{\hat{G}(\tilde{Y}_i \wedge t | X_i, Z_i)}$$

where $\hat{G}(u \mid X_i, Z_i) = \Pr(C_i \geqslant u \mid X_i, Z_i) \times \Pr(S_i \geqslant u \mid X_i, Z_i)$ computed by two Cox proportional hazards models. This estimator (referred to as the *extended Zeng's method*) generalizes those described in Zeng et al. [2021] to corporate two covariate-dependent censoring in survival settings.

With the increasing development of more advanced causal inference methods, it is important to be able to evaluate method performance in different scenarios and make comparisons between methods to guide their use in practice. The idea behind the next Section is that the plausibility of the estimated treatment effects will increase if the inferences are insensitive over a wide range of relevant scenarios. **Simulation studies** are a key tool for such investigations and can be used to assess properties such as bias, efficiency, and coverage of confidence intervals. The results help analysts to choose which methods are most appropriate for answering research questions using their data. As a result, in the next section, we assess the performance of the estimator under the different assumptions and test the sensitivities of the adjustment methods to changes in key scenario assumptions.

3 Simulation Study

To assess the sensitivity of key assumptions of the censoring mechanism, in this section, we conducted comprehensive simulation studies under the scenarios described in Table 2. Subsections 3.1 and 3.3 illustrate, respectively, the simulations design and the results. We review briefly methods to estimate survival function in Subsection 3.2. Especially, according to different censoring assumptions, we apply the Marginal Structural Cox Model to the different weighted samples.

Table 2: The overview of different scenarios in Simulation Studies

`Scenario 1`	Both types of censoring are non-informative/completely ignorable
`Scenario 2`	Censoring due to treatment switching is ignorable conditional on the covariates and administrative censoring is completely ignorable
`Scenario 3`	Both types of censoring are ignorable conditional on covariates

3.1 Simulation's Design

We generate four pre-treatment covariates: $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})^\top$ where

$$\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix} \right\},$$

$X_{i4} \sim \text{Bernoulli}(0.5)$ and $X_{i3} \sim \text{Bernoulli}(0.6X_{i4} + 0.4(1 - X_{i4}))$. We consider two treatment groups, with the true propensity score model given by $\text{logit}(\mathrm{e_i}) = \tilde{X}_i^\top \beta$ where $\tilde{X}_i = (1, X_i^\top)^\top$. Set $\beta = (-0.1\Psi, -0.9\Psi, -0.3\Psi, -0.1\Psi, -0.2\Psi)^\top$ where $\Psi = 1$ and $\Psi = 5$ represent good and poor overlap between groups, respectively. Distribution of the true propensity scores under each specification is presented in Figure 2. The model to generate potential survival times is Cox-Weibull model with hazard rate $h(t|X_i) = \lambda \nu t^{\nu-1} \exp\{L_i\}$ where

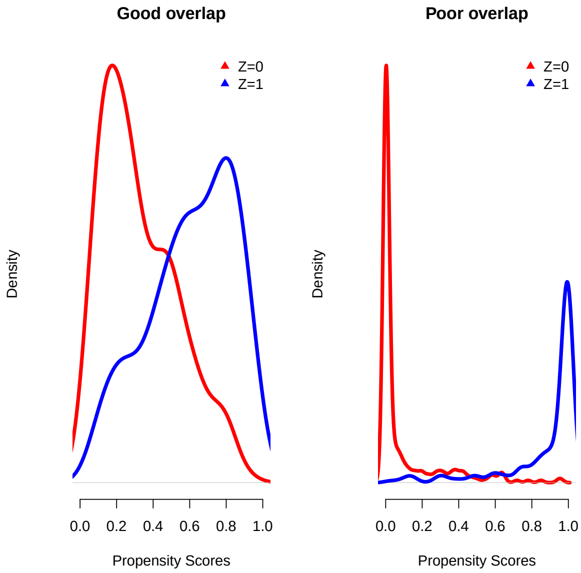
$$L_i = Z_i \gamma + X_{i1}\alpha_1 + X_{i2}\alpha_2 + X_{i3}\alpha_3 + X_{i4}\alpha_4.$$

We specify $(\gamma, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \lambda, \nu) = (1, 2, 1.5, -1, 1, 0.0001, 3)$. The potential survival time Y_i is then drawn using

$$Y_i = \{\frac{-\log(U)}{\lambda \exp(L_i)}\}^{\frac{1}{\nu}}$$

where $U \sim \text{Uniform}(0, 1)$. For each simulation setting, we simulated 500 Monte Carlo repetitions and averaged the results. While the IPW-based estimators focus on the combined population, the OW-based estimators focus on the overlap population [Zeng et al., 2021]. When comparing treatments, the true values of target estimands can be different between OW-based methods and IPW-based methods (albeit very similar under good overlap) and are computed via Monte Carlo integration. We vary the study sample size $n = (100, 300, 500)$ and fix the evaluation

Figure 2: The propensity score distribution under different overlap conditions across two treatments in the simulation studies



point $t = 10$ for estimating SPCE and RACE. We evaluate the performance of the models in terms of the absolute bias, root mean squared error (RMSE), and The 95% empirical coverage corresponding to each estimator. To obtain the empirical coverage for Marginal Structural Cox Models (with and without covariate), Bootstrap CIs are used. For Zeng's method, we used variance estimators suggested by Zeng et al. [2021]. For each scenario, we generate different administrative censoring and switch the treatment based on pre-specified assumptions. Table 3 prepared to show how to generate a time for censoring based on different scenarios. The true values of causal estimands were calculated from these simulated potential outcomes as weighted averages among the simulated subjects.

3.2 Methods to estimate survival function

The marginal structural cox models are powerful tools to control for both observed confounding and selection bias due to censoring. Thus, in this section, we review all necessary steps to use Marginal structural models in the simulation study to assess the performance of the estimator under different assumptions of the censoring mechanism. As an alternative, we consider Zeng's method to be able to compare the performance of different estimators.

Table 3: Generate time to censoring and time to switching the treatment based on different scenarios.

`Scenario 1` We generate administrative censoring and censoring due to treatment switching as

$$\{C_i, S_i\} \sim \text{Uniform}(0, K), \;\; K \in \mathbb{N}$$

where different values to K make different percentage of censoring; e.g. $K = 60$ makes censoring rates as %50 and $K = 220$ makes censoring rates as %25

`Scenario 2` To generate administrative censoring, we consider $C_i \sim \text{Uniform}(0, K'), \;\; K' \in \mathbb{N}$ and under covariate-dependent censoring due to treatment switching, S_i generate from a Weibull survival model with hazard rate

$$h_S(t|\mathbf{X_i}) = \lambda_S \nu_S t^{\nu_S - 1} \exp\left\{\mathbf{X_i}^\top \alpha_{\mathbf{S}} + \mathbf{Z_i}\gamma\right\}$$

The parameters are specified so that censoring rate varies. Specifically, for censoring rate 50%, we set $K' = 1000$, $\alpha_{\mathbf{S}} = (1, 0.5, -0.5, 1)^\top$, $\lambda_S = 0.0001$ and $\nu_S = 3$ and for censoring rate 25%, we set $K' = 50$, $\alpha_{\mathbf{S}} = (-11.5, 3.5, -15.5, -11.5)^\top$, $\lambda_S = 0.0001$ and $\nu_S = 3$. Also, we set $\gamma = 1$ in both cases.

`Scenario 3` Under covariate-dependent censoring due to both censoring mechanisms, generate administrative censoring (C_i) from a Weibull survival model with hazard rate

$$h_c(t|\mathbf{X_i}) = \lambda_c \nu_c t^{\nu_c - 1} \exp\{\mathbf{X_i}^\top \alpha_{\mathbf{c}} + \mathbf{Z_i}\gamma\}$$

and S_i is generated from a Weibull survival model with hazard rate

$$h_s(t|\mathbf{X_i}) = \lambda_S \nu_S t^{\nu_S - 1} \exp\{\mathbf{X_i}^\top \alpha_{\mathbf{S}} + \mathbf{Z_i}\gamma\}.$$

Different values of λ_c, λ_s, ν_c, ν_s and $\alpha_{\mathbf{s}}, \alpha_{\mathbf{c}}$ are specified so that the censoring rate varies. In detail, for censoring rate 50%, we set $\alpha_c = (0.5, 0.25, -0.25, 0.75)$ and $\alpha_s = (1, 0.5, -0.5, 1)$. Also, $\lambda_c = \lambda_s = 0.0001$ and $\nu_c = \nu_s = 3$. For censoring rate 25%: we set $\alpha_c = (-10, 0.25, -15, -12)$ and $\alpha_s = (-11.5, 3.5, -15.5, -11.5)$. Also, $\lambda_c = \lambda_s = 0.0001$ and $\nu_c = \nu_s = 3$. We consider $\gamma = 1$ in both cases.

The Marginal Structural Cox Model without covariates:

- In the present of the observed confounding, it is necessary to compute weights of treatment. In practice, fit a logistic regression model for the PS to compute weights $(\omega_i^{ipw}, \omega_i^{ow})$.
- Assign weights for each unit as

1. Under the completely ignorable/ completely independent assumption of both censoring mechanisms due to administrative and switching the treatment, in **Scenario 1**, they are $\hat{\omega}_i = \omega_i^{ipw}$ or $\hat{\omega}_i = \omega_i^{ow}$.

2. In **Scenario 2**, due to covariate-dependent of switching censoring, we compute $\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \qquad \text{or} \qquad \hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss}$

3. In the present of the oberved confounding and dependent censoring due to treatment switching and administrative censoring in **Scenario 3**, compute $\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \times \omega_i^{sc} \qquad \text{or} \qquad \hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss} \times \omega_i^{sc}$

- Fit a Cox proportional hazard model with a hazard rate

$$h(t|Z_i) = h_0(t) \exp\left\{\gamma Z_i\right\}$$

- Calculate the survival probability $\hat{\mathbb{S}}(t|Z_i)$ specific to each treatment.

The Marginal Structural Cox Model with covariates:

To assess the performance of the estimators under different assumptions, in simulation studies, we also, consider another form of Marginal Structural Cox Model with a hazard rate $h(t|X_i, Z_i)$ as

$$h(t|X_i, Z_i) = h_0(t) \exp\left\{X_i \alpha^\top + \gamma Z_i\right\} \tag{6}$$

We assumed in **Scenario 1** that both censoring mechanisms (i.e. administrative and switching the treatment) are completely ignorable/ completely independent. As a result, in order using the Marginal Structural Cox Model with covariates (`COX.MSM.COV.IPW`, `COX.MSM.COV.OW`), one should compute ω_i^{ipw} and ω_i^{ow} at first step. Then, for each unit, a weight is assigned, such as

$$\hat{\omega}_i = \omega_i^{ipw}$$

and

$$\hat{\omega}_i = \omega_i^{ow}.$$

Next, the Cox proportional hazard model with the hazard rate $h(t|X_i, Z_i)$ (Equation 6) is fitted and the conditional survival probability function $\hat{\mathbb{S}}(t|X_i, Z_i)$ is computed to estimate interested estimands.

To adjust for observed confounders and dependent censoring due to treatment switching based on **Scenario 2** (`COX2.MSM.COV.IPW`, `COX2.MSM.COV.OW`), we estimate propensity scores by fitting logistic regression to construct weights $(\omega_i^{ipw}, \omega_i^{ow})$. Then, due to covariate-dependent of switching censoring, stabilized IPCW (ω_i^{ss}) is calculated. After that, $\hat{\omega}_i$ is assigned to each unit as

$$\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss}$$

and

$$\hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss}.$$

Finally, a Cox proportional hazard model with a hazard rate $h(t|X_i, Z_i)$ (Equation 6) is fitted to calculate the survival probability $\hat{\mathbb{S}}(t|X_i, Z_i)$.

The **Scenario 3** is based on the covariate-dependent assumptions of both censoring mechanisms due to administrative and switching the treatment. To accommodate Marginal Structural Cox Model with covariates in this Scenario (`COX3.MSM.COV.IPW`, `COX3.MSM.COV.OW`), first, we fit a logistic regression model to estimate propensity scores computing weights $(\omega_i^{ipw}, \omega_i^{ow})$. Then, the stabilized IPCW weights for both administrative and switching censoring (ω_i^{sc} and ω_i^{ss}) are calculated. After that, weights $\hat{\omega}_i$ is assigned to each unit as

$$\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \times \omega_i^{sc}$$

and

$$\hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss} \times \omega_i^{sc}.$$

Finally, we fit the Cox proportional hazard model with the hazard rate $h(t|X_i, Z_i)$ (Equation 6). Based on the estimated hazard rate, we can calculate the conditional survival probability function $\hat{\mathbb{S}}(t|X_i, Z_i)$ and then compute estimands.

Zeng's method:

Zeng et al. [2021] proposed the nonparametric Hajek-type estimator (as mentioned in equation 5) as

$$\hat{\tau}_{j,j'}^{k,h}(t) = \frac{\sum_{i=1}^{n} \mathbf{1}\{Z_i = j\}\hat{\theta}_i^k(t) w_j^h(\mathbf{X_i})}{\sum_{i=1}^{n} \mathbf{1}\{Z_i = j\} w_j^h(\mathbf{X_i})} - \frac{\sum_{i=1}^{n} \mathbf{1}\{Z_i = j'\}\hat{\theta}_i^k(t) w_{j'}^h(\mathbf{X_i})}{\sum_{i=1}^{n} \mathbf{1}\{Z_i = j'\} w_{j'}^h(\mathbf{X_i})}, \quad j \neq j'$$

where the $w_j^h(X)$ for $j = 0, 1$ corresponds to the balancing weight $(\omega^{ipw}, \omega^{ow})$. According to the key assumption of **Scenario 1** (noted as `Zeng.IPW` and `Zeng.OW`), this approach will be used as an alternative method.

To accommodate covariate-dependent censoring (based on **Scenario 2**), estimator ($\hat{\tau}^{k,h}_{j,j'}(t)$) can be extended by considering inverse probability of censoring weighted pseudo-observation as [Robins and Finkelstein, 2000, Binder et al., 2014]:

$$\hat{\theta}^k_i(t) = \frac{\nu_k(\tilde{Y}_i; t)\, 1\{C_i \geq \tilde{Y}_i \wedge t\}}{\hat{G}(\tilde{Y}_i \wedge t \mid X_i, Z_i)}$$

where $\hat{G}(u \mid X_i, Z_i)$ is computed by the Cox proportional hazards regression (ω^{ss}_i). `Zeng2.IPW` and `Zeng2.OW` are the Zeng's approach referring to the assumptions of **Scenario 2**.

In the presence of two different types of covariate-dependent censoring (**Scenario 3**), the $\hat{\theta}^k_i(t)$ is developed as

$$\hat{\theta}^k_i(t) = \frac{\nu_k(\tilde{Y}_i; t)\mathbf{1}\{\min(C_i, S_i) \geqslant \tilde{Y}_i \wedge t\}}{\hat{G}(\tilde{Y}_i \wedge t|X_i, Z_i)}$$

where $\hat{G}(u \mid X_i, Z_i) = \Pr(C_i \geqslant u \mid X_i, Z_i) \times \Pr(S_i \geqslant u \mid X_i, Z_i)$ computed by two Cox proportional hazards models (i.e. ω^{sc}_i and ω^{ss}_i). `Zeng3.IPW` and `Zeng3.OW` are the *extended* Zeng's approach referring to the assumptions of **Scenario 3**.

3.3 Simulation Results

3.3.1 Scenario 1:

We analyze a simulated data set using the Marginal Structural Cox Model *without* covariates, Zeng's method, and the Marginal Structural Cox Model *with* covariates. On a deeper level, we evaluate how well the estimators under different assumptions perform on the generated data set. In order to accomplish this, it is essential to bear in mind the following: under the assumption of both types of censoring are completely ignorable, we denote the estimators of the Marginal Structural Cox Model *without* covariates as (`COX.MSM.IPW`, `COX.MSM.OW`), the estimators of the Marginal Structural Cox Model *with* covariates as (`COX.MSM.COV.IPW`, `COX.MSM.COV.OW`) and Zeng's methods as (`Zeng.IPW`, `Zeng.OW`). Furthermore, the Marginal Structural Cox Model *without* covariates (denoted as `COX2.MSM.IPW`, `COX2.MSM.OW`), the Marginal Structural Cox Model *with* covariates (denoted as `COX2.MSM.COV.IPW`, `COX2.MSM.COV.OW`) and Zeng's methods (denoted as `Zeng2.IPW`, `Zeng2.OW`) were calculated under assumption that switching censoring is ignorable conditional on the covariates and administrative censoring is completely ignorable. Moreover, under assumption of both types of censoring are ignorable conditional on covariate, the estimators of the Marginal Structural Cox Model *without* covariates (denoted as `COX3.MSM.IPW`, `COX3.MSM.OW`), the Marginal Structural Cox

Model *with* covariates (denoted as `COX3.MSM.COV.IPW`, `COX3.MSM.COV.OW`) and Zeng's methods (denoted as `Zeng3.IPW`, `Zeng3.OW`) were computed.

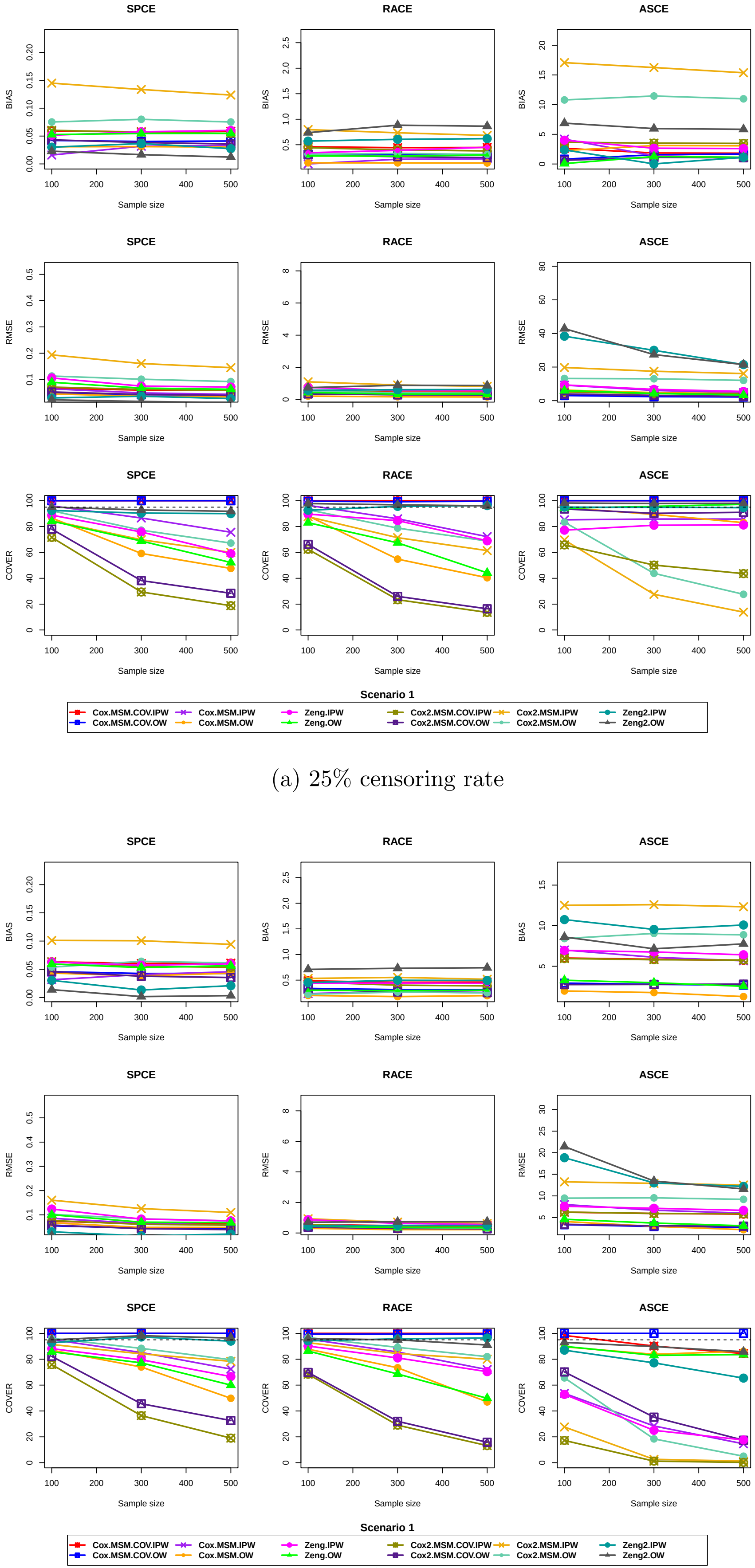
The absolute bias, root mean square error, and coverage probability of the 95 percent confidence interval for the OW and IPW estimators are presented in Table 4 based on the assumptions that both types of censoring are completely ignorable. In general, OW estimators outperform others across SPCE and RACE with a reduced absolute bias and RMSE, and they get closer to nominal coverage. Under poor overlap, the IPW estimator leads to larger bias, variance, and low coverage. The absolute bias and RMSE are reduced as the sample size increases.

The figures 3–6 display the absolute bias, RMSE, and coverage probability of the 95% confidence interval for the OW-based and IPW-based estimators. Across all three estimands (SPCE, RACE, and ASCE), estimators computed under the assumption of completely ignorable both types of censoring outperform other estimators with a lower absolute bias and RMSE. In addition, under good overlap, in the absence of non-informative both types of censoring assumption, estimators computed by Marginal Structural Cox Model *with* covariates have a lower 95% coverage probability, whereas estimators computed by Marginal Structural Cox Model *without* covariates and Zeng's method report more bias, and RMSE. Regarding the performance of Zeng's method when estimating ASCE, `Zeng.OW` is closer to nominal coverage. Under poor overlap, the OW is more robust than the IPW. This is because the IPW estimator is susceptible to the lack of overlap due to extreme propensity scores. As predicted, `COX.MSM.COV.OW` outperforms all other estimators regardless of the degree of overlap and censoring rate.

3.3.2 Scenario 2:

Figures 7-10 illustrates the comparison of different estimators in the generated data based on varying degrees of overlap between two treatment arms. We run the Marginal Structural Cox Model *without* covariates, the Marginal Structural Cox Model *with* covariates, and Zeng's methods under different assumptions. Moreover, Table 5 depicts the performance of the estimators in the presence of key assumptions of Scenario 2.

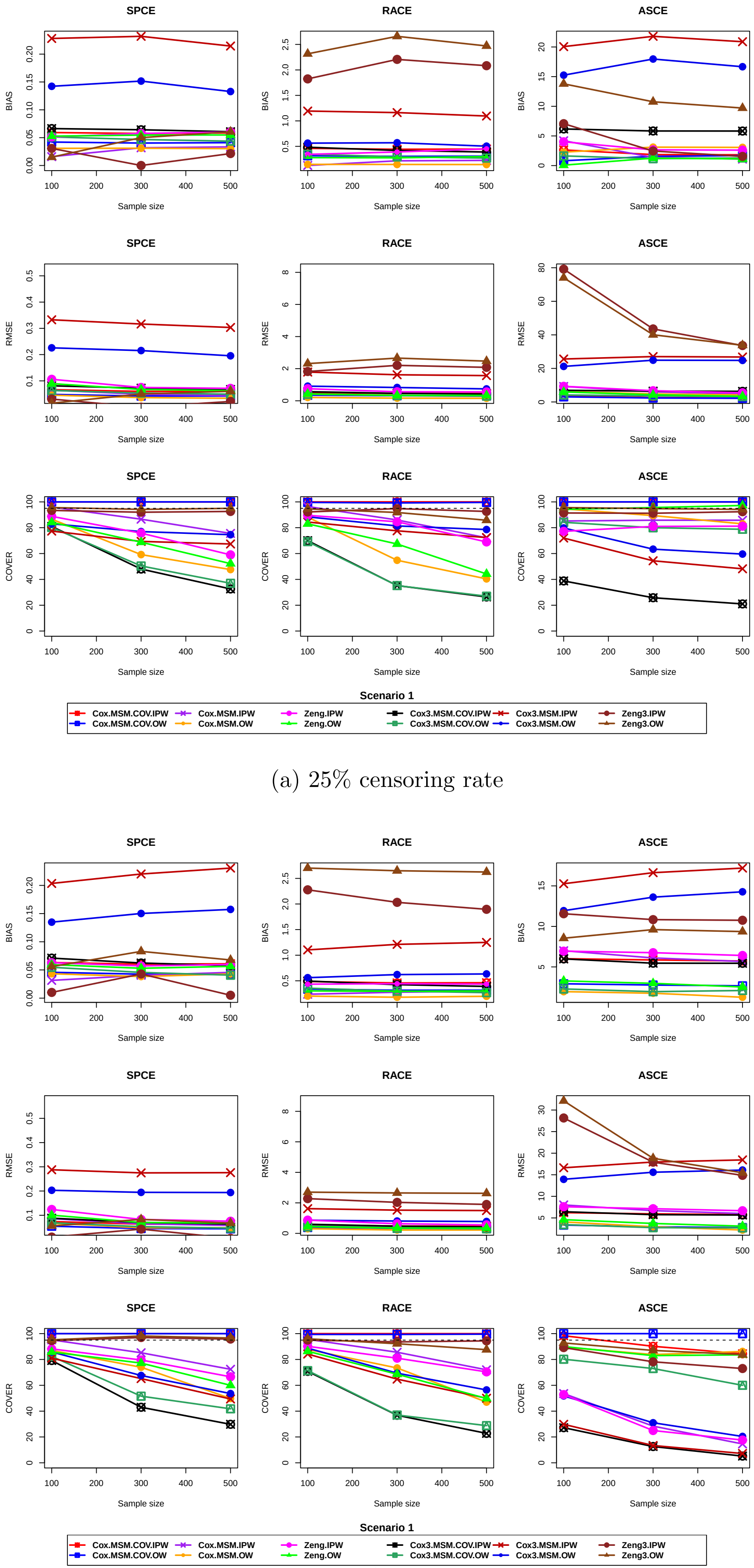
Under good overlap, `Cox2.MSM.COV.OW-Cox2.MSM.COV.IPW` outperforms Zeng's methods even though `Zeng2.OW` and `Zeng2.IPW` are constructed based on covariate-dependent switching censoring assumption. This is probably because the data on hand are generated based on proportionality assumption. Besides, the OW is more robust under poor overlap than the IPW. When the censoring rate is small (i.e., 25%), marginal structural models with covariates (`Cox2.MSM.COV.OW`, `Cox2.MSM.COV.IPW`) achieve lower bias and RMSE compared with other estimators in most cases as the outcome model is correctly specified. By increasing the



(a) 25% censoring rate

(b) 50% censoring rate

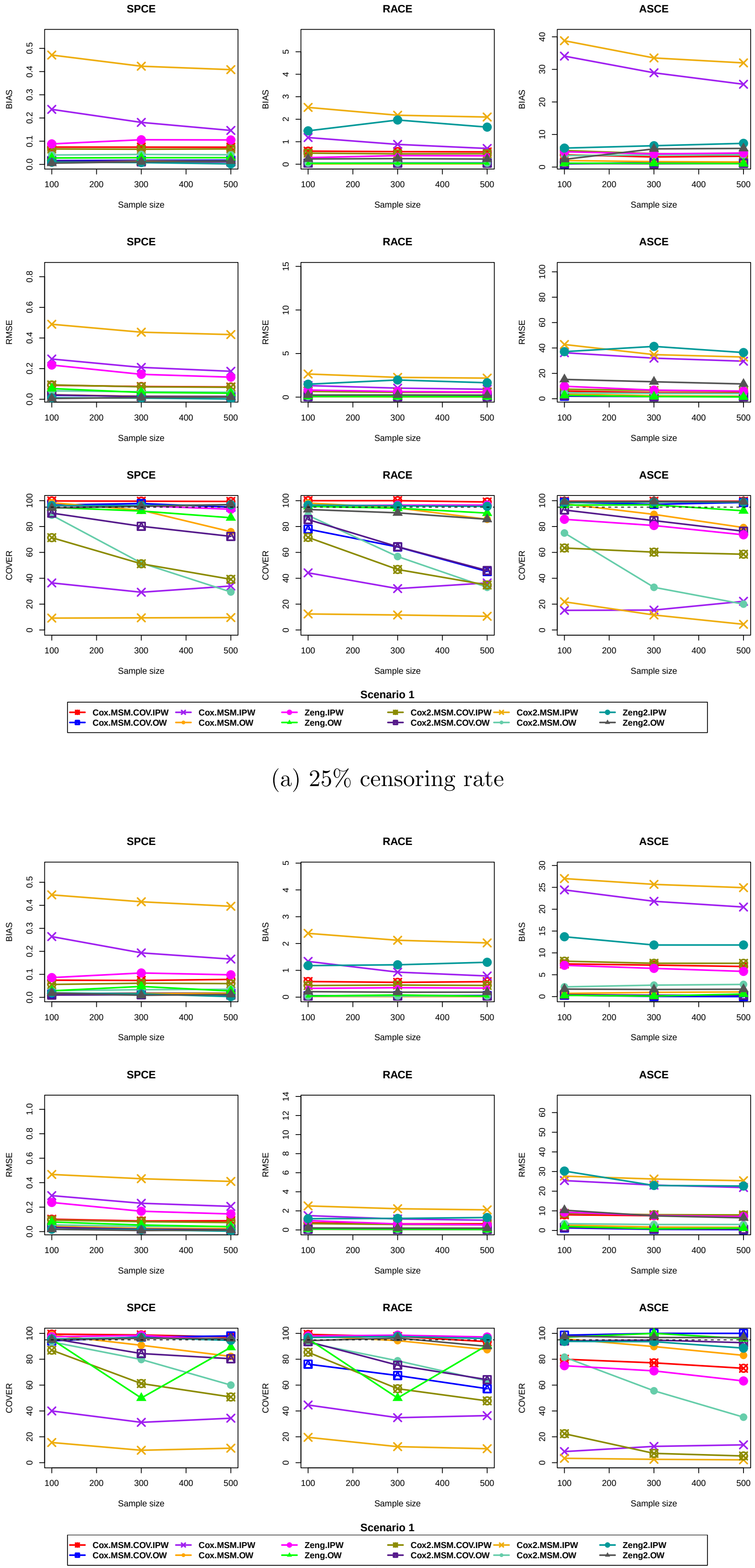
Figure 3: **Scenario 1**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatments under *good overlap* for Scenario 1 and Scenario 2.



(a) 25% censoring rate

(b) 50% censoring rate

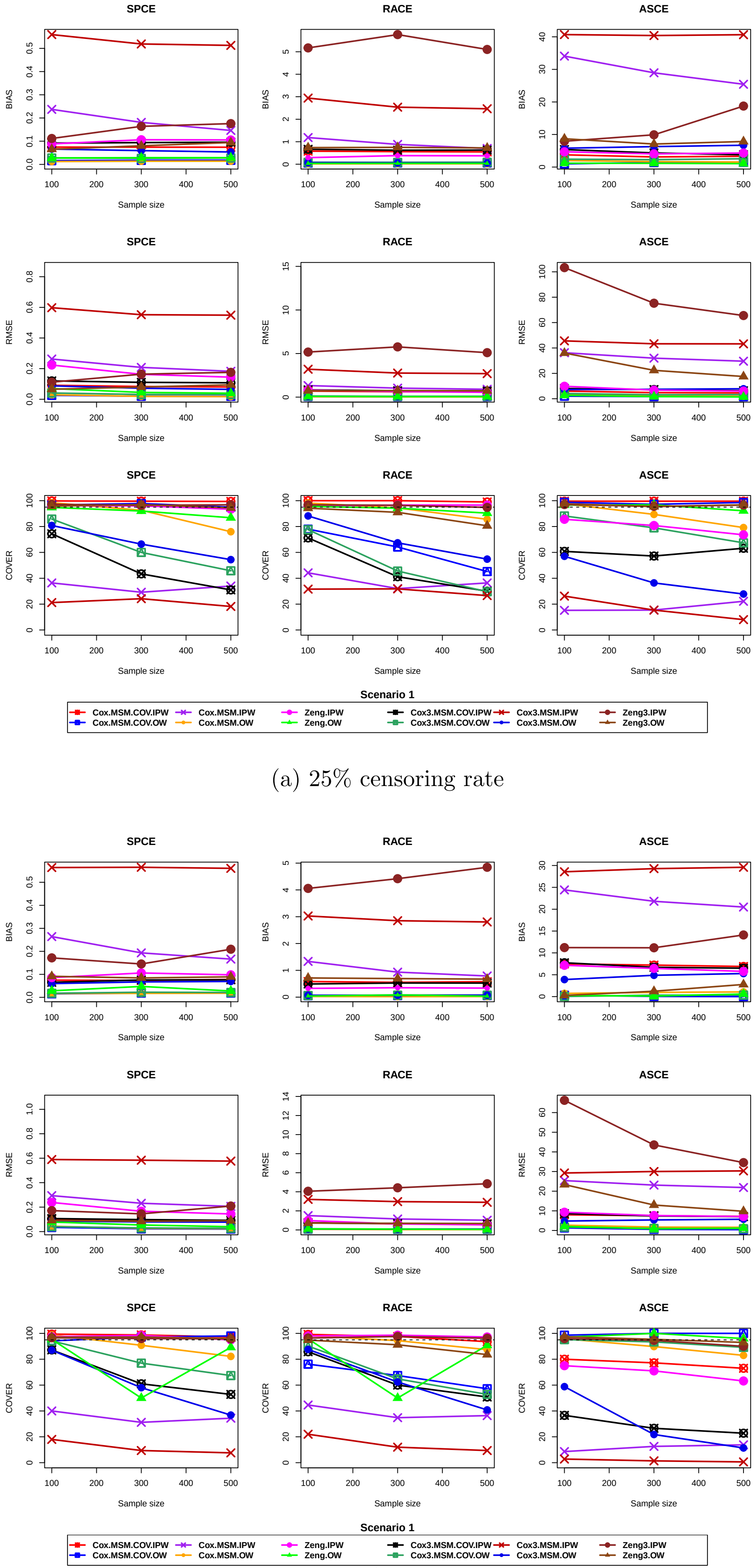
Figure 4: **Scenario 1**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatments under *good overlap* for Scenario 1 and Scenario 3.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 5: **Scenario 1**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 1 and Scenario 2.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 6: **Scenario 1**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 1 and Scenario 3.

Table 4: **Scenario 1**: Absolute bias, root mean squared error (RMSE) and coverage for comparing two treatments under different degrees of overlap (good, poor), different sample size (100, 300, 500) and various censoring rate (25%, 50%) when both types of censoring are completely independent

		MSM+COV						MSM						Zeng					
	sample	Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage	
	size	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW
Good overlap-25% censoring rate																			
RACE	100	0.46	0.32	0.50	0.35	100.00	99.60	0.12	0.15	0.40	0.21	96.40	88.20	0.34	0.28	0.75	0.43	89.60	83.00
	300	0.44	0.30	0.46	0.31	100.00	99.20	0.22	0.14	0.32	0.17	86.00	54.80	0.39	0.28	0.56	0.33	84.60	67.40
	500	0.45	0.31	0.46	0.31	100.00	99.60	0.22	0.14	0.28	0.16	72.20	40.40	0.45	0.30	0.55	0.33	69.00	44.20
Good overlap-50% censoring rate																			
RACE	100	0.49	0.34	0.56	0.39	100.00	99.40	0.24	0.20	0.56	0.30	95.60	87.80	0.43	0.30	0.87	0.46	90.20	86.37
	300	0.46	0.32	0.48	0.33	100.00	99.40	0.28	0.18	0.43	0.22	85.60	73.40	0.44	0.29	0.63	0.36	81.00	68.60
	500	0.46	0.32	0.47	0.32	100.00	99.60	0.30	0.20	0.37	0.22	72.00	47.00	0.43	0.30	0.56	0.35	70.40	49.80
Poor overlap-25% censoring rate																			
RACE	100	0.58	0.05	0.70	0.07	100.00	78.00	1.19	0.01	1.33	0.04	44.20	98.20	0.29	0.04	0.81	0.10	96.00	95.60
	300	0.56	0.06	0.62	0.06	100.00	64.20	0.88	0.02	1.03	0.02	32.00	94.40	0.38	0.04	0.65	0.06	96.60	94.20
	500	0.55	0.06	0.59	0.06	99.00	45.20	0.70	0.02	0.90	0.02	36.40	85.60	0.37	0.04	0.55	0.05	96.60	90.40
Poor overlap-50% censoring rate																			
RACE	100	0.58	0.05	0.75	0.08	99.20	76.20	1.33	0.02	1.50	0.06	44.60	98.20	0.33	0.04	0.98	0.11	98.00	95.80
	300	0.55	0.06	0.64	0.07	97.60	67.40	0.93	0.02	1.15	0.03	34.80	94.40	0.35	0.08	0.63	0.08	98.60	50.00
	500	0.58	0.06	0.65	0.07	93.60	57.20	0.79	0.02	1.02	0.03	36.40	87.40	0.34	0.04	0.53	0.05	97.20	90.80
Good overlap-25% censoring rate																			
SPCE	100	0.06	0.04	0.07	0.05	100.00	100.00	0.02	0.03	0.06	0.05	96.00	86.60	0.05	0.05	0.11	0.09	88.80	84.20
	300	0.06	0.04	0.06	0.04	100.00	100.00	0.03	0.03	0.05	0.04	86.60	59.20	0.06	0.05	0.07	0.07	75.80	68.80
	500	0.06	0.04	0.06	0.04	100.00	100.00	0.03	0.03	0.04	0.03	75.60	47.60	0.06	0.05	0.07	0.06	59.00	52.20
Good overlap-50% censoring rate																			
SPCE	100	0.06	0.05	0.07	0.05	100.00	100.00	0.03	0.04	0.09	0.07	95.20	87.00	0.06	0.06	0.12	0.10	88.20	85.77
	300	0.06	0.04	0.06	0.05	100.00	100.00	0.04	0.04	0.06	0.05	85.20	74.00	0.06	0.05	0.08	0.07	79.60	77.20
	500	0.06	0.04	0.06	0.04	100.00	100.00	0.05	0.04	0.06	0.05	72.40	49.80	0.06	0.06	0.08	0.07	66.60	60.00
Poor overlap-25% censoring rate																			
SPCE	100	0.07	0.02	0.09	0.03	99.80	96.60	0.24	0.01	0.26	0.03	36.40	98.40	0.09	0.03	0.22	0.07	96.20	94.80
	300	0.07	0.02	0.08	0.02	99.60	97.80	0.18	0.01	0.21	0.02	29.20	92.60	0.11	0.03	0.16	0.04	96.00	92.00
	500	0.07	0.02	0.08	0.02	99.40	95.00	0.15	0.01	0.18	0.02	34.00	76.00	0.11	0.03	0.14	0.04	93.40	86.80
Poor overlap-50% censoring rate																			
SPCE	100	0.07	0.02	0.10	0.03	99.40	94.20	0.26	0.02	0.29	0.04	40.00	98.80	0.09	0.03	0.24	0.08	97.40	95.40
	300	0.07	0.02	0.09	0.02	98.80	97.20	0.19	0.02	0.23	0.03	31.20	90.80	0.11	0.05	0.17	0.05	98.40	50.00
	500	0.08	0.02	0.09	0.02	97.20	98.00	0.17	0.02	0.21	0.02	34.40	82.20	0.10	0.03	0.14	0.04	95.00	89.20

censoring rate, `Zeng2.OW` performs better for estimating ASCE. The same pattern appears in Table 5. Consequently, in terms of SPCE and RACE, OW consistently outperforms IPW with a decreased absolute bias and RMSE, as well as coverage closer to nominal at all sample size levels.

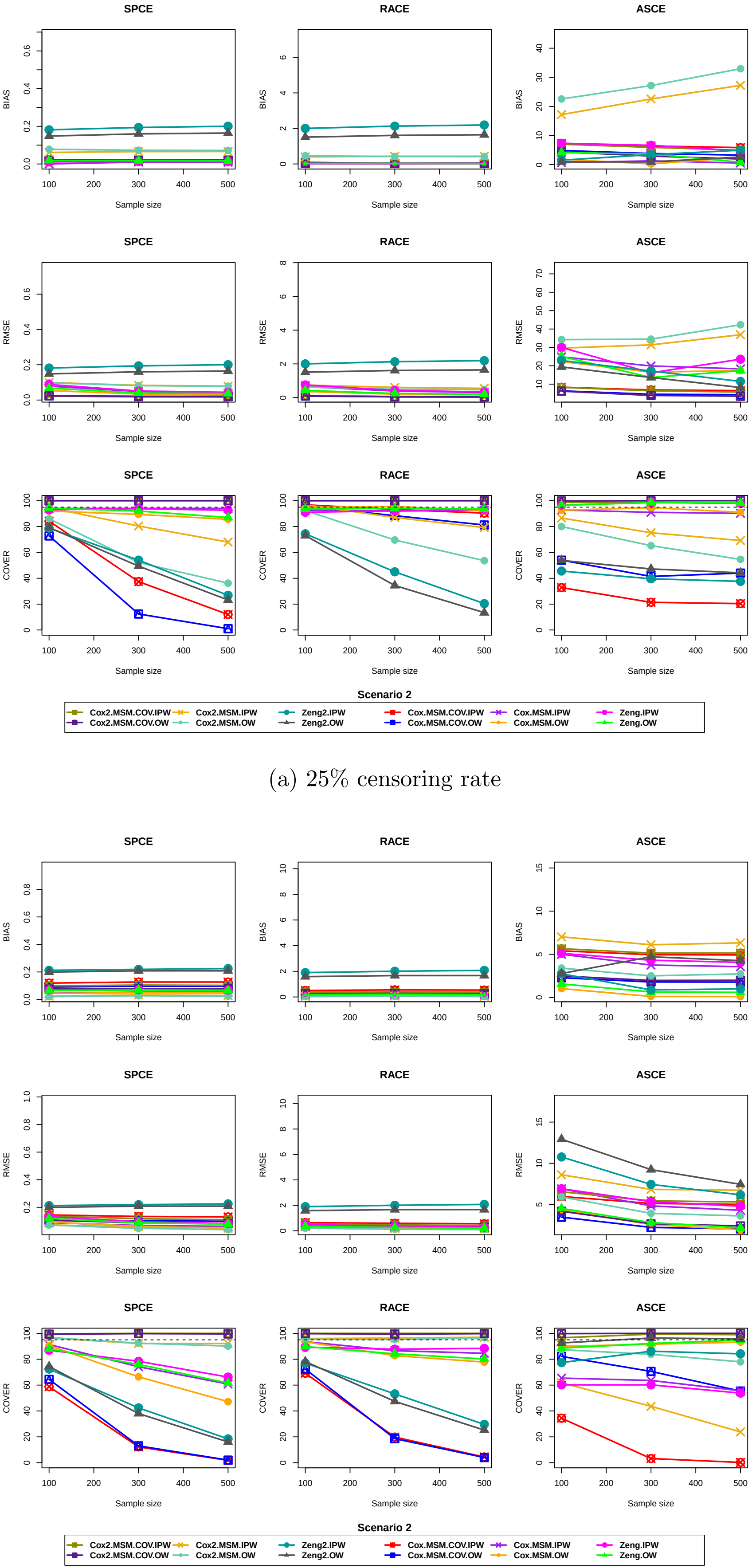
Under the assumption that censoring due to treatment switching is ignorable conditional on the covariates and administrative censoring is completely ignorable, `Cox2.MSM.COV.IPW` and `Cox2.MSM.COV.OW` maintains the smallest bias, largest efficiency, and closest to nominal coverage across all types of estimands compared to all other estimators. All in all, the MSMs with covariates are more efficient than other estimators regardless of the degree of overlap or the value of censoring rates because survival outcomes are correctly specified.

3.3.3 Scenario 3:

Under covariate-dependent censoring due to both administrative and treatment switching, in this thesis, two censoring weights are considered. The final weight (i.e., $\hat{\omega}$) is accomplished by multiplying the treatment and two censoring weights. Using the Marginal Structural Cox Model, one can estimate all estimands (ASCE, RACE, and SPCE). Furthermore, the non-parametric Zeng's methods considering both types of covariate-dependent censoring extended to adjust selection bias due to the covariate-dependent censoring.

Figures 11–14 depict the absolute bias, RMSE, and coverage probability of the 95 percent confidence interval for the OW-based and IPW-based estimators, where both types of censoring are covariate-dependent. Although MSMs that include only treatment's weight (i.e. `Cox.MSM.COV.IPW` and `Cox.MSM.COV.OW`) and MSMs that consider treatment and switching weight (i.e. `Cox2.MSM.COV.IPW` and `Cox2.MSM.COV.OW`) have been developed depending on different censoring assumptions, we use them to compare the performance of the all estimators.

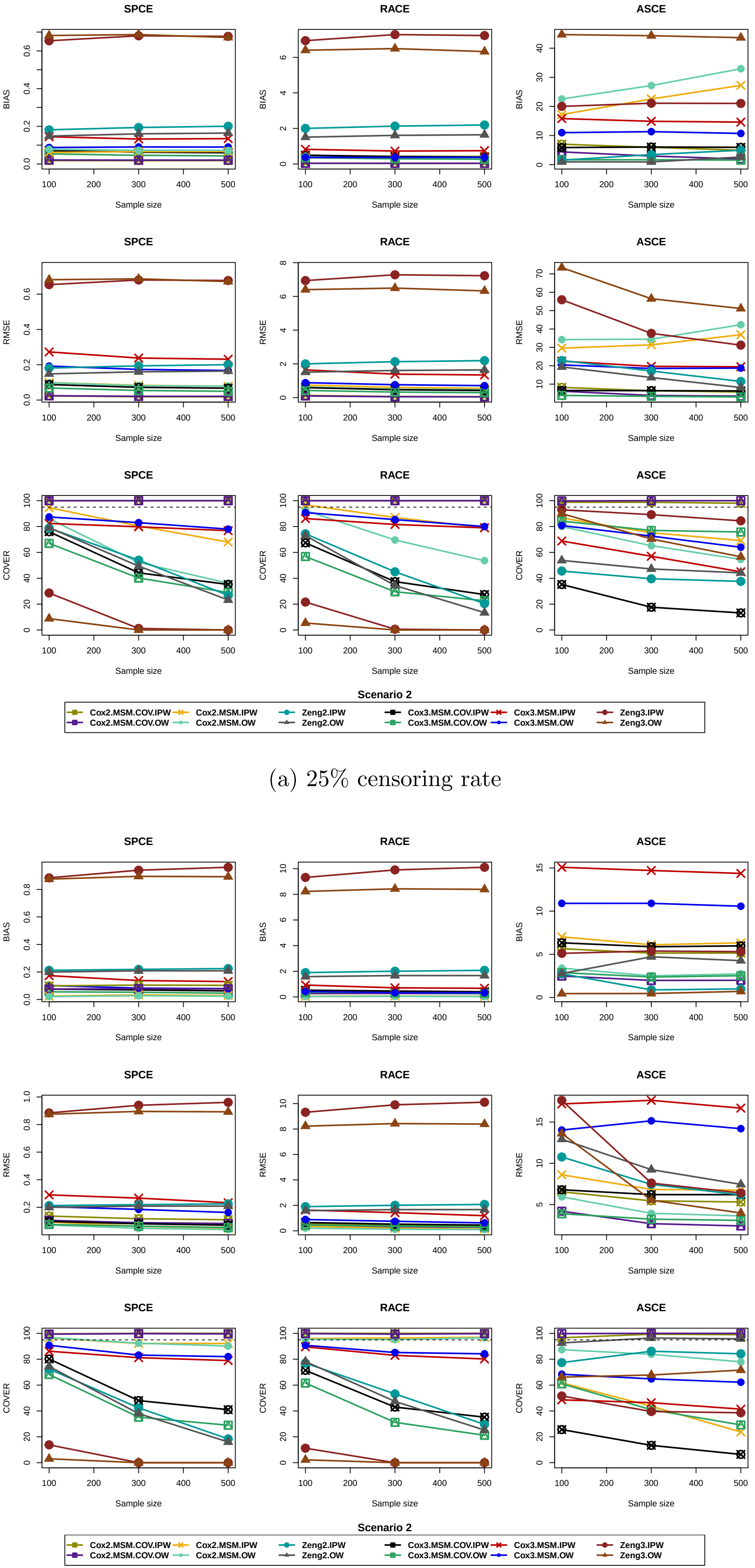
As expected, `Cox3.MSM.COV.OW` across all three estimands (SPCE, RACE, and ASCE), consistently outperforms alternative estimators with less absolute bias and RMSE, as well as coverage that is closer to the nominal value, regardless of the degree of overlap. This is because `Cox3.MSM.COV.OW` computes employing two censoring weights, and also because all models are correctly specified. While the empirical coverage of IPW-based approaches declines in the absence of overlap, the empirical coverage of OW-based methods remains robust. Although extended Zeng's method considers two censoring weights, it results in a higher absolute bias, RMSE for estimating ASCE and RACE when the censoring rate is small. In contrast, by increasing sample size and censoring rates (from 25 percent to 50 percent), the performance of `Zeng3.OW` in estimating ASCE improves, whereas it is not efficient in estimating RACE. Compared to the proposed weighted estimators



(a) 25% censoring rate

(b) 50% censoring rate

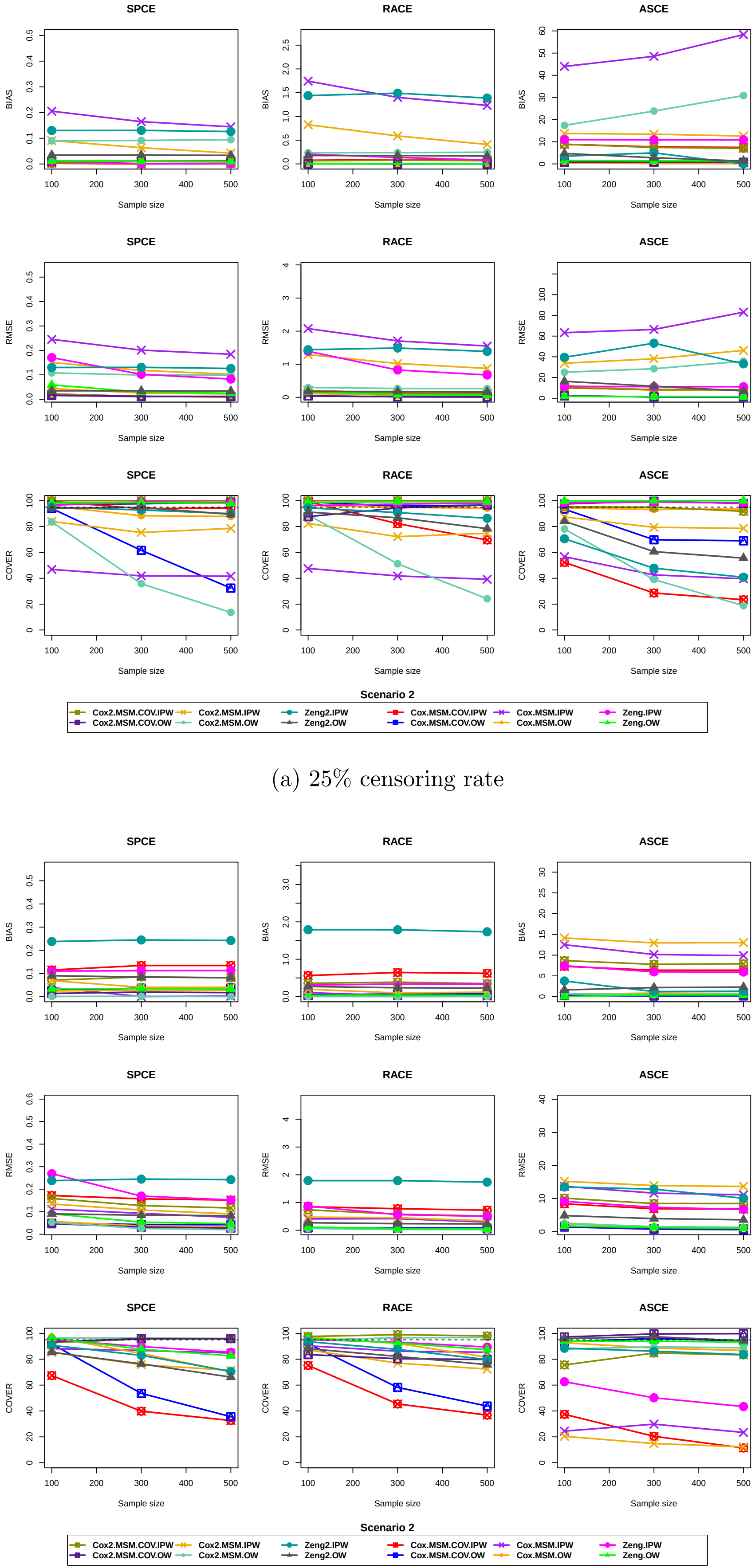
Figure 7: **Scenario 2**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatments under *good overlap* for Scenario 2 and Scenario 1.



(a) 25% censoring rate

(b) 50% censoring rate

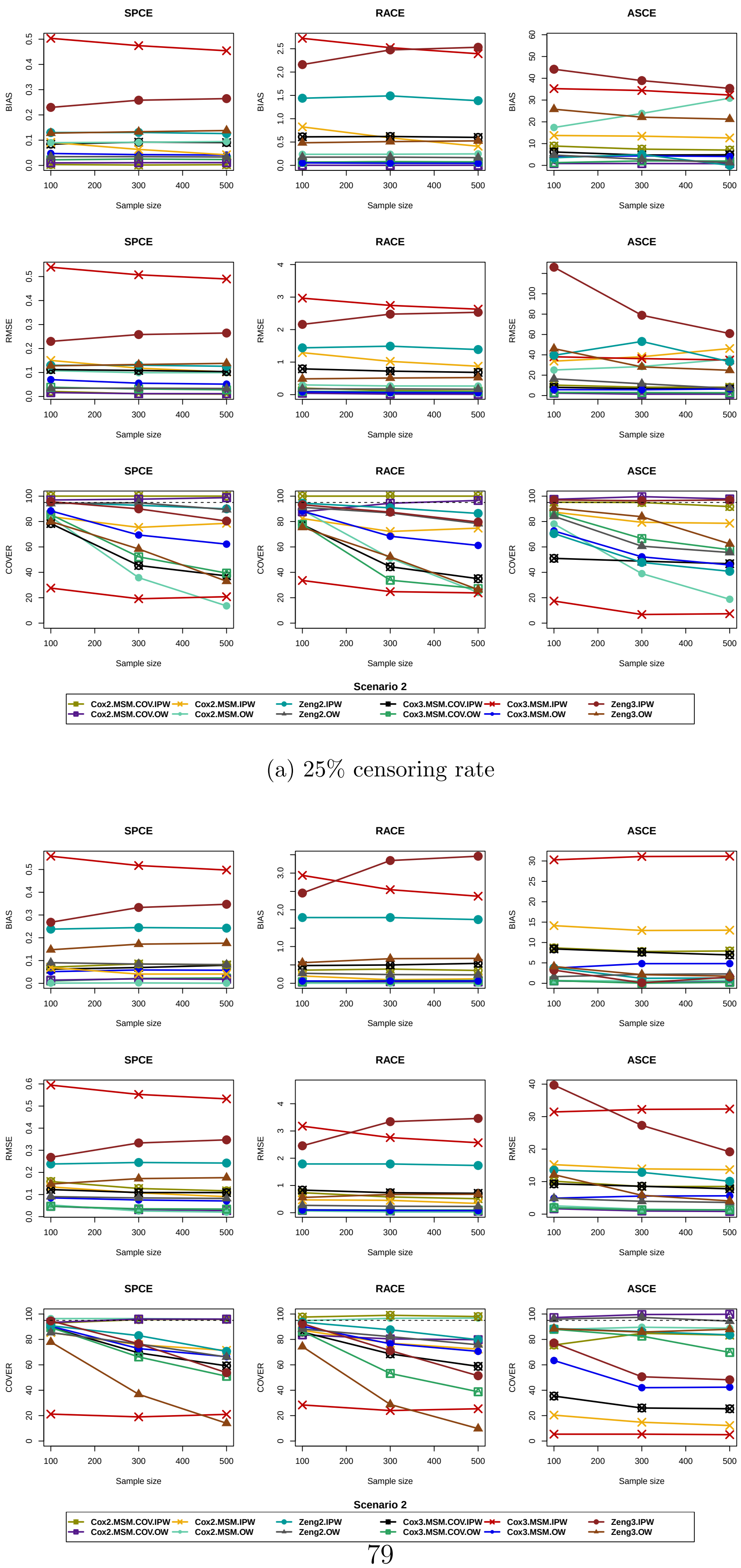
Figure 8: **Scenario 2**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatments under *good overlap* for Scenario 2 and Scenario 3.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 9: **Scenario 2**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 2 and Scenario 1.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 10: **Scenario 2**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 2 and Scenario 3.

Table 5: **Scenario 2**: Absolute bias, root mean squared error (RMSE) and coverage for comparing two treatments under different degrees of overlap (good, poor), different sample size (100, 300, 500) and various censoring rate (25%, 50%) when switching censoring is covariate-dependent and administrative censoring is completely independent

		MSM+COV						MSM						Zeng					
	sample	Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage	
	size	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW
Good overlap-25% censoring rate																			
RACE	100	0.03	0.04	0.13	0.10	100.00	100.00	0.40	0.46	0.75	0.63	96.80	92.40	2.00	1.51	2.57	1.86	74.40	73.00
	300	0.03	0.03	0.06	0.05	100.00	100.00	0.43	0.42	0.61	0.50	87.00	69.60	2.13	1.61	2.35	1.74	45.00	34.40
	500	0.03	0.04	0.05	0.05	100.00	100.00	0.43	0.41	0.55	0.47	79.20	53.60	2.19	1.64	2.33	1.73	20.40	13.40
Good overlap-50% censoring rate																			
RACE	100	0.38	0.26	0.56	0.39	100.00	99.80	0.05	0.03	0.36	0.21	96.00	95.40	1.90	1.58	2.74	2.20	77.00	78.20
	300	0.40	0.27	0.47	0.32	100.00	99.40	0.07	0.05	0.20	0.12	96.40	95.20	2.00	1.67	2.29	1.86	53.20	47.40
	500	0.38	0.26	0.41	0.28	100.00	99.80	0.05	0.03	0.15	0.09	97.00	96.60	2.07	1.67	2.24	1.78	29.60	25.20
Poor overlap-25% censoring rate																			
RACE	100	0.07	0.00	0.20	0.04	100.00	87.66	0.82	0.24	1.29	0.30	82.37	88.92	1.44	0.18	2.97	0.37	94.46	91.18
	300	0.08	0.00	0.13	0.02	100.00	94.40	0.59	0.24	1.02	0.27	72.20	51.20	1.49	0.18	2.20	0.24	90.80	86.80
	500	0.07	0.00	0.11	0.01	100.00	96.60	0.41	0.25	0.87	0.26	74.80	24.20	1.39	0.17	1.86	0.21	86.40	78.40
Poor overlap-50% censoring rate																			
RACE	100	0.36	0.03	0.73	0.09	97.60	83.60	0.20	0.00	0.47	0.07	87.20	94.80	1.79	0.27	4.29	0.61	93.60	88.18
	300	0.39	0.04	0.58	0.06	99.00	80.40	0.10	0.00	0.45	0.03	77.00	96.60	1.79	0.24	2.77	0.35	87.60	82.20
	500	0.35	0.04	0.51	0.05	98.00	79.80	0.12	0.00	0.34	0.02	72.40	96.80	1.73	0.23	2.44	0.30	79.80	75.80
Good overlap-25% censoring rate																			
SPCE	100	0.02	0.02	0.02	0.02	100.00	100.00	0.06	0.08	0.10	0.10	94.60	86.00	0.18	0.15	0.24	0.19	78.80	79.20
	300	0.01	0.02	0.02	0.02	100.00	100.00	0.07	0.07	0.08	0.08	80.40	52.20	0.19	0.16	0.22	0.18	54.00	49.40
	500	0.01	0.02	0.02	0.02	100.00	100.00	0.07	0.07	0.08	0.08	68.00	36.20	0.20	0.16	0.22	0.18	26.80	23.20
Good overlap-50% censoring rate																			
SPCE	100	0.10	0.08	0.14	0.11	99.60	99.40	0.02	0.02	0.09	0.07	96.40	96.80	0.21	0.20	0.28	0.26	72.40	74.40
	300	0.10	0.08	0.12	0.09	100.00	99.80	0.03	0.03	0.06	0.05	92.20	92.40	0.22	0.21	0.24	0.23	42.40	38.00
	500	0.10	0.08	0.11	0.08	100.00	99.60	0.03	0.02	0.05	0.04	92.20	90.20	0.22	0.21	0.24	0.22	18.40	16.00
Poor overlap-25% censoring rate																			
SPCE	100	0.00	0.01	0.02	0.02	100.00	96.98	0.09	0.09	0.15	0.11	83.88	83.63	0.13	0.03	0.30	0.09	94.96	94.21
	300	0.00	0.01	0.01	0.01	100.00	97.60	0.06	0.09	0.12	0.10	75.40	35.80	0.13	0.03	0.22	0.06	92.80	94.80
	500	0.00	0.01	0.01	0.01	100.00	98.80	0.04	0.09	0.10	0.10	78.60	13.60	0.13	0.03	0.18	0.05	90.00	89.40
Poor overlap-50% censoring rate																			
SPCE	100	0.07	0.01	0.16	0.05	92.80	93.40	0.07	0.00	0.13	0.05	85.40	96.40	0.24	0.09	0.46	0.18	90.80	85.37
	300	0.09	0.02	0.13	0.03	95.60	96.00	0.04	0.00	0.11	0.03	75.80	96.40	0.24	0.08	0.32	0.11	83.00	76.40
	500	0.08	0.02	0.12	0.03	95.80	96.00	0.04	0.00	0.09	0.02	71.40	96.00	0.24	0.08	0.30	0.10	70.60	66.20

of MSMs (i.e. `Cox3.MSM.COV.OW`, `Cox3.MSM.COV.IPW`), the rest estimators, as expected, are frequently less efficient and have less than nominal coverage.

Based on covariate-dependent assumptions of both censoring mechanisms, the absolute bias, RMSE, and coverage probability of the 95 percent confidence interval for the OW, and IPW estimators are presented in Table 6. OW performs better than IPW in terms of SPCE and RACE, with reduced absolute bias and RMSE and coverage closer to nominal for all sample sizes. Due to extreme propensity scores, the IPW estimator is susceptible to having a lack of overlap in general. Consequently, this leads to extremely large bias and variation and low coverage in most cases. In this regard, OW is superior to IPW in terms of both bias and variance, demonstrating that it is the more effective method. Compared to the IPW estimator, the coverage provided by the OW estimator is less sensitive to the presence or absence of overlap. The same as Figures 11–14, `Cox3.MSM.COV.OW` across SPCE and RACE, consistently outperforms alternative methods in Table 6. Then, the MSMs *without* covariate perform well in terms of RACE when compared to extended Zeng's method. However, extended Zeng's methods improve SPCE results when the overlap is poor. In general, extended Zeng's methods have higher RMSE despite the degree of overlap and censoring rate.

3.3.4 Misspecification

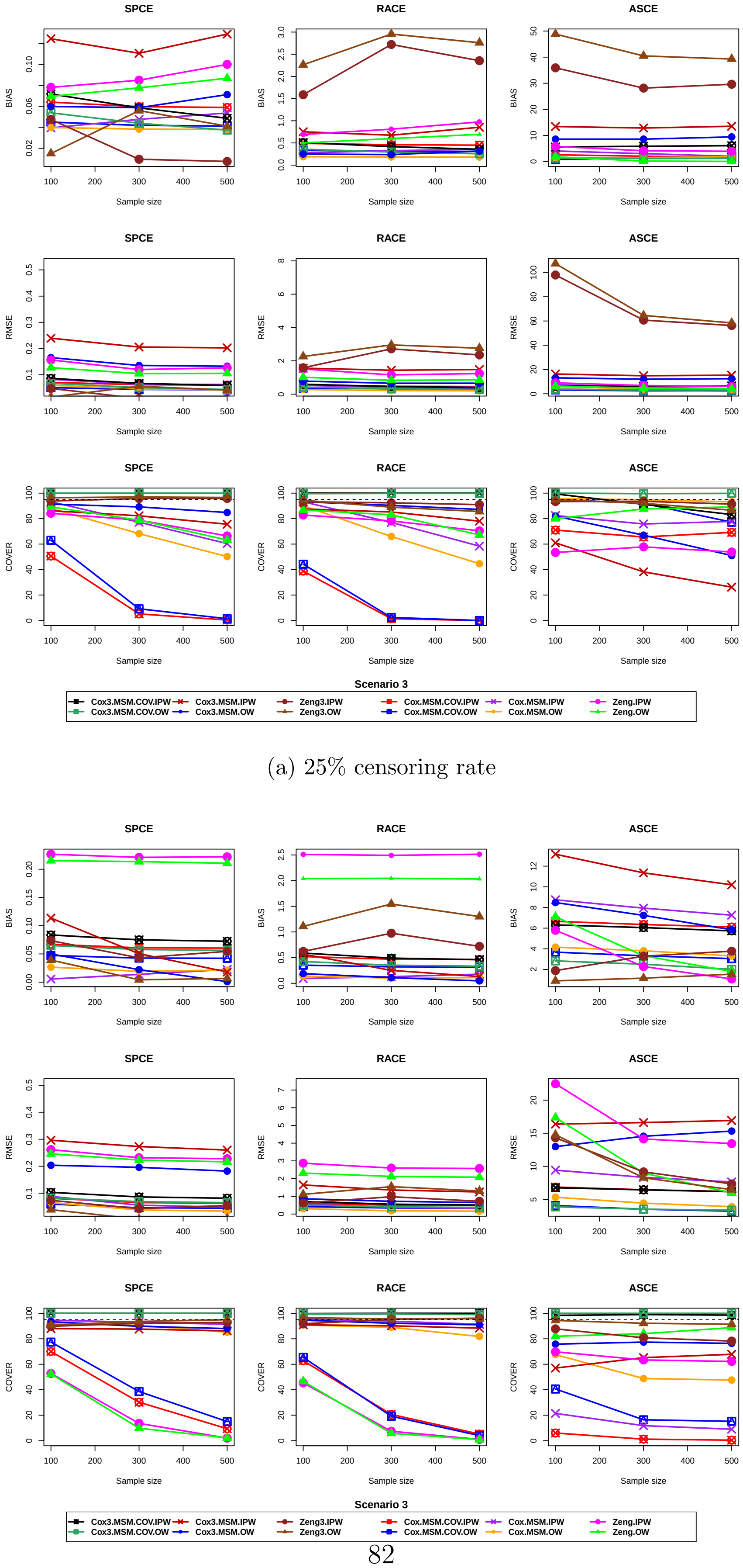
When conducting a simulation study, it is desirable to generate the data in such a way that the correct form of any analysis model to be fitted to those data is known so that we know that the analysis model is correctly specified. This is important since we wished to use a simulation study to assess the performance of causal estimands when the models for the censoring weights are misspecified in some way. It would be essential to know that the MSM itself is correctly specified so that any bias in the estimates can be attributed to the misspecification of the models used for the censoring weights. Building upon this investigation, our objective is to investigate how robust are the estimators in the presence of misspecification of censoring weights under the assumption of **Scenario 3** (i.e., both types of censoring are covariate-dependent) and various degrees of overlap (good, poor), censoring rate (25%, 50%) and sample sizes (100,300,500).

The process of generating data with two covariate-dependent censoring mechanisms is the same as **Scenario 3** explained in Subsection 3.1. We explore a framework for assessing bias and RMSE of causal estimands with omitted some covariates and different specifications in IPCW. To do so, we consider

(a) omission (X_1, X_3) in constructing IPCW.

(b) different specification form of main effects in constructing IPCW:

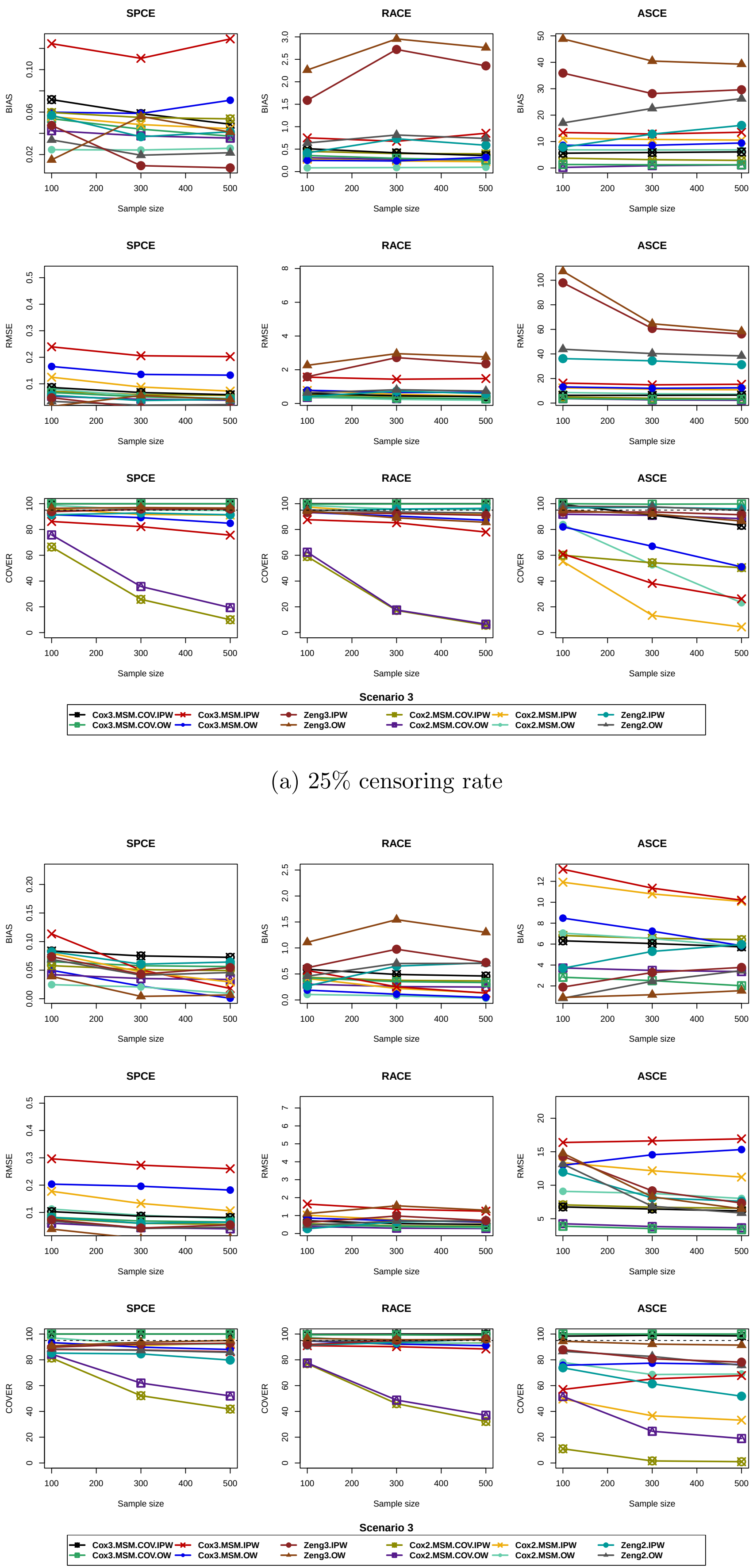
$$X_1 + X_2 + X_3 + X_4 + {X_1}^2 + X_4 \times X_2$$



(a) 25% censoring rate

(b) 50% censoring rate

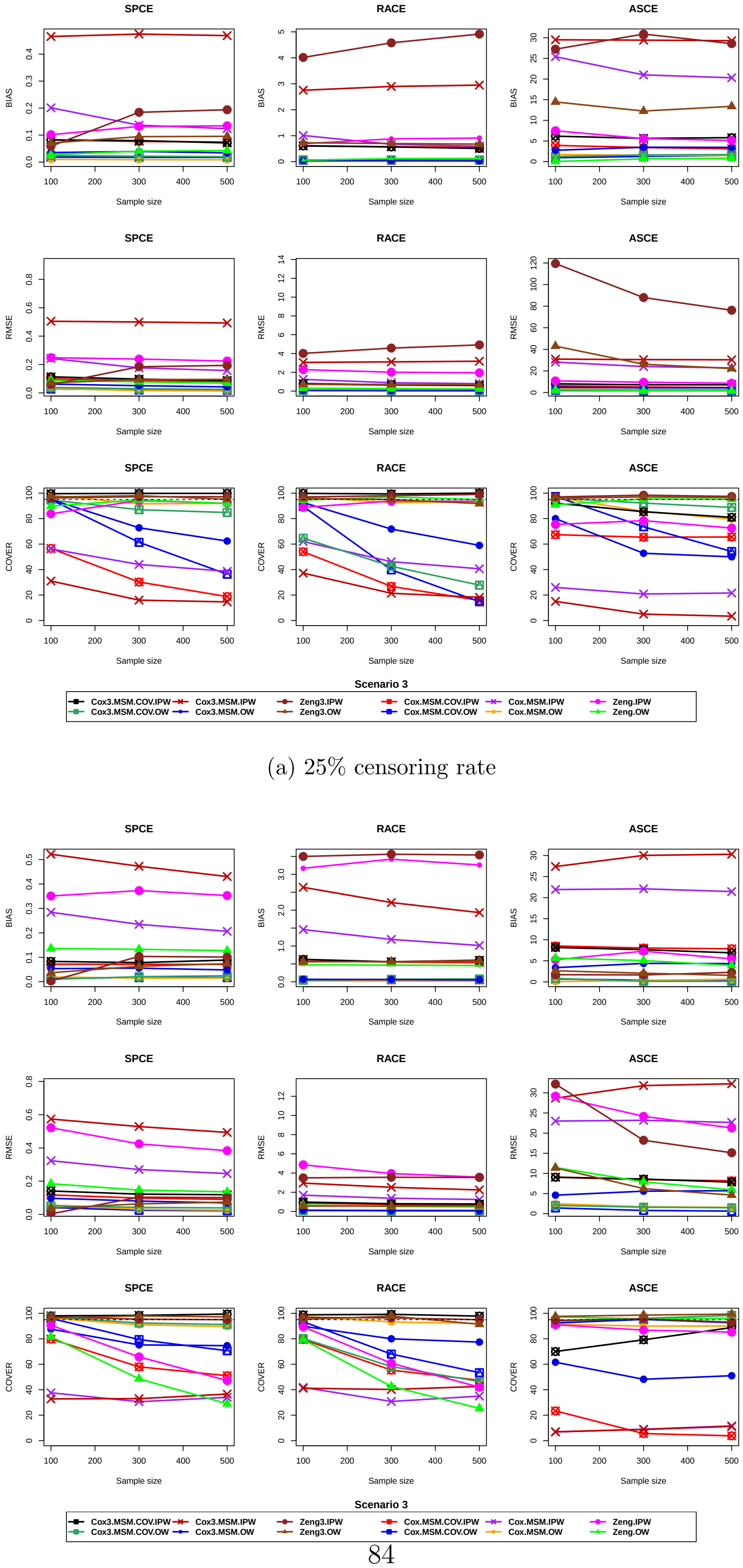
Figure 11: **Scenario 3**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *good overlap* for Scenario 3 and Scenario 1.



(a) 25% censoring rate

(b) 50% censoring rate

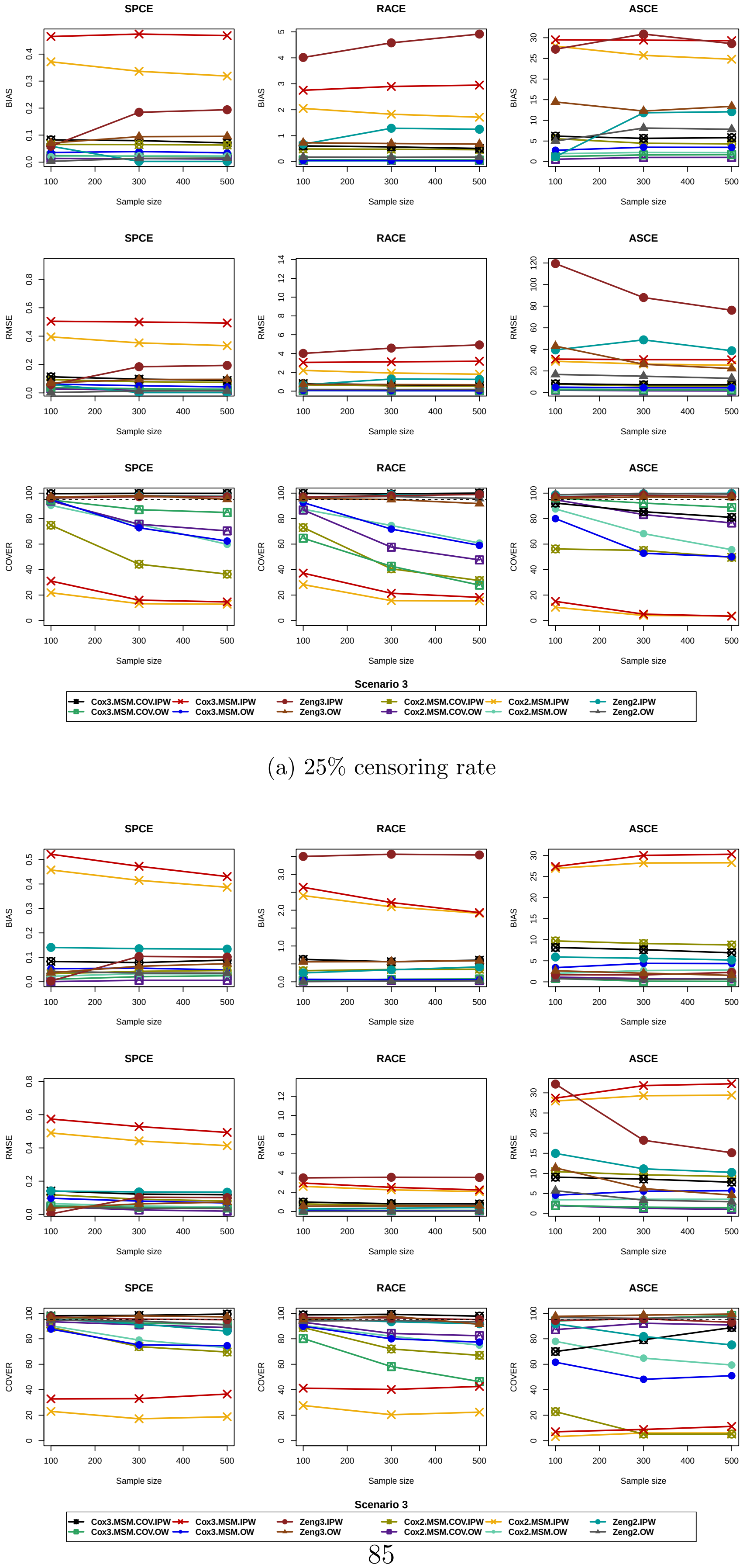
Figure 12: **Scenario 3**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *good overlap* for Scenario 3 and Scenario 2.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 13: **Scenario 3**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 3 and Scenario 1.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 14: **Scenario 3**: Absolute bias, root mean squared error (RMSE) and coverage of the 95% confidence interval for comparing two treatment under *poor overlap* for Scenario 3 and Scenario 2.

Table 6: **Scenario 3**: Absolute bias, root mean squared error (RMSE) and coverage for comparing two treatments under different degrees of overlap (good, poor), different sample size (100, 300, 500) and various censoring rate (25%, 50%) when both types of censoring are covariate-dependent.

		MSM+COV						MSM						Zeng					
	sample	Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage		Absolute Bias		RMSE		95% Coverage	
	size	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW	IPW	OW
Good overlap-25% censoring rate																			
RACE	100	0.51	0.36	0.59	0.42	100.00	99.60	0.75	0.25	1.56	0.78	87.60	93.20	1.59	2.26	7.84	6.36	93.20	94.40
	300	0.42	0.30	0.46	0.33	100.00	99.80	0.67	0.24	1.44	0.65	85.20	90.40	2.72	2.96	5.32	4.40	92.40	89.00
	500	0.36	0.26	0.40	0.28	100.00	99.80	0.85	0.32	1.48	0.66	78.00	87.20	2.35	2.76	4.25	3.87	91.00	85.60
Good overlap-50% censoring rate																			
RACE	100	0.59	0.42	0.70	0.51	99.80	99.40	0.57	0.19	1.64	0.86	91.00	94.60	0.62	1.11	7.35	4.96	91.80	96.60
	300	0.49	0.35	0.55	0.40	100.00	99.40	0.25	0.11	1.35	0.72	90.20	92.20	0.98	1.54	4.19	3.36	95.20	95.60
	500	0.46	0.33	0.51	0.37	100.00	99.00	0.13	0.05	1.25	0.65	88.40	91.00	0.72	1.30	3.02	2.45	96.20	95.80
Poor overlap-25% censoring rate																			
RACE	100	0.61	0.07	0.80	0.10	99.80	64.60	2.75	0.04	3.05	0.08	37.20	92.60	4.01	0.73	13.57	1.71	97.00	95.80
	300	0.57	0.07	0.68	0.09	99.40	42.60	2.90	0.04	3.11	0.06	21.40	71.80	4.58	0.70	9.20	1.04	98.00	95.00
	500	0.51	0.07	0.60	0.08	100.00	27.80	2.95	0.04	3.18	0.04	18.20	59.00	4.92	0.68	8.32	0.88	99.00	91.80
Poor overlap-50% censoring rate																			
RACE	100	0.63	0.04	0.97	0.11	98.80	80.20	2.64	0.07	2.96	0.12	41.20	89.80	3.50	0.56	13.31	1.72	96.79	95.20
	300	0.56	0.07	0.80	0.11	99.20	58.20	2.21	0.06	2.51	0.09	40.20	80.00	3.56	0.56	9.77	0.97	96.40	97.80
	500	0.60	0.08	0.78	0.11	97.80	46.40	1.93	0.05	2.24	0.07	42.60	77.40	3.54	0.59	7.17	0.87	95.00	91.40
Good overlap-25% censoring rate																			
SPCE	100	0.07	0.05	0.09	0.07	100.00	100.00	0.12	0.06	0.24	0.17	86.20	91.40	0.05	0.02	0.52	0.48	93.80	96.40
	300	0.06	0.04	0.07	0.05	100.00	100.00	0.11	0.06	0.21	0.14	82.20	89.20	0.01	0.06	0.33	0.29	95.80	97.00
	500	0.05	0.04	0.06	0.04	100.00	100.00	0.13	0.07	0.20	0.13	75.60	84.80	0.01	0.04	0.25	0.23	95.80	96.60
Good overlap-50% censoring rate																			
SPCE	100	0.08	0.06	0.10	0.08	100.00	100.00	0.11	0.05	0.30	0.20	88.00	93.40	0.07	0.04	0.51	0.42	89.80	90.80
	300	0.07	0.06	0.09	0.07	100.00	100.00	0.05	0.02	0.27	0.20	87.60	89.80	0.04	0.00	0.27	0.25	92.20	93.60
	500	0.07	0.06	0.08	0.07	100.00	100.00	0.02	0.00	0.26	0.18	86.20	88.00	0.05	0.01	0.20	0.18	92.80	95.00
Poor overlap-25% censoring rate																			
SPCE	100	0.08	0.02	0.11	0.04	99.60	94.60	0.47	0.04	0.50	0.06	31.00	95.00	0.06	0.07	0.91	0.32	96.20	97.00
	300	0.08	0.02	0.10	0.03	99.80	87.00	0.47	0.04	0.50	0.05	16.00	72.80	0.18	0.09	0.71	0.20	97.20	98.00
	500	0.07	0.02	0.09	0.02	99.80	84.80	0.47	0.03	0.49	0.04	14.60	62.40	0.19	0.10	0.51	0.16	97.20	95.20
Poor overlap-50% censoring rate																			
SPCE	100	0.08	0.01	0.14	0.05	98.00	97.20	0.52	0.05	0.57	0.10	32.80	87.60	0.00	0.04	0.79	0.32	97.19	96.20
	300	0.08	0.02	0.12	0.04	98.40	92.40	0.47	0.06	0.53	0.08	33.00	75.20	0.10	0.06	0.77	0.18	95.40	98.00
	500	0.09	0.02	0.12	0.04	99.40	91.20	0.43	0.05	0.49	0.07	36.60	74.60	0.10	0.07	0.51	0.15	95.00	97.20

as a *exploratory* sensitivity analysis using our simulation structure to assess how bias and RMSE of causal estimands vary under various forms of IPCW.

Figures 15–20 display the performance of estimators computed by MSMs with and without covariates and Zeng's methods in the presence of omission and/or misspecification of main effects. Furthermore, in Figures 21–24, we present absolute bias and RMSE comparing two treatments with various degrees of overlap and a censoring rate of 50 percent. The results of the 25 percent censoring rate are reported in the Supplementary Material (SM.2). Specifically, Z1-Z2 are `Zeng3.IPW` and `Zeng3.OW` under the correct specification of IPCW. Z3-Z4 are Zeng's methods (IPW-OW) computed through omission of (X_1, X_3). Z5-Z6 are Zeng's methods (IPW-OW) under a different specification form of IPCW. In addition, MC1-MC2 are displayed as `Cox3.MSM.COV....` under IPW and OW *with* the correct IPCW specification. In MC3-MC4, we utilized MSMs *with* covariates under omission, however in MC5-MC6, we used MSMs with covariates under a different IPCW specification. Furthermore, M1 and M2 are MSMs *without* covariates giving the correct form of IPCW. M3-M4 compute MSMs *without* covariates under omission, whereas M5-M6 are MSMs *without* covariates under a different IPCW form.

As expected, by omitting some main effects and different specification forms of IPCW, the performance of all methods deteriorated (see Figures 15–20), demonstrated by the larger bias and RMSE regardless of the degree of overlaps, censoring rates and same sizes. However, MSMs with covariates are more robust and Zeng's methods are so sensitive to misspecification assumptions.

3.3.5 Proportionality

To do a sensitivity analysis in terms of proportionality assumption, we generate four covariates: $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})^\top$ where

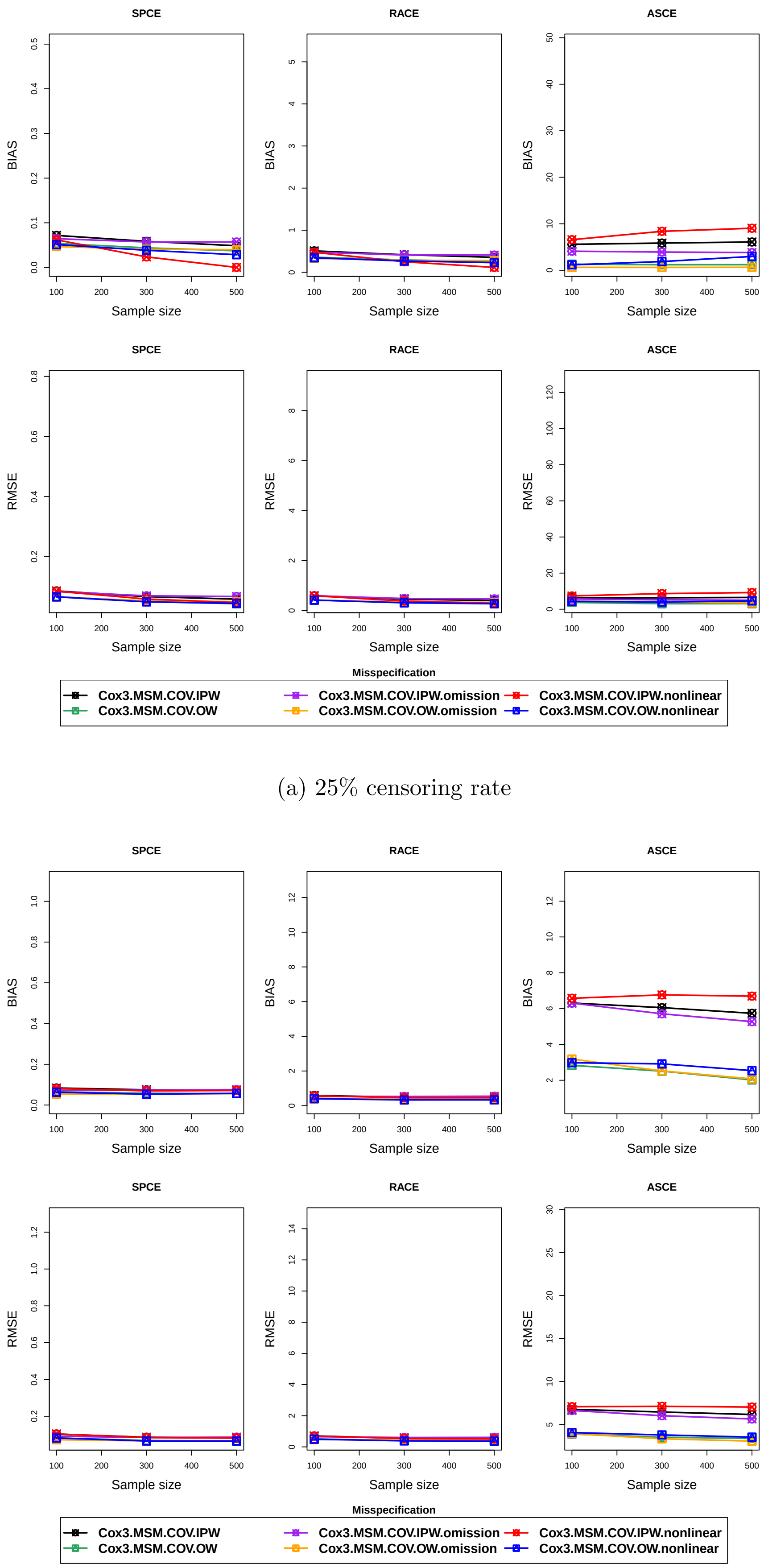
$$\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix} \right\},$$

$X_{i4} \sim \text{Bernoulli}(0.5)$ and $X_{i3} \sim \text{Bernoulli}(0.6X_{i4} + 0.4(1 - X_{i4}))$. We consider two treatment groups, with the true propensity score model given by $\text{logit}(\mathrm{e_i}) = \tilde{X}_i^\top \beta$ where $\tilde{X}_i = (1, X_i^\top)^\top$. Set $\beta = (-0.1\Psi, -0.9\Psi, -0.3\Psi, -0.1\Psi, -0.2\Psi)^\top$ where $\Psi = 1$ and $\Psi = 5$ represent good and poor overlap between groups, respectively.

The model to generate potential survival times is an accelerated failure time model that violates the proportional hazards assumption. Specifically, Y_i is drawn from a log-normal distribution

$$\log\{\mathrm{Y_i}\} \sim \mathrm{N}(\mu, \sigma^2 = (0.81)^2),$$

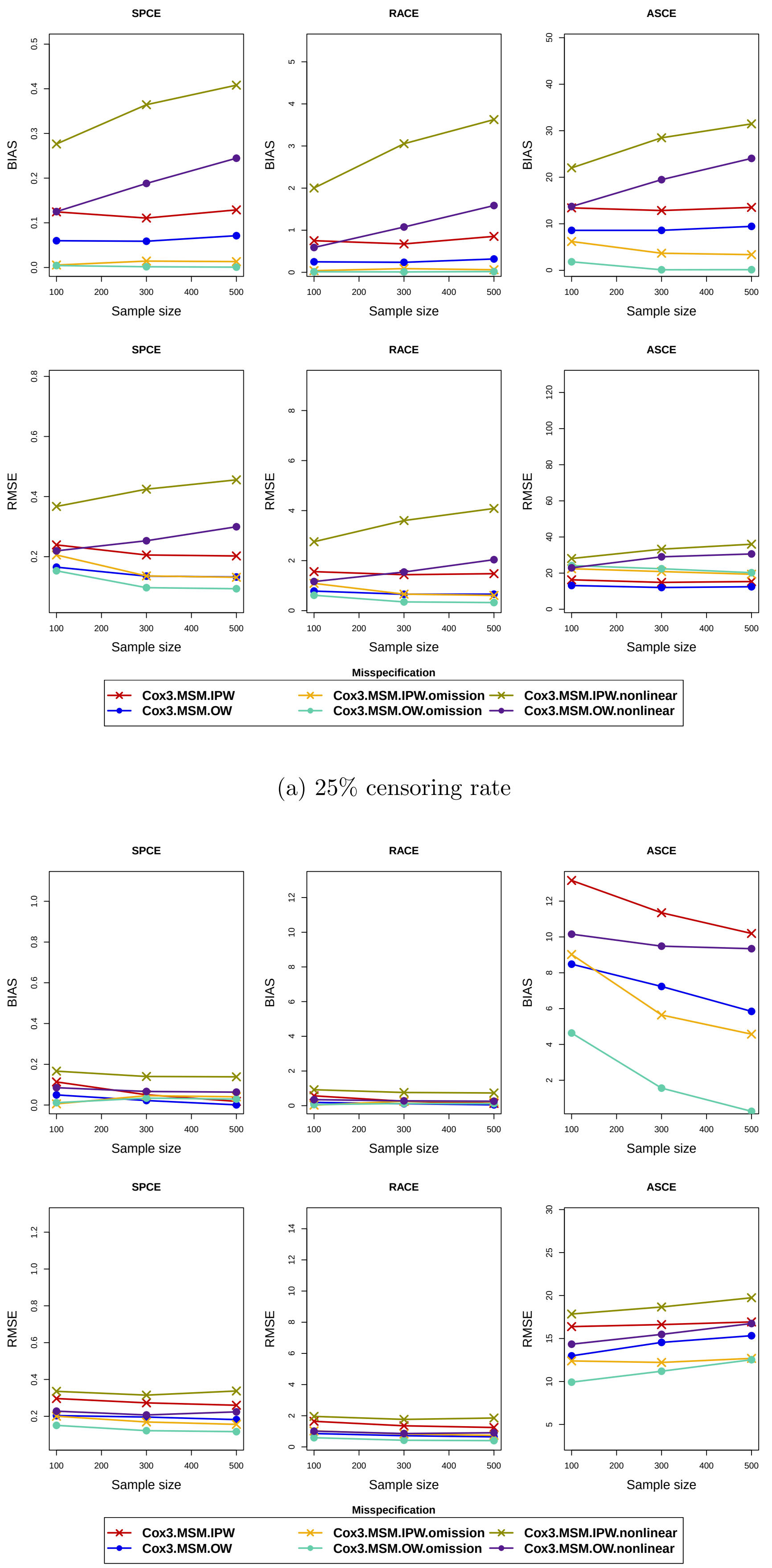
where $\mu = 2.25 - \gamma Z_i - \mathbf{X_i}^\top \alpha$ with $\gamma = 1$ and $\alpha = (0, 2, 1.5, -1, 1)^\top$. We simulated 500 Monte Carlo repetitions for each simulation setting and averaged the results.



(a) 25% censoring rate

(b) 50% censoring rate

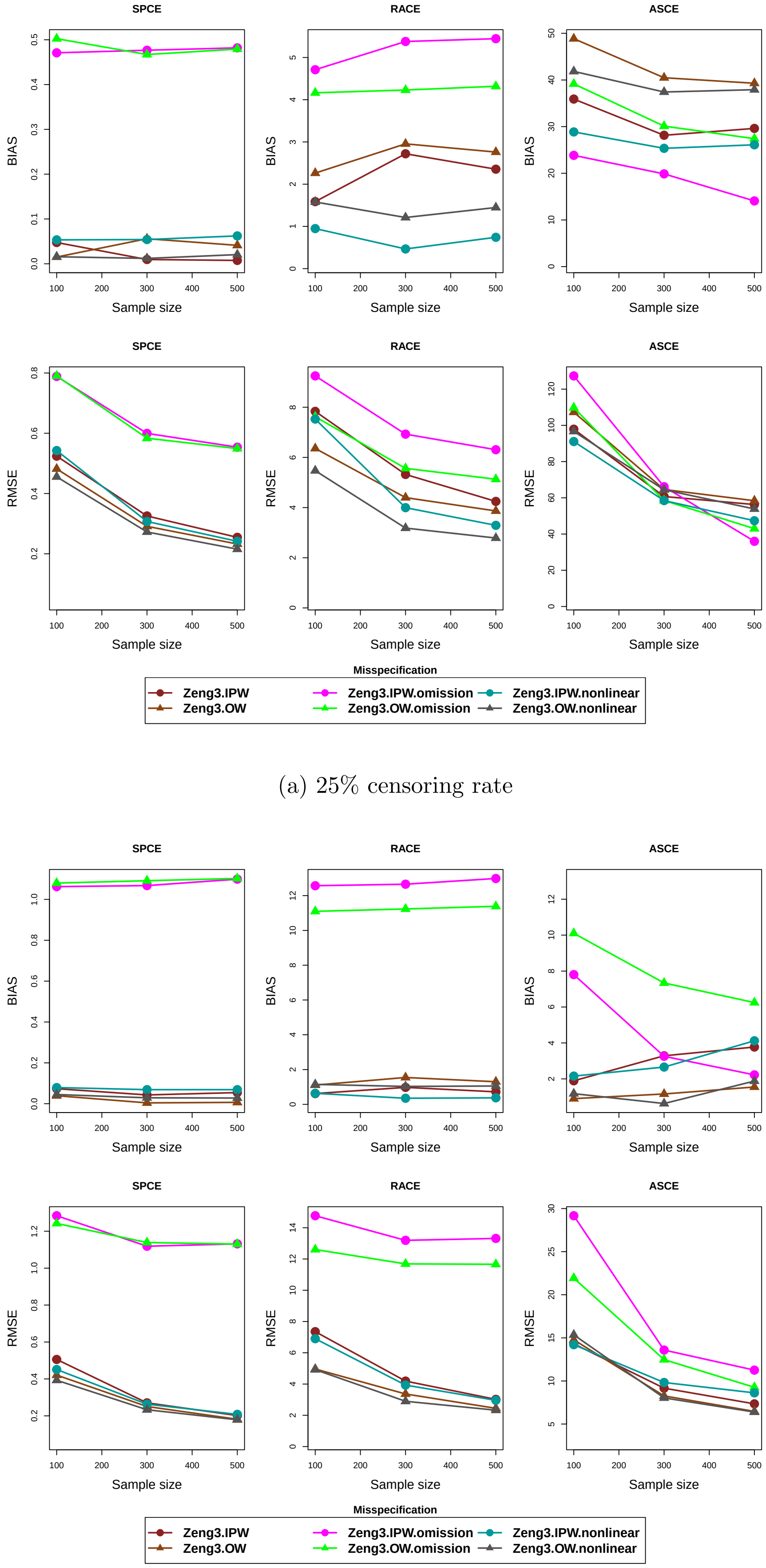
Figure 15: Absolute bias, root mean squared error (RMSE) using MSMs with covariates under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *good overlap*.



(a) 25% censoring rate

(b) 50% censoring rate

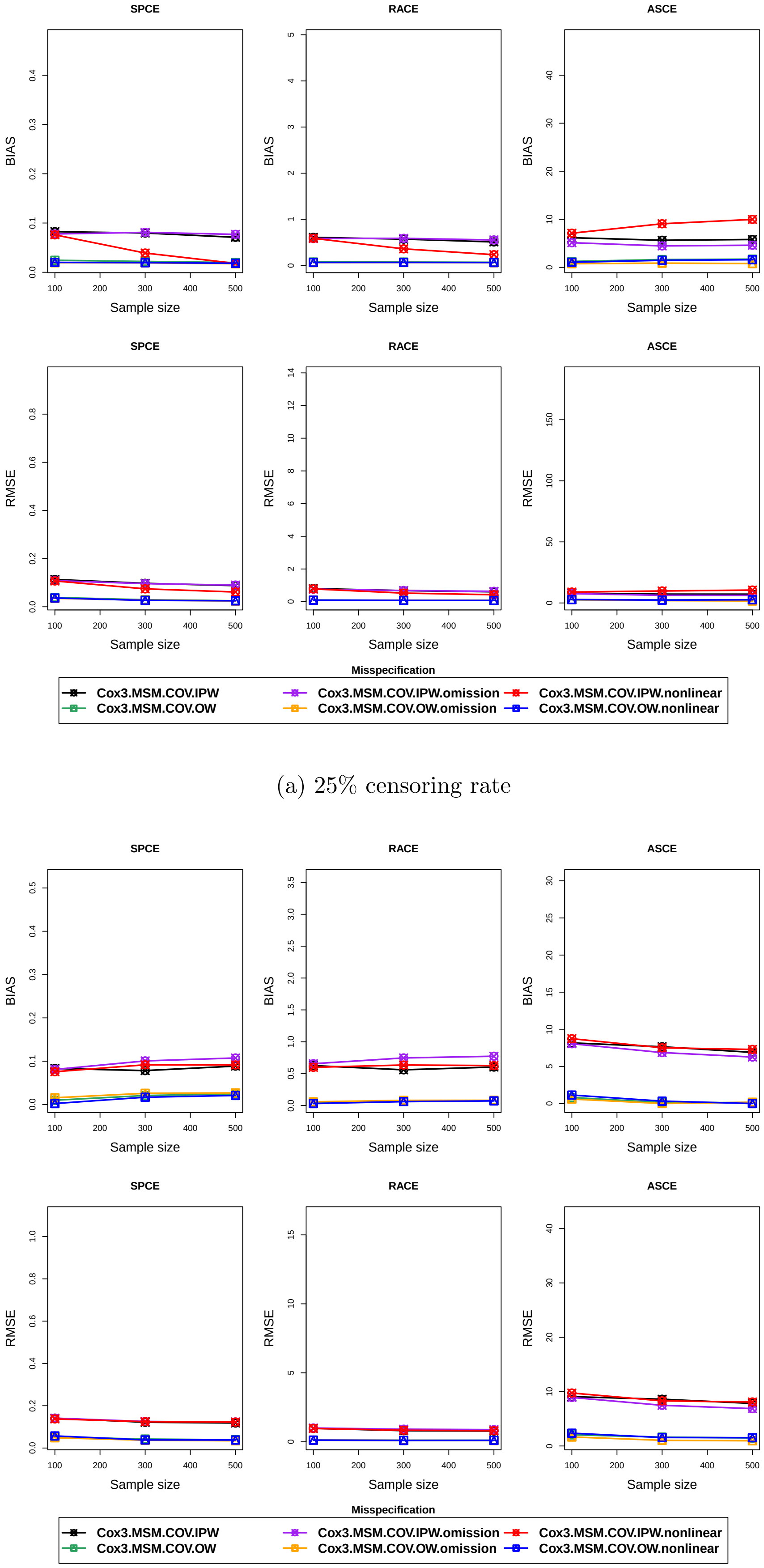
Figure 16: Absolute bias, root mean squared error (RMSE) using MSMs without covariates under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *good overlap*.



(a) 25% censoring rate

(b) 50% censoring rate

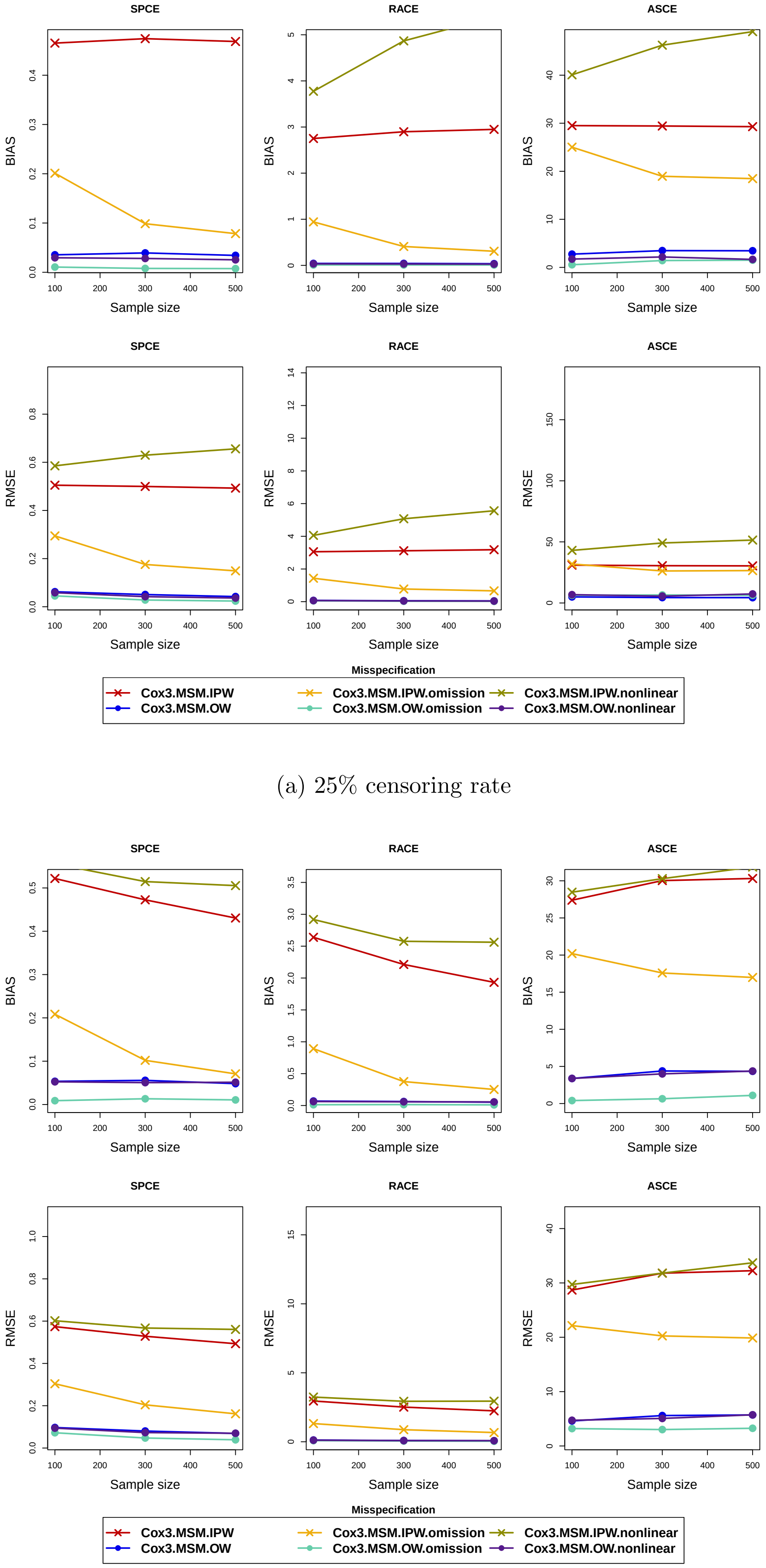
Figure 17: Absolute bias, root mean squared error (RMSE) using Zeng's method under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *good overlap*.



(a) 25% censoring rate

(b) 50% censoring rate

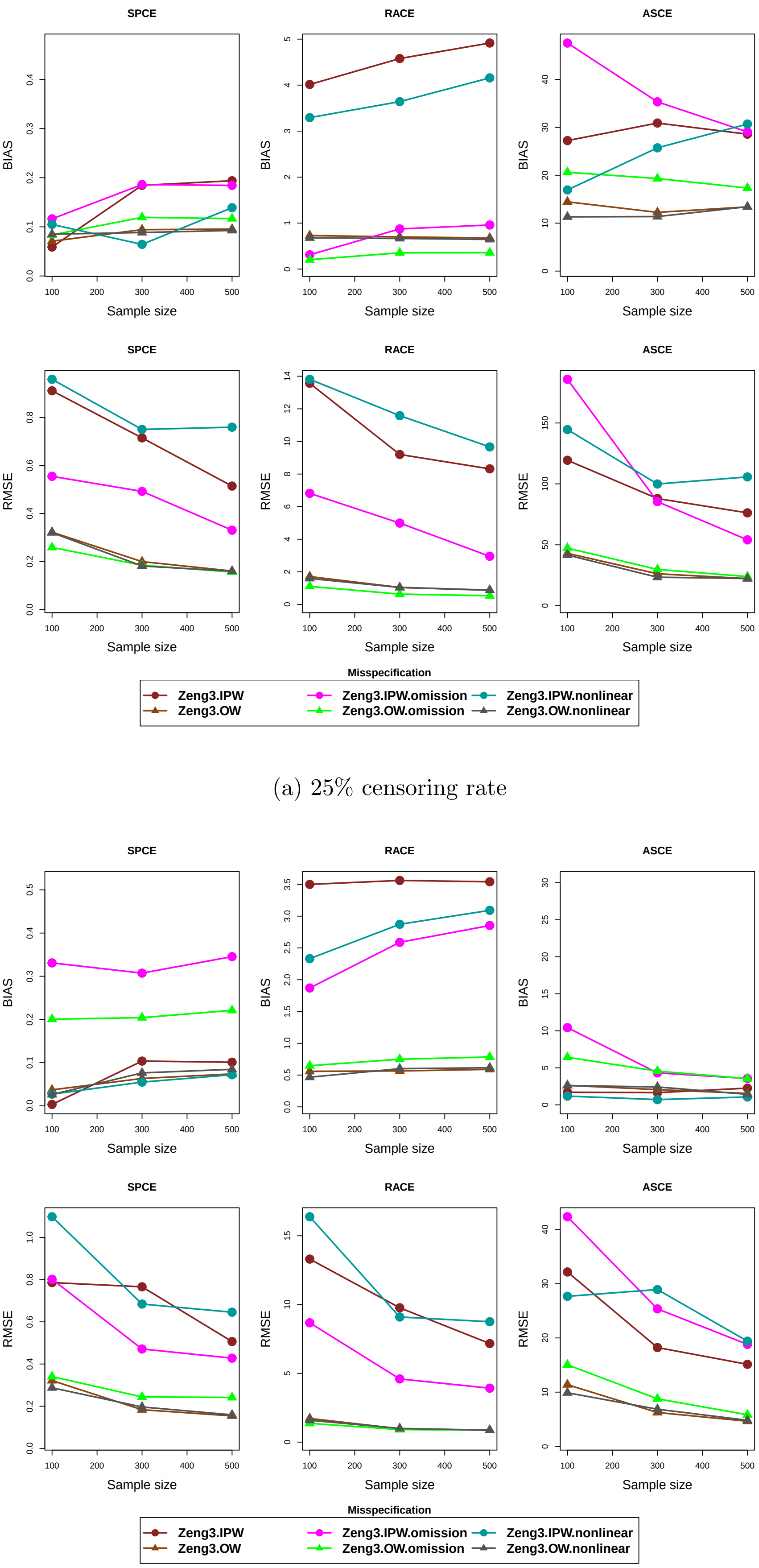
Figure 18: Absolute bias, root mean squared error (RMSE) using MSMs with covariates under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *poor overlap*.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 19: Absolute bias, root mean squared error (RMSE) using MSMs without covariates under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *poor overlap*.



(a) 25% censoring rate

(b) 50% censoring rate

Figure 20: Absolute bias, root mean squared error (RMSE) using Zeng's method under correct form, omission and different specification of covariates in IPCW for comparing two treatment under *poor overlap*.

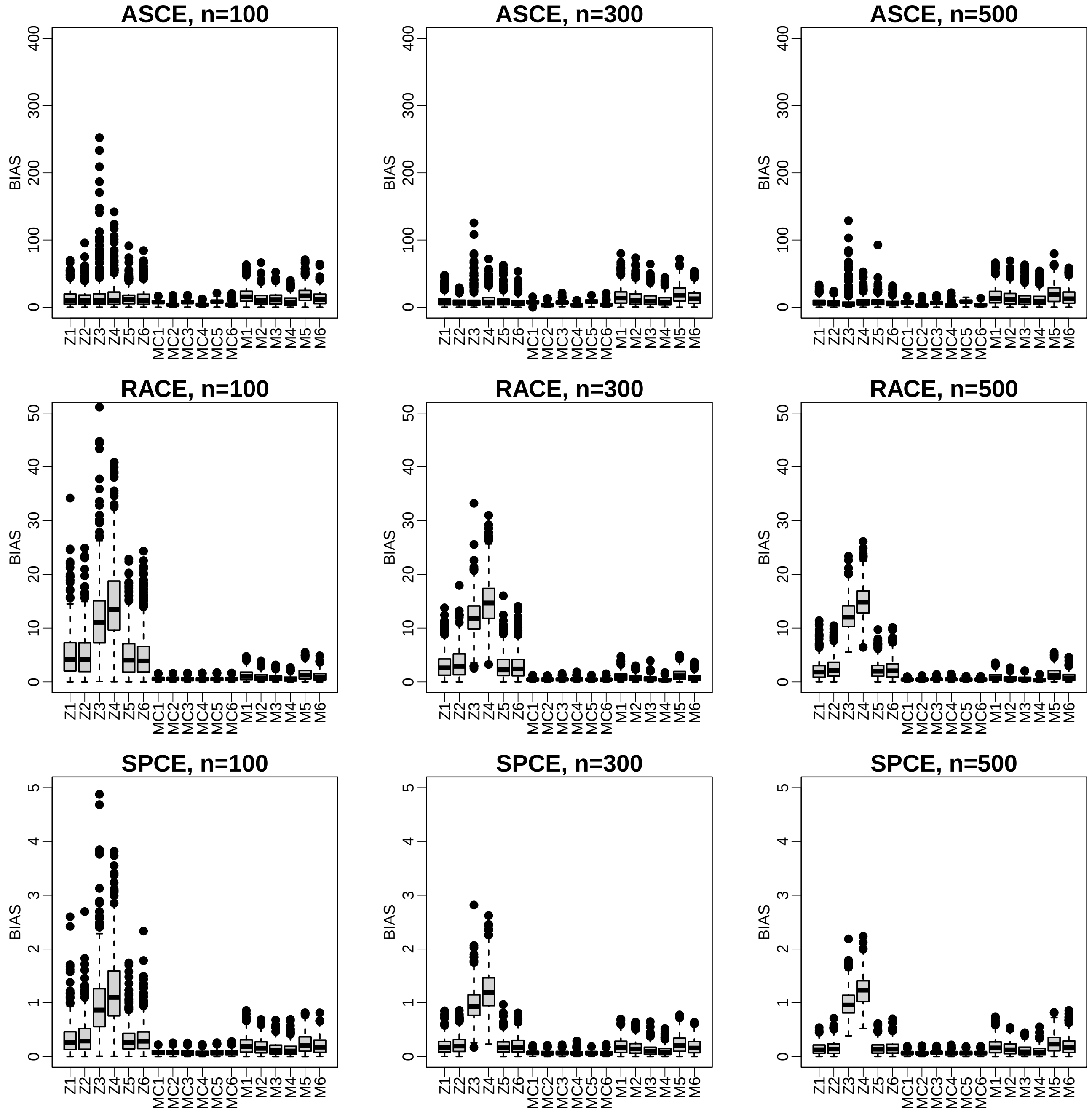


Figure 21: Absolute bias comparing two treatment under *good overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

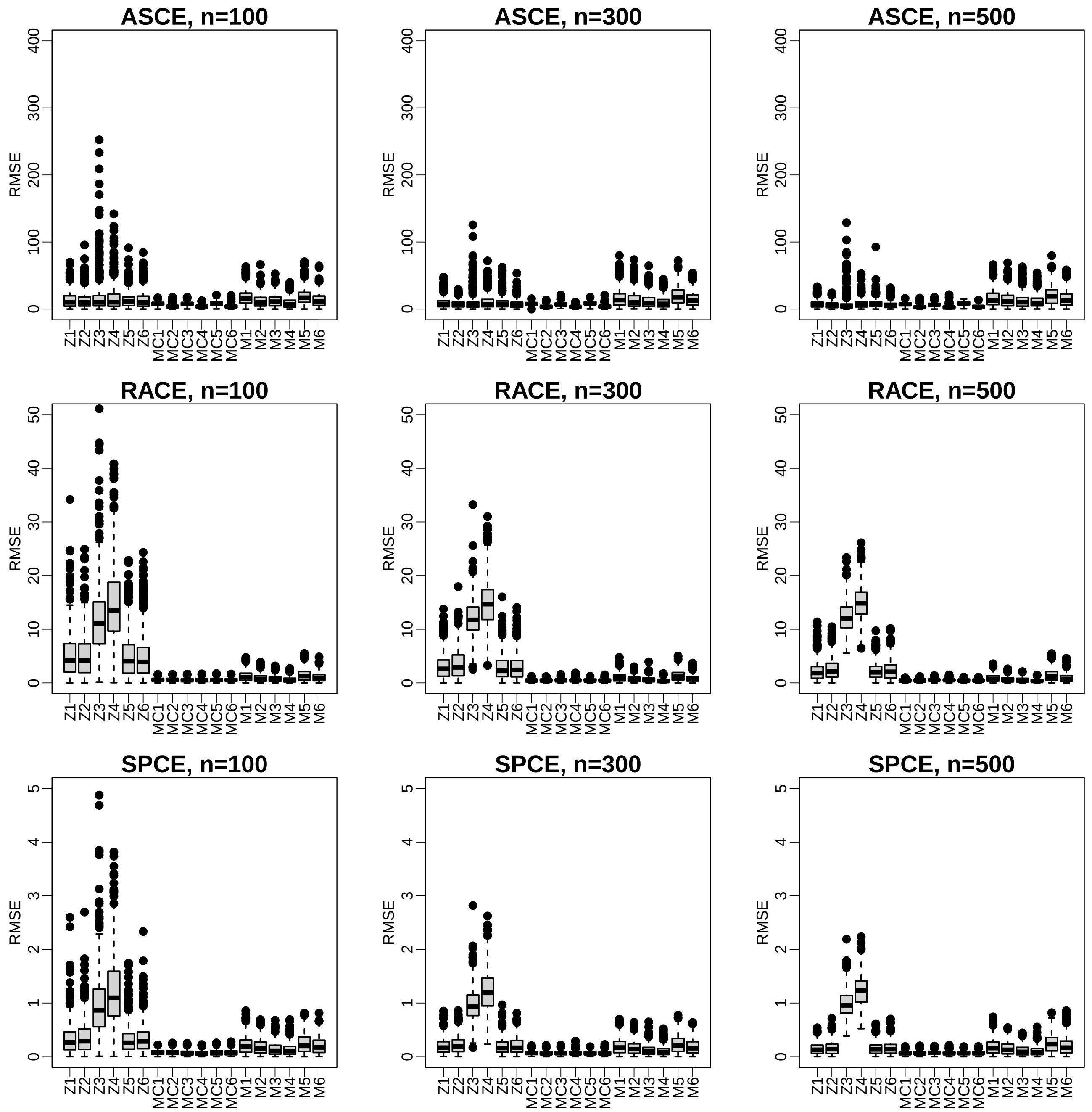


Figure 22: RMSE comparing two treatment under *good overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

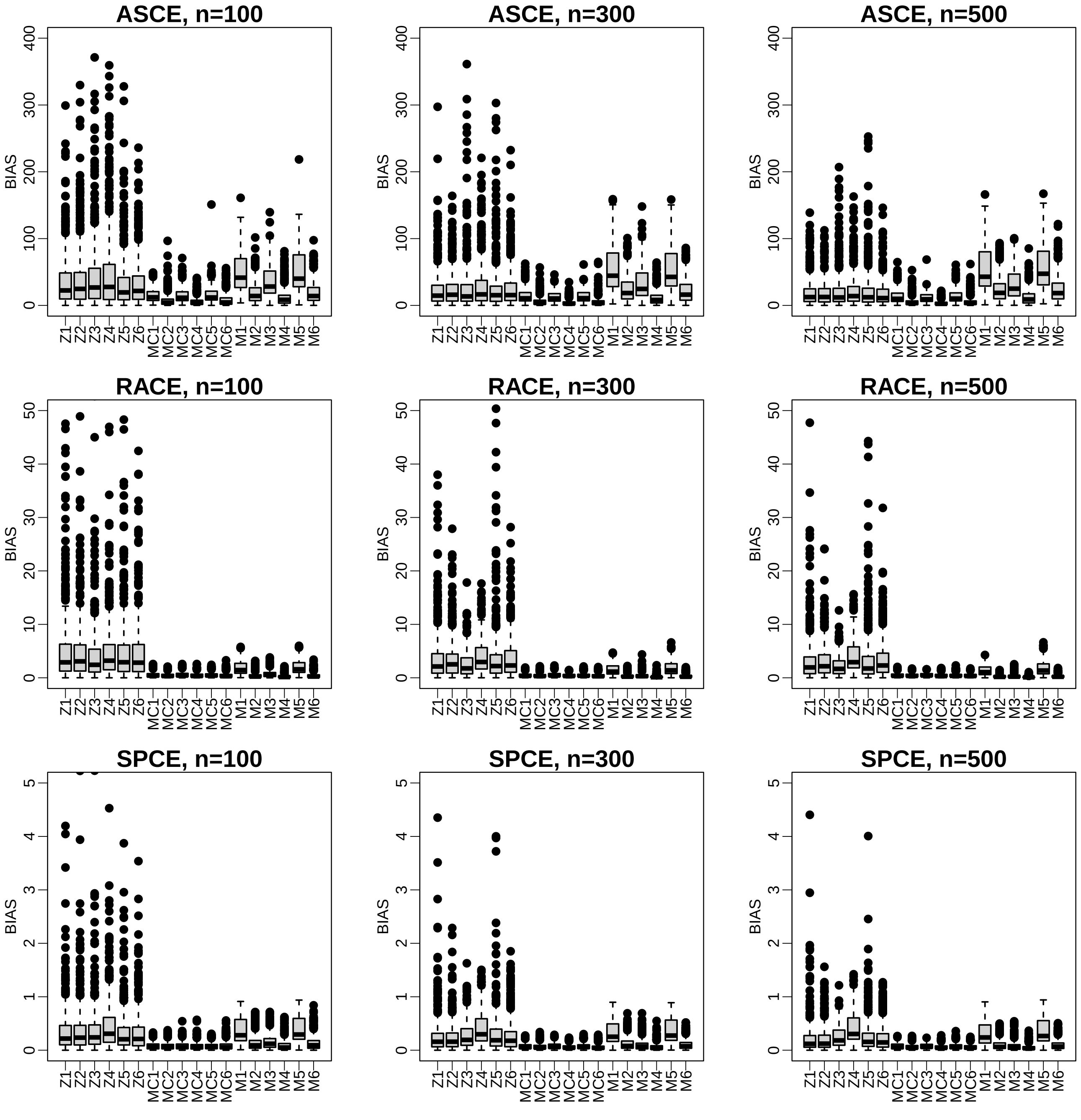


Figure 23: Absolute bias comparing two treatment under *poor overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

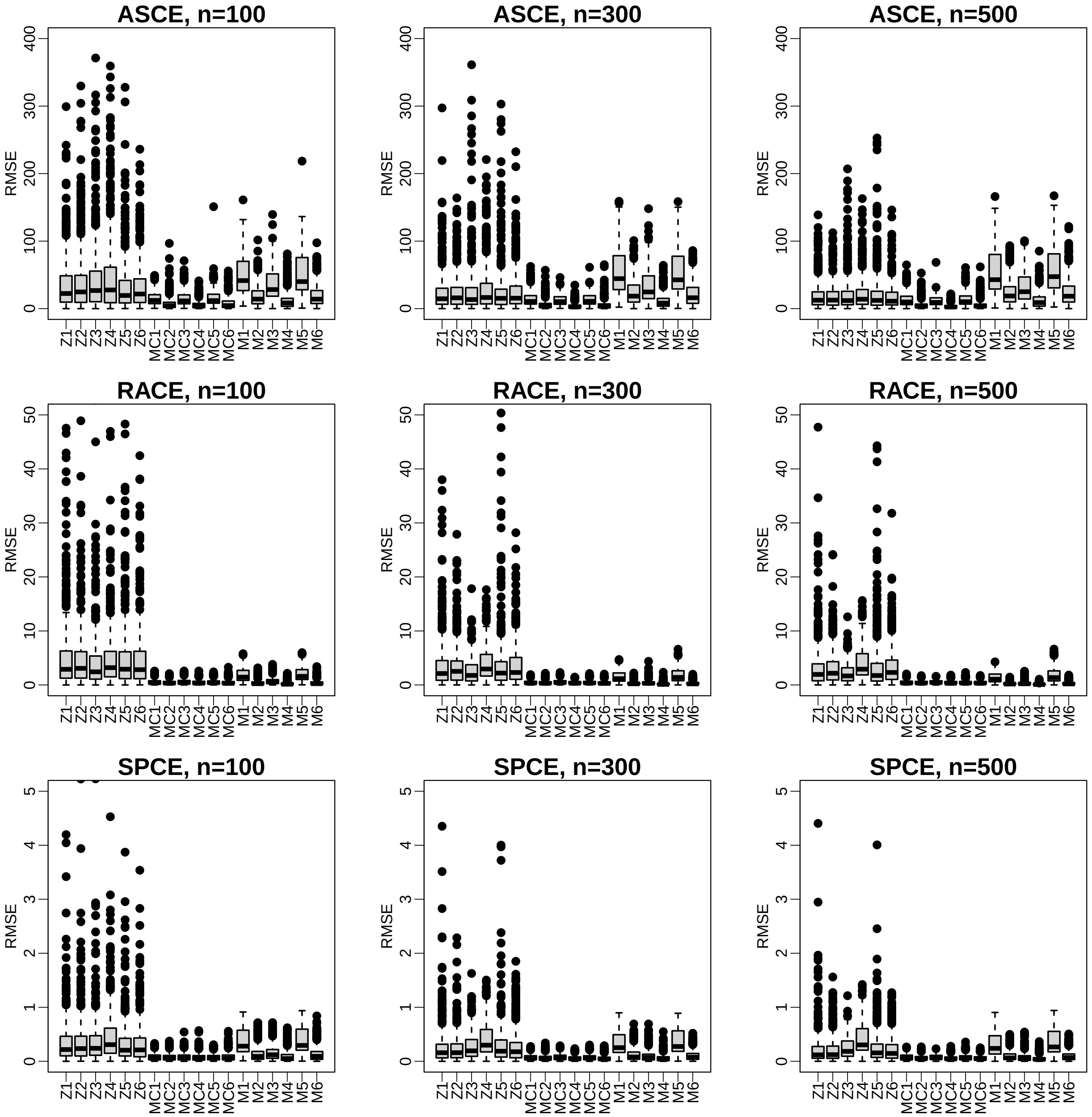


Figure 24: RMSE comparing two treatment under *poor overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

The sample size varies $n = (100, 300, 500)$ and we fix the evaluation point $t = 10$ for estimating SPCE and RACE. Furthermore, we evaluate the performance of the models in terms of the absolute bias and root mean squared error (RMSE) corresponding to each estimator. In addition, under covariate-dependent censoring, due to both censoring mechanisms, we generate administrative censoring (C_i) from a Weibull survival model with a hazard rate

$$h_c(t|\mathbf{X_i}) = \lambda_c \nu_c t^{\nu_c - 1} \exp\{\mathbf{X_i}^{\top} \alpha_{\mathbf{c}} + \mathbf{Z_i}\gamma\}$$

and S_i is generated from a Weibull survival model with hazard rate

$$h_s(t|\mathbf{X_i}) = \lambda_S \nu_S t^{\nu_S - 1} \exp\{\mathbf{X_i}^{\top} \alpha_{\mathbf{S}} + \mathbf{Z_i}\gamma\}.$$

According to the design of Simulation study in **Scenario 3** (Subsection 3.1), we consider same values for λ_c, λ_s, ν_c, ν_s and $\alpha_{\mathbf{s}}, \alpha_{\mathbf{c}}$ so that censoring rate is roughly 25% or 50%.

In Figure 25–28, we compare two treatments with good overlap and a censoring rate of 50 percent under the assumption of non-proportional hazards, illustrating absolute bias and RMSE. Interestingly, the `Cox3.MSM.COV.OW` has a lower bias and RMSE for all three estimands, despite the proportional hazards assumption no longer being valid. In terms of RMSE, Zeng's methods for estimating SPCE and RACE are effective. Nevertheless, these nonparametric estimators are frequently less accurate in bias, regardless of sample size, censoring rate, or overlap degree. Across all types of estimands, OW-based estimators maintain the smallest bias and highest efficiency.

Figures SM–5–SM–12 (in Supplementary Material SM.2) provide general overview across 500 iteration of MSMs with and without covariates as well as Zeng's method in estimating three estimands. Regarding some notations in Figures SM–9–SM–12, Z1-Z2 are Zeng's methods (i.e. `Zeng3.IPW-Zeng3.OW`). Furthermore, MC1-MC2 are shown as `Cox3.MSM.COV...` under IPW and OW. Moreover, MSMs *without* covariates (i.e. `Cox3.MSM.IPW-Cox3.MSM.OW`) are noted as M1-M2. In general, MSMs *with* covariates continue to be robust and efficient across all types of estimands in terms of bias and variance despite the proportional hazards assumption no longer be valid.

3.4 Discussion:

By running simulation studies in different Scenarios, we assess the sensitivity of all estimators under the censoring mechanism's key assumptions. To deal with both observed confounding and covariate-dependent censoring, Marginal Structural Cox models work well. Mainly, in the presence of two different covariate-dependent

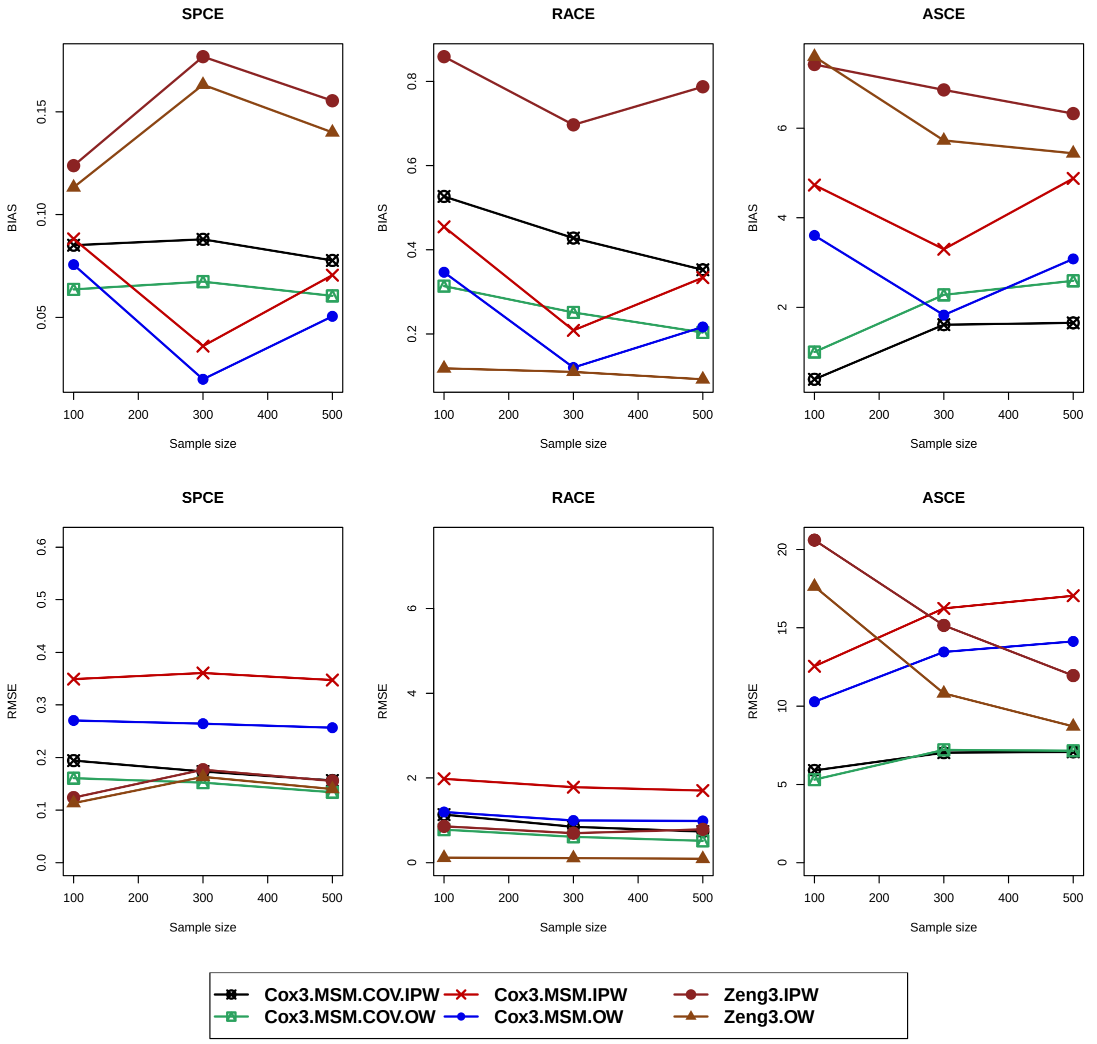


Figure 25: Absolute bias and RMSE comparing two treatment under *good overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

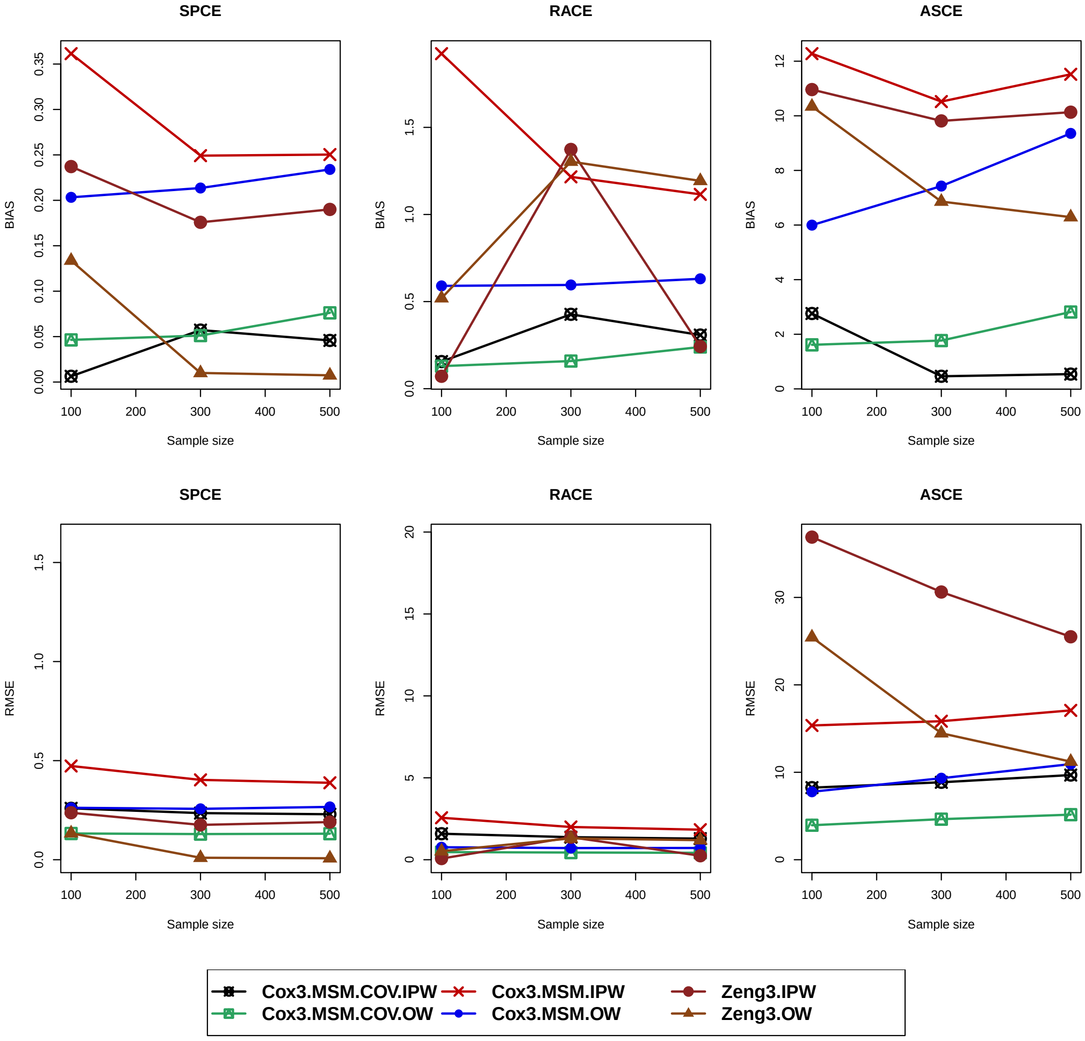


Figure 26: Absolute bias and RMSE comparing two treatment under *poor overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

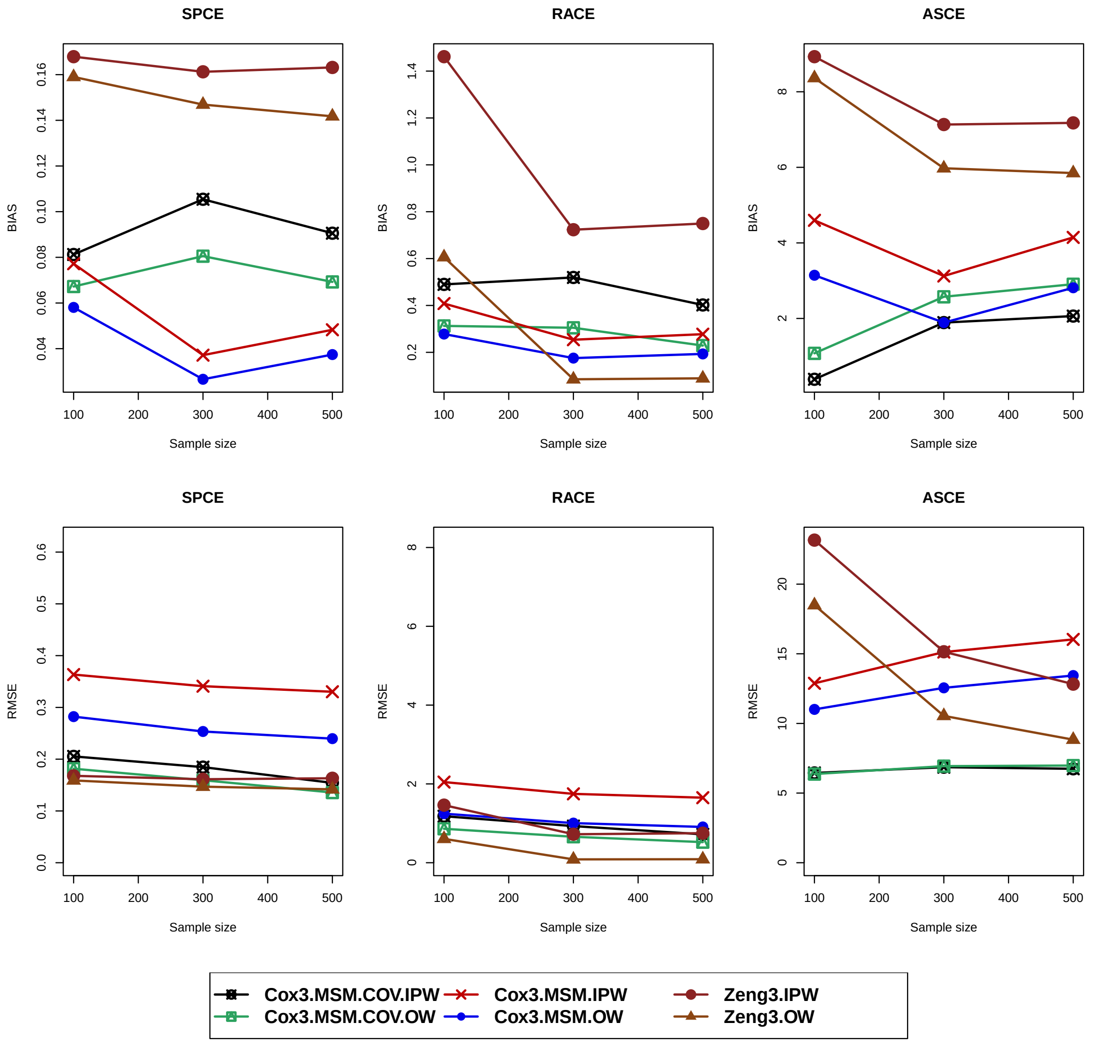


Figure 27: Absolute bias and RMSE comparing two treatment under *good overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

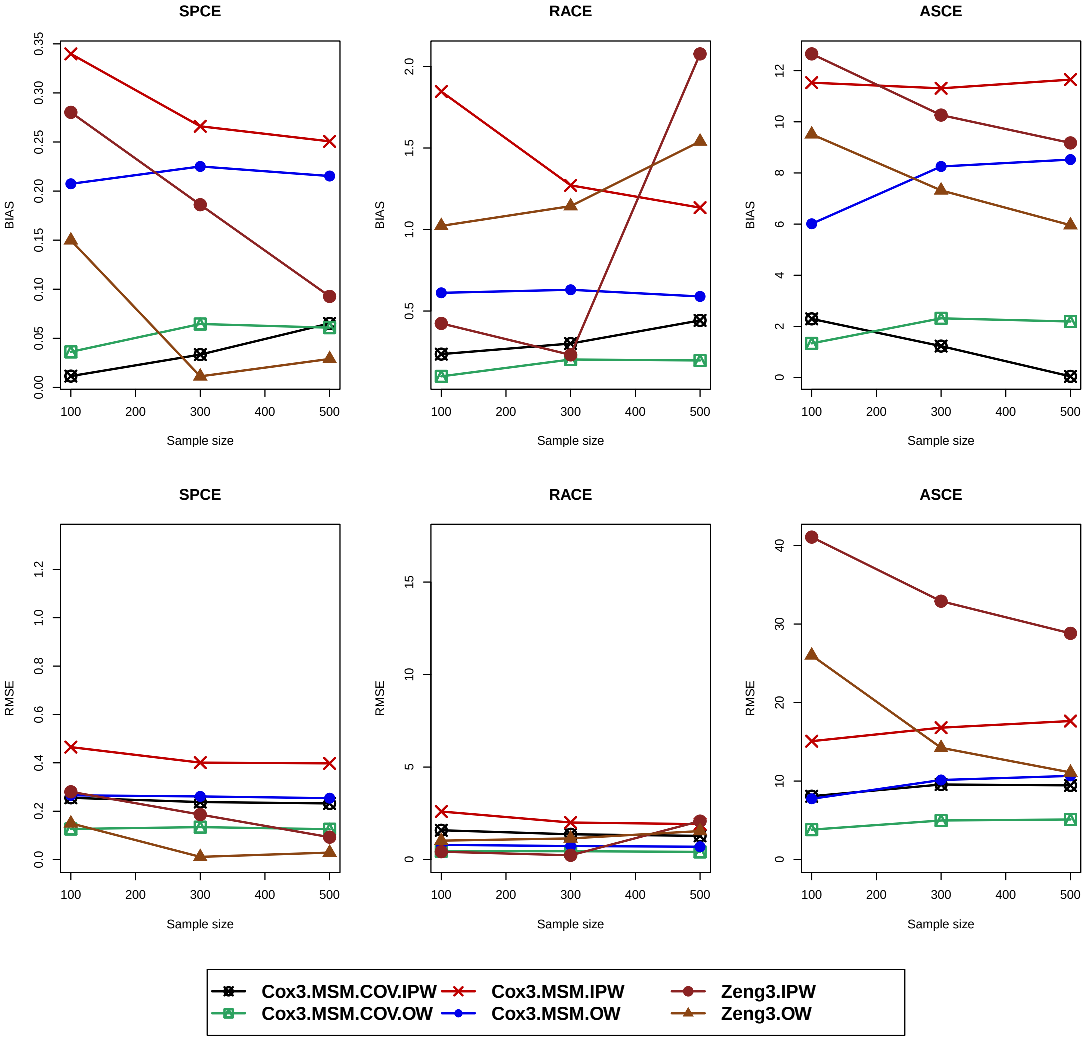


Figure 28: Absolute bias and RMSE comparing two treatment under *poor overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

censoring mechanisms, Cox3.MSM.COV.OW is more efficient than the alternative estimators across all three estimands (SPCE, RACE, and ASCE). Even in the presence of a non-proportionality assumption and misspecification of censoring weights, the estimator computed by MSM with all main effects remains more efficient than alternative approaches, as determined by two *exploratory* sensitivity analyses utilizing our simulation study. In general, estimators based on OW perform well across all criteria. The results help us to determine which method is most effective in resolving research questions with the MS data set.

In our data set, there are two types of censoring: administrative censoring and censoring due to switching treatment. In the presence of staggered entry (i.e., with a varying start of follow-up date) and not-fixed end of follow-up, administrative censoring is unlikely to be CCAR (censoring completely at random). This is because censoring time for all patients is not fixed, and it depends on patient characteristics in our case. For instance, young adults are at higher risk of suffering from MS and need treatment than older adults. This implies that the event is often observed in younger than in old patients. In addition, censored event times due to non-administrative reasons such as treatment switching is unlikely to be CCAR. This is because, in our data set, lack of efficacy, side effects, risk of long-term adverse events, and pregnancy may cause individuals to switch treatments. Consequently, the two censoring mechanisms may be driven by different covariates or are more or less likely ignorable. For instance, the higher EDSS, the more disabled a patient is. Thus, it is likely that the type of treatment is changed according to the values of EDSS. On the other hand, patients with low EDSS at the beginning of the study and having stable EDSS during follow-up are likely to neither worsen the disease nor switch the treatment. Besides, all other main effects (Age,ARR_pre, Disease_durat, Dummy_EDSS, Gender, PI_pre, Relaps_pre, Relapse_Dummy, Year) make likely that the two censoring mechanisms are ignorable suggested to consider different specifications for the weights for the two censoring mechanisms. Thus, it seems that using the proposed weights in the Marginal Structural Cox model is the reasonable option to assess the effectiveness of two treatments in real application. Due to the proper performance of Marginal Structural Cox models with all main effects, we will consider it in the real application. It is worth noting that using censoring weights (both administrative and switching) generally makes the censoring assumption more plausible, thereby leading to a higher chance of obtaining the best fit via MSMs based on the proportionality assumption. Therefore, in the real application, we must assess this assumption while analyzing the MS data set.

Chapter 6

Real Application (Part I)

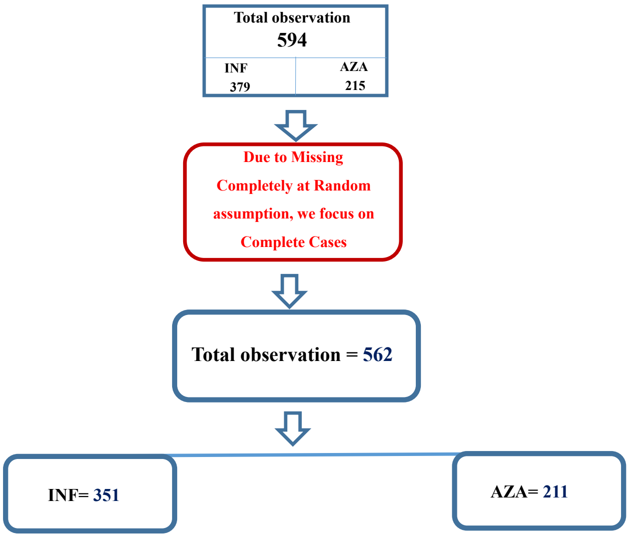
In this chapter, we are interested in assessing the effectiveness of two treatments (INF and AZA) on time to the first worsening of the MS disease based on the Marginal Structural Cox Model. We consider all main effects in constructing censoring weights to prevent overfitting or improve model fit in a meaningful way. After considering the main effects, functional forms of covariates are examined by Martingale and Deviance residuals as well as backward stepwise regression to include in the Cox model constructing censoring weights. Finally, we apply weights in the Marginal structural cox model to assess the efficiency of Interferon (INF) and Azathioprine (AZA) on time to the first worsening of the MS disease.

1 Descriptive and preliminary analysis

Multiple Sclerosis (MS) is a potentially disabling disease of the brain and spinal cord (central nervous system). MS symptoms vary mainly depending on the extent of nerve damage and which nerves are impacted. Multiple sclerosis has no known cure. Treatment primarily focuses on accelerating recovery after attacks, reducing disease progression, and controlling MS symptoms.

The goal of this study is **to examine the efficacy of** two treatments: Interferon (INF) and Azathioprine (AZA) on time to the first worsening of the disease. **594** patients were enrolled between **1981** and **2019** in an observational study in Italy. The description of the data set is presented in Chapter 2. As shown in Chapter 2 and Figure 1, some patients switched from one treatment to another multiple times due to lack of efficacy, side effects, risk of long-term adverse events, and pregnancy. Some covariates are recorded for each unit in this study, and their summary statistics are presented in Table 1. Furthermore, some box plots for covariate are prepared in Figure 2. Clearly, there are some statistical differences between the two treatment arms in some covariates. A primary outcome in this

Figure 1: The schema of data set


Total observation
594
INF
379
AZA
215
Due to Missing Completely at Random assumption, we focus on Complete Cases
Total observation = 562
INF= 351
AZA= 211

study is the "`time to the first worsening of the disease`". In detail, we focus on `Progression-Free Survival (PFS)`, which is the time from treatment initiation until disease progression or worsening. Summary statistics for PFS are illustrated in Table 2.

The staggered entry in our study indicates that the start of follow-up for individuals is varied, leading the time to the first worsening to be subject to `censoring` due to the end of the follow-up (i.e. administrative censoring). Therefore, administrative censoring is unlikely to be CCAR. This is because censoring time is not fixed for all patients, and it depends on patient characteristics. Additionally, it is doubtful that censored event times due to non-administrative reasons, such as treatment switching, are CCAR. This is due to the fact that, according to our data set, lack of efficacy, side effects, the risk of a long-term adverse event, and pregnancy may cause patients to switch treatments. As a result, the two censoring mechanisms may be influenced by covariates. In the presence of two dependent-censoring mechanisms in the observational study, in this thesis, we are interested in assessing the effectiveness of the treatments as initially administered. Therefore we consider all patients' *first* records.

Unadjusted survival curves of the two treatments of Multiple Sclerosis (MS) disease are shown in Figure 3. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (here, a worsening of the disease) at any time. According to the output of Table 3, there is no reason to reject the null hypothesis.

Table 1: Summary statistics of covariates for all units ($N = 562$)

Covariates	*median*	*mean*	*sd*	*description*	X_{AZA}	X_{INF}	*Z.test*	*P.value*
`Age`	34.00	35.5	10.20	Age	36.45	34.92	-1.71	0.08
`ARR_pre`	0.53	1.31	4.59	Annualized Relapse Rate (ARR): average number of attacks per year	0.83	1.60	2.46	0.01
`Disease_durat`	48.00	73.80	73.2	Disease duration (months)	85.62	66.67	-2.83	0.00
`Dummy_EDSS`	0.00	0.08	0.27	0 if Baseline. EDSS < 4 and 1 otherwise	0.08	0.08	0.00	1
`Baseline.EDSS`	1.50	1.64	1.34	Expanded Disability Status Scale (EDSS $\in (0, 10)$)	1.80	1.54	32479	0.01
`Gender`	1.00	0.71	0.46	0 for Male and 1 for Female	0.71	0.70	0.00	0.93
`PI_pre`	0.28	0.99	2.81	The progression index	0.75	1.14	1.81	0.07
`Relaps_pre`	2.00	2.71	2.37	Number of relapse before therapy	2.85	2.63	-1.08	0.27
`Relapse_Dummy`	0.00	0.03	0.17	1 if `Relapse_pre` is missing and 0 otherwise	0.03	0.03	0.06	0.79
`Year`	2005	2004.30	5.86	The year of treatment started	2002.68	2005.44	5.29	0.00

Table 2: Summary statistics for Progression Free Survival (PFS) in months

Outcome	*min*	*max*	25th *quantile*	*median*	75th *quantile*	*administrative censoring rate*	*switching censoring rate*	*median* INF	*median* AZA
`PFS`	2	196	18.25	41	72	0.64	0.12	37	48

Table 3: Log rank test

	Treatment	
	INF	AZA
N	351	211
Observed (O)	83	49
Expected (E)	77.4	54.6
$(O - E)^2/E$	0.412	0.584
$(O - E)^2/V$	1.02	1.02

p-value = 0.3

Figure 2: Box plots of Covariates

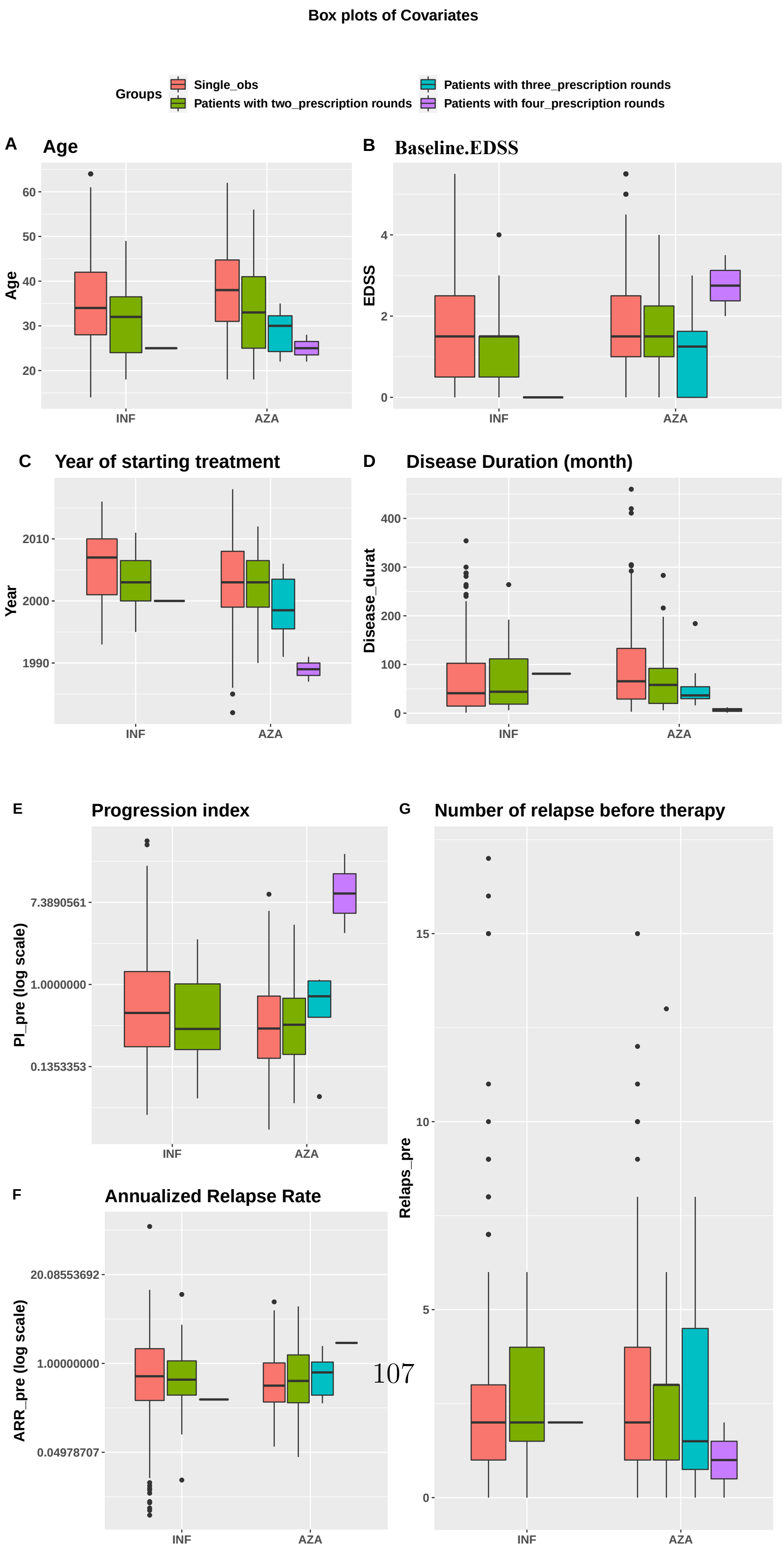
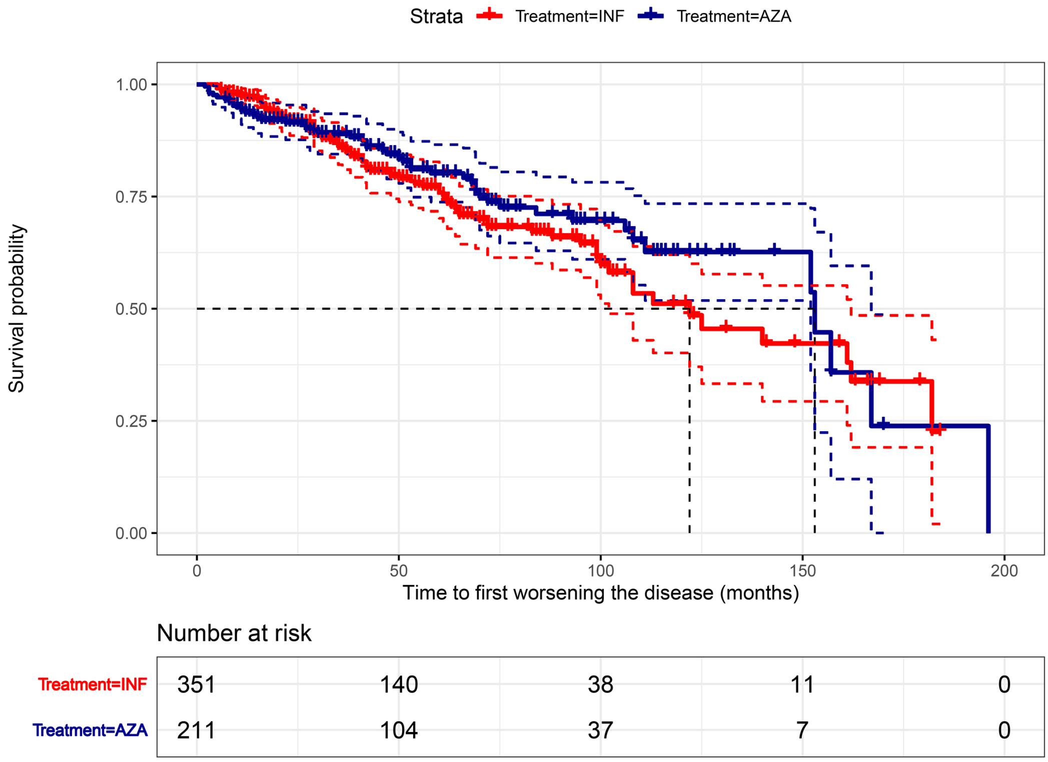


Figure 3: **Survival curves of the two treatments of Multiple Sclerosis (MS) disease (unadjusted)**



Treatment=INF	351	140	38	11	0
Treatment=AZA	211	104	37	7	0

2 Construct weights and diagnostics for assessing the modeling assumptions

Introducing novel weights in Marginal Structural Cox models to adjust for both observed confounding and dependent-censoring is the core of this thesis. The parameters of a marginal structural model can be consistently estimated using weighted estimators. To do so, this section describes how to construct weights and diagnostics for assessing assumptions in the assignment mechanism (Subsection 2.1) and the censoring mechanism (Subsection 2.2).

2.1 The weights for causal effect estimation

In the context of weighting methods, it is common to use logistic regression to estimate the propensity score that is later used to form the weights. In practice, we usually do not know the proper form of the propensity score model. The typical work is to start with logistic regression with the main effects of all the covariates. To choose which interaction and higher-order should be included, it is common to use computationally faster *forward, backward, and hybrid stepwise* approaches. All available criteria have in common that they prevent overfitting by penalizing models that contain many predictors unless these improve model fit in a meaningful way. Where stepwise regression must be used, backward elimination is generally preferable to forward selection as it has been shown to perform better (particularly in the presence of collinearity) and forces the researcher to start with a fully fitted model [Harrell, 2001]. Therefore, the backward selection based on AIC is performed for choosing the nonlinear terms to include in the models for the assignment mechanism. The stepwise approach is useful because it is computationally fast to find the "best" model across all possible models, reducing the multicollinearity problem, and it is one of the ways to resolve overfitting.

Since the objective is to adjust for all observed confounding variables, it is desirable not to exclude any main effects. However, some selection may be required to consider higher-order terms and interactions. The results of the backward stepwise regression based on AIC are shown in Table 4. Accordingly, the "best" PS model for constructing treatment weight includes the following:

- all main effects: `Age, Gender, Baseline.EDSS, Dummy_EDSS, Relapse_pre, Relapse_Dummy, Disease_durat, ARR_pre, PI_pre, Year`
- and some interaction and higher order terms: $(\texttt{Year})^2$, $(\texttt{Relapse\_pre})^2$, $(\texttt{Year})^3$, $(\texttt{Baseline.EDSS})^3$, $(\texttt{Baseline.EDSS})^2$, $(\texttt{Disease\_durat})^3$, $(\texttt{Disease\_durat})^2$, $(\texttt{Relapse\_pre})^3$, `Dummy_EDSS:Baseline.EDSS`, `Dummy_EDSS:ARR_pre`)

After selecting the specific form of the propensity score model, the balance must be checked. Utilizing the propensity score is one of the most common methods to

Table 4: The backward stepwise regression for treatment model

Start: AIC = 715.13 Stop: AIC = 693.55			
	Df	Deviance	AIC
< none >		667.55	693.55
− I(Relaps_pre^3)	1	669.62	693.62
− I(Baseline.EDSS^3)	1	669.93	693.93
− I(Disease_durat^3)	1	670.07	694.07
− I(Relaps_pre^2)	1	670.21	694.21
−Dummy_EDSS:Baseline.EDSS	1	670.74	694.74
− I(Year^3)	1	670.88	694.88
− I(Year^2)	1	670.89	694.89
−Year	1	670.91	694.91
−Dummy_EDSS:ARR_pre	1	676.89	700.89

assess the balance of a multivariate distribution. Any imbalance in the population covariate distributions, whether in expectation, dispersion, or shape, leads to a difference in the population distributions of correct propensity scores by treatment status. Thus, we proceed by estimating the propensity score through logistic regression. After estimating the PS, the overlap (region of common support) of the distributions of the estimated PS for the treatment and comparison group should be examined graphically. Figure 5 is provided to check whether there is sufficient overlap between the two treatment groups or not. It seems there is a good overlap between the two treatment arms (left side of Figure 5). According to the Love plot (right side of Figure 5), the standardized differences (mean difference divided by pooled standard deviation) are reduced substantially after propensity score weighting adjustment.

To diagnostics for assessing the assumptions of treatment weights, Cole and Hernán [2008] proposed to check model misspecifications by exploring the distribution of weights. In particular, positivity and no model misspecifications can be explored by evaluating the sensitivity of inferences to truncating extreme weights, as illustrated in Tables 6–5 and Figure 4. According to Table 5, the mean stabilized weight was equal to 0.99, while the standard deviation of the stabilized weights was equal to 0.44. The minimum and maximum weights were 0.38 and 5.95, respectively. There was no evidence of non-positivity or misspecification of the propensity score model based on an examination of the distribution of the weights derived from the specification of the propensity score model. Furthermore, Table 6 illustrates that all mean are close to 1.00, and there are no extreme minimum and maximum values. Moreover, Figure 4 confirms a good overlap between

Table 5: Summary statistics of estimated PS ($\hat{e}$) for each treatment arms and diagnostics based on the stabilized IPW.

$\hat{e}_{\text{AZA}}$		$\hat{e}_{\text{INF}}$		Stabilized IPW			
mean	sd	mean	sd	mean	sd	min	max
0.46	0.19	0.33	0.14	0.99	0.44	0.38	5.95

Table 6: Check assumptions based on the stabilized IPW

Truncation	**Estimated weights**			
percentiles	mean	sd	min	max
0-100	0.99	0.44	0.38	5.95
1-99	1.00	0.44	0.38	5.97
5-95	1.00	0.42	0.41	5.60
10-90	1.00	0.36	0.44	4.08
25-75	1.00	0.29	0.42	3.08
45-55	1.01	0.90	0.50	7.27

the two treatment arms, and these assumptions are held.

By choosing the propensity score model specification and checking the balance, one can adjust for observed confounders by weighting estimators for causal effects. The following section addresses the covariate-dependent censoring assumption.

2.2 Censoring weights

The covariate-dependent assumption outlined in Chapter 5 is one of the most critical assumptions underlying censoring mechanisms. To accomplish this, a multivariate Cox proportional hazards model must be applied to administrative and switching censoring. Nevertheless, prior to doing so, the *proportional hazards assumption* must be held.

Statistical tests and graphical diagnostics can be used based on the scaled Schoenfeld residuals to check the proportional hazards assumption. The results are shown in Table 7 and Figures 6–7. From the output of the test, all covariates are not statistically significant, and the global test is also not statistically significant. Therefore, we can assume proportional hazards. Theoretically, the Schoenfeld residuals [Grambsch and Therneau, 1994] are independent of time. We plot the Schoenfeld residuals against survival times to graphically assess proportionality. If the assumption is satisfied, the Schoenfeld residuals should approximately scatter

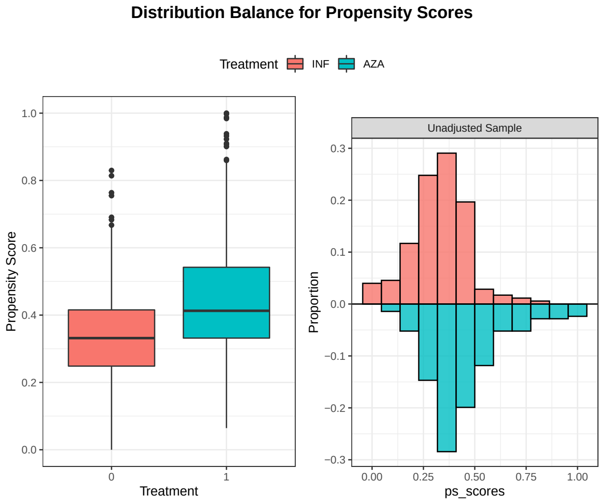


Figure 4: Distribution of balance for Propensity Scores

around 0 [Grambsch and Therneau, 1994]. From the visual inspection (Figures 6–7), there is no pattern with time. The assumption of proportional hazards appears to be supported by the covariates.

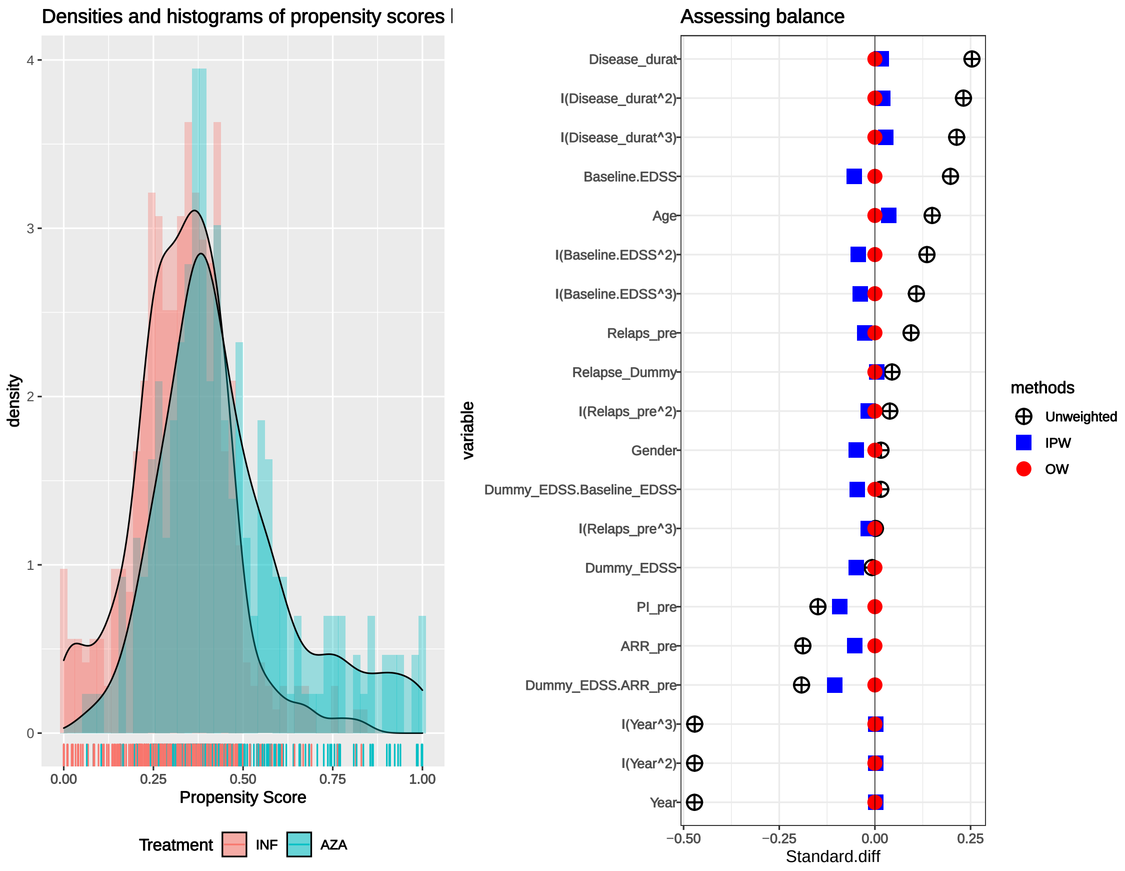


Figure 5: Left: check overlap based on the propensity scores for treatment groups. Right: (love plot) standardized differences between treatment groups before and after propensity score weighting adjustment. IPW: inverse probability weighting, OW: overlap weight

Table 7: Test proportional Hazards assumption

	administrative censoring			switching censoring		
Baseline Covariates	*chisq*	*df*	*p*	*chisq*	*df*	*p*
Age	0.02	1.00	0.88	0.11	1.00	0.74
ARR_pre	1.06	1.00	0.30	0.30	1.00	0.58
Disease_durat	1.25	1.00	0.26	0.50	1.00	0.48
Dummy_EDSS	2.21	1.00	0.14	0.49	1.00	0.48
Baseline.EDSS	1.89	1.00	0.17	1.30	1.00	0.25
Gender	0.54	1.00	0.46	0.60	1.00	0.44
PI_pre	0.86	1.00	0.35	2.34	1.00	0.13
Relaps_pre	0.29	1.00	0.59	0.04	1.00	0.84
Relapse_Dummy	0.84	1.00	0.36	0.06	1.00	0.81
Year	0.00	1.00	1.00	3.23	1.00	0.07
GLOBAL	14.39	10.00	0.16	9.51	10.00	0.48

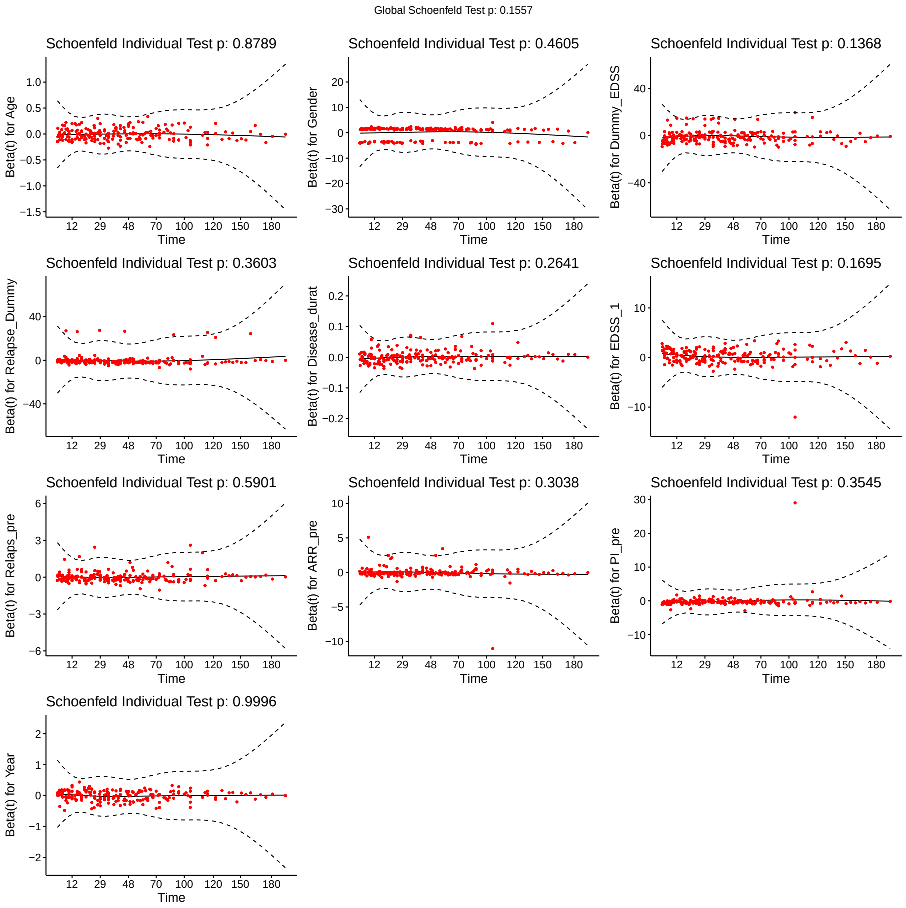


Figure 6: The Schoenfeld residuals for administrative censoring.

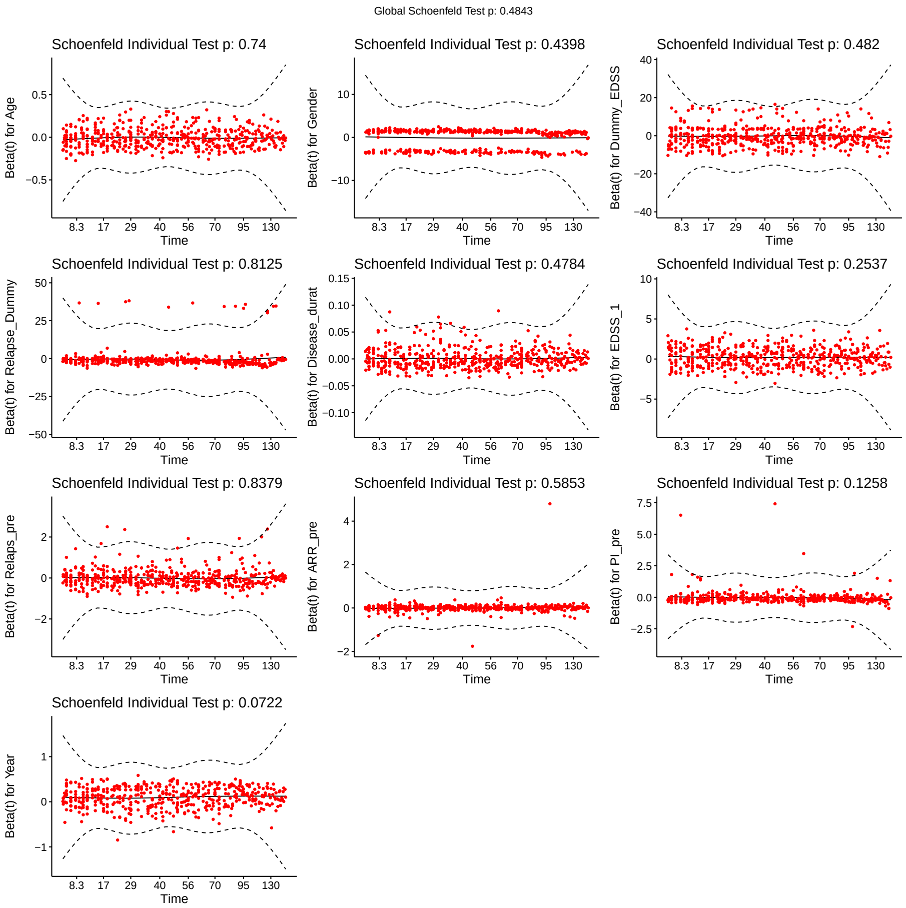


Figure 7: The Schoenfeld residuals for switching censoring.

After holding the proportionality hazard assumption, a multivariate Cox model could be utilized for administrative and switching censoring. According to Tables 8 and 9, both censoring mechanisms are ignorable to different sets of covariates. In the presence of covariate-dependent censoring, as described in Chapter 2, the most well-known method to deal with is the Inverse Probability of Censoring Weighted (IPCW). This thesis focuses mostly on correcting selection bias caused by covariate-dependent censoring by assigning additional weight to individuals who are not censored for a considerable time. In practice, a Cox proportional hazard model is assumed, and an inverse probability of censoring weight is applied to the Cox model score equation. Consequently, using two separate weights to address assumptions dependent on covariates seems plausible.

Table 8: covariate-dependent assumption for administrative censoring

	beta	HR(95% CI)	wald.test	p.value
Age	-0.00	1.00 (0.98-1.00)	-0.55	0.58
Gender1	0.08	1.10 (0.79-1.50)	0.48	0.63
Disease_durat	0.00	1.00 (1.00-1.00)	0.29	0.77
Baseline.EDSS	0.24	1.30 (1.10-1.50)	2.90	0.00**
Dummy_EDSS1	-0.69	0.50 (0.25-0.99)	-2.00	0.05*
Relapse_Dummy1	-0.11	0.90 (0.43-1.90)	-0.28	0.77
Relaps_pre	0.01	1.00 (0.95-1.10)	0.45	0.66
ARR_pre	-0.01	0.98 (0.88-1.10)	-0.28	0.77
PI_pre	-0.15	0.86 (0.74-1.00)	-1.80	0.06$^{\cdot}$
Year	0.00	1.00 (0.97-1.00)	-0.01	0.99

Table 9: covariate-dependent assumption for switching censoring

	beta	HR(95% CI)	wald.test	p.value
Age	-0.00	0.99 (0.98-1.00)	-1.60	0.10
Gender1	-0.09	0.91 (0.75-1.10)	-0.91	0.37
Disease_durat	0.00	1.00 (1.00-1.00)	0.98	0.33
Baseline.EDSS	0.22	1.20 (1.10-1.40)	4.10	0.00**
Dummy_EDSS1	-0.24	0.79 (0.51-1.20)	-1.00	0.29
Relapse_Dummy1	-0.52	0.59 (0.34-1.00)	-1.80	0.06$^{\cdot}$
Relaps_pre	-0.00	1.00 (0.96-1.00)	-0.11	0.92
ARR_pre	0.00	1.00 (0.98-1.00)	0.32	0.75
PI_pre	-0.03	0.96 (0.92-1.00)	-1.70	0.09
Year	0.10	1.10 (1.10-1.10)	11.00	0.00**

Generally, the cox model is used to calculate IPCW for each censoring mechanism, as described in Chapter 5 (Section 2.2). Diagnostics for the Cox model are required for this purpose (as mentioned in Chapter 4, Section 3).

Martingale residuals are very useful for diagnostics for the Cox model to identify outliers, choose a functional form for the covariate, etc. Since non-linearity is not an issue for categorical variables, we only examine plots of martingale residuals against a continuous variable. Figures SM–13-SM–14 (in Supplementary Material) represent Martingale residuals to choose a functional form for the continuous covariate based on administrative and switching censoring. Moreover, we plot the martingale and deviance residuals to check for possible outliers as discussed in Therneau et al. [1990]. The large outlier (See Figure 8) is a woman with Baseline.EDSS equal to 3 assigned to AZA and never switched the treatment, yet survived 196 months. Nevertheless, the primary drawback to the martingale residual is its clear asymmetry (its upper bound is 1, but it has no lower bound). Therefore, it is necessary to detect outliers by deviance residuals.

A technique for creating symmetric, normalized residuals widely used in generalized linear modelling is “deviance residual”. According to the deviance residual's plots (Figures 9), there are no extreme outliers; the largest residuals are only 3 standard deviations away from zero. Furthermore, Figures SM–15-SM–16 (in Supplementary Material) represent Deviance residuals to choose a functional form for the continuous covariate based on administrative and switching censoring.

In the plot of delta-beta residuals (Chapter 4, Section 3), the estimated changes in the regression coefficients upon deleting each observation are determined. Comparing the magnitudes of the dfbeta values to the regression coefficients (Figures SM–17-SM–18 in Supplementary Material), none of the observations is influential individually, even though some of the dfbeta values are rather large.

Although there is some visual inspection of finding the best functional form of covariates such as Martingale and Deviance residuals (Figures SM–13-SM–16 in Supplementary Material) in constructing IPCW, criteria for model fit, such as estimates of the expected prediction error obtained via cross-validation, or information criteria like the AIC are more useful to compare the extent to which different regression models fit the data well [Harrell, 2001]. In this regard, some computationally efficient forward, backward, and hybrid stepwise algorithms are used to temper concerns about misspecification. It has been demonstrated that backward elimination performs better than forward selection (especially in collinearity) and drives the researcher to start with a fully fitted model [Harrell, 2001]. In this regard, the first possibility is to consider all of the main effects in addition to some nonlinear terms computed by backward selection. The backward selection based on AIC is performed for both censoring models. Tables 10–11 are prepared to display the selected nonlinear terms for both censoring mechanisms. Although some main effects are eliminated due to backward regression in Tables 10—11, we maintain all main effects when creating IPCW. It is worth mentioning that the IPCW is constructed using the Cox model. As a result, the proportional hazards assumption of all covariates (main effects and nonlinear terms) must be examined. Table 12 prepared for this manner. All covariates satisfy the individual proportionality test at the 0.05 level, and the model meets the global proportionality test according to the test results. Consequently, depending on the best-selected form of covariates, we create the IPCW for each censoring mechanism.

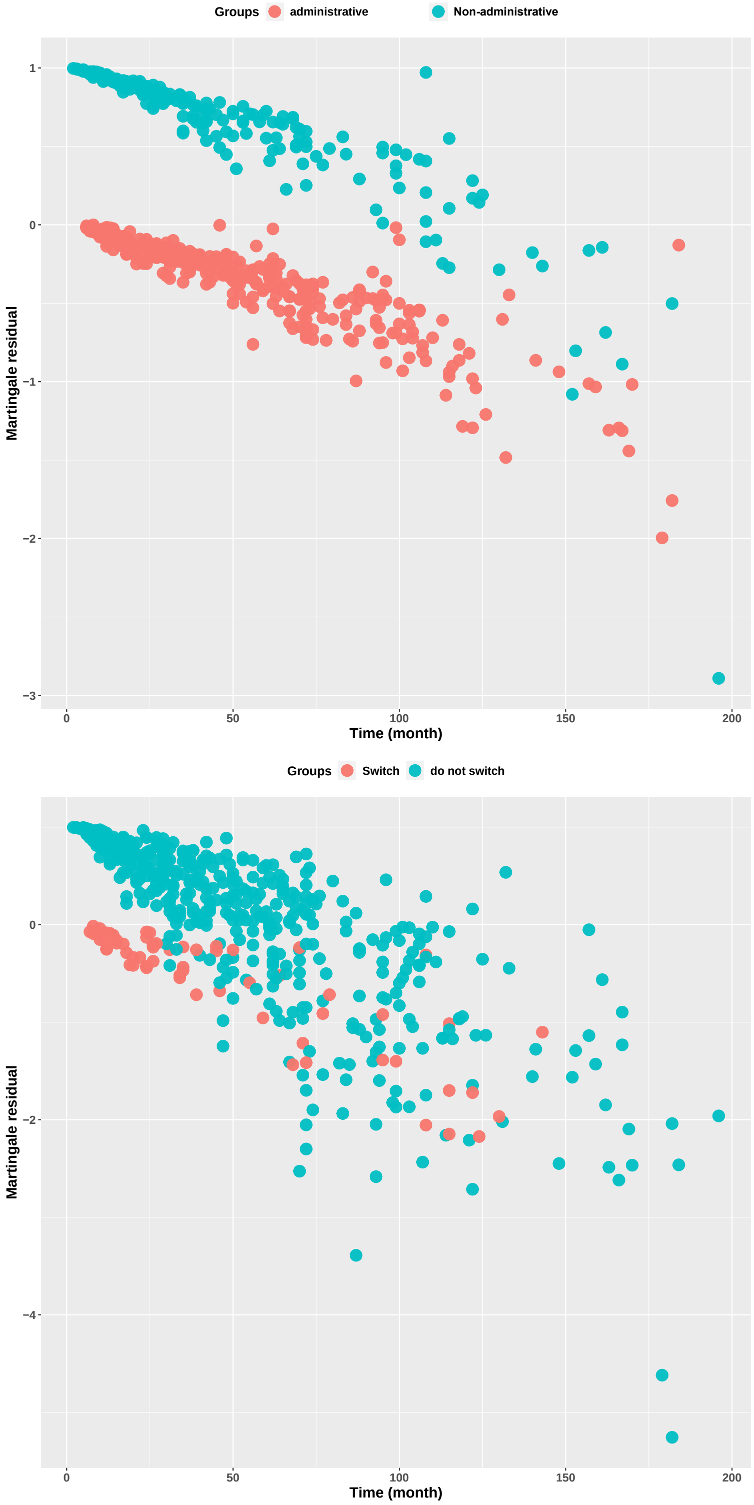


Figure 8: The Martingale residuals to detect outlier based on (above) administrative censoring and (below) switching the treatment.

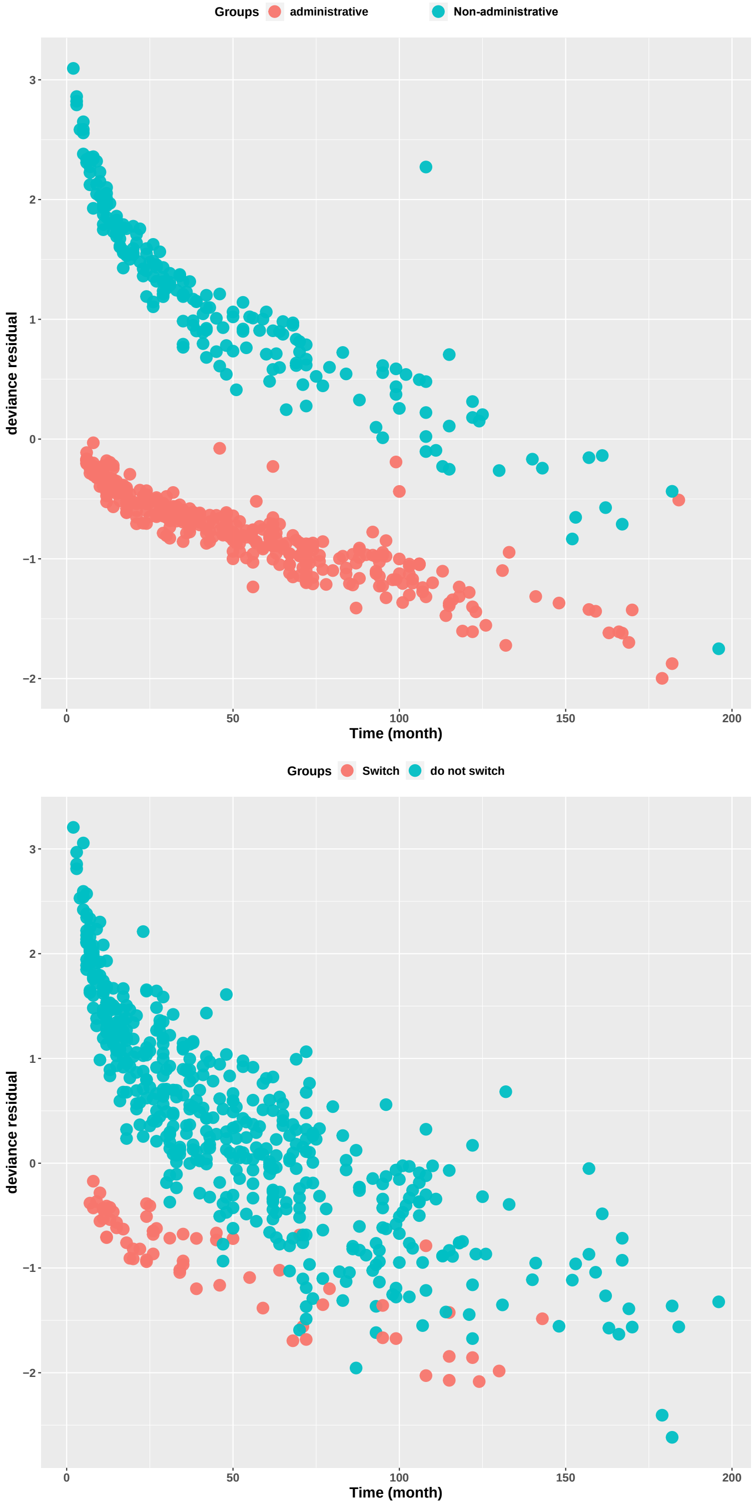


Figure 9: The deviance residuals to detect outlier based on (above) administrative censoring and (below) switching the treatment.

Table 10: The backward stepwise regression for administrative censoring model based on all covariates.

Start: AIC = 2140.521

Stop: AIC = 2124.37

Selected Model:

`surv_object` $\sim$ `Baseline.EDSS` + `PI_pre` + `Dummy_EDSS`+
`Dummy_EDSS:Baseline.EDSS`

	Df	AIC
< `none` >		2124.4
−`Dummy_EDSS:Baseline.EDSS`	1	2127.6
−`PI_pre`	1	2138.1

Table 11: The backward stepwise regression for switching censoring model based on all covariates.

Start: AIC = 5147.26

Stop: AIC = 5135.97

Selected Model:

`surv_object` $\sim$ `Relapse_Dummy+Baseline.EDSS` + `PI_pre + Year +`
`Baseline.EDSS:Year + PI_pre:Year`

	Df	AIC
< `none` >		5136.0
−`Baseline.EDSS:Year`	1	5136.9
−`Relapse_Dummy`	1	5137.8
−`PI_pre:Year`	1	5140.1

Table 12: Proportional hazards assumption of all covariates

Baseline Covariates	*chisq*	*df*	*p*
Administrative censoring			
Age	0.00	1.00	0.99
Gender	0.64	1.00	0.42
Dummy_EDSS	3.18	1.00	0.07
Relapse_Dummy	0.82	1.00	0.37
Disease_durat	1.68	1.00	0.20
Baseline.EDSS	2.81	1.00	0.09
Relaps_pre	0.24	1.00	0.63
ARR_pre	0.96	1.00	0.33
PI_pre	0.85	1.00	0.36
Year	0.00	1.00	0.99
Dummy_EDSS:Baseline.EDSS	3.46	1.00	0.06
GLOBAL	17.51	11.00	0.09
Switching censoring			
Age	0.04	1.00	0.83
Gender	0.89	1.00	0.35
Dummy_EDSS	1.50	1.00	0.22
Relapse_Dummy	0.05	1.00	0.83
Disease_durat	0.01	1.00	0.92
Baseline.EDSS	2.69	1.00	0.10
Relaps_pre	0.01	1.00	0.91
ARR_pre	1.66	1.00	0.20
PI_pre	0.04	1.00	0.84
Year	3.67	1.00	0.06
Baseline.EDSS:Year	2.67	1.00	0.10
PI_pre:Year	0.04	1.00	0.84
GLOBAL	8.97	12.00	0.71

3 Application of Marginal Structural Model to the MS dataset

The marginal structural model is an effective method for providing consistent causal effect estimators. To calculate the survival probability to assess the relative effectiveness of two `Treatments`: Interferon (INF) and Azathioprine (AZA) on Progression-Free Survival (PFS), consider the following steps:

1. Fit a logistic regression model with

$$\operatorname{logit}(\Pr(Z=1 \mid X)) = \delta_0 + \delta_1' X$$

   with X= (`Age, Gender, Baseline.EDSS, Dummy_EDSS, Relapse_pre, Relapse_Dummy, Disease_durat, ARR_pre, PI_pre, Year, (Year)`^{2}`, (Relapse_pre)`^{2}`, (Baseline.EDSS)`^{3}`, (Baseline.EDSS)`^{2}`, (Disease_durat)`^{2}`, (Disease_durat)`^{3}`, (Relapse_pre)`^{3}`, (Year)`^{3}`, Dummy_EDSS:Baseline.EDSS, Dummy_EDSS:ARR_pre`) to estimate the propensity score and then compute assignment weights $(\omega_i^{ipw}, \omega_i^{ow})$ as

   Stabilized IPW weights: $$\omega_i^{ipw} = \frac{Z_i}{e(X_i)} + \frac{1-Z_i}{1-e(X_i)}$$

   Overlap weights: $$\omega_i^{ow} = (1-Z_i)\, e(X_i) + Z_i\, (1-e(X_i))$$

2. Fit a Cox proportional hazard model for administrative censoring time

$$\lambda^C(t \mid Z, X) = \lambda_0^C(t) e^{\gamma_1 Z + \gamma_2' X}$$

   We consider all main effects as well as some nonlinear terms according to Table 10 as X=(`Age`, `Gender`, `Baseline.EDSS`, `Dummy_EDSS`, `Relapse_pre`, `Relapse_Dummy`, `Disease_durat`, `ARR_pre`, `PI_pre`, `Year`, `Dummy_EDSS:Baseline.EDSS`) to compute $\omega_i^{sc} = \dfrac{\Pr(C_i > t)}{\Pr(C_i > t \mid X_i)}$.

3. Fit a Cox proportional hazard model for switching censoring time

$$\lambda^S(t|Z, X) = \lambda_0^S(t) e^{\nu_1 Z + \nu_2' X}$$

   We consider all main effects as well as some nonlinear terms according to Table 11 as X=(`Age`, `Gender`, `Baseline.EDSS`, `Dummy_EDSS`, `Relapse_pre`, `Relapse_Dummy`, `Disease_durat`, `ARR_pre`, `PI_pre`, `Year`, `Baseline.EDSS:Year` `PI_pre:Year`) to calculate $\omega_i^{ss} = \dfrac{\Pr(S_i > t)}{\Pr(S_i > t \mid X_i)}$

4. Assign weights for each unit, i.e. $\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \times \omega_i^{sc}$ or $\hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss} \times \omega_i^{sc}$

5. Fit Marginal Structural Cox model

$$h(t|Z_i) = h_0(t) \exp\left\{X_i \alpha^\top + \gamma Z_i\right\}$$

6. Calculate the survival probability $\hat{\mathbb{S}}(t|Z_i)$ specific to each treatment.

7. Finally, for each pre-specified time point (t^*), calculate **causal estimands** as

$$\Delta_{\text{RACE}} = \int_0^{t^*} \hat{\mathbb{S}}_1(t)dt - \int_0^{t^*} \hat{\mathbb{S}}_0(t)dt$$

$$\Delta_{\text{SPCE}} = \hat{\mathbb{S}}_1(t^*) - \hat{\mathbb{S}}_0(t^*)$$

4 Results

We illustrate the performance of the proposed weighting estimators by comparing two treatments for MS observational data set. The median and maximum follow-up times are 41 and 196 months, respectively.

Figures 10 depict estimated causal survival curves using the Marginal Structural Cox Model. Importantly, Figure 10(c) is shown based on unadjusted samples and Figures 10(a)–10(b) illustrate the survival curves for two treatments after adjustment. According to the results, the AZA shows a slightly more survival benefit during the follow-up. Nonetheless, the estimated causal survival curves among the two target populations are generally similar. Furthermore, Figures 11 characterized the SPCE and RACE as a function of time t with the associated 95% confidence intervals in the pseudo-population (corresponding to IPW) and the overlap population (corresponding to OW). We have chosen 196 grid points equally spaced by a month for this evaluation. Figures 11(a) provide some evidence of the possible beneficial effect of AZA over INF in terms of SPCE and RACE before being adjusted. After adjusting for observed confounding and dependent-censoring assumption, both SPCE and RACE show some evidence of a possible beneficial effect of AZA over INF (See Figures 11(b)). However, all adjusted cases confirm that the difference is never statistically significant. This is because the confidence intervals of SPCE and RACE from IPW and OW straddle zero across the entire follow-up period. Therefore, we can not infer much from the sign of the estimated coefficients since the confidence intervals cover zero. In general, there is no significant causal survival benefit of AZA over INF at the 0.05 level. This analysis is essential because it shows no superiority of INF over AZA when properly adjusting for confounding and censoring.

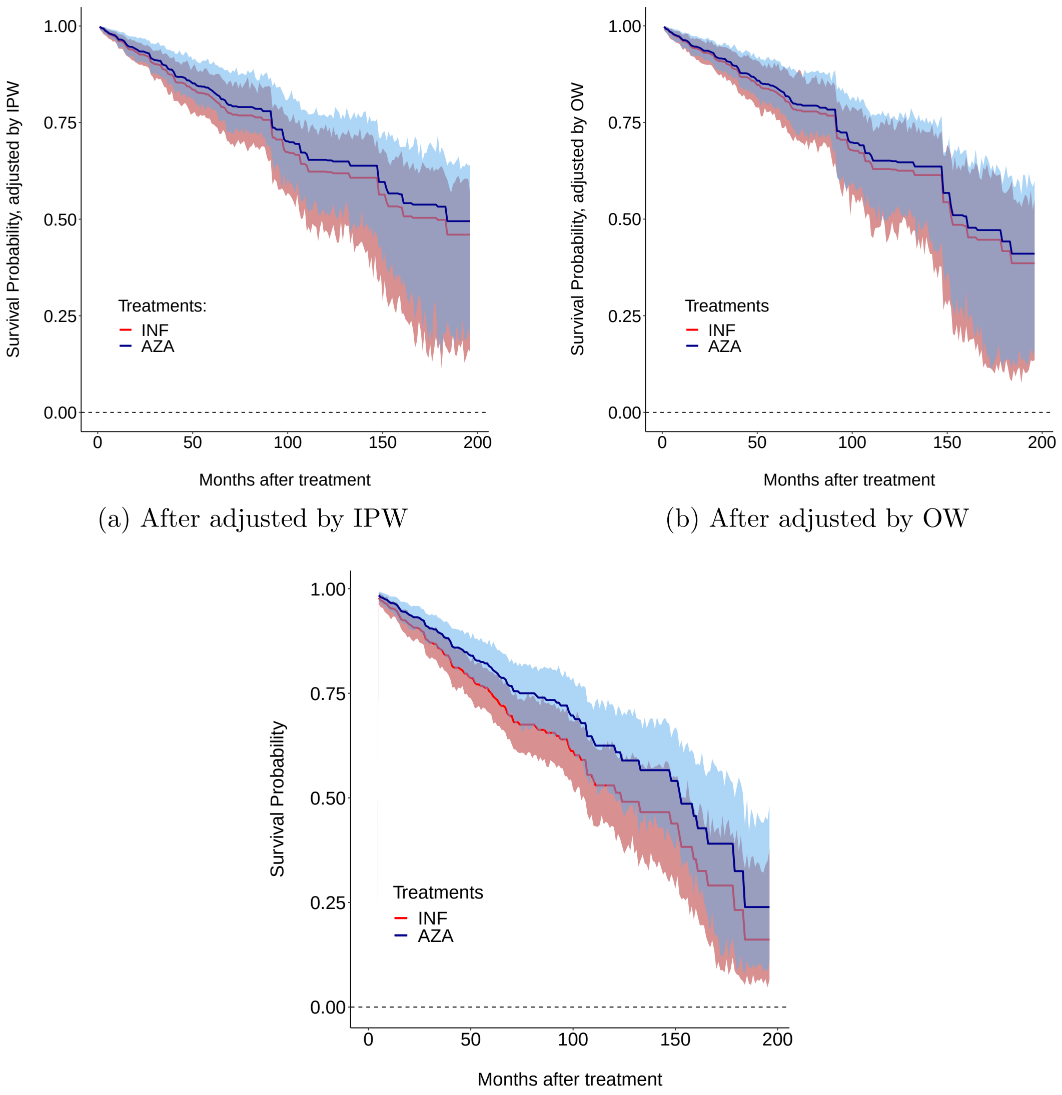
In Table 13, we also reported the SPCE and RACE using the (`unadjusted / adjusted`) MSMs at $t = 24$ months, $t = 41$ months, i.e. the median of follow-up and $t = 72$ months, i.e. the 75^{th} quantile of the follow-up time. Before adjusted, AZA increased the restricted average causal effect of the population at 24, 41 and 72 months by 0.182, 0.523 and 1.608 months, respectively, improving the survival probability from 24 to 72 months, around 1.8%-4.5%. However, after adjusting the pseudo-population (corresponding to IPW), AZA also has expanded the restricted average causal effect on 24, 41 and 72 months by 0.1, 0.282 and 0.829 months, respectively. It seems that the survival probability increased around 0.8%-2.2% in 24 and 72 months of follow-up. Although there is some benefit of AZA in decreasing the probability of worsening the disease rate at 24, 41 and 72 months than INF, this difference is not statistically significant. Furthermore, AZA increased the restricted average causal effect on overlap population at 24, 41 and 72 months by 0.073, 0.195 and 0.566 months, respectively, improving the survival probability from 24 to 72 months, around 0.5%-1.5%. Since all confidence intervals of SPCE and RACE from IPW and OW straddle zero, this difference is not statistically significant. Generally speaking, all methods conclude that there is **no statistically significant difference** between AZA and INF in terms of time to the first worsening of the disease.

4.1 Non-inferiority test

In this section, we would like to assess non-inferiority test when adjusting for confounding and censoring. Let π_{INF} and π_{AZA} denote the probability that a patient experiences a progression of disability within two years under INF and AZA respectively. We are interested in assessing whether the AZA treatment is not unacceptably less efficacious than the INF treatment. We deal with this issue by testing the following statistical hypotheses:

$$\begin{aligned}\mathbb{H}_0 &: \pi_{\text{INF}} - \pi_{\text{AZA}} \geq \mathrm{M}\\ \mathbb{H}_1 &: \pi_{\text{INF}} - \pi_{\text{AZA}} < \mathrm{M}\end{aligned}$$

where M is the non-inferiority margin, the maximum acceptable extent of clinical noninferiority of the AZA treatment. The margin must be prospectively defined. To calculate the Margin, we consider the results of Jacobs et al. [1996], which assess the efficacy of INF versus placebo on time to sustained progression of disability for 104 weeks. According to Jacobs et al. [1996], the timing of beneficial effects of INF was explored by determining the probability of sustained progression onset occurring in year one and year 2 for patients in the study. Therefore, the Margin is the difference in survival up to that specific time point (i.e. two years) which is

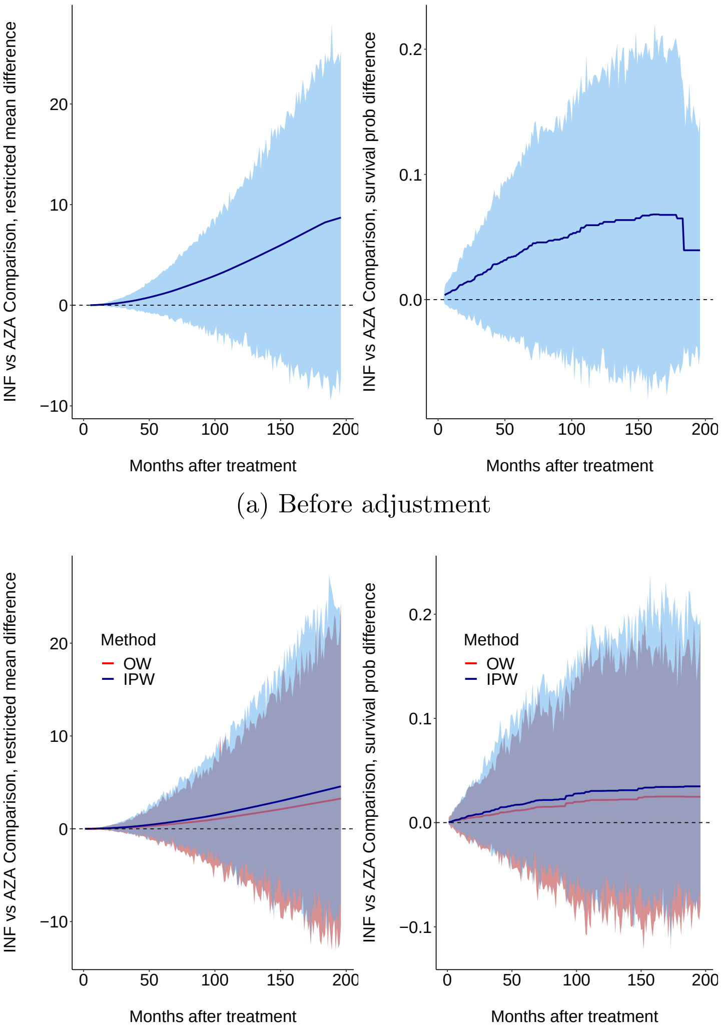


(a) After adjusted by IPW

(b) After adjusted by OW

(c) Before adjustment

Figure 10: Estimates of the survival curves for two treatments using Marginal Structural Cox Model: (first row) after adjusted by (a) IPW and (b) OW; (second row) (c) before adjustment.



(a) Before adjustment

(b) After adjustment

Figure 11: Point estimates and 95% confidence intervals of SPCE and RACE as a function of time using MSMs (a) before adjustment (b) after adjustment

Table 13: Estimates of the treatment effect using two methods: restricted average causal effect (RACE) and survival probability causal effect (SPCE) at **24** months, **41** months (the median of the follow-up time) and **72** months (the 75^{th} quantile of the follow-up time).

	Method	Estimate	Standard error	%95 Confidence interval	P.value
24 months					
		Restricted average causal effect			
	`Unadjusted`	0.182	0.176	**(-0.183,0.536)**	0.284
	`OW`	0.073	0.198	**(-0.324,0.529)**	0.540
	`IPW`	0.100	0.204	**(-0.251,0.515)**	0.480
		Survival probability causal effect			
	`Unadjusted`	0.014	0.014	**(-0.015,0.042)**	0.284
	`OW`	0.005	0.015	**(-0.025,0.036)**	0.540
	`IPW`	0.008	0.016	**(-0.019,0.042)**	0.480
41 months					
		Restricted average causal effect			
	`Unadjusted`	0.523	0.537	**(-0.646,1.487)**	0.334
	`OW`	0.195	0.543	**(-0.811,1.356)**	0.914
	`IPW`	0.282	0.608	**(-0.801,1.660)**	0.824
		Survival probability causal effect			
	`Unadjusted`	0.028	0.029	**(-0.032,0.081)**	0.334
	`OW`	0.010	0.028	**(-0.039,0.067)**	0.902
	`IPW`	0.015	0.032	**(-0.041,0.089)**	0.786
72 months					
		Restricted average causal effect			
	`Unadjusted`	1.608	1.615	**(-1.379,4.895)**	0.290
	`OW`	0.566	1.555	**(-1.962,3.826)**	0.804
	`IPW`	0.829	1.676	**(-1.780,4.940)**	0.748
		Survival probability causal effect			
	`Unadjusted`	0.045	0.045	**(-0.037,0.141)**	0.290
	`OW`	0.015	0.042	**(-0.055,0.103)**	0.828
	`IPW`	0.022	0.044	**(-0.044,0.127)**	0.726

computed as

$$\mathrm{M} = \frac{0.349 - 0.219}{2} = \frac{0.13}{2} = 0.065.$$

Hence,

$$\begin{aligned} \mathbb{H}_0 &: \pi_{\mathrm{INF}} - \pi_{\mathrm{AZA}} \geq 0.065 \\ \mathbb{H}_1 &: \pi_{\mathrm{INF}} - \pi_{\mathrm{AZA}} < 0.065 \end{aligned} \tag{1}$$

If the upper bound of the 95% confidence interval for $\pi_{\mathrm{INF}} - \pi_{\mathrm{AZA}}$ is less than M, we find evidence that AZA is non *inferior* to INF at 2.5% level. It can be seen from Table 13 (24 months) that the upper bound of the 95% confidence interval of SPCE is less than 0.065 for both weighting methods (`IPW`, `OW`). So, we conclude that there is no inferiority of AZA over INF at 2.5% level.

Chapter 7

Real Application (Part II)

The socioeconomic costs associated with neurodegenerative disorders are a significant concern for all patients. Depending on the type of treatment and patient demographic, follow-up periods might range from a few months to over ten years, and a screening or preventive study could take more than a decade. Although the cost per patient per year increases dramatically with increasing disability, another concern may be that AZA is usually not administered for more than ten years. Some patients may have exceeded this time limit during the follow-up period. Specifically, the approval of two more effective MS medicines in 2006 and 2011 during the follow-up period may have prompted an earlier switch from first-line therapy (such as AZA or INF) to more effective ones, hence determining an earlier treatment switch in the case of inefficacy. These events motivated us to investigate the impact of two treatments on PFS during shorter follow-up periods. It is important to note that due to outliers and different patients monitored for longer periods, the results of the previous chapter based on longer follow-ups cannot be considered for shorter periods. More specifically, if patients either experience an event of interest or switch the treatment for a longer period, they are administratively censored due to the end of the new follow-up. As a result, the data must be reorganized before applying weights in the Marginal structural cox model to analyze the efficacy of Interferon (INF) and Azathioprine (AZA) on PFS during a shorter period (for example, 60 months).

1 Descriptive analysis

To prepare data to analyze, it is necessary to re-organize it based on the new follow-up period. To do so, if patients either experience an event of interest or switch the treatment after 60 months, they are administratively censored due to the end of follow-up. Summary statistics for updated PFS are illustrated in Table 1. As

Table 1: Summary statistics for Progression Free Survival (PFS) in 5 years of follow-up.

Outcome	*min*	*max*	25^{th} *quantile*	*median*	75^{th} *quantile*	*administrative censoring rate*	*switching censoring rate*	*median* INF	*median* AZA
PFS	2	60	18.25	41	60	0.75	0.08	37	48

Table 2: Log rank test

	Treatment	
	INF	AZA
N	351	211
Observed (O)	55	32
Expected (E)	52.8	34.2
$(O-E)^2/E$	0.095	0.146
$(O-E)^2/V$	0.243	0.243

p-value = 0.6

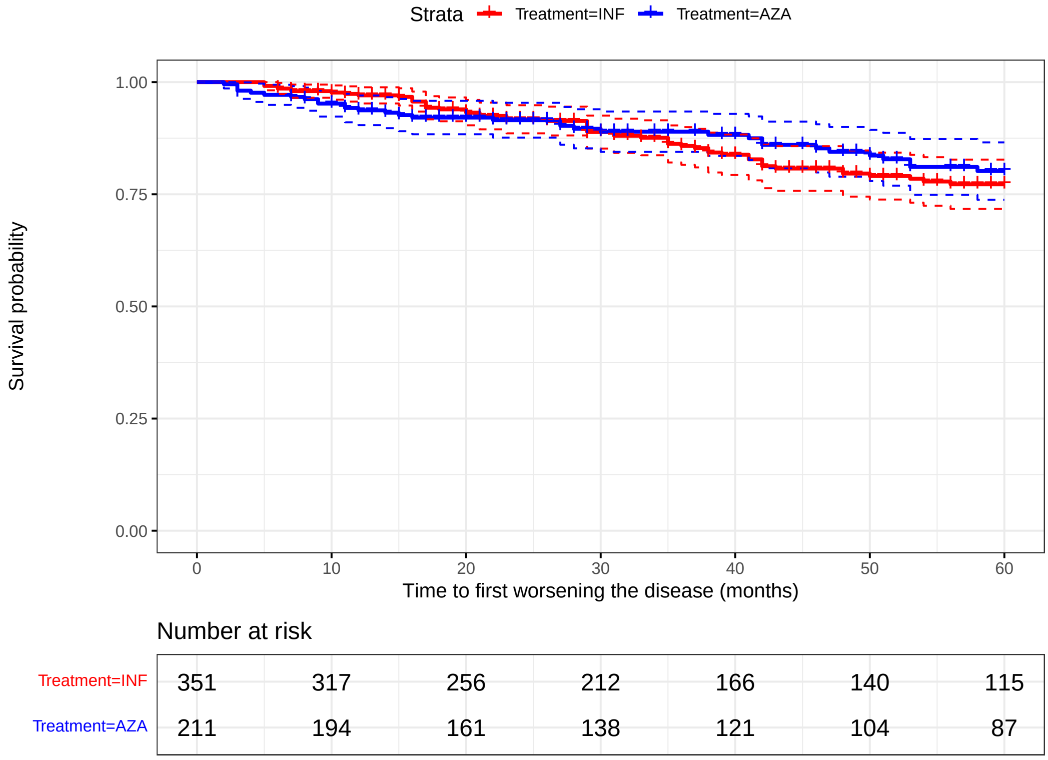
shown, data reorganization modifies the proportions of both types of censoring. The Unadjusted survival curves of the two treatments of Multiple Sclerosis (MS) disease up to 5 years are shown in Figure 1. According to the output of Table 2, there is no reason to reject the null hypothesis, which is no difference between the populations in the probability of an event (here, a worsening of the disease) at any time.

2 Construct weights and diagnostics for assessing the assumptions

By setting the propensity score model specification and checking the balance, one can correct for observed confounders by weighting causal effect estimators [Li et al., 2018]. As the design phase does not include any outcome, it is not necessary to re-analyze it at this new follow-up. In other words, one can adjust for observed confounders by computing the weight estimators and checking the balance as described in Chapter 6 and Section 2.1.

IPCW via the cox proportional hazards model [Jackson et al., 2014, Willems et al., 2018] is the most well-known method for addressing selection bias caused by covariate-dependent censoring mechanisms, and the proportional hazards assumption must be satisfied prior to its computation. Statistical tests and graphical

Figure 1: Unadjusted Survival curves of the two treatments of Multiple Sclerosis (MS) disease based on 5 years follow-up.



Treatment=INF	351	317	256	212	166	140	115
Treatment=AZA	211	194	161	138	121	104	87

Table 3: Test proportional Hazards assumption

	administrative censoring			switching censoring		
Baseline Covariates	*chisq*	*df*	*p*	*chisq*	*df*	*p*
`Age`	0.15	1.00	0.70	0.01	1.00	0.92
`ARR_pre`	0.14	1.00	0.71	0.36	1.00	0.55
`Disease_durat`	0.56	1.00	0.45	0.66	1.00	0.42
`Dummy_EDSS`	0.03	1.00	0.86	0.15	1.00	0.70
`Baseline.EDSS`	0.59	1.00	0.44	0.95	1.00	0.33
`Gender`	2.30	1.00	0.13	0.00	1.00	0.97
`PI_pre`	0.09	1.00	0.77	0.01	1.00	0.91
`Relaps_pre`	0.41	1.00	0.52	0.54	1.00	0.46
`Relapse_Dummy`	0.15	1.00	0.70	0.76	1.00	0.38
`Year`	1.56	1.00	0.21	**13.46**	**1.00**	**0.00**
`GLOBAL`	12.40	11.00	0.33	**20.92**	**11.00**	**0.03**

diagnostics based on scaled Schoenfeld residuals can be used to assess the proportional hazards assumption. The results are depicted in Table 3 and Figures 2-3. The test output indicates that neither the variables nor the global test are statistically significant for administrative censoring. Nonetheless, under switching censoring, the variable `Year` is statistically significant, hence violating the proportional hazards assumption. The test results are supported by graphical inspection (Figures 2-3).

It is common to discretize the variable `Year` based on certain thresholds to fix the proportionality assumption. In 2006 and 2011, two additional effective MS treatments were approved, according to expert knowledge. Specifically, Natalizumab was approved within the current MS data set's follow-up. Natalizumab is normally well tolerated; however, a protocol restricting its distribution was approved in 2006 due to unpredictable and potentially fatal side effects. The European Medical Agency recommends Natalizumab for people with relapsing variants of MS who have failed early disease-modifying therapies [Rudick et al., 2013, Tintore et al., 2019]. In 2011, a second effective treatment was authorized. These events may have prompted a switch from first-line therapies (such as AZA or INF) to more current foundational therapies [Tintore et al., 2019]. As a result, the following periods are used to discretize `Year` for switching censoring:

- From 1981 (start of follow-up) to 2006,
- From 2007 to 2010,
- After 2011 to 2019 (end of follow-up).

and we named the new variable as `dis_Year`. The statistical test (Table 4) and

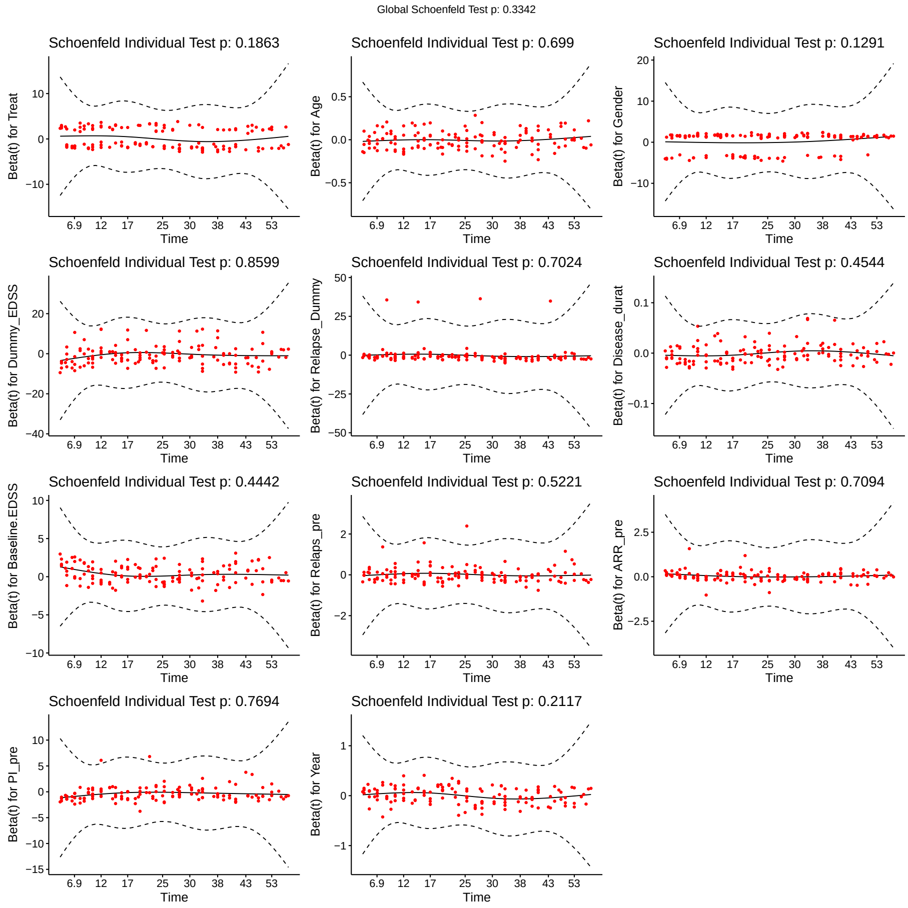


Figure 2: The Schoenfeld residuals for administrative censoring.

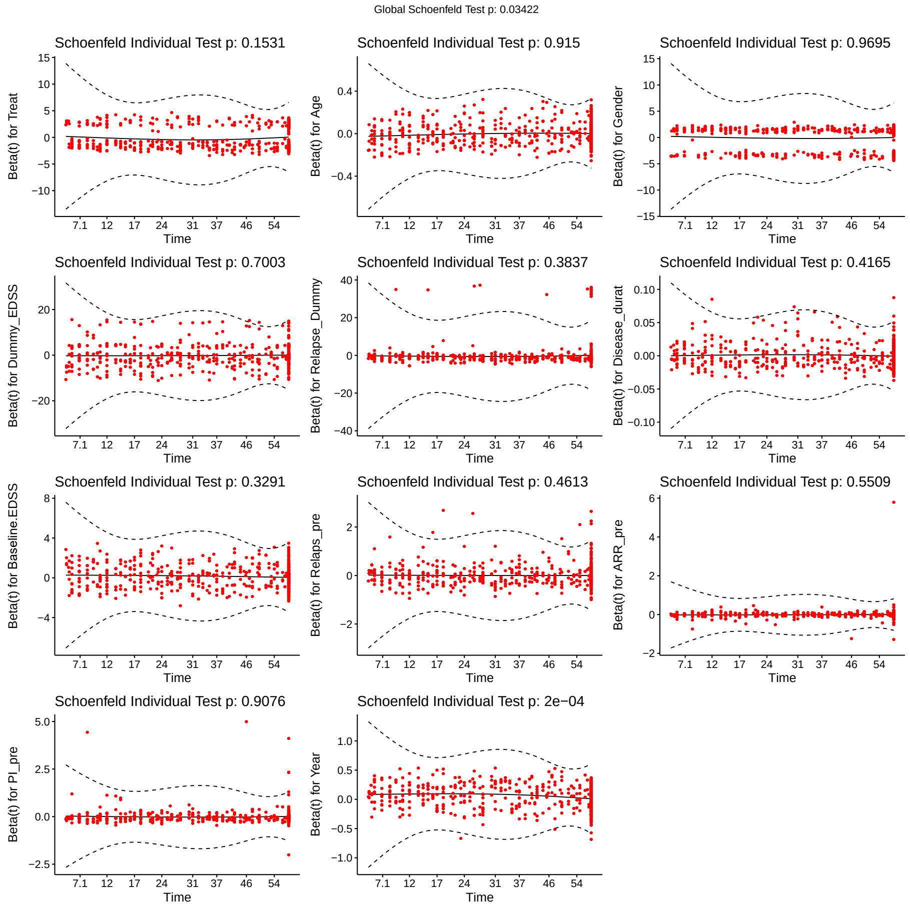


Figure 3: The Schoenfeld residuals for switching censoring.

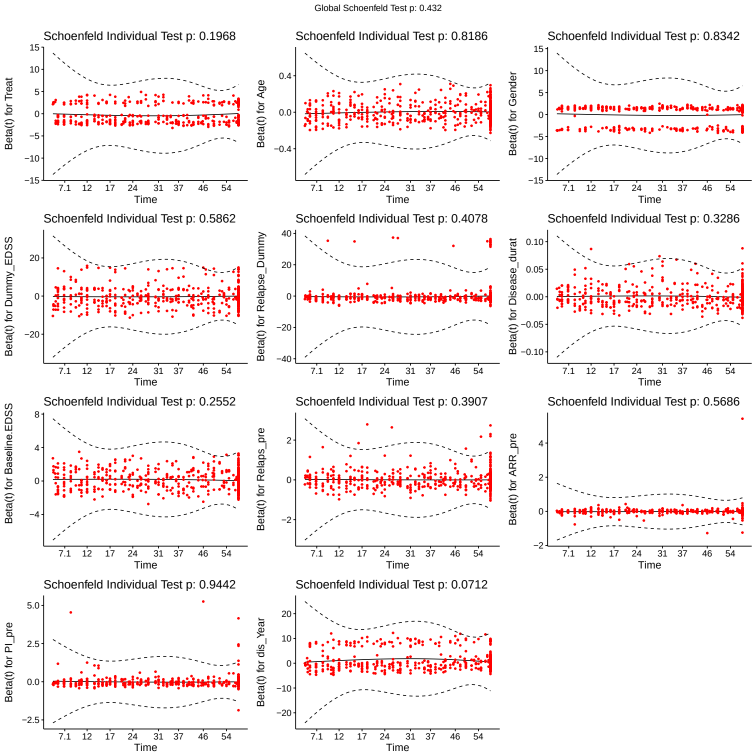


Figure 4: The Schoenfeld residuals for switching censoring after discretizing.

graphical diagnostics based on scaled Schoenfeld residuals (Figure 4) confirm the proportional hazards assumption satisfied for switching censoring. Now, we should assess the covariate-dependent censoring assumption using a multivariate Cox model. According to Tables 5 and 6, both censoring mechanisms are ignorable to different sets of covariates. Some diagnostics for the Cox model in both censoring mechanisms are presented in the Supplementary Material. In short, Martingale residuals and Deviance residuals (Figures SM–19-SM–20 in Supplementary Material) confirm that there is no outlier. Besides, the plot of Martingale residuals and Deviance residuals against covariates (see Figures SM–21–SM–24 in Supplementary Material) show the relationship between a covariate and unexplained variation. In the delta-beta residuals plot (Figures SM–25- SM–26 in Supplementary Material), we figure out that none of the observations is influential individually, even though some of the dfbeta values are large compared with the others.

As described in Chapter 5 (Section 2.2), following diagnostics for the Cox model, we calculate IPCW for each censoring mechanism to adjust for selection bias.

Administrative and switching censoring satisfied the proportionality assumption based on different covariates (`Year` for administrative and `dis_Year` for switch-

Table 4: Test proportional Hazards assumption of switching censoring.

Baseline covariates	chisq	df	p
Age	0.04	1.00	0.85
Gender	0.04	1.00	0.84
Dummy_EDSS	0.26	1.00	0.61
Relapse_Dummy	0.75	1.00	0.39
Disease_durat	1.09	1.00	0.30
Baseline.EDSS	1.29	1.00	0.26
Relaps_pre	0.76	1.00	0.38
ARR_pre	0.37	1.00	0.54
PI_pre	0.00	1.00	0.95
dis_Year	5.10	2.00	0.08
GLOBAL	11.86	12.00	0.46

Table 5: covariate-dependent assumption for administrative censoring

	beta	HR(95% CI)	wald.test	p.value
Age	-0.00	1.00 (0.98-1.01)	-0.36	0.71
Gender1	0.18	1.20 (0.81-1.77)	0.92	0.36
Disease_durat	0.00	1.00 (0.99-1.00)	-0.75	0.45
Baseline.EDSS	0.35	1.43 (1.15-1.76)	3.33	0.00**
Dummy_EDSS1	-0.64	0.52 (0.23-1.17)	-1.57	0.11
Relapse_Dummy1	-0.11	0.90 (0.32-2.52)	-0.21	0.83
Relaps_pre	0.00	1.00 (0.93-1.09)	0.18	0.85
ARR_pre	0.04	1.04 (0.95-1.14)	0.92	0.36
PI_pre	-0.38	0.68 (0.50-0.93)	-2.40	0.01**
Year	0.00	1.00 (0.97-1.03)	0.25	0.80

Table 6: covariate-dependent assumption for switching censoring

	beta	HR(95% CI)	wald.test	p.value
Age	0.00	1.00 (0.99-1.01)	0.41	0.68
Gender1	-0.06	0.94 (0.77-1.14)	-0.61	0.54
Disease_durat	0.00	1.00 (0.99-1.00)	0.30	0.76
Baseline.EDSS	0.12	1.13 (1.02-1.20)	2.46	0.01**
Dummy_EDSS1	-0.19	0.83 (0.53-1.28)	-0.84	0.40
Relapse_Dummy1	-0.22	0.80 (0.46-1.37)	-0.81	0.41
Relaps_pre	0.00	1.00 (0.96-1.05)	0.20	0.84
ARR_pre	-0.01	0.99 (0.97-1.01)	-0.64	0.52
PI_pre	-0.01	0.99 (0.95-1.03)	-0.31	0.76
dis_Year2	0.22	1.24 (0.99-1.54)	1.95	0.05*
dis_Year3	0.75	2.11 (1.64-2.73)	5.79	0.00**

ing). Thus, we consider all main effects (based on the proportionality assumption) when constructing IPCW, followed by AIC-based backward selection. Tables 7–8 depict the nonlinear terms considered for both censoring mechanisms. Although some main effects are eliminated due to backward regression in Tables 7–8, we maintain all main effects when creating IPCW. The proportional hazards assumption must then be tested for all covariates (main effects and nonlinear terms). Table 9 has been prepared accordingly. At the 0.05 level, all covariates and the global test satisfy the proportionality assumption. Consequently, we compute the IPCW for each censoring mechanism.

3 Application of Marginal Structural Model to the MS dataset

To calculate the survival probability to assess the relative effectiveness of two Treatments: Interferon (INF) and Azathioprine (AZA) on Progression-Free Survival (PFS) during 5 years of follow-up, one should consider the following steps:

1. Fit a logistic regression model with

$$\text{logit}(\Pr(Z = 1 \mid X)) = \delta_0 + \delta_1' X$$

with X= (Age, Gender, Baseline.EDSS, Dummy_EDSS, Relapse_pre, Relapse_Dummy, Disease_durat, ARR_pre, PI_pre, Year, $(\texttt{Year})^2$, $(\texttt{Relapse\_pre})^2$, $(\texttt{Baseline.EDSS})^3$, $(\texttt{Baseline.EDSS})^2$, $(\texttt{Disease\_durat})^3$, $(\texttt{Disease\_durat})^2$, $(\texttt{Relapse\_pre})^3$, $(\texttt{Year})^3$, Dummy_EDSS:Baseline.EDSS,

Table 7: The backward stepwise regression for administrative censoring model based on all covariates in 5 years follow-up.

Start: AIC = 1601.89

Stop: AIC = 1583.07

Selected Model:

$$\texttt{surv\_object} \sim \texttt{Baseline.EDSS} + \texttt{PI\_pre} + \texttt{Dummy\_EDSS} + \texttt{I}(\texttt{PI\_pre}^2) + \texttt{Dummy\_EDSS:Baseline.EDSS}$$

	Df	AIC
$<$ none $>$		1583.1
$-$Dummy_EDSS:Baseline.EDSS	1	1586.7
$-$ I(PI_pre^2)	1	1597.4

Dummy_EDSS:ARR_pre) to estimate the propensity score and then compute assignment weights $(\omega_i^{ipw}, \omega_i^{ow})$ as

Stabilized IPW weights: $$\omega_i^{ipw} = \frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)}$$

Overlap weights: $$\omega_i^{ow} = (1 - Z_i)\, e(X_i) + Z_i\,(1 - e(X_i))$$

This step is the same as we have done in Chapter 6 since the design phase does not include outcome data meaning it is not needed to re-compute it at this new follow-up.

2. Fit a Cox proportional hazard model for administrative censoring time

$$\lambda^C(t \mid Z, X) = \lambda_0^C(t) e^{\gamma_1 Z + \gamma_2' X}$$

X=(Age, Gender, Baseline.EDSS, Dummy_EDSS, Relapse_pre, Relapse_Dummy, Disease_durat, ARR_pre, PI_pre, Year, (PI_pre)^2+ Dummy_EDSS:Baseline.EDSS) to compute $\omega_i^{sc} = \dfrac{\Pr(C_i > t)}{\Pr(C_i > t \mid X_i)}$.

3. Fit a Cox proportional hazard model for switching censoring time

$$\lambda^S(t|Z, X) = \lambda_0^S(t) e^{\nu_1 Z + \nu_2' X}$$

Table 8: The backward stepwise regression for switching censoring model based on all covariates in 5 years of follow-up.

Start: AIC = 5406.09

Stop: AIC = 5393.73

Selected Model:

$$\texttt{surv\_object} \sim \texttt{dis\_Year} + \mathrm{I}(\texttt{Baseline.EDSS}^2) + \mathrm{I}(\texttt{Age}^2)$$
$$\mathrm{I}(\texttt{PI\_pre}^2) + \mathrm{I}(\texttt{Baseline.EDSS}^3) + \mathrm{I}(\texttt{Age}^3) + \mathrm{I}(\texttt{PI\_pre}^3)$$

	Df	AIC
$< \texttt{none} >$		5393.7
$-\,\mathrm{I}(\texttt{PI\_pre}^3)$	1	5394.1
$-\,\mathrm{I}(\texttt{PI\_pre}^2)$	1	5394.4
$-\,\mathrm{I}(\texttt{Age}^3)$	1	5395.6
$-\,\mathrm{I}(\texttt{Age}^2)$	1	5395.6
$-\,\mathrm{I}(\texttt{Baseline.EDSS}^3)$	1	5398.4
$-\,\mathrm{I}(\texttt{Baseline.EDSS}^2)$	1	5400.8
$-\texttt{dis\_Year}$	2	5430.9

Table 9: Proportional hazards assumption of all covariates based on 5 years follow-up

Baseline Covariates	*chisq*	*df*	*p*
Administrative censoring			
Age	0.24	1.00	0.62
Gender	1.98	1.00	0.16
Dummy_EDSS	0.01	1.00	0.94
Relapse_Dummy	0.20	1.00	0.66
Disease_durat	0.61	1.00	0.44
Baseline.EDSS	0.45	1.00	0.50
Relaps_pre	0.46	1.00	0.50
ARR_pre	0.53	1.00	0.47
PI_pre	0.16	1.00	0.68
Year	1.70	1.00	0.19
I(PI_pre^2)	0.02	1.00	0.90
Dummy_EDSS:Baseline.EDSS	0.00	1.00	0.96
GLOBAL	9.39	12.00	0.67
Switching censoring			
Age	0.13	1.00	0.72
Gender	0.04	1.00	0.83
Dummy_EDSS	0.20	1.00	0.65
Relapse_Dummy	0.64	1.00	0.42
Disease_durat	0.55	1.00	0.46
Baseline.EDSS	1.09	1.00	0.30
Relaps_pre	0.60	1.00	0.44
ARR_pre	0.30	1.00	0.58
PI_pre	0.09	1.00	0.77
dis_Year	5.10	2.00	0.08
I(Age^2)	0.31	1.00	0.58
I(EDSS^2)	1.11	1.00	0.29
I(PI_pre^2)	0.15	1.00	0.70
I(Age^3)	0.52	1.00	0.47
I(Baseline.EDSS^3)	0.74	1.00	0.39
I(PI_pre^3)	0.29	1.00	0.59
GLOBAL	17.09	17.00	0.45

X=$\big($ `Age`, `Gender`, `Baseline.EDSS`, `Dummy_EDSS`, `Relapse_pre`, `Relapse_Dummy`, `Disease_durat`, `ARR_pre`, `PI_pre`, `dis_Year`, `(Age)`^{2}, `(Baseline.EDSS)`^{2}, `(PI_pre)`^{2}, `(Age)`^{3}, `(Baseline.EDSS)`^{3}, `(PI_pre)`^{3} $\big)$ to calculate

$$\omega_i^{ss} = \frac{\Pr(S_i > t)}{\Pr(S_i > t \mid X_i)}.$$

4. Assign weights for each unit, i.e.

$$\hat{\omega}_i = \omega_i^{ipw} \times \omega_i^{ss} \times \omega_i^{sc}$$

or

$$\hat{\omega}_i = \omega_i^{ow} \times \omega_i^{ss} \times \omega_i^{sc}$$

5. Fit Marginal Structural Cox model

$$h(t|Z_i) = h_0(t)\exp\left\{X_i\alpha^\top + \gamma Z_i\right\}$$

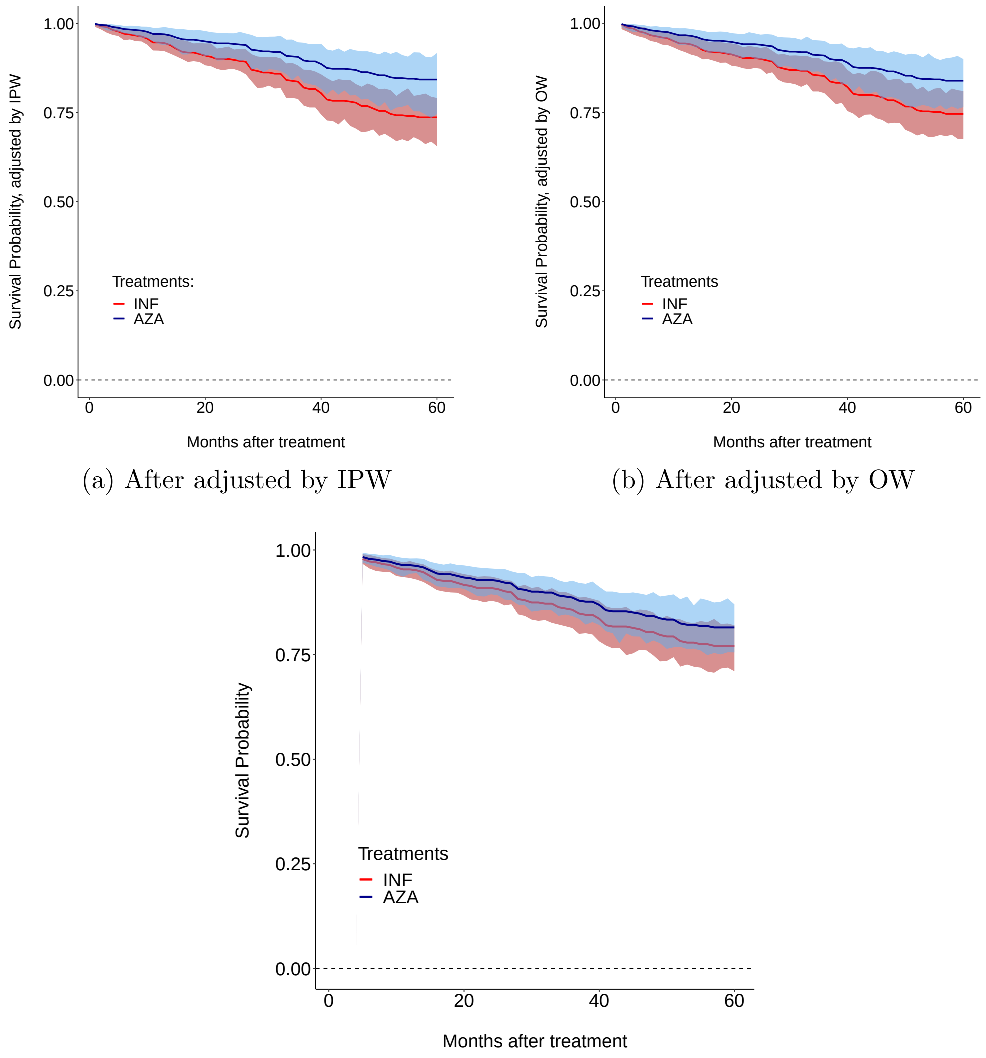
6. Calculate the survival probability $\hat{\mathbb{S}}(t|Z_i)$ specific to each treatment.

7. Finally, for each pre-specified time point (t^*), calculate **causal estimands** as

$$\Delta_{\text{RACE}} = \int_0^{t^*} \hat{\mathbb{S}}_1(t)dt - \int_0^{t^*} \hat{\mathbb{S}}_0(t)dt$$
$$\Delta_{\text{SPCE}} = \hat{\mathbb{S}}_1(t^*) - \hat{\mathbb{S}}_0(t^*)$$

4 Results

Figures 5 present the estimated causal survival curves for each treatment using the Marginal Structural Cox Model. Figures 5(a)–(b) illustrate the survival curves for two treatments after adjustment, and Figures 5(c) show the survival curves before adjustment. Besides, Figures 6 illustrate the SPCE and RACE as a function of time t with the associated 95% confidence intervals in the IPW population and the overlap population. Although the AZA shows a slightly larger survival benefit during the five years follow-up, the difference is never statistically significant in any of the adjusted situations. This is because the confidence intervals for SPCE and RACE from IPW and OW cover zero throughout the five years of the follow-up period. Regarding the SPCE and RACE, there is no significant causal survival benefit of AZA over INF at the 0.05 level. In Table 10, we also report the SPCE and RACE using the (`adjusted`) MSMs at $t = 12, 24$ and 36 months which are one-year, two-year and three-year of follow-up, respectively. Table 10 also confirms the results of Figures 5 and Figures 6. As a result, this analysis indicates that when confounding and censoring are correctly adjusted, INF is not superior to AZA even in the shorter period of follow-up (5 years). This result follows the result of the previous chapter for a longer period, which shows there is **no statistically significant difference** between AZA and INF in terms of time to the first worsening of the disease.

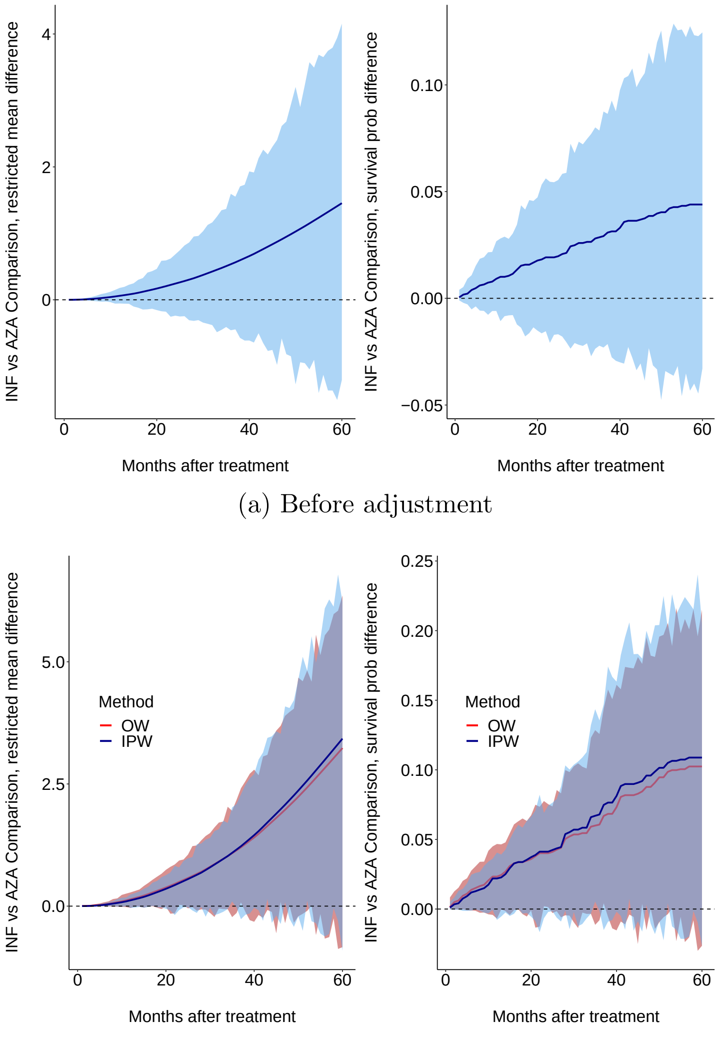


(a) After adjusted by IPW

(b) After adjusted by OW

(c) Before adjustment

Figure 5: Estimates of the survival curves for two treatments using Marginal Structural Cox Model: (first row) after adjusted by (a) IPW and (b) OW ; (second row) (c) before adjustment



(a) Before adjustment

(b) After adjustment

Figure 6: Point estimates and 95% confidence intervals of SPCE and RACE as a function of time using MSMs (a) before adjustment (b) after adjustment

Table 10: Estimates of the treatment effect using two methods: restricted average causal effect (RACE) and survival probability causal effect (SPCE) at **12, 24** and **36** months of 5 years of follow-up.

	Method	Estimate	Standard error	%95 Confidence interval	P.value
12 months					
		Restricted average causal effect			
	OW	0.119	0.068	**(-0.021,0.247)**	0.122
	IPW	0.110	0.063	**(-0.009,0.232)**	0.086
		Survival probability causal effect			
	OW	0.020	0.011	**(-0.003,0.039)**	0.086
	IPW	0.020	0.011	**(-0.002,0.039)**	0.122
24 months					
		Restricted average causal effect			
	OW	0.446	0.265	**(-0.107,0.899)**	0.126
	IPW	0.465	0.254	**(-0.080,0.892)**	0.100
		Survival probability causal effect			
	OW	0.033	0.020	**(-0.007,0.070)**	0.100
	IPW	0.037	0.020	**(-0.006,0.070)**	0.126
36 months					
		Restricted average causal effect			
	OW	0.941	0.570	**(-0.150,2.029)**	0.126
	IPW	1.032	0.550	**(-0.196,1.970)**	0.086
		Survival probability causal effect			
	OW	0.049	0.031	**(-0.008,0.112)**	0.086
	IPW	0.059	0.033	**(-0.010,0.122)**	0.126

Chapter 8

Conclusion and Further research

This thesis aimed to propose a method for estimating causal effects in observational MS data sets subject to two different censoring mechanisms (administrative censoring and switching censoring). The thesis discussed some theoretical background, in which we addressed an unknown assignment mechanism and two different covariate-dependent censoring mechanisms. Then, utilizing Marginal Structural Cox models to estimate three estimands focusing on the survival outcome, we proposed a new weighting method to adjust for both observed confounders and selection bias due to censoring. Finally, we tested the sensitivities of the adjustment methods to changes in key assumptions in simulation studies before implementing them in the Multiple Sclerosis (MS) data set in Italy. In terms of time to the first worsening of the disease, our results demonstrated no statistically significant difference between the two treatments. This is because the confidence intervals of SPCE and RACE from IPW and OW straddle zero across the entire follow-up period. Consequently, when confounding and censoring are adequately adjusted, INF is not superior to AZA. To assess the non-inferiority of INF over AZA, we accomplished a test and concluded that there is no inferiority of AZA over INF at 2.5% level. The shorter-term follow-up result confirms the longer-term result, which concluded no statistically significant difference between AZA and INF in terms of time to the first worsening of the disease.

For further research, one of the interesting areas is whether the proposed method can be applied to more complicated cases with time-varying confounders influenced by prior treatment. Because of the clinical decisions, the treatment will likely be switched based on the current (and previous) EDSS values. A high EDSS indicating poor control would likely lead to switching to a different treatment. However, high EDSS is also thought to lead to an increased risk of worsening the disease, making EDSS at a particular time a confounder of the relationship between treatment and the outcome. Thus, considering EDSS as a time-vary confounder is a topic for further research.

This thesis uses the inverse probability of censoring weighting (IPCW) to correct the selection bias resulting from dependent censoring. IPCW requires a correctly specified censoring model. Robins et al. [1994], Scharfstein et al. [1999] introduced augmented inverse probability of censoring weighting (AIPCW) estimators specifically to improve the efficiency of IPCW estimators. However, this approach is not straightforward to apply to the cox model due to the non-collapsibility of the Cox model [Tchetgen Tchetgen and Robins, 2012, Martinussen and Vansteelandt, 2013]. Recently, [Luo and Xu, 2022] proposed an AIPCW estimator based on data-adaptive machine learning methods for the cox model. Combining this with our proposed weighting methods will be considered for further research.

In this thesis, the analysis of the observational study relies on the "no unmeasured confounders" assumption to identify the estimand. Sensitivity analysis plays a key role in assessing the validity of statistical inference results. Doing a sensitivity analysis to investigate how the results could vary if the unverifiable assumption is broken to a certain degree is an excellent topic for further research.

Chapter 9

Supplementary Material

The Supplementary Material contains some theoretical background and some outputs of the methods presented in the thesis for assessing effectiveness of AZA and INF on the PFS.

SM.1 Appendix for Chapter 3

Balancing Weights

Proof of Theorem 3: Under the regularity conditions on v_z and $\mathbb{E}[\boldsymbol{Y}(z) \mid \boldsymbol{X}]$, the WATE for the population with density proportional to $f(x)h(x)$ with respect to base measure μ is defined as

$$\begin{aligned}\tau_h &= \int \tau(x)f(x)h(x)\mu(dx)\bigg/\int f(x)h(x)\mu(dx)\\ &= \frac{\int \mathbb{E}_{\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X}}[\{\boldsymbol{Y}(1)\boldsymbol{Z}[h(x)/e(x)] - \boldsymbol{Y}(0)(1-\boldsymbol{Z})[h(x)/(1-e(x))]\}]f(x)\mu(dx)}{\int f(x)h(x)\mu(dx)}\\ &= \frac{\int \mathbb{E}_{\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X}}\boldsymbol{Y}(1)\boldsymbol{Z}[h(x)/e(x)]f(x)\mu(dx)}{\int \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X}}\boldsymbol{Z}[h(x)/e(x)]f(x)\mu(dx)} - \frac{\int \mathbb{E}_{\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X}}\boldsymbol{Y}(0)(1-\boldsymbol{Z})[h(x)/e(x)]f(x)\mu(dx)}{\int \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X}}(1-\boldsymbol{Z})[h(x)/e(x)]f(x)\mu(dx)}\end{aligned} \tag{SM–1}$$

where $\tau(x) = \mathbb{E}[\boldsymbol{Y}(1) - \boldsymbol{Y}(0) \mid \boldsymbol{X} = x]$, and using the unconfoundedness assumption that $\boldsymbol{Y}(1), \boldsymbol{Y}(0) \perp \boldsymbol{Z}, \boldsymbol{X}$. The terms of SM–1 can be read as expectations of weighted means of $\boldsymbol{Y}(z)$ in samples drawn from the population with density $f(x)$, respectively for the strata with $z = 0$ or $z = 1$. Replacing expectations by sample means, and substituting weight expressions from 10, we obtain the following estimator for the sample WATE:

$$\hat{\tau}_h = \frac{\sum_i Y_i(1) Z_i w_1(x_i)}{\sum_i Z_i w_1(x_i)} - \frac{\sum_i Y_i(0)(1 - Z_i) w_0(x_i)}{\sum_i (1 - Z_i) w_0(x_i)} \tag{SM–2}$$

where each summation (divided by n) is an unbiased estimator of the corresponding integral in SM–1; therefore by Slutsky's theorem $\hat{\tau}_h$ is a consistent estimator of τ_h.

Proof of Theorem 5: The score functions of the logistic propensity score model, $\text{logit}\{e(X_i)\} = \beta_0 + X_i \boldsymbol{\beta}^\top$ with $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_p)$, are

$$\frac{\partial \log L}{\partial \beta_k} = \sum_i x_{ik}(Z_i - \hat{e}_i), \quad \text{for} \quad k = 0, 1, \ldots, p.$$

where $x_{0k} \equiv 1$ and $\hat{e}_i = [1 + \exp(-X_i \boldsymbol{\beta}^\top)]^{-1}$. Equating to 0 and solving, the MLE $\hat{\beta}$ satisfies

$$\sum Z_i = \sum \hat{e}_i, \qquad \text{and} \quad \sum x_{ik} Z_i = \sum x_{ik} \hat{e}_i$$

It follows that

$$\sum_i Z_i(1 - \hat{e}_i) = \sum_i \hat{e}_i - \sum_i Z_i \hat{e}_i = \sum_i \hat{e}_i (1 - Z_i),$$
$$\sum_i x_{ik} Z_i (1 - \hat{e}_i) = \sum x_{ik} \hat{e}_i - \sum x_{ik} Z_i \hat{e}_i = \sum x_{ik} \hat{e}_i (1 - Z_i), \quad \text{for} \quad k = 1, \ldots, p.$$

Therefore, for any $k = 1, \ldots, p$, we have

$$\frac{\sum_i x_{ik} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i x_{ik} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}, \quad \text{for} \quad k = 1, \ldots, p$$

SM.2 Appendix for Simulation Studies

In this Section, we present some results of a sensitivity analysis in terms of misspecification and proportionality assumption.

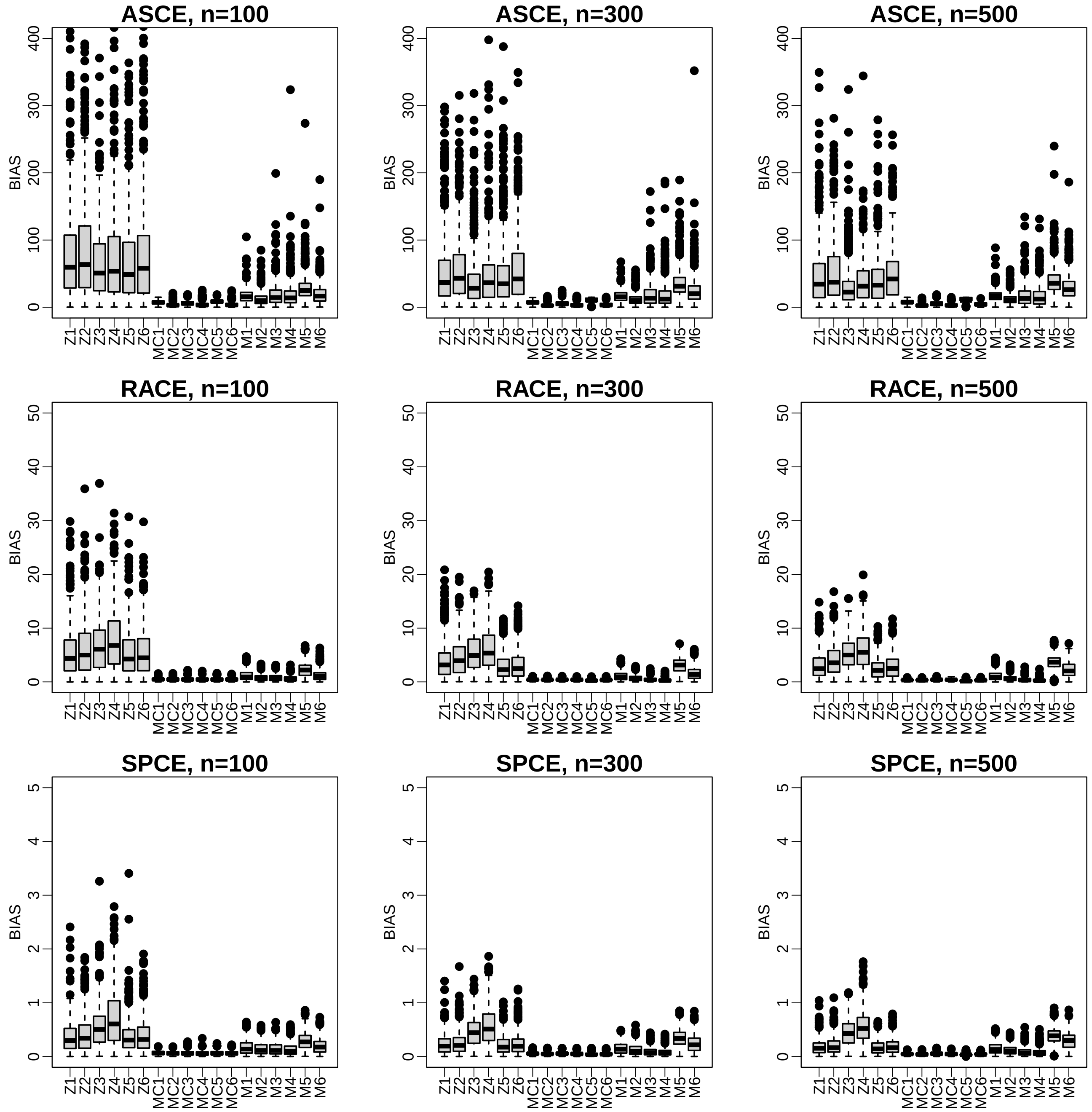


Figure SM–1: Absolute bias comparing two treatment under *good overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

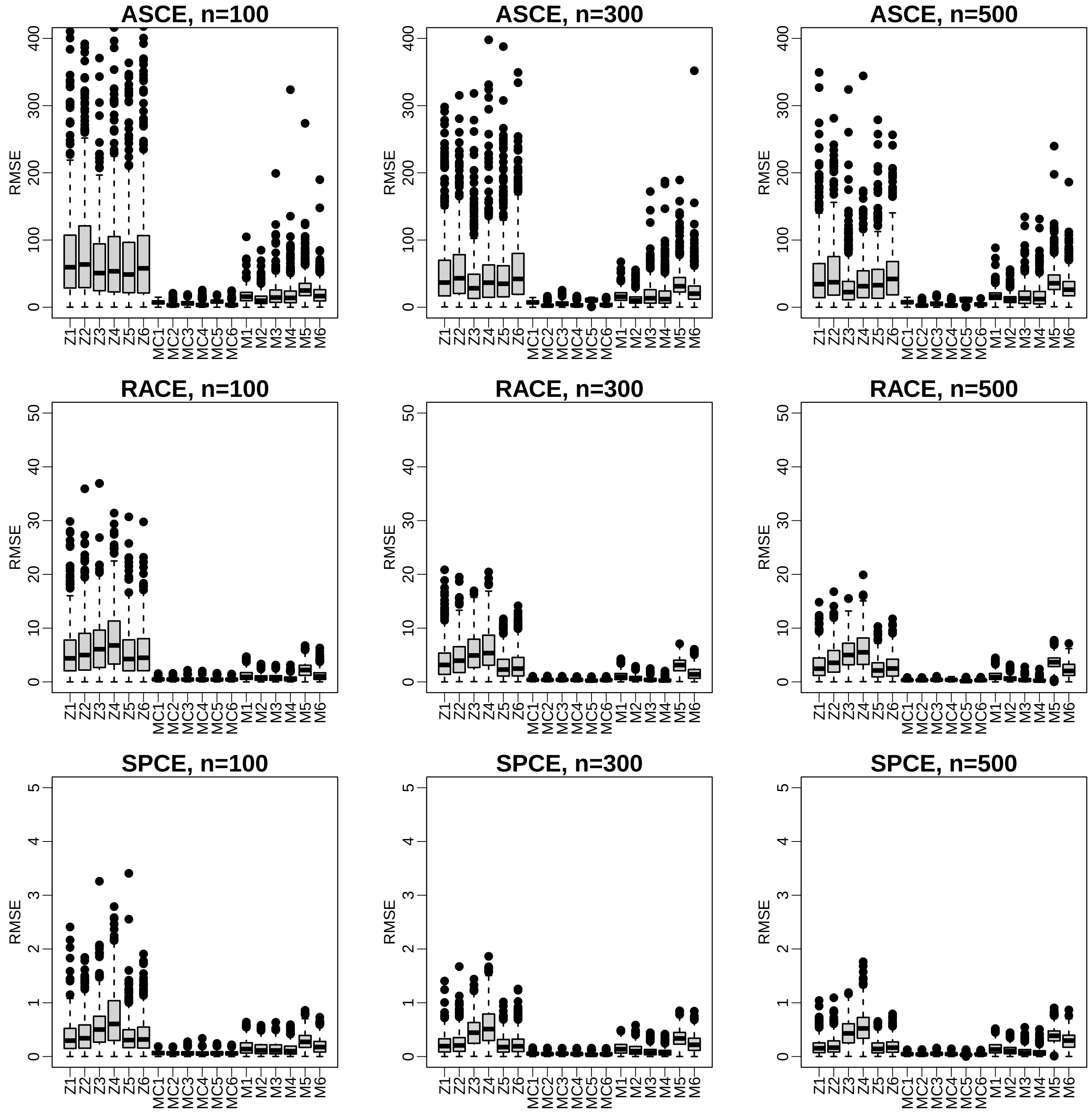


Figure SM–2: RMSE comparing two treatment under *good overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

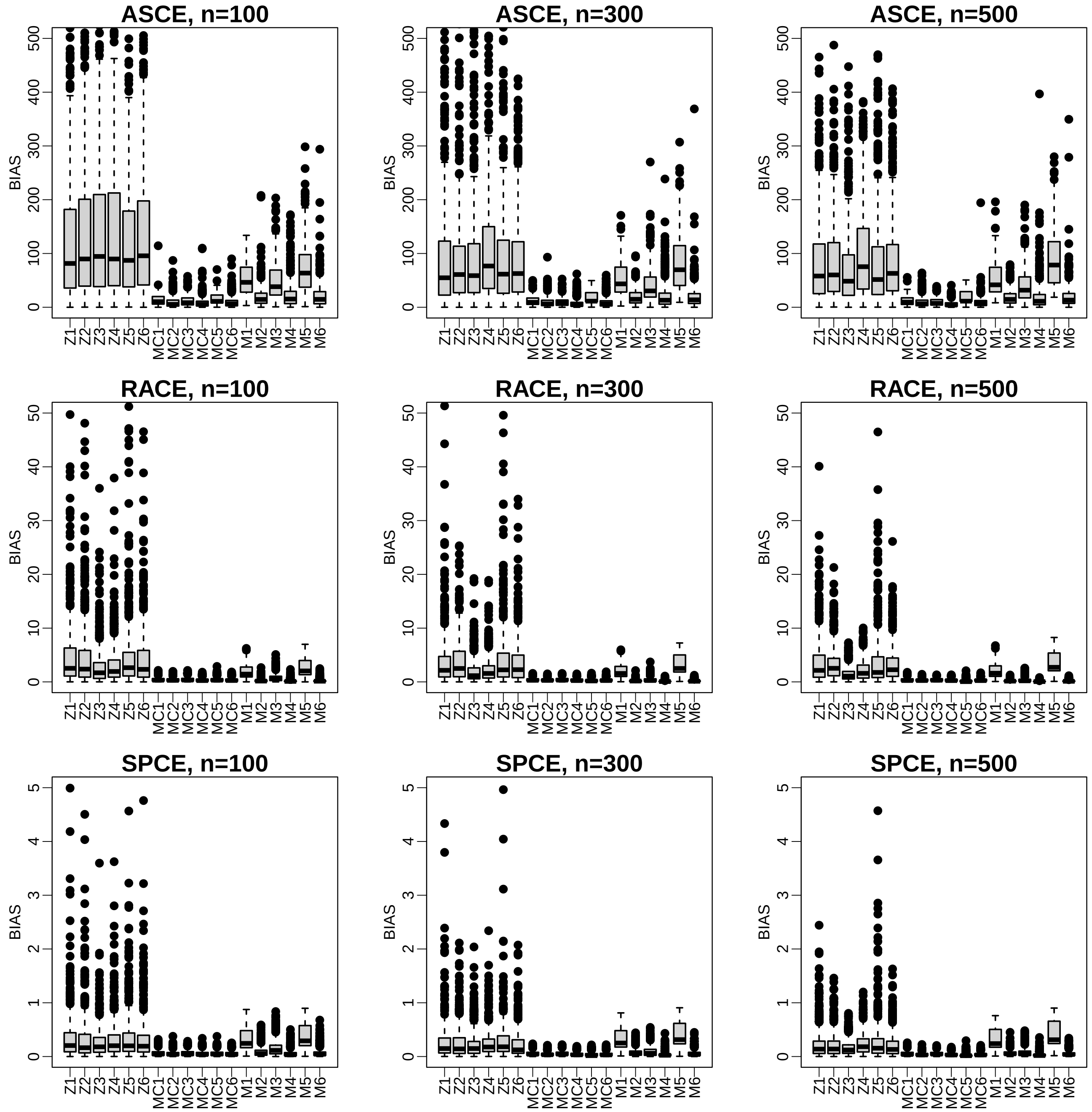


Figure SM–3: Absolute bias comparing two treatment under *poor overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

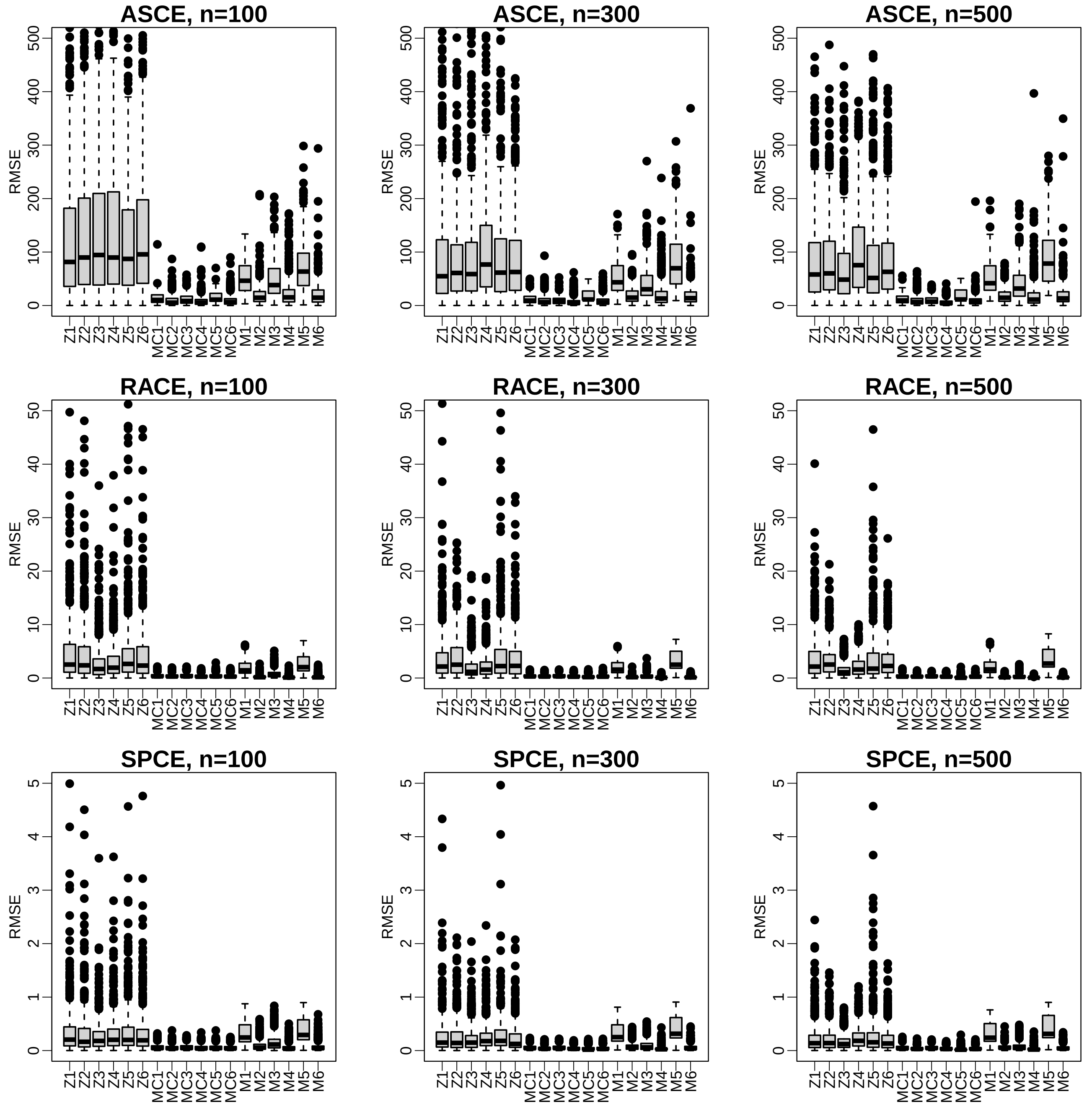


Figure SM–4: RMSE comparing two treatment under *poor overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under correct form, omission and different specification of IPCW.

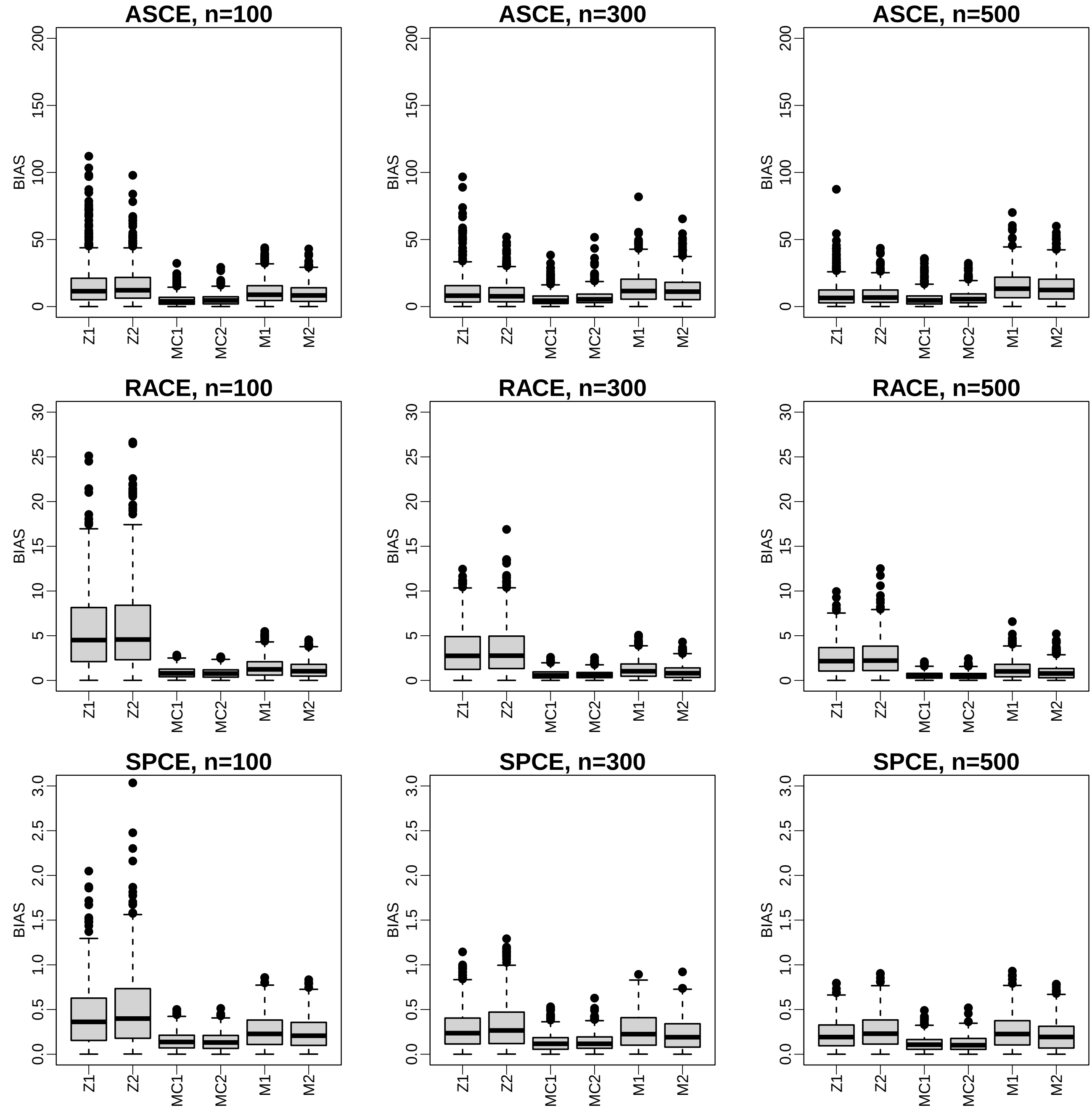


Figure SM–5: Absolute bias comparing two treatment under *good overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

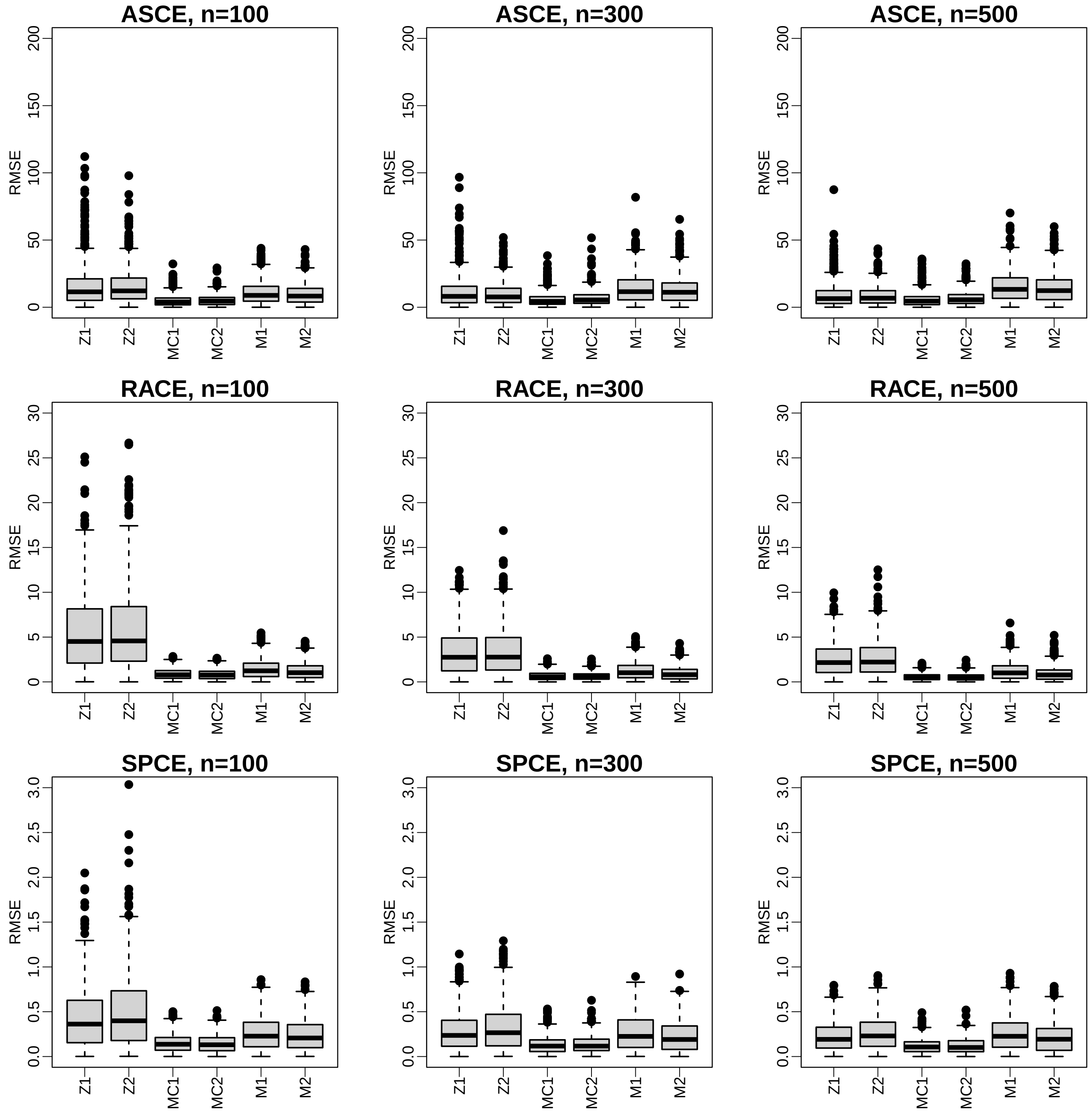


Figure SM–6: RMSE comparing two treatment under *good overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

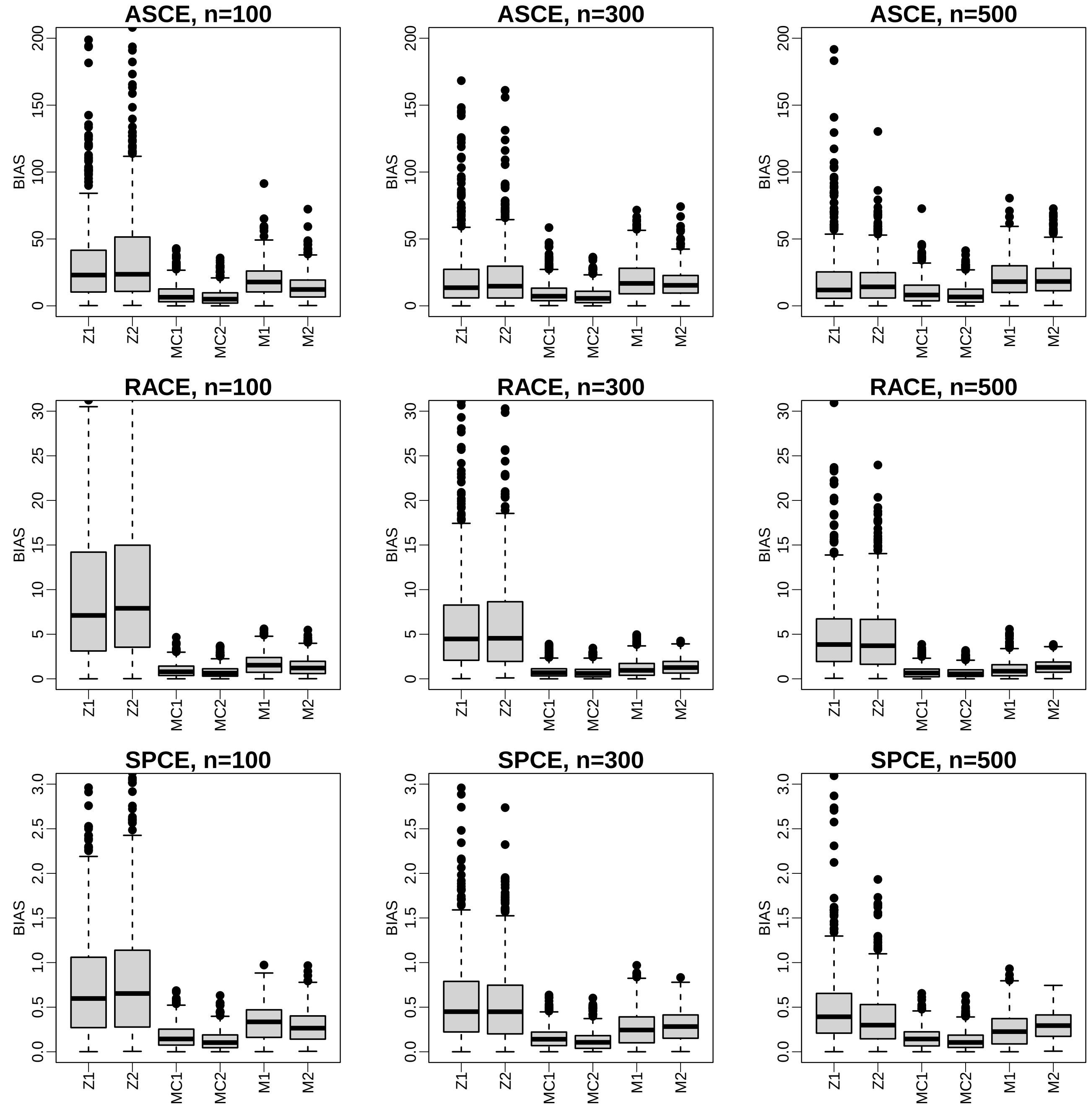


Figure SM–7: Absolute bias comparing two treatment under *poor overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

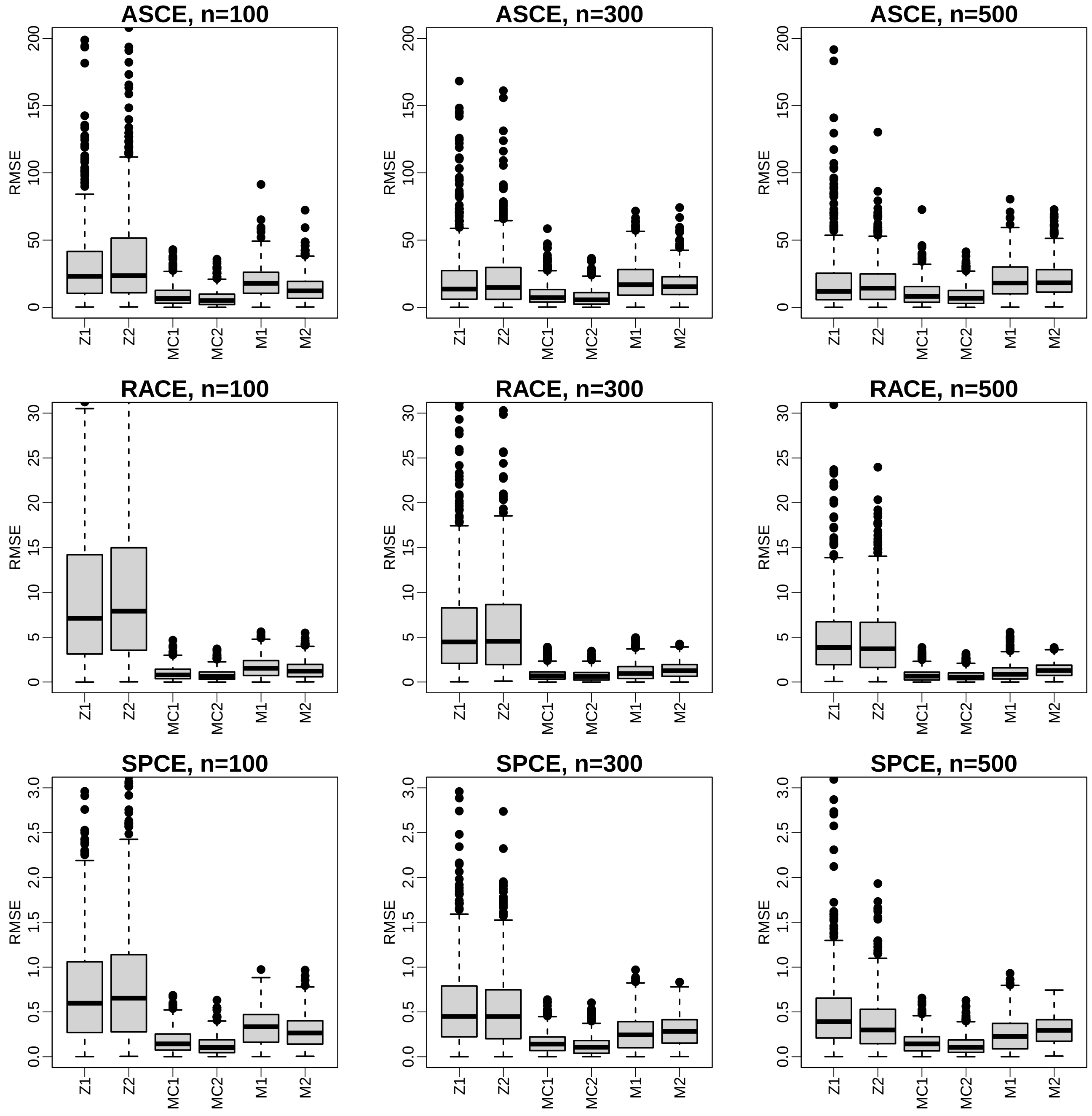


Figure SM–8: RMSE comparing two treatment under *poor overlap* and 25% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

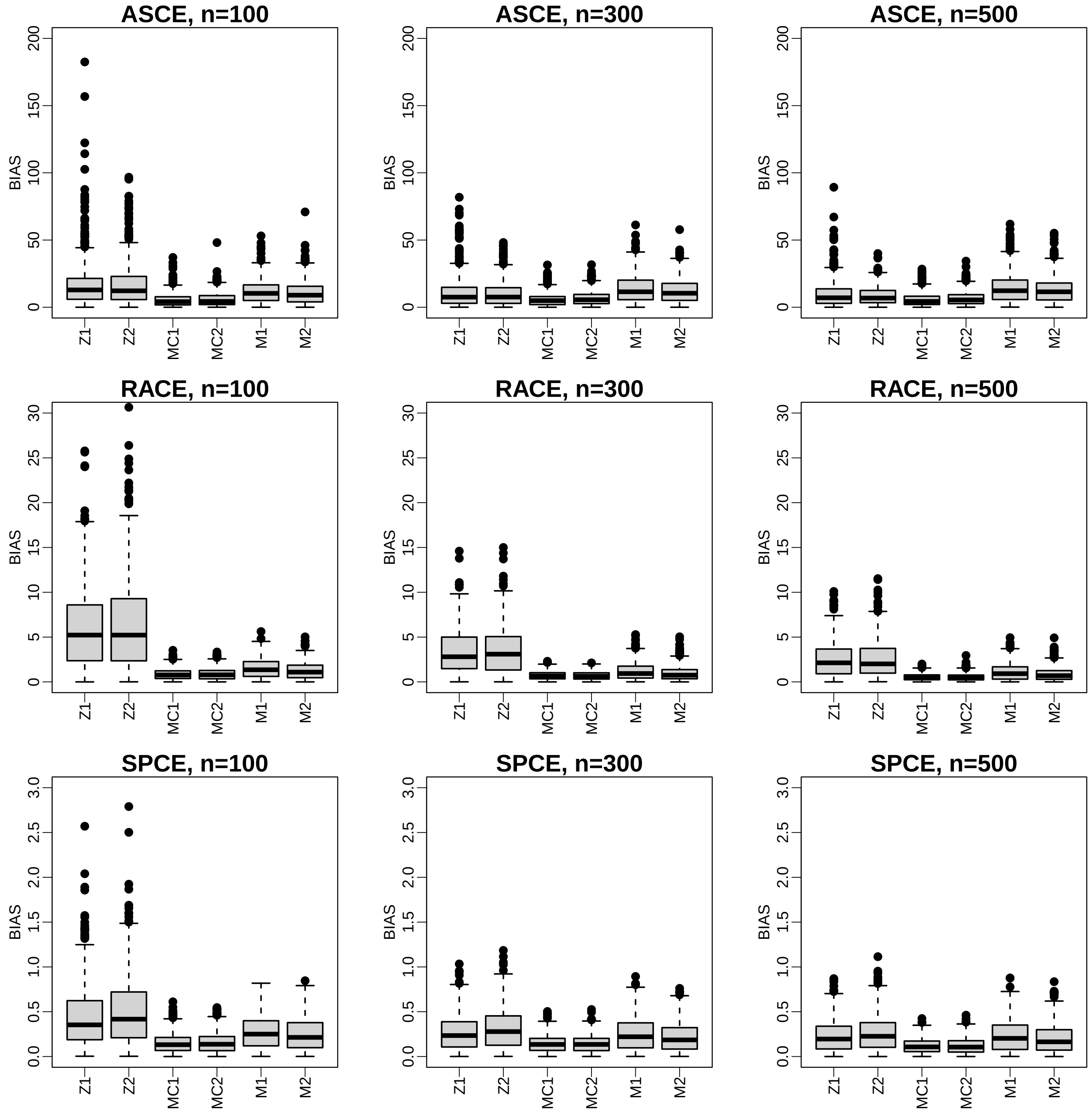


Figure SM–9: Absolute bias comparing two treatment under *good overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

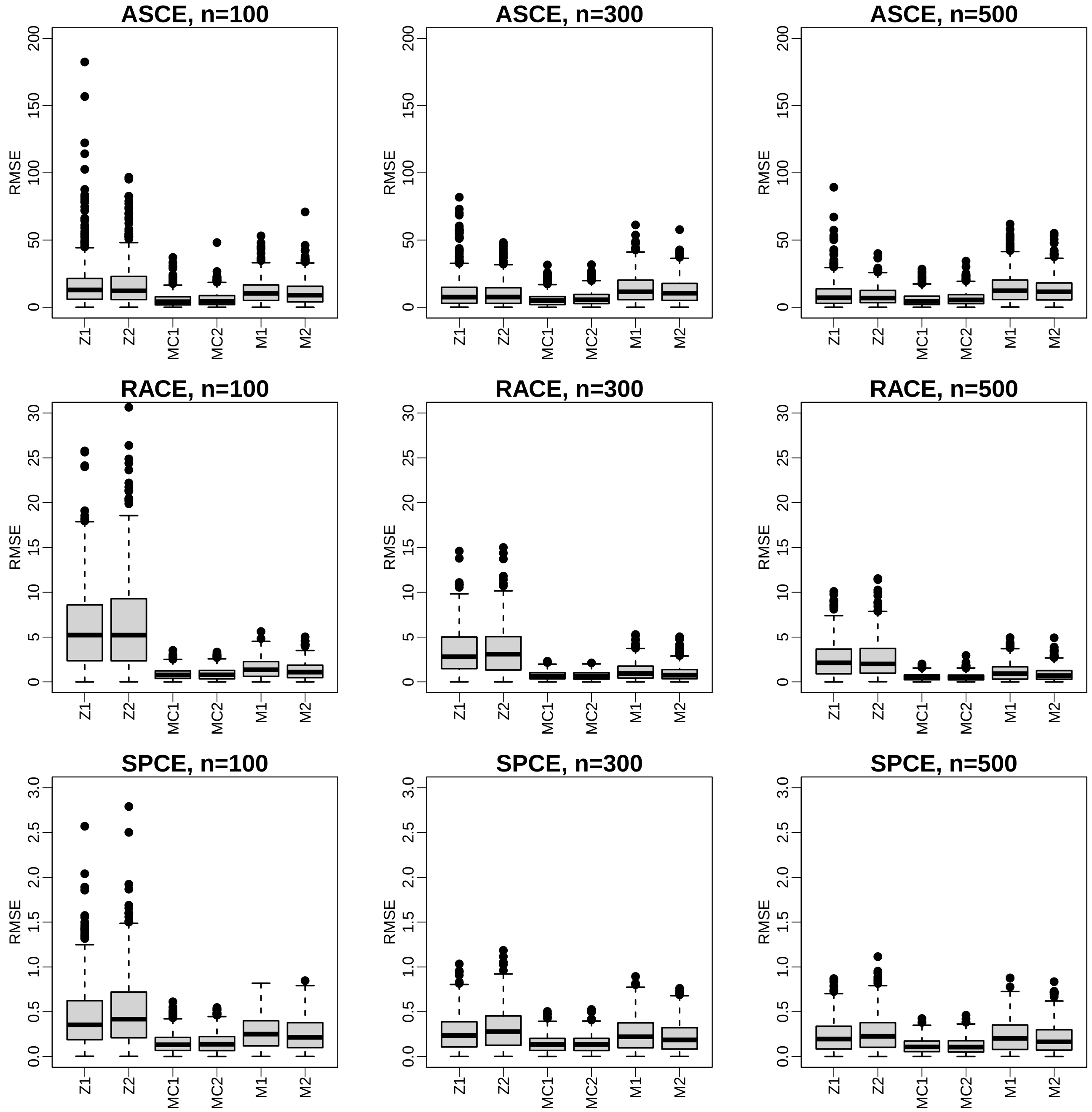


Figure SM–10: RMSE comparing two treatment under *good overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

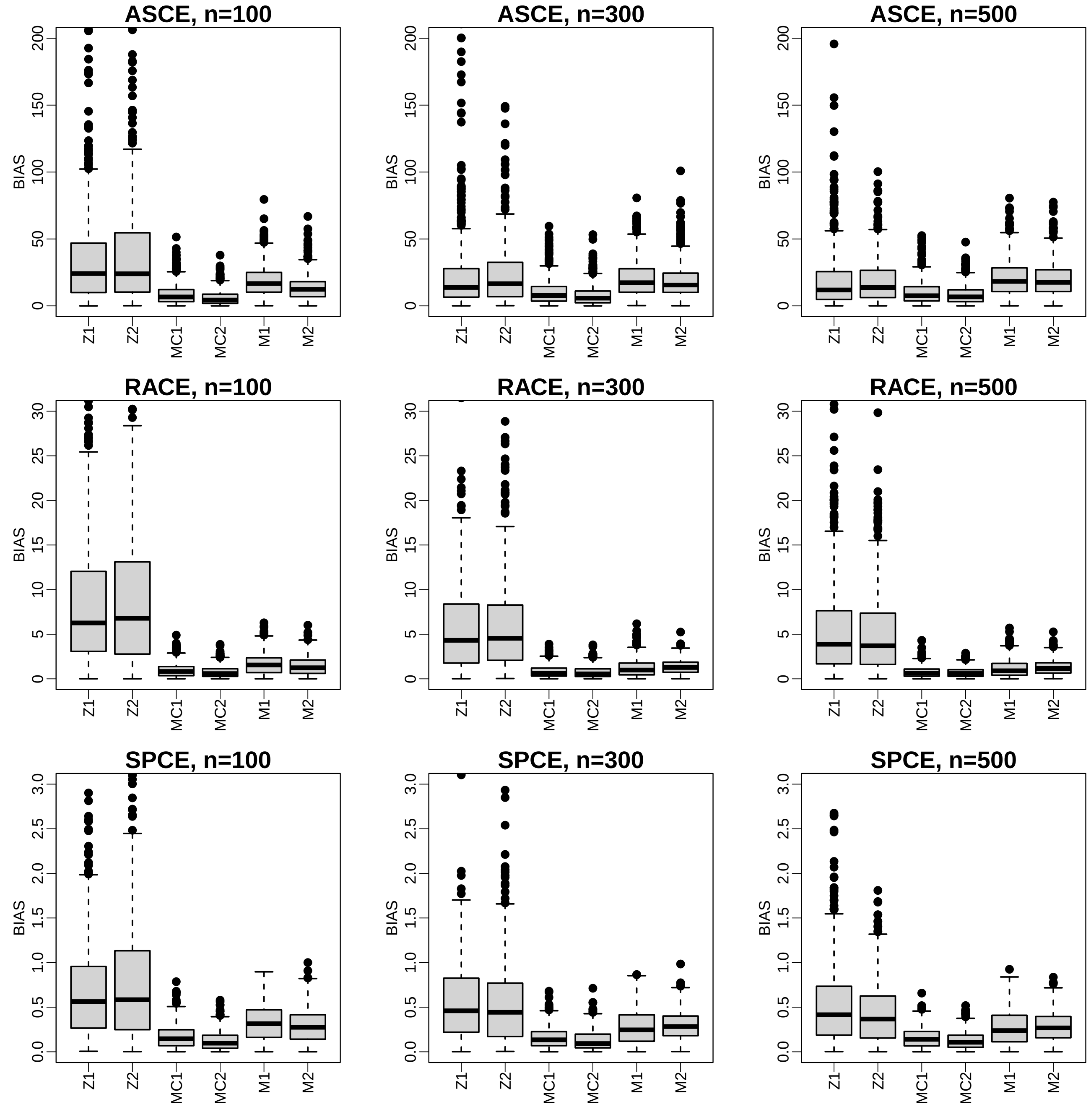


Figure SM–11: Absolute bias comparing two treatment under *poor overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

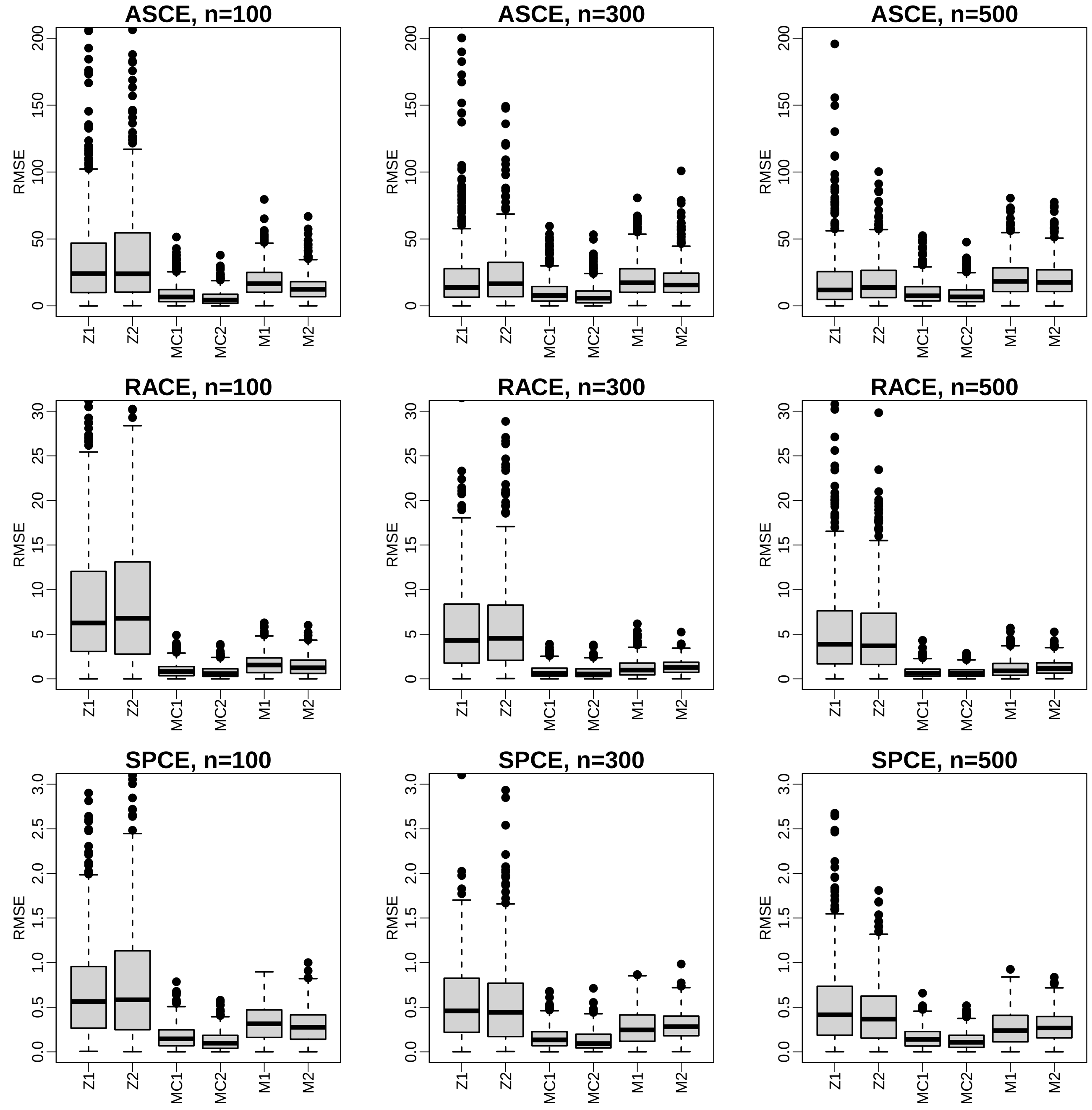


Figure SM–12: RMSE comparing two treatment under *poor overlap* and 50% censoring rate using Zeng's method, MSMs with and without covariates under non-proportional hazards assumption.

SM.3 Appendix for Real Application (part I)

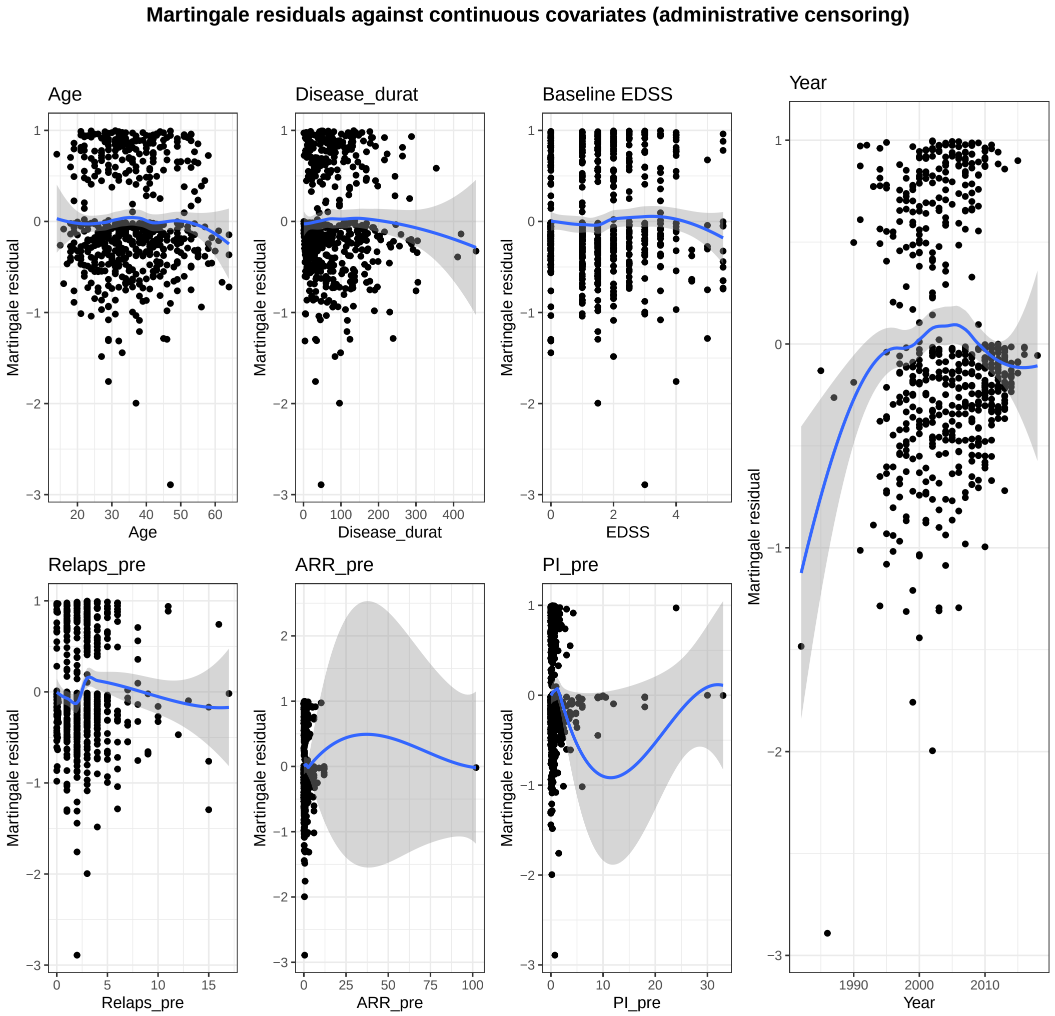


Figure SM–13: The Martingale residuals to choose a functional form for the covariate of administrative censoring.

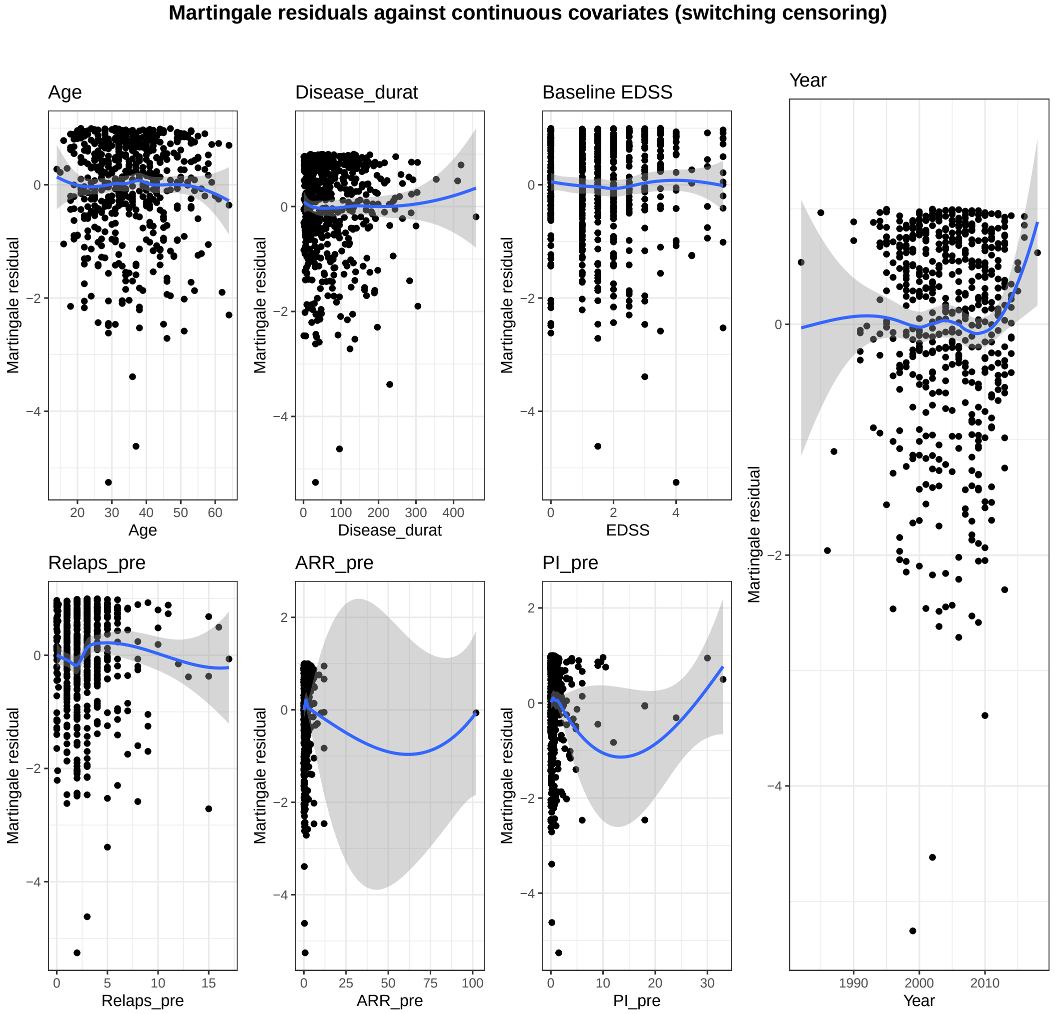


Figure SM–14: The Martingale residuals to choose a functional form for the covariate of switching censoring.

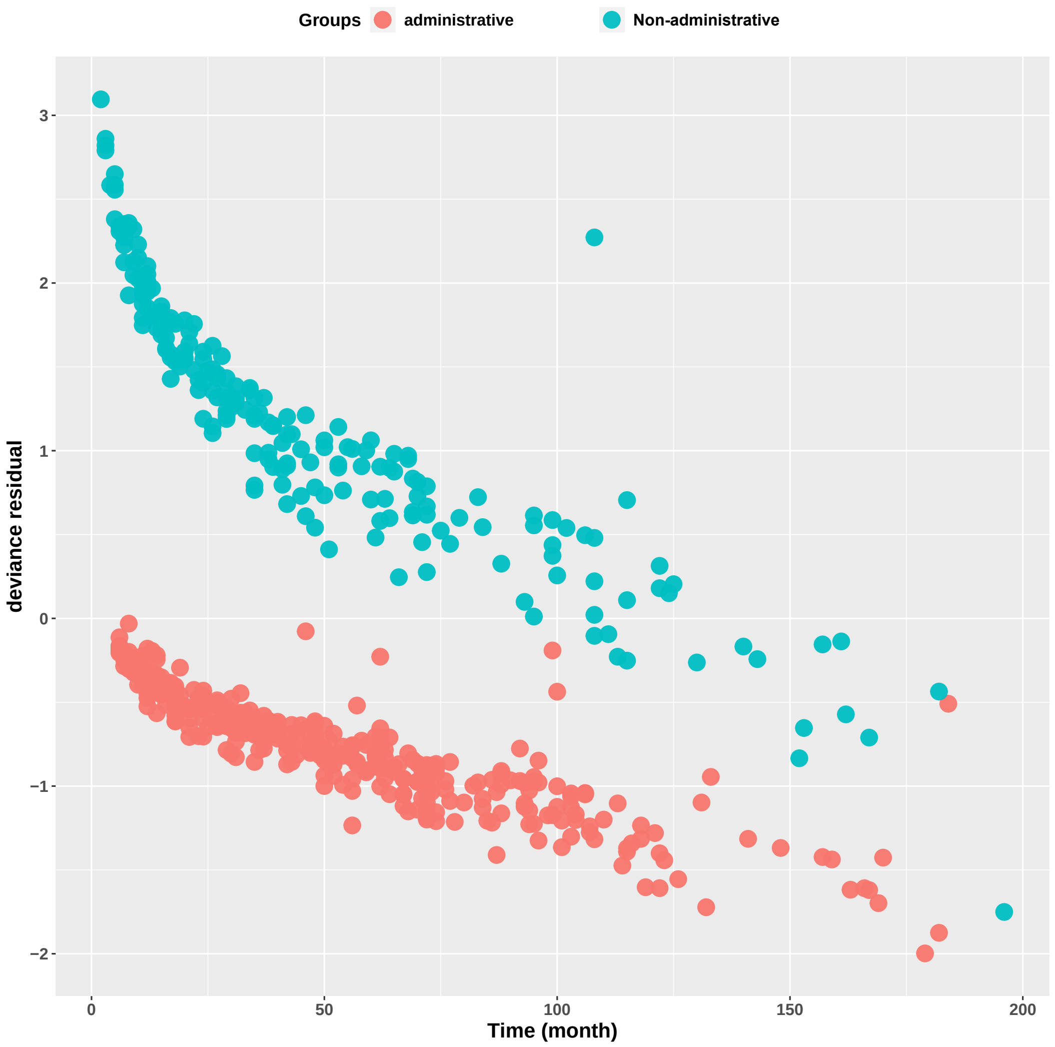


Figure SM–15: The Deviance residuals to choose a functional form for the covariate of administrative censoring.

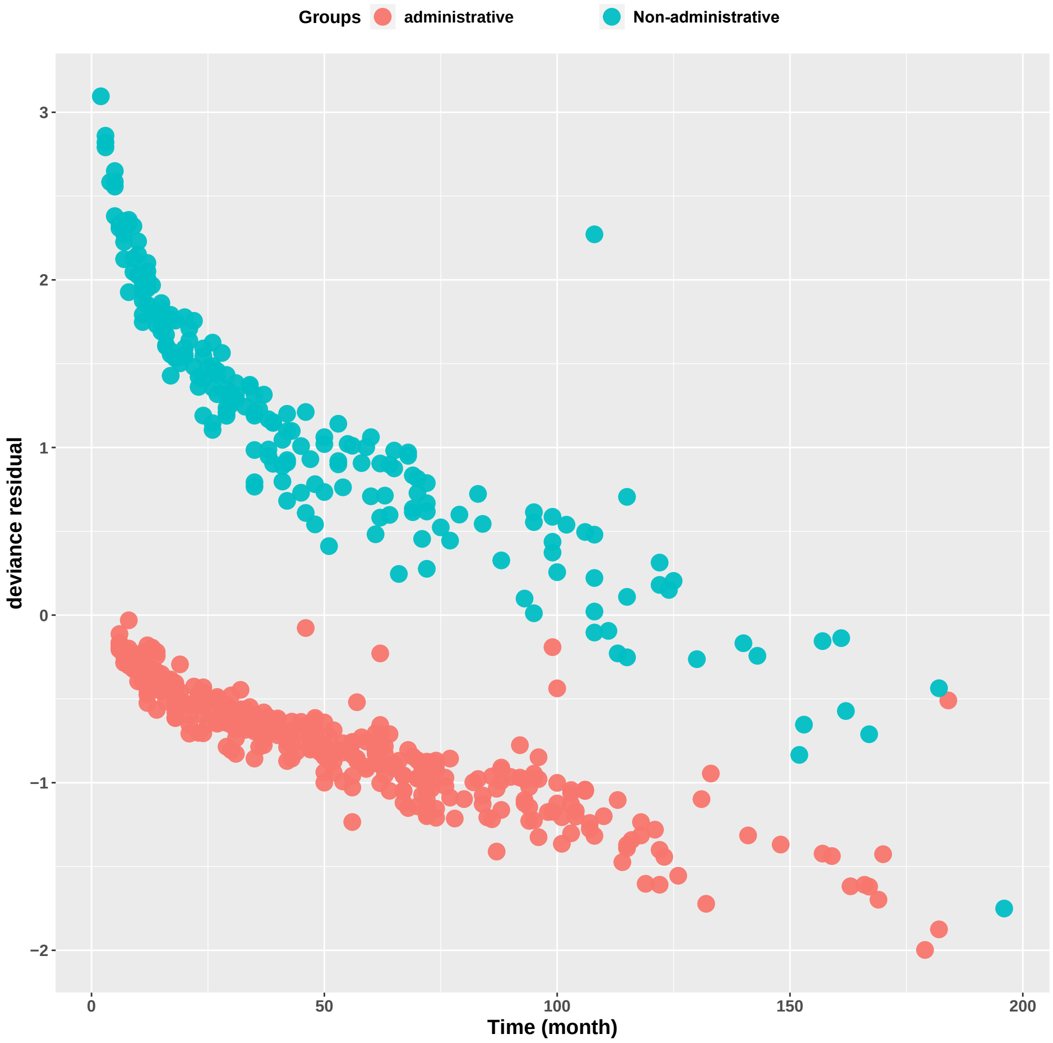


Figure SM–16: The Deviance residuals to choose a functional form for the covariate of switching censoring.

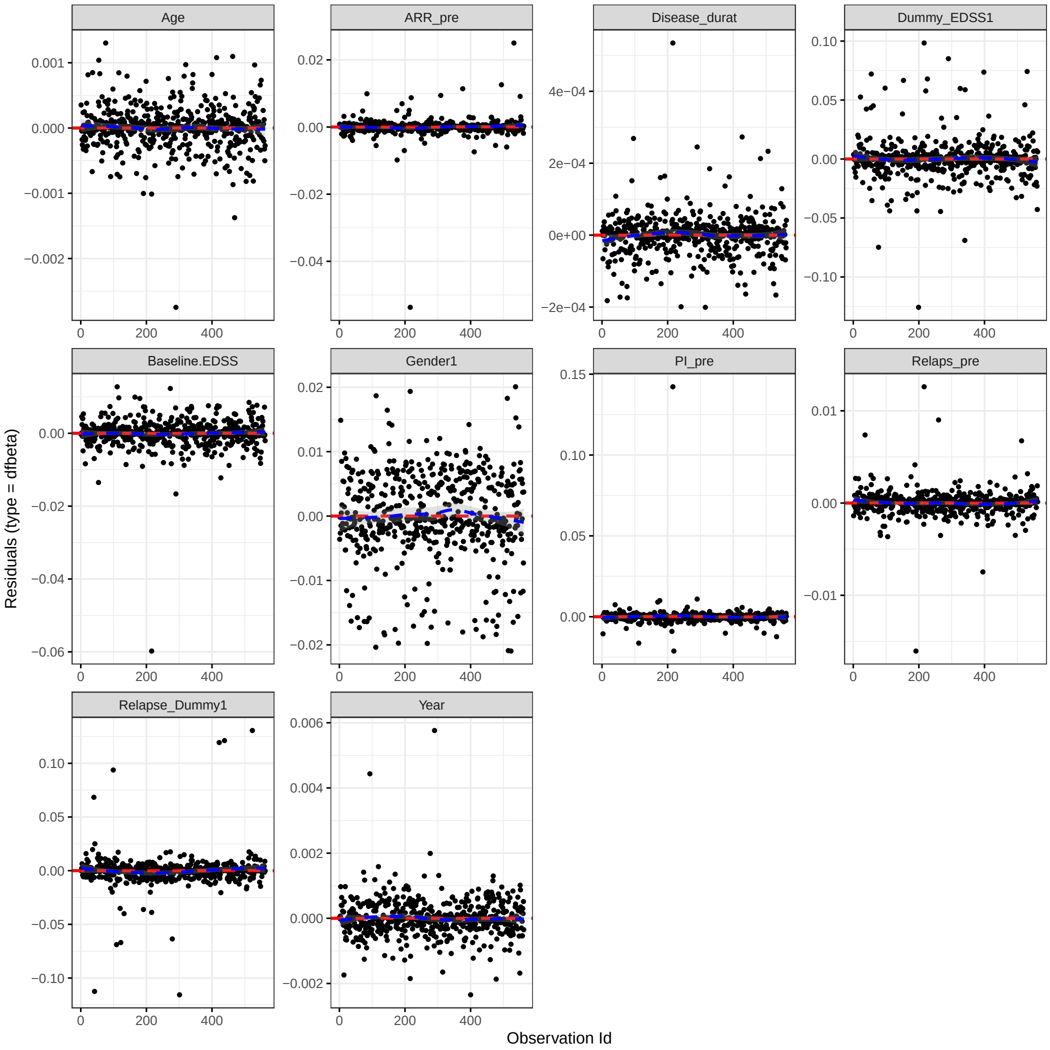


Figure SM–17: Testing Influential Observations by delta-beta residuals based on administrative censoring.

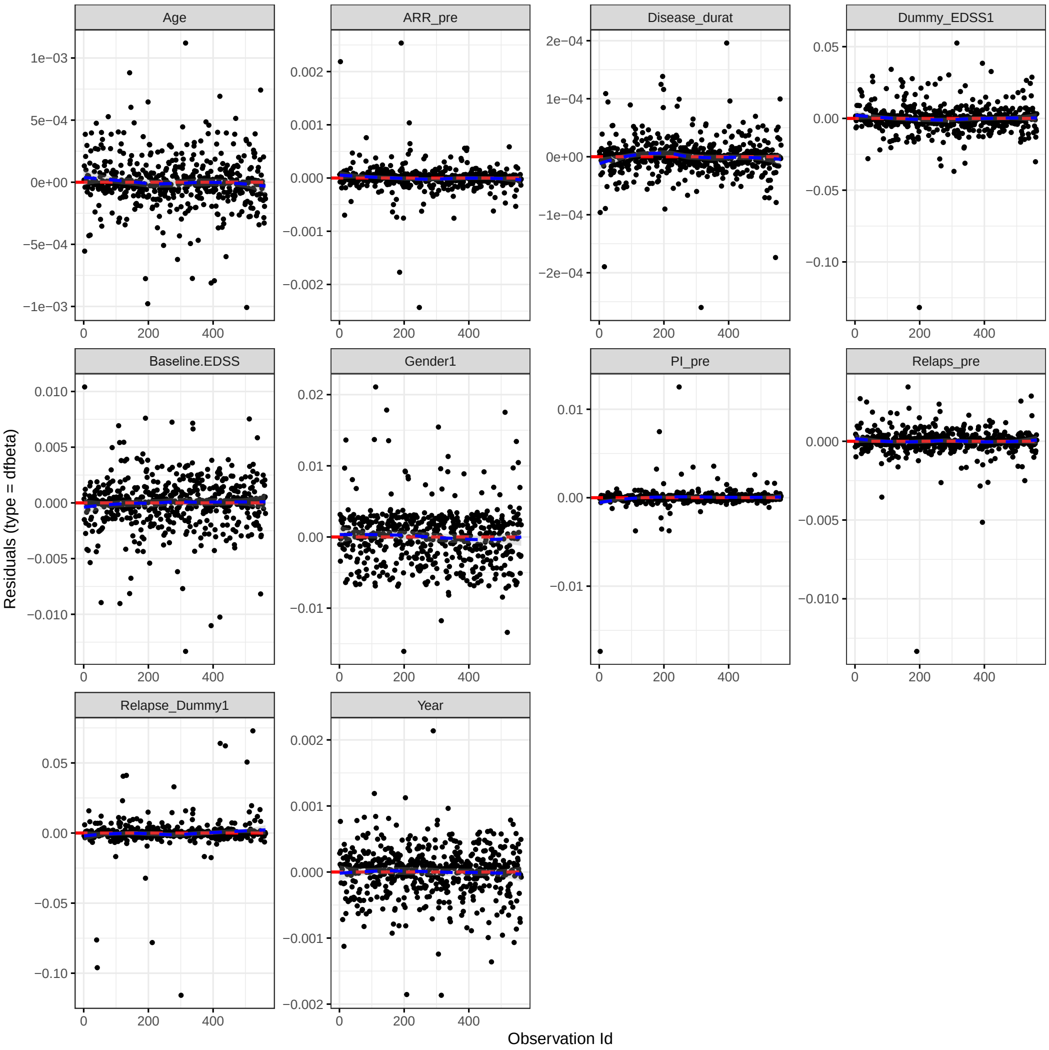


Figure SM–18: Testing Influential Observations by delta-beta residuals based on switching censoring.

SM.4 Appendix for Real Application (part II)

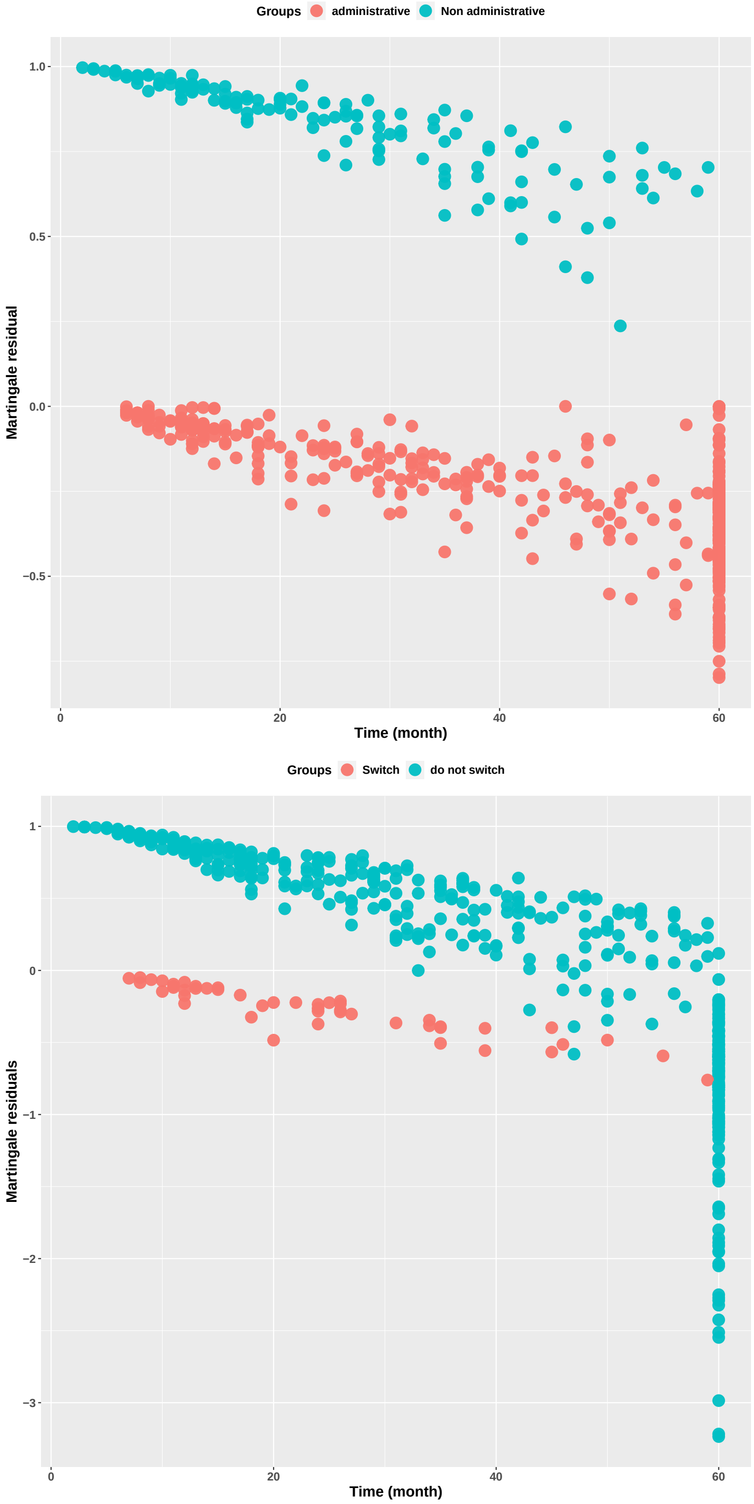


Figure SM–19: The Martingale residuals to detect outlier based on (above) administrative censoring and (below) switching the treatment.

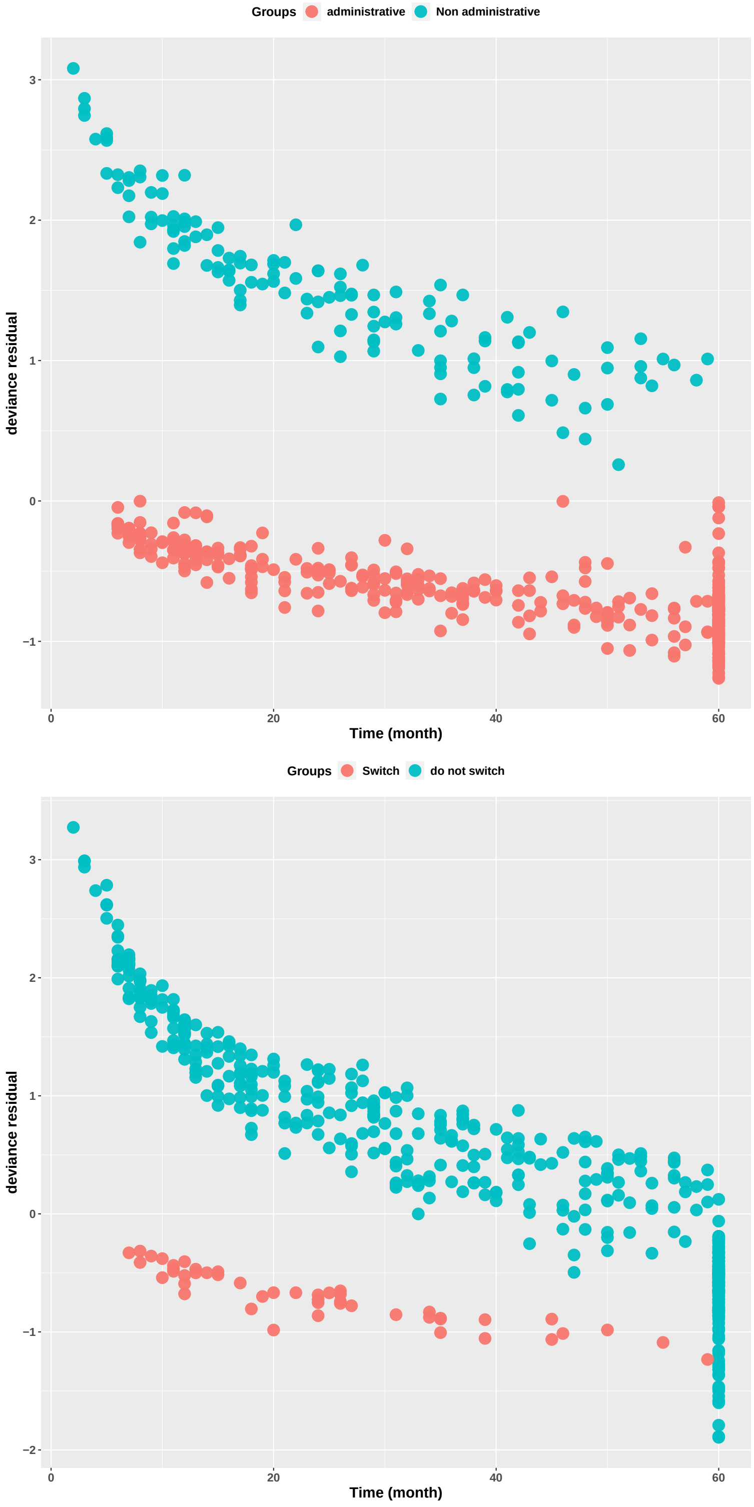


Figure SM–20: The deviance residuals to detect outlier based on (above) administrative censoring and (below) switching the treatment.

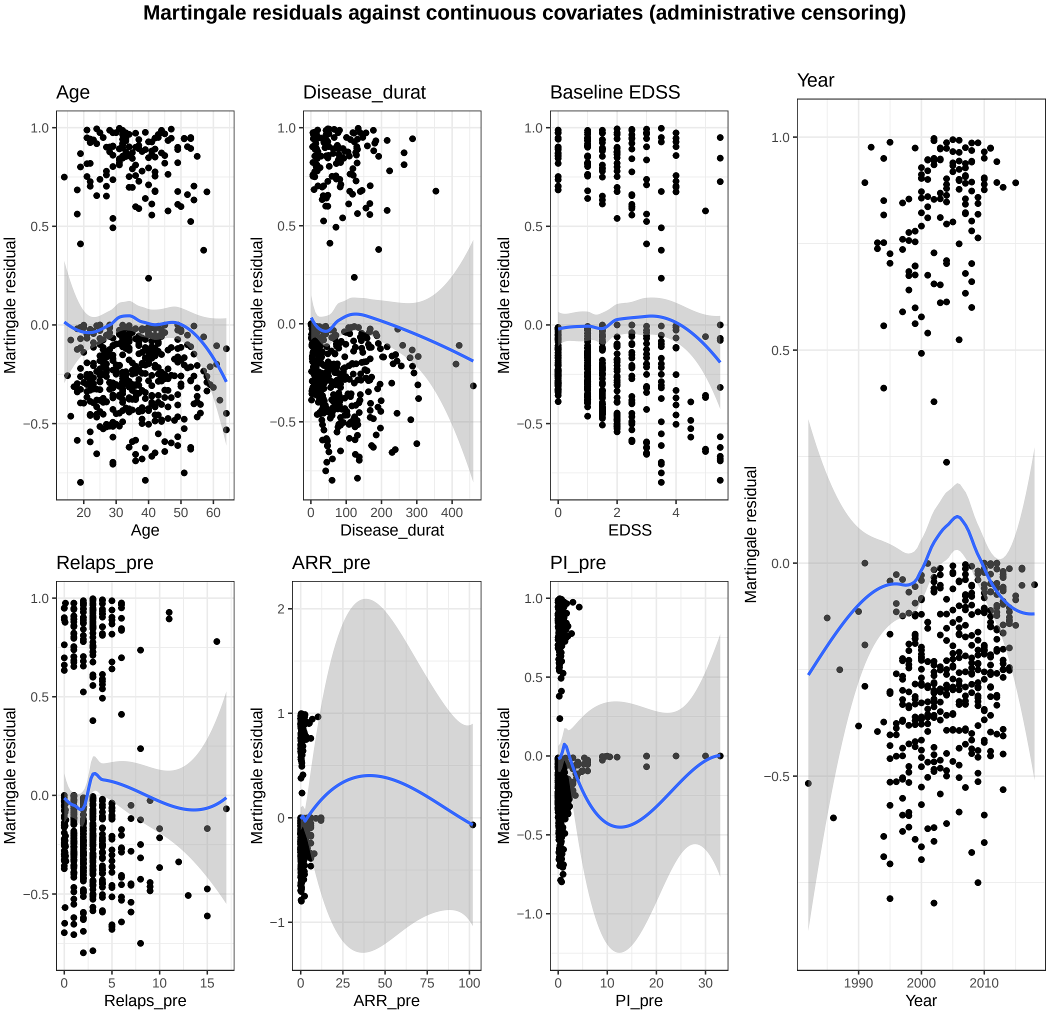


Figure SM–21: The Martingale residuals to choose a functional form for the covariate of administrative censoring.

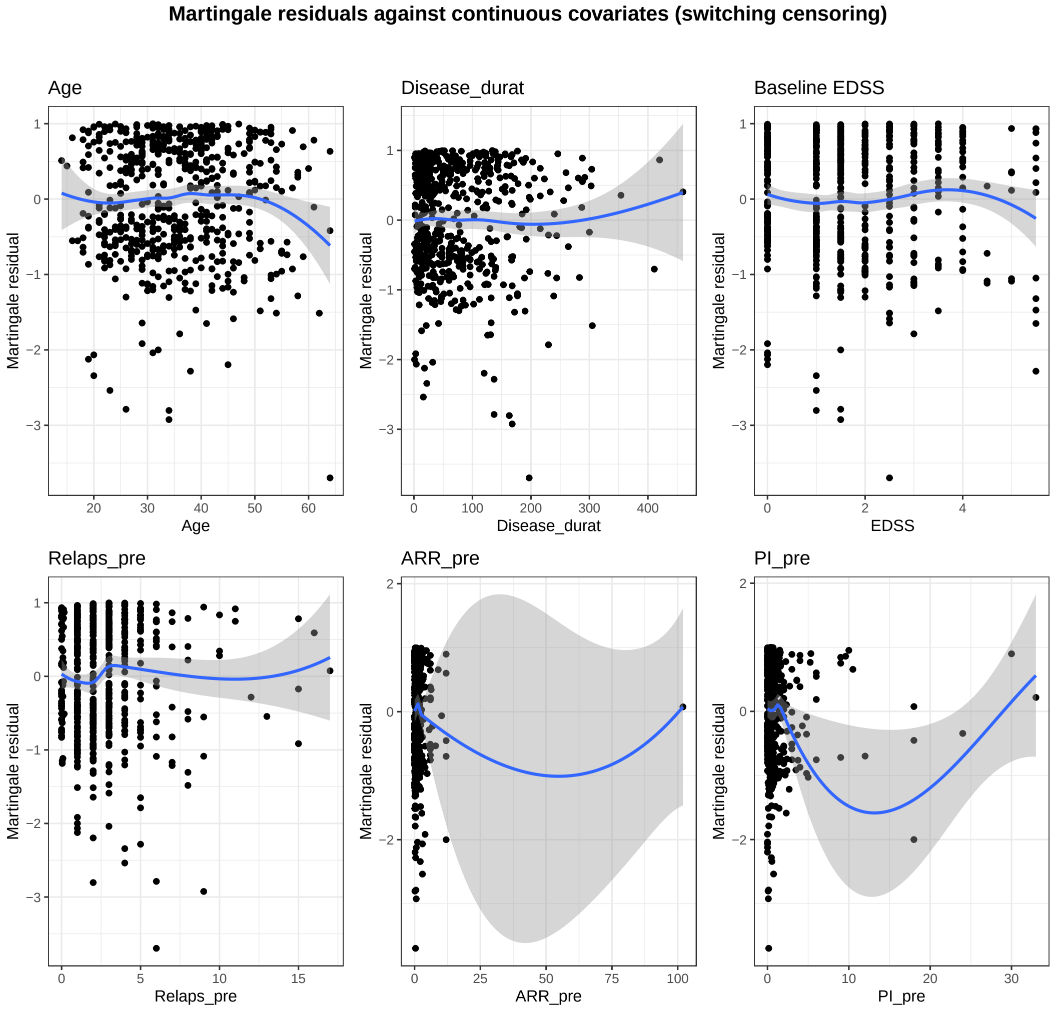


Figure SM–22: The Martingale residuals to choose a functional form for the covariate of switching censoring.

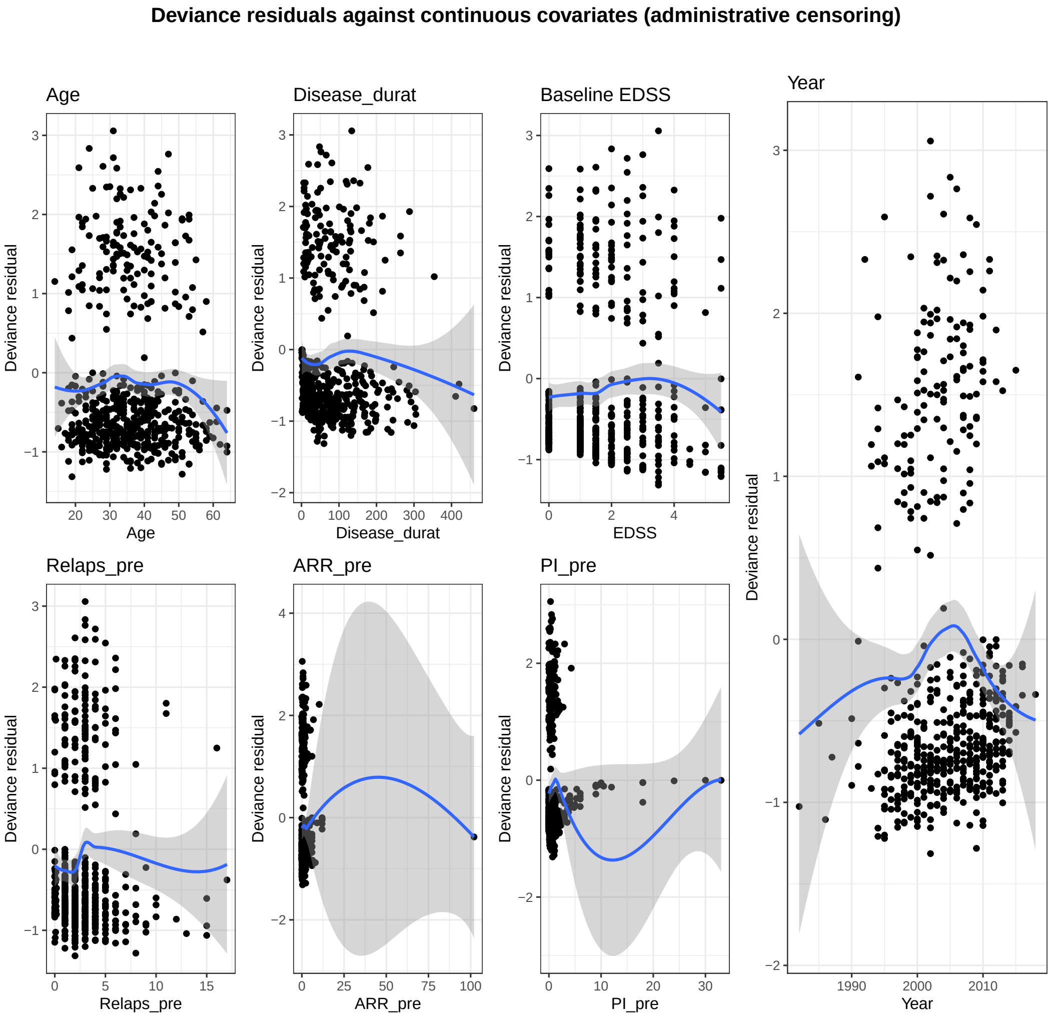


Figure SM–23: The Deviance residuals to choose a functional form for the covariate of administrative censoring.

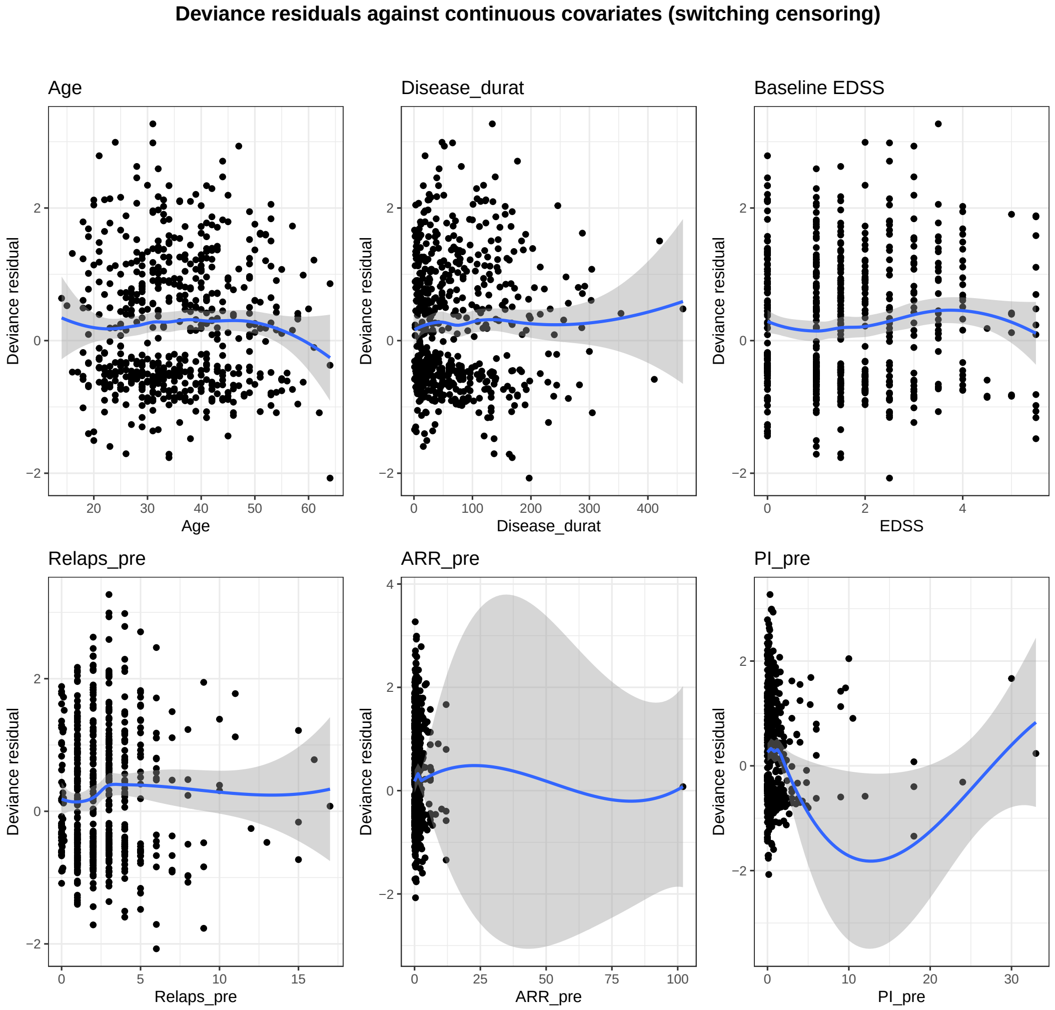


Figure SM–24: The Deviance residuals to choose a functional form for the covariate of switching censoring.

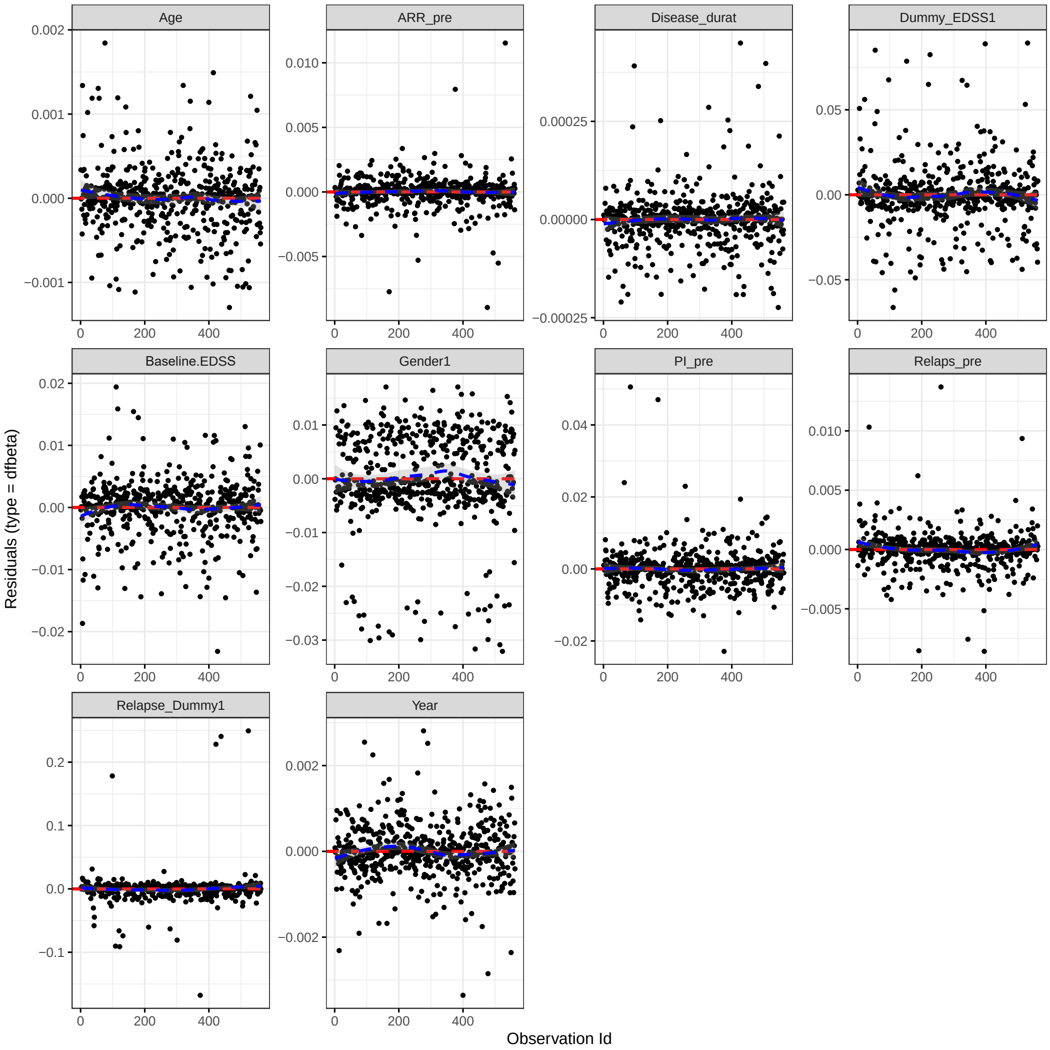


Figure SM–25: Testing Influential Observations by delta-beta residuals based on administrative censoring.

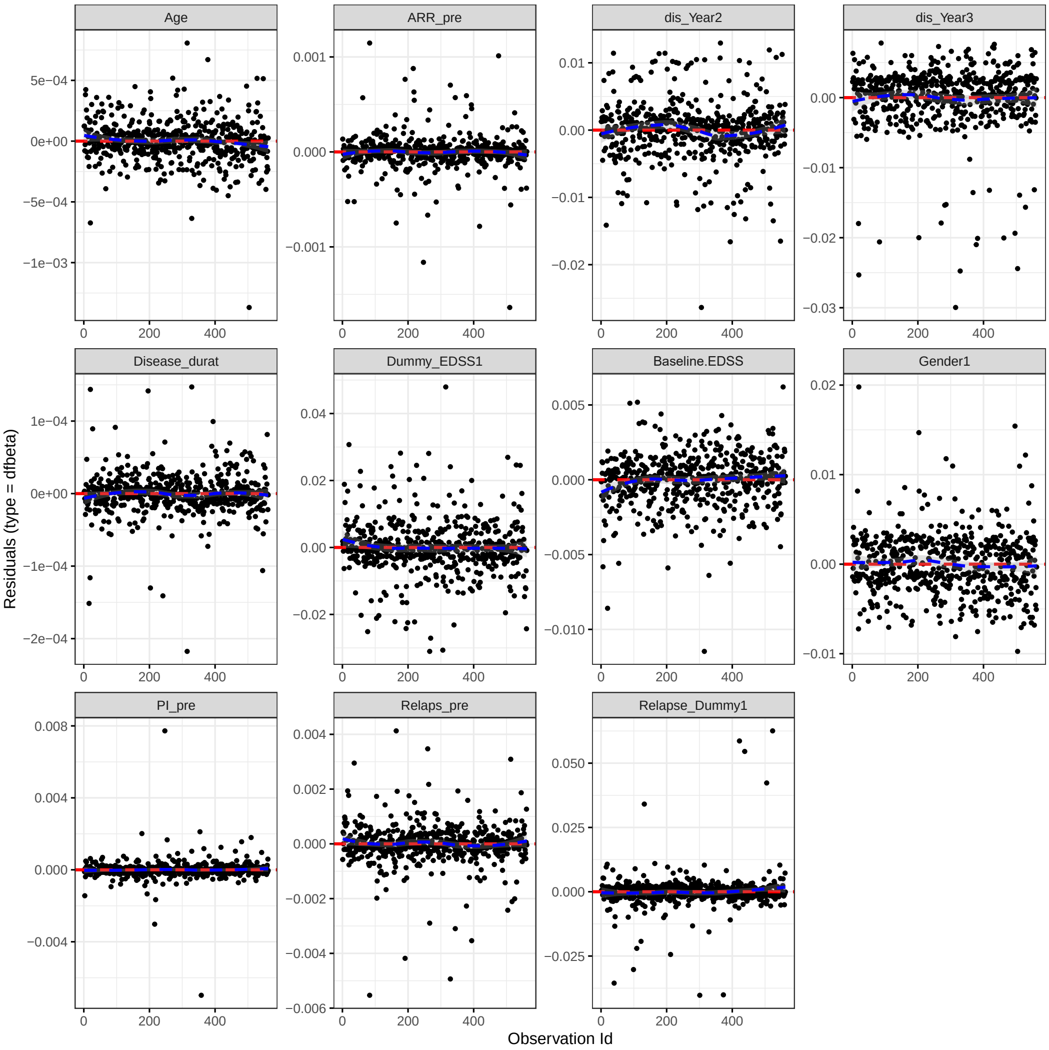


Figure SM–26: Testing Influential Observations by delta-beta residuals based on switching censoring.

Bibliography

A. Abadie and G. Imbens. Simple and bias-corrected matching estimators for average treatment effects, 2002.

A. Ahmed, M. W. Rich, P. W. Sanders, G. J. Perry, G. L. Bakris, M. R. Zile, T. E. Love, I. B. Aban, and M. G. Shlipak. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *The American journal of cardiology*, 99(3):393–398, 2007.

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

P. K. Andersen, E. Syriopoulou, and E. T. Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681, 2017.

T. Arbizu, J.C. Álvarez-Cermeño, G. Decap, O. Fernandez, D.F. Uria, A. Garcia Merino, G. Izquierdo, and X. Montalban. Interferon beta-1b treatment in patients with relapsing–remitting multiple sclerosis under a standardized protocol in spain. *Acta neurologica scandinavica*, 102(4):209–217, 2000.

P. C. Austin. The performance of different propensity-score methods for estimating relative risks. *Journal of clinical epidemiology*, 61(6):537–545, 2008a.

P. C. Austin. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and drug safety*, 17(12):1202–1217, 2008b.

P. C. Austin. Type i error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The international journal of biostatistics*, 5(1), 2009a.

P. C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009b.

P. C. Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29(6):661–677, 2009c.

P. C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in medicine*, 29(20):2137–2148, 2010.

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

P. C. Austin and M. M. Mamdani. A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics in medicine*, 25(12):2084–2106, 2006.

P. C. Austin and T. Schuster. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical methods in medical research*, 25(5):2214–2237, 2016.

P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28): 3661–3679, 2015.

P. C. Austin, P. Grootendorst, and G. M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4):734–753, 2007.

W. E. Barlow and R. L. Prentice. Residuals for relative risk regression. *Biometrika*, 75(1):65–74, 1988.

M.D. Benedetti, L. Massacesi, I. Tramacere, G. Filippini, L. La Mantia, A. Solari, et al. Non-inferiority of azathioprine versus interferon beta for relapsing remitting multiple sclerosis: A multicenter randomized trial. *Neuroepidemiology*, 39 (3-4):218–9, 2012.

N. Binder, T. A. Gerds, and P. K. Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis*, 20(2):303–315, 2014.

R. Braekers and N. Veraverbeke. Cox's regression model under partially informative censoring. *Communications in Statistics—Theory and Methods*, 34(8): 1793–1811, 2005.

N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics*, pages 437–453, 1974.

M.G. Brown, S. Kirby, C. Skedgel, J.D. Fisk, T.J. Murray, V. Bhan, and I. S. Sketris. How effective are disease-modifying drugs in delaying progression in relapsing-onset ms? *Neurology*, 69(15):1498–1507, 2007.

F. S. Chapin. *Experimental Designs in Sociological Reserach.* Harper & Brothers, 1947.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

P. J. Clare, T. A. Dobbins, and R. P. Mattick. Causal models adjusting for time-varying confounding—a systematic review of the literature. *International journal of epidemiology*, 48(1):254–265, 2019.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

W. G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968.

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

S. R. Cole and C. E. Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

S. R. Cole and M. A. Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

S. R. Cole, M. A. Hernán, J. M. Robins, K. Anastos, J. Chmiel, R. Detels, C. Ervin, J. Feldman, R. Greenblatt, L. Kingsley, Lai. S., Cohen M. Young, M., and A. Munoz. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American journal of epidemiology*, 158(7):687–694, 2003.

A. J. Coles, D.A. Compston, K. W. Selmaj, S. L. Lake, S. Moran, D. H. Margolin, K. Norris, P.K. Tandon, CAMMS223 Trial Investigators, et al. Alemtuzumab vs. interferon beta-1a in early multiple sclerosis. *The New England journal of medicine*, 359(17), 2008.

G. Coppola, R. Lanzillo, C. Florio, G. Orefice, P. Vivo, S. Ascione, V. Schiavone, A. Pagano, G Vacca, G. De Michele, et al. Long-term clinical experience with weekly interferon beta-1a in relapsing multiple sclerosis. *European journal of neurology*, 13(9):1014–1021, 2006.

D. R. Cox. Planning of experiments. 1958.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

R. Crump, V.J. Hotz, G.W. Imbens, and O. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

J. R. B. D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17 (19):2265–2281, 1998.

R. M. Daniel, B. L. De Stavola, and S. N. Cousens. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479–517, 2011.

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

R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

J. B. Dimick and E. H. Livingston. Comparing treatments using observational study designs: what can we do about selection bias? *Archives of surgery*, 145 (10):927–927, 2010.

F. Dominici, F. J. Bargagli-Stoffi, and F. Mealli. From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework. *arXiv preprint arXiv:2012.06865*, 2020.

G. C. Ebers, A. Traboulsee, D. Li, D. Langdon, A.T. Reder, D.S. Goodin, T. Bogumil, K. Beckmann, C. Wolf, A. Konieczny, et al. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal ifnb-1b trial. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(8):907–912, 2010.

M. C Elze, J. Gregson, U. Baber, E. Williamson, S. Sartori, R. Mehran, M. Nichols, G. W. Stone, and S. Pocock. Comparison of propensity score methods and covariate adjustment: Evaluation in four cardiovascular studies. *Journal of the American college of cardiology*, 69(3):345–357, 2017.

M. Etemadifar, M. Janghorbani, and V. Shaygannejad. Comparison of interferon beta products and azathioprine in the treatment of relapsing-remitting multiple sclerosis. *Journal of neurology*, 254(12):1723–1728, 2007.

G. Filippini, L. Munari, B. Incorvaia, G. C. Ebers, C. Polman, R. D'Amico, and G. P.A. Rice. Interferons in relapsing remitting multiple sclerosis: a systematic review. *The Lancet*, 361(9357):545–552, 2003.

O. Gout. Confounders in natural history of interferon-β–treated relapsing multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 63(1):126–126, 2008.

P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

S. Greenland, J. Pearl, and J. M. Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

E. Greenwood. Experimental sociology. In *Experimental Sociology.* Columbia University Press, 1945.

M. Greenwood. The natural duration of cancer (report on public health and medical subjects no 33). *London: Stationery Office*, 1926.

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

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

V. S. Harder, E. A. Stuart, and J. C. Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3):234, 2010.

F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

E. Havrdova, R. Zivadinov, J. Krasensky, M.G. Dwyer, I. Novakova, O. Dolezal, V. Ticha, L. Dusek, E. Houzvickova, J.L. Cox, N. Bergsland, S. Hussein, A. Svobodnik, Z. Seidl, Vaneckova, and M. D. Horakova. Randomized study of interferon beta-1a, low-dose azathioprine, and low-dose corticosteroids in multiple sclerosis. *Multiple Sclerosis Journal*, 15(8):965–976, 2009.

J. J. Heckman, H. Ichimura, and P. Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998.

M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

M. A. Hernán and J. M. Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International journal of obesity*, 32(3):S8–S14, 2008.

M. A. Hernán, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570, 2000.

M. A. Hernán, B. Brumback, and J. M. Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

M. A. Hernán, S. R. Cole, J. Margolick, M. Cohen, and J. M. Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety*, 14(7):477–491, 2005.

K. Hirano and G. W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Out. Res. Meth.*, 2:259–278, 2001.

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

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

P. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986a.

P. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986b.

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

X. Huang and R. A. Wolfe. A frailty model for informative censoring. *Biometrics*, 58(3):510–520, 2002.

MS Study Group. IFNB et al. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis: I. clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology*, 43(4):655–655, 1993.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Statist.*, 86:1–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction.* Cambridge University Press, New York, USA, 2015.

D. Jackson, I.R. White, S. Seaman, H. Evans, K. Baisley, and J. Carpenter. Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in medicine*, 33(27):4681–4694, 2014.

L. D. Jacobs, D. L. Cookfair, R. A. Rudick, R. M. Herndon, J. R. Richert, A. M. Salazar, J. S. Fischer, D. E. Goodkin, C. V. Granger, J. H. Simon, et al. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 39(3):285–294, 1996.

P. M. Jones, R. A. Cherry, B. N. Allen, K. M. B. Jenkyn, S. Z Shariff, S. Flier, K. N. Vogt, and D. N. Wijeysundera. Association between handover of anesthesia care and adverse postoperative outcomes among patients undergoing major surgery. *Jama*, 319(2):143–153, 2018.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data.* John Wiley & Sons, 1980.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

M. E. Karim, P. Gustafson, J. Petkau, Y. Zhao, A. Shirani, E. Kingwell, C. Evans, M. Van Der Kop, J. Oger, and H. Tremlett. Marginal structural cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *American journal of epidemiology*, 180(2):160–171, 2014.

L. Keele. Proportionally difficult: testing for nonproportional hazards in cox models. *Political Analysis*, 18(2):189–205, 2010.

A. P. Keil, J. K. Edwards, D. R. Richardson, A. I. Naimi, and S. R. Cole. The parametric g-formula for time-to-event data: towards intuition with a worked example. *Epidemiology (Cambridge, Mass.)*, 25(6):889, 2014.

M. Koch, J. Mostert, J. De Keyser, H. Tremlett, and G. Filippini. Interferon-β treatment and the natural history of relapsing-remitting multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 63(1):125–126, 2008.

N. R. Latimer, K.R. Abrams, P.C. Lambert, M.J. Crowther, A.J. Wailoo, J.P. Morden, R.L. Akehurst, and M.J. Campbell. Adjusting for treatment switching in randomised controlled trials–a simulation study and a simplified two-stage method. *Statistical methods in medical research*, 26(2):724–751, 2017.

N. R. Latimer, K. R. Abrams, P. C. Lambert, J. P. Morden, and M. J. Crowther. Assessing methods for dealing with treatment switching in clinical trials: a follow-up simulation study. *Statistical methods in medical research*, 27(3):765–784, 2018.

N.R. Latimer, K.R. Abrams, and U. Siebert. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. *BMC medical research methodology*, 19(1):1–19, 2019.

B. K. Lee, J. Lessler, and E. A. Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

L. Li and T. Greene. A weighting analogue to pair matching in propensity score analysis. *International Journal of Biostatistics*, 9:1–20, 2013.

F. D. Lublin, S. C. Reingold, J. A. Cohen, G. R. Cutter, P. S. Sørensen, A. J. Thompson, J. S. Wolinsky, L. J. Balcer, B. Banwell, F. Barkhof, et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3): 278–286, 2014.

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

J. Luo and R. Xu. Doubly robust inference for hazard ratio under informative censoring with machine learning. *arXiv preprint arXiv:2206.02296*, 2022.

study group MAGNIMS. Magnims consensus guidelines on the use of mri in multiple sclerosis—establishing disease prognosis and monitoring patients. *Nature Reviews Neurology*, 11(10):597–606, 2015.

H. Mao, L. Li, W. Yang, and Y. Shen. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in medicine*, 37(26):3745–3763, 2018.

T. Martinussen and S. Vansteelandt. On collapsibility and confounding bias in cox and aalen regression models. *Lifetime data analysis*, 19(3):279–296, 2013.

L. Massacesi, A. Parigi, A. Barilaro, A. M. Repice, G. Pellicanò, A. Konze, G. Siracusa, R. Taiuti, and L. Amaducci. Efficacy of azathioprine on multiple sclerosis new brain lesions evaluated using magnetic resonance imaging. *Archives of neurology*, 62(12):1843–1847, 2005.

L. Massacesi, I. Tramacere, M. D. Benedetti, G. Filippini, L. Lamantia, A. Solari, S. Amoroso, M. Battaglia, G. Tedeschi, and C. Milanese. Direct comparison of azathioprine and beta interferon efficacy in multiple sclerosis, 2013.

L. Massacesi, I. Tramacere, S. Amoroso, M. A. Battaglia, M. D. Benedetti, G. Filippini, L. La Mantia, A. Repice, A. Solari, G. Tedeschi, and C. Milanese. Azathioprine versus beta interferons for relapsing-remitting multiple sclerosis: a multicentre randomized non-inferiority trial. *PLoS One*, 9(11):e113371, 2014.

L. Massacesi, M. Grammatico, L. Vuolo, A. Barilaro, M.D. Benedetti, L. La Mantia, C. Milanese, A.M. Repice, A. Solari, G. Tedeschi, et al. Comparison of azathioprine and of beta interferon efficacy on measures of brain damage evaluated by mri in relapsing-remitting multiple sclerosis. In *EUROPEAN JOURNAL OF NEUROLOGY*, volume 23, pages 405–405, 2016.

A. Mattei, F. Mealli, and A. Nodehi. Design and analysis of experiments. In *Handbook of Labor, Human Resources and Population Economics*, pages 1–41. Springer, 2022.

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

P. McCullagh and J. A. Nelder. *Generalized linear models.* London: Chapman & Hall, 1983.

F. Mealli and D. B. Rubin. Clarifying missing at random and related definitions and implications when coupled with exchangeability. *Biometrika*, 102:995–1000, 2015.

G.J. Melendez-Torres, X. Armoiry, J. Patterson, A. Kan, P. Auguste, J. Madan, C. Counsell, O. Ciccarelli, A. Clarke, et al. Comparative effectiveness of beta-interferons and glatiramer acetate for relapsing-remitting multiple sclerosis: systematic review and network meta-analysis of trials including recommended dosages. *BMC neurology*, 18(1):1–17, 2018.

C. Milanese, L. La Mantia, R. Palumbo, V. Martinelli, A. Murialdo, M. Zaffaroni, D. Caputo, R. Capra, and R. Bergamaschi. A post-marketing study on interferon β 1b and 1a treatment in relapsing-remitting multiple sclerosis: different response in drop-outs and treated patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(12):1689–1692, 2003.

D. Moher, K. F. Schulz, D. Altman, Consort Group, CONSORT Group, et al. The consort statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Jama*, 285(15):1987–1991, 2001.

J. P. Morden, P. C. Lambert, N. Latimer, K. R. Abrams, and A. J. Wailoo. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC medical research methodology*, 11(1):1–20, 2011.

K. L. Morgan. Balancing covariates via propensity score weighting. Stochastic Modeling and Computational Statistics Seminar, 2014.

J. Neyman and K. Iwaszkiewicz. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society (series A)*, 2 (2):107–180, 1935.

A. Paolillo, C. Pozzilli, E. Giugni, V. Tomassini, C. Gasperini, M. Fiorelli, C. Mainero, M. Horsfield, S. Galgani, S. Bastianello, et al. A 6-year clinical and mri follow-up study of patients with relapsing–remitting multiple sclerosis treated with interferon-beta. *European Journal of Neurology*, 9(6):645–655, 2002.

S. M. Perkins, W. Tu, M. G. Underhill, X. Zhou, and M. D. Murray. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety*, 9(2):93–101, 2000.

J. A. V. Peterson. Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, 72(360a): 854–858, 1977.

S. Picciotto, M. A. Hernán, J. H. Page, J. G. Young, and J. M. Robins. Structural nested cumulative failure time models to estimate the effects of interventions. *Journal of the American Statistical Association*, 107(499):886–900, 2012.

C. Pozzilli, L. Prosperini, E. Sbardella, L. De Giglio, E. Onesti, and V. Tomassini. Post-marketing survey on clinical response to interferon beta in relapsing multiple sclerosis: the roman experience. *Neurological Sciences*, 26(4):s174–s178, 2005.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

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

J. M. Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.

J. M. Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, volume 24, page 3. San Francisco, USA, 1993.

J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000a.

J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000b.

J. M. Robins and D. M. Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

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

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129, 1995.

J. M. Robins, D. Blevins, G. Ritter, and M. Wulfsohn. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. *Epidemiology*, pages 319–336, 1992.

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

J.M. Robins and A. A. Tsiatis. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in statistics-Theory and Methods*, 20(8):2609–2631, 1991.

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

P. R. Rosembaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516—-524, 1984.

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

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

P. R. Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002.

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

P. Royston and M. K. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine*, 30(19):2409–2421, 2011.

P. Royston and M. K. Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, 13(1):1–15, 2013.

D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29: 159–183, 1973a.

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

D. B Rubin. Bayesian inference for causality: The importance of randomization. *Proceedings of the Social Statistics Section of the American Statistical Association*, pages 233–239, 1975.

D. B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34–58, 1978.

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318—-324, 1979.

D. B Rubin. Discussion of "randomization analysis of experimental data in the fisher randomization test" by basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

D. B Rubin. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292, 1990.

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

D. B. Rubin. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety*, 13(12):855, 2004.

D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

D. B. Rubin and N. Thomas. Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, pages 1079–1093, 1992a.

D. B. Rubin and N. Thomas. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, 79(4):797–809, 1992b.

D. B. Rubin and N. Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264, 1996.

D. B. Rubin and R. P. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, pages 206–222, 2006.

R. Rudick, C. Polman, D. Clifford, D. Miller, and L. Steinman. Natalizumab: bench to bedside and beyond. *JAMA neurology*, 70(2):172–182, 2013.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

D. Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980.

D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

N. Scolding, D. Barnes, S. Cader, J. Chataway, A. Chaudhuri, A. Coles, G. Giovannoni, D. Miller, W. Rashid, K. Schmierer, et al. Association of british neurologists: revised (2015) guidelines for prescribing disease-modifying treatments in multiple sclerosis. *Practical neurology*, 15(4):273–279, 2015.

S. Seaman, O. Dukes, R. Keogh, and S. Vansteelandt. Adjusting for time-varying confounders in survival analysis using structural nested cumulative survival time models. *Biometrics*, 76(2):472–483, 2020.

S. Setoguchi, S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

A. Shirani, Y. Zhao, M. E. Karim, C. Evans, E. Kingwell, M. L. Van Der Kop, J. Oger, P. Gustafson, J. Petkau, and H. Tremlett. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. *Jama*, 308(3):247–256, 2012.

L. A. Stefanski and D. D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

J. A.C. Sterne, M. A. Hernán, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J. M. Robins, and M. Egger. Long-term effectiveness of potent antiretroviral therapy in preventing aids and death: a prospective cohort study. *The Lancet*, 366(9483):378–384, 2005.

E. A. Stuart. 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 austin, statistics in medicine. *Statistics in medicine*, 27(12):2062–2065, 2008.

E. A. Stuart. 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, 2010.

E. A. Stuart and D. B. Rubin. Best practices in quasi-experimental designs. *Best practices in quantitative methods*, pages 155–176, 2008.

E. A. Stuart, G. King, K. Imai, and D. Ho. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of statistical software*, 2011.

E. J. Tchetgen Tchetgen and J. Robins. On parametrization, robustness and sensitivity analysis in a marginal structural cox proportional hazards model for point exposure. *Statistics & Probability Letters*, 82(5):907–915, 2012.

T. M. Therneau. *A Package for Survival Analysis in R*, 2021. URL `https://CRAN.R-project.org/package=survival`. R package version 3.2-13.

T. M. Therneau and P. M. Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.

T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

M. Tintore, A. Vidal-Jordana, and J. Sastre-Garriga. Treatment of multiple sclerosis—success from bench to bedside. *Nature Reviews Neurology*, 15(1):53–58, 2019.

M. Trojano, F. Pellegrini, A. Fuiani, D. Paolicelli, V. Zipoli, G. B. Zimatore, E. Di Monte, E. Portaccio, V. Lepore, P. Livrea, et al. New natural history of interferon-β–treated relapsing multiple sclerosis. *Annals of neurology*, 61(4):300–306, 2007.

A. A. Tsiatis. *Semiparametric theory and missing data.* Springer, 2006.

H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, H. Skali, S. Solomon, S. Jacobus, M. Hughes, M. Packer, and L. J. Wei. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of clinical Oncology*, 32 (22):2380, 2014.

S. Vansteelandt and R. M. Daniel. On regression adjustment for the propensity score. *Statistics in medicine*, 33(23):4053–4072, 2014.

S. Vansteelandt and M. Joffe. Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

C. Watkins, X. Huang, N. Latimer, Y. Tang, and E. J. Wright. Adjusting overall survival for treatment switches: commonly used methods and practical application. *Pharmaceutical statistics*, 12(6):348–357, 2013.

S.J.W. Willems, A. Schat, M.S. van Noorden, and M. Fiocco. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical methods in medical research*, 27(2):323–335, 2018.

S. Zeng, F. Li, and L. Hu. Propensity score weighting analysis of survival outcomes using pseudo-observations. *arXiv preprint arXiv:2103.00605*, 2021.

M. Zheng and J. P. Klein. A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula. *Communications in Statistics-Theory and Methods*, 23(8):2299–2311, 1994.

Y. Zhou, R. A. Matsouaka, and L. Thomas. Propensity score weighting under limited overlap and model misspecification. *arXiv preprint arXiv:2006.04038*, 2020.