Design and Analysis of Clinical Trials with Restricted Mean Survival Time

by

Kaiyuan Hua

Department of Biostatistics and Bioinformatics
Duke University

Defense Date: March 27, 2024

Approved:

Xiaofei Wang, Supervisor

Hwanhee Hong, Supervisor

Susan Halabi

Sin-Ho Jung

Christopher Granger

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2024

Abstract

Design and Analysis of Clinical Trials with Restricted Mean Survival Time

by

Kaiyuan Hua

Department of Biostatistics and Bioinformatics
Duke University

Defense Date: March 27, 2024

Approved:

Xiaofei Wang, Supervisor

Hwanhee Hong, Supervisor

Susan Halabi

Sin-Ho Jung

Christopher Granger

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2024

Abstract

Restricted mean survival time (RMST), a summary of survival time up to a pre-specified clinically relevant truncation time, is increasingly recognized as a measure for treatment effect in recent biomedical studies with time-to-event endpoints. The difference or ratio of RMST between two groups (e.g., treatment versus control) measures the relative treatment effect concerning a gain or loss of survival time. The RMST offers greater flexibility than the hazard ratio (HR), which is often estimated from the Cox proportional hazards model under the proportional hazards (PH) assumption. Due to delayed treatment effects or other biomedical reasons, the PH assumption is often violated in oncology and cardiovascular trials, leading to biased estimation and misleading interpretations of treatment effects. Compared to HR, RMST requires no PH assumption and offers a more straightforward interpretation of treatment effects. In this dissertation, we propose novel RMST-based methodologies for clinical trials with time-to-event endpoints in three research areas including 1) individual participant data network meta-analysis, 2) inference in multi-regional clinical trials, and 3) biomarker-guided adaptive and enrichment design.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations

AD	Aggregate data.
CW	Calibration weight/weighting/weighted.
HR	Hazard ratio.
IPD	Individual participant data.
IPCW	Inverse probability of censoring weight/weighting/weighted.
IPSW	Inverse probability of sampling weight/weighting/weighted.
KM	Kaplan-Meier
MRCT	Multi-regional clinial trials.
MSE	Mean square error.
NMA	Network meta-analysis.
NSCLC	Non-small-cell lung cancer.
OS	Overall survival.
PFS	Progression-free survival.
PH	Proportional hazard.
PQL	Penalized quasi-likelihood.
RCT	Randomized controlled trials.
RMST	Restricted mean survival time.

Acknowledgements

It has been an incredible seven-year journey since the first time I set foot at Duke University. Reflecting on this span of time, it is both impressive and humbling to see how much has transpired. The experiences, challenges, and growth encountered along the way have made it a truly transformative and enriching chapter of my life. In this dissertation, I extend my deepest gratitude to all who have walked with me on this challenging but fulfilling journey.

To begin with, words cannot express my gratitude to my two advisors, Dr. Hwanhee Hong and Dr. Xiaofei Wang, for their invaluable guidance, support and expertise throughout my PhD research. I would like to thank Hwanhee for creating numerous opportunities for collaborative work with clinicians at Duke Hospital and for her dedicated efforts to enhance my professional research and writing skills. Hwanhee's mentorship and encouragement have played a pivotal role in my development as a more proficient scientist and collaborator. I express my appreciation to Xiaofei for his invaluable insights and expertise in shaping the methodology and theoretical aspects of my dissertation and for his financial support for my PhD research. His connections facilitated my collaboration with researchers in the industry, broadening my professional horizons. Xiaofei's guidance and support have been instrumental in my growth as a more proficient independent researcher and statistician.

Next, I would like to express my gratitude to my committee members, Dr. Susan Halabi, Dr. Sin-Ho Jung, and Dr. Christopher Granger, for their unwavering support

and invaluable insights throughout my dissertation. I extend a special thanks to Dr. Granger for our collaboration efforts and his financial support. I also could not have undertaken this journey without the DGS of the Biostatistics PhD program, Dr. Sean O'Brien and Dr. Huiman Barnhart, the former DGS of the Biostatistics Master program, Dr. Megan Neely, the advisor of my research rotation, Dr. Sheng Luo, and the program coordinators, Michelle Evans, Tasha Allison, LaDonna Huseman, and Kendall Mincey. I would like to extend my sincere thanks to them for their unwavering support on my academic and personal pursuits.

Furthermore, I am grateful to the people I have collaborated with over the past five years at DCRI, Biogen and Sanofi: Dr. Anthony Carnicelli, Dr. Josephine Harrington, Dr. Manasi Tannu, Dr. Adam Nelson, Dr. Sydney Collins, Dr. Yiwei Zhang, Dr. George Kong, and Dr. Xiaodong Luo. Additionally, I appreciate the help of my course instructors, classmates, cohort members, and senior students.

Finally, I cannot overlook the immense support of my family, without whom achieving a PhD would have appeared insurmountable. I extend my heartfelt thanks to my parents for their support. Additionally, a special mention goes to my cat, Tirami, for providing both entertainment and emotional support throughout this journey. In particular, I want to convey my profound gratitude to my partner, Lei. She has been like a star, bringing brightness and joy into my life. I appreciate her kindness, patience, understanding, and the constant smiles she brings to every aspect of my life.

1

Introduction, Motivation and Background

1.1 Background

Hazard ratio (HR) is the most commonly used statistic for analyzing treatment effects in biomedical studies with time-to-event endpoints. It represents the ratio of instantaneous event rates between two groups, which is often derived from the Cox proportional hazards model (Cox, 1972). The Cox regression model characterizes the hazard as a function of explanatory covariates, which does not impose parametric assumptions on the baseline hazard function. A fundamental assumption for the Cox model is the proportional hazard (PH) assumption, which assumes that the HR between two groups remains constant over time. However, violation of the PH assumption renders HR a misleading and inappropriate summary of treatment effect, and the interpretation and inference of HRs become inaccurate and invalid (Lin and Wei, 1989). This occurs because HR is typically estimated as a weighted average of HR function over time, with these weights dependent on the censoring distribution. Consequently, when the PH assumption is violated, the use of these weights may complicate the use of HR for effective clinical decision-making (Tian et al., 2018).

Moreover, averaging HRs over the follow-up time introduces a conceptual dilemma when HRs in two or more follow-up time intervals have opposite directions (Finn et al., 2020).

The log-rank test (Bland and Altman, 2004) is the most commonly used statistical testing approach when comparing the survival distributions of two or more groups, such as different treatment groups in a clinical trial. The null hypothesis of the log-rank test is that there is no difference between groups in the probability of an event at any time point. While the log-rank test is known to be robust to non-PH, violations of the PH assumption may still considerably impact its statistical power to detect the treatment differences (Mukhopadhyay et al., 2022).

1.1.1 PH Assumption Violation in Clinical Trials

The PH assumption might face violation in clinical trials with prolonged follow-up and the presence of a time-dependent treatment effect (Gregson et al., 2019). Recently, non-constant HRs have been observed in many clinical trials of oncology or cardiovascular diseases due to delayed treatment effects or other biomedical reasons (Kantoff et al., 2010; Rajkumar et al., 2010; Brahmer et al., 2015; Robert et al., 2015; Borghaei et al., 2015; Reck et al., 2016; Fehrenbacher et al., 2016; Rittmeyer et al., 2017; Barlesi et al., 2018).

One motivation example is a biomarker-guided randomized Phase III oncology trial, JAVELIN Lung 200 (Barlesi et al., 2018). This trial aimed to compare the efficacy and safety of avelumab, an anti-PD-L1 immune checkpoint inhibitor antibody, with docetaxel, a standard care for patients with non-small-cell lung cancer (NSCLC). PD-L1 expression on tumor cells is a well-studied biomarker in NSCLC (Sankar et al., 2022), and previous studies have indicated a positive association between high PD-L1 expression levels and progression-free survival (PFS) and overall survival (OS)

after treatment with PD-L1 inhibitors for NSCLC (Shi and Zhao, 2020). In this trial, 792 patients were enrolled and randomized to avelumab (n=396) or docetaxel (n=396). Among them, 264 participants in the avelumab group and 265 in the docetaxel group had PD-L1-positive tumors, defined as PD-L1 expression $\geqslant 1\%$. In this study, avelumab was not found to have lower OS than docetaxel within PD-L1-positive patients, but exploratory analyses revealed that avelumab was associated with improved OS and PFS compared to docetaxel in patients with higher PD-L1 expression at $\geqslant 50\%$ and $\geqslant 80\%$ cutoffs (Barlesi et al., 2018).

However, the PH assumption was violated in this trial. As indicated in Supplemental Figure 1 in Barlesi et al. (2018), the survival curves of OS between two treatments crossed in the full analysis dataset. As shown in Supplemental Figure 3 in Barlesi et al. (2018), the survival curves of PFS between two treatments intersected among PD-L1-positive patients, and Avelumab exhibited a strong delayed effect relative to docetaxel for PFS in patients with PD-L1 expression $\geqslant 50\%$ and $\geqslant 80\%$. As such, the characterization of treatment effects based on the HR and its confidence interval estimated from the Cox PH model might not be valid, and the hypothesis testing based on the suboptimal log-rank test was less powerful.

1.1.2 Assessment of PH Assumption

There are two widely used PH assessment strategies: statistical test-based methods (Persson and Khamis, 2007) and graphical methods (Ng'andu, 1997). Commonly used test-based methods include testing time-dependency effects (Bellera et al., 2010), Schoenfeld's residual goodness-of-fit test (Schoenfeld, 1980; Grambsch and Therneau, 1994; Xue and Schifano, 2017), and a score process test based on cumulative martingale residual (Lin et al., 1993). As an alternative to the test-based approaches, visually examining survival curves allows researchers to closely monitor

the data patterns that might be deemed a violation of the PH assumption. Some options of the graphical approaches include the plot of log-cumulative hazard versus log-time for each subgroup of a covariate or the scatterplot of Schoenfeld's residuals over time.

However, criticisms surround the assessment of the PH assumption (Stensrud and Hernán, 2020). The test-based approaches may exhibit low statistical power in detecting a weak violation of the PH assumption (Stensrud and Hernán, 2020), even with a substantial number of observed events (Austin, 2018), and these tests are sensitive to the sample size. For example, a gross violation of PH might not be significant with a small sample, while a slight violation of PH could be deemed highly substantial with a large sample (Kleinbaum and Klein, 2012). In contrast, although the graphical methodologies are more straightforward to comprehend than the test-based approaches, decisions based on these methods are subjective and highly dependent on the researchers. Consequently, these approaches may only prove helpful in identifying a strong violation of the PH assumption.

Furthermore, assessing the PH assumption becomes more complicated in analyzing multiple clinical trials, such as in network meta-analysis (NMA) with individual participant data (IPD). A recent study (Hua et al., 2024) has addressed the limitations of the methods for PH assumption evaluation when combining multiple trials, leading to various potential. For example, when the PH assumption is violated in the pooled dataset but not in each trial or when the violation occurs in some trials but not in others. These concerns are exemplified by the COMBINE-AF study (Carnicelli et al., 2021, 2022), which incorporates IPD from four phase III randomized controlled trials of direct oral anticoagulants versus warfarin for stroke prevention in patients with atrial fibrillation. For example, when using Schoenfeld's residual test with a significance level of 0.05 to evaluate the PH assumption of treatment effect

for the event of bleeding, the PH assumption is violated in two trials but not in the other two (Table 1.1). In this case, using the Cox model may be problematic, and alternative approaches are necessary.

Table 1.1: The result from the Schoenfeld's residual test for evaluating PH assumption of treatment effect for the event of bleeding in COMBINE-AF.

Trials	ARISTOTLE	ENGAGE	RELY	ROCKET
P-value	0.10	0.06	<0.01	<0.01

1.2 Restricted Mean Survival Time

There are several alternatives to summarize the treatment effect from a time-to-event endpoint without relying on the PH assumption. Examples include median survival time difference, event rate ratio, absolute risk difference, and restricted mean survival time difference (Kim et al., 2017). Additionally, to address issues for the log-rank test under non-PH, various alternative tests have been proposed (Lin et al., 2020), such as weighted Kaplan-Meier test (Pepe and Fleming, 1989), weighted log-rank test (Harrington and Fleming, 1982), Max-Combo test (Karrison, 2016), and restricted mean survival time test (Royston and Parmar, 2013; Uno et al., 2014). In this dissertation, we emphasize proposing methodologies based on a compelling measurement: Restricted Mean Survival Time (RMST).

Estimating RMSTs does not require model assumptions as they are calculated non-parametrically (see Equations (1.1) and (1.2) below), and their interpretations are more straightforward than HRs across any distribution of time-to-event outcomes (Perego et al., 2020). Recently, RMST has been acknowledged in the FDA guidance for industry as an alternative approach to analyzing time-to-event data when non-PHs are expected (Food and Drug Administration (FDA), 2020), and used in many

trials reviewed by FDA (Food and Drug Administration (FDA), 2019a,b).

1.2.1 RMST Definitions

RMST was proposed by Irwin (1949) and popularized by Andersen et al. (2004); Royston and Parmar (2013); Uno et al. (2014); Tian et al. (2014) as a measure of treatment effect for time-to-event endpoints. RMST summarizes the survival time up to a clinically relevant (usually pre-specified) truncation time. Let T denote the event time and t^* denote the truncation time for evaluating the treatment effect. The RMST, $\mu(t^*)$, is defined as the mean of the truncated event time $Y = \min(T, t^*)$. The RMST can be calculated by the area under the survival curve $S(t) = P[T > t]$ from $t = 0$ to $t = t^*$ (Panel A in Figure 1.1), which is written as follows (Royston and Parmar, 2013):

$$\mu(t^*) = E(Y) = E[\min(T, t^*)] = \int_0^{t^*} S(t)dt.$$

The difference or ratio of RMST between two groups (e.g., treatment versus control) measures relative treatment effect concerning a gain or loss of event-free survival time up to the pre-defined truncation time (Kim et al., 2017). For example, as illustrated in Panel B in Figure 1.1, the RMST difference can be calculated by the area between two survival curves.

1.2.2 RMST Estimators

Suppose we have a random sample of n patients. For the i^{th} patient, let T_i denote the event time and be subject to right censoring with censoring time C_i. In the presence of right censoring, we observe $U_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I[T_i \leqslant C_i]$. The observed data are written as $\{(U_i, \delta_i), i = 1, \ldots, n\}$.

Typically, one can use the Kaplan-Meier (KM) estimator of the survival function

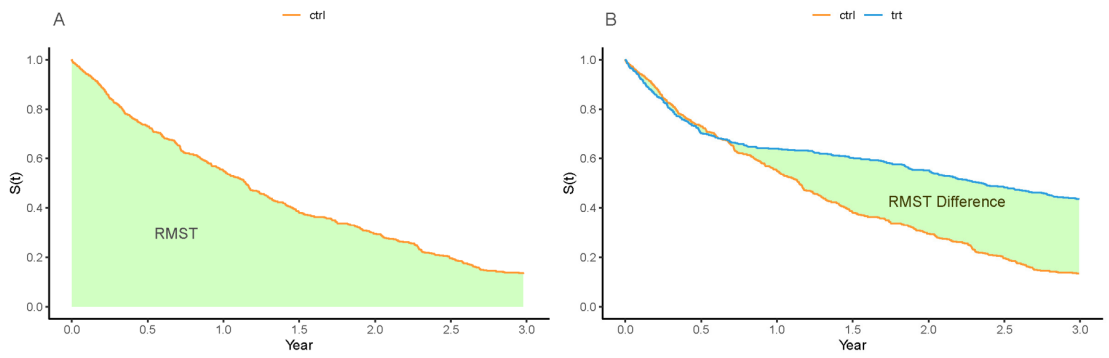

FIGURE 1.1: (A) Restricted mean survival time (RMST), and (B) RMST difference between treatment and control groups.

to estimate $\mu(t^*)$ nonparametrically. For $i = 1, ..., n$, let $N_i(t) = I[U_i \leqslant t; \delta_i = 1]$ denote the individual counting process and $Y_i(t) = I[U_i \geqslant t]$ denote the individual at risk process. The KM estimator of the survival function is written as

$$\hat{S}(t) = \prod_{u \leqslant t} \left\{ 1 - \frac{dN(u)}{Y(u)} \right\},$$

where $N(u) = \sum_{i=1}^{n} N_i(u)$ and $Y(u) = \sum_{i=1}^{n} Y_i(u)$. Consequently, the KM estimator of $\mu(t^*)$ is written as:

$$\hat{\mu}_{KM}(t^*) = \int_0^{t^*} \hat{S}(t)dt, \tag{1.1}$$

where $\hat{S}(t)$ is the KM estimator of the survival function for the event time. According to Zhao et al. (2016); Tian et al. (2018), as the sample size $n \to \infty$, $\hat{\mu}_{KM}(t^*)$ is consistent to $\mu(t^*)$ and $\sqrt{n}\{\hat{\mu}_{KM}(t^*) - \mu(t^*)\} \xrightarrow{d} N\left(0, \sigma^2(t^*)\right)$, where $\sigma^2(t^*)$ is estimated by:

$$\hat{\sigma}^2(t^*) = \int_0^{t^*} \left\{ \int_t^{t^*} \hat{S}(u)du \right\}^2 \frac{d\hat{\Lambda}(t)}{Y(t)/n}.$$

Here, $\hat{\Lambda}(t) = -\log\{\hat{S}(t)\}$ is the Nelson-Aalen estimator of the cumulative hazard rate

function.

Another estimator of RMST was proposed by Bang and Tsiatis (2000) by using the inverse probability of censoring weight (IPCW). For $i = 1, \ldots, n$, let $Y_i = \min(T_i, t^*)$ denote the individual truncated time-to-event outcome at t^*. In the absence of censoring, an obvious estimator for $\mu(t^*)$ is $n^{-1}\sum_{i=1}^{n} Y_i$. In the presence of right censoring, the IPCW estimator of $\mu(t^*)$ is

$$\hat{\mu}_{IPCW}(t^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{G}(Y_i)}Y_i, \tag{1.2}$$

where $\delta_i^* = I[C_i \geqslant Y_i]$, and $\hat{G}(Y_i) = P(C_i > Y_i)$ is the KM estimator of the survival function for the censoring time based on $\{(U_i, 1-\delta_i), i = 1, \ldots, n\}$. To understand the validity of $\hat{\mu}_{IPCW}(t^*)$, note that

$$E\left[\frac{\delta_i^*}{\hat{G}(Y_i)}Y_i|T_i\right] = Y_i\frac{P(C_i \geqslant Y_i|T_i)}{P(C_i > Y_i)} = Y_i.$$

1.2.3 RMST Regression

The regression models for RMST allow flexibility in incorporating individual covariates. Here, we introduce the IPCW RMST regression model proposed by Tian et al. (2014). Suppose $\tilde{\mathbf{X}}$ is a $p \times 1$ covariate vector for a study participant, where $\tilde{\mathbf{X}}$ may contain any exploratory variables, including intercept, treatment indicators, baseline covariates, and treatment-by-covariate interactions. We assume conditionally independent censoring, formally written as $C \perp T|\tilde{\mathbf{X}}$. In a study with n participants, the observed data are written as $\{(U_i, \delta_i, \tilde{\mathbf{X}}_i), i = 1, \ldots, n\}$. We define the conditional

RMST as $\mu(t^*|\tilde{\mathbf{X}}) = E[\min(T, t^*)|\tilde{\mathbf{X}}] = E[Y|\tilde{\mathbf{X}}]$.

The IPCW RMST regression model is written as follows:

$$g(\mu(t^*|\tilde{\mathbf{X}})) = \tilde{\mathbf{X}}^T\boldsymbol{\beta}, \tag{1.3}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and $g(\cdot)$ is the link function. The log or identity link is often used in the IPCW RMST regression. Since the distribution of Y is unknown, a quasi-likelihood method is applied to estimate the model parameters (Tian et al., 2014; Gardiner, 2021). For the i^{th} patients, let $\mu_i = E[Y_i|\tilde{\mathbf{X}}_i]$ and assume its variance is $\text{Var}(\mu_i) = \phi v(\mu_i)$, where ϕ is a dispersion parameter and $v(\cdot)$ is a specified variance function. By assuming the link function $g(\cdot)$ is a canonical link (McCullagh, 2019), we have $v(\cdot) = 1/g'(\cdot)$. The parameters $\hat{\boldsymbol{\beta}}$ can be estimated from the following weighted quasi-score estimating equation:

$$U_n(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} w_i \frac{\partial \mu_i}{\partial \hat{\boldsymbol{\beta}}} \frac{Y_i - \mu_i}{\phi v(\mu_i)} = \sum_{i=1}^{n} w_i \left\{ Y_i - g^{-1}(\tilde{X}_i^T \hat{\boldsymbol{\beta}}) \right\} \tilde{X}_i^T = 0,$$

where $w_i = \delta_i^*/\hat{G}(Y_i)$ is the inverse probability of censoring weight for the i^{th} participant, as defined in Equation (1.2). Note that although Y_i may not be observable for all participants due to censoring, the parameter $\boldsymbol{\beta}$ is still estimable in this model (Tian et al., 2014). The large sample properties for the estimated coefficient $\hat{\boldsymbol{\beta}}$ have been discussed in Tian et al. (2014); Gardiner (2021). We provide their proofs in Appendix A.

There is another RMST regression model using pseudo observations proposed by Andersen et al. (2004). As our proposed methods in the following chapters are primarily based on the IPCW RMST regression, we briefly introduce the pseudo-value RMST regression in Appendix A.

1.3 Objectives

As RMST demonstrates potential benefits and advantages in analyzing time-to-event data in clinical studies, novel RMST-based methodologies are actively developed across various research areas in clinical trials. In this dissertation, we aim to address several methodological and practical gaps in RMST-based approaches within the following three research areas:

1.3.1 Individual Participant Data Network Meta-Analysis

Network meta-analysis (NMA) compares multiple treatments simultaneously by combining direct and indirect comparisons. The availability of individual participant data (IPD) makes it possible to evaluate treatment effect moderation and to draw inferences about treatment effects by taking the full utilization of individual covariates from multiple clinical trials. In IPD-NMA, RMST models have gained popularity when analyzing time-to-event outcomes because they offer more straightforward interpretations of treatment effects with fewer assumptions than HRs commonly estimated from Cox models. However, current approaches need to incorporate individual covariates properly to evaluate the effect moderation. In this study, we propose advanced models that fully utilize individual-level covariates to adjust RMST by covariates and estimate treatment-by-covariate interaction effects in IPD-NMA. Our models enable us to study treatment effect moderation and provide a comprehensive understanding of the comparative effectiveness of treatments and subgroup effects. The methods are evaluated by an extensive simulation study and illustrated using a real NMA example about treatments for patients with atrial fibrillation.

1.3.2 Inference in Multi-Regional Clinical Trials

Multi-regional clinical trials (MRCTs) are increasingly pivotal in the global development of pharmaceutical products by accelerating the process of data gathering and regulatory approval for needed patients worldwide. In MRCTs, the baseline characteristics of study participants often differ across regions due to the differences in recruitment practice and genetic and demographic profiles of the relevant regions. Existing methods have ignored the incomparability of baseline covariates, inevitably yielding biased assessments of treatment effects and misleading conclusions on the consistency of treatment effects across regions. Moreover, limited attention has been given to time-to-event outcomes. In this study, we employ RMSTs and propose inference methods for assessing regional treatment effect heterogeneity with consideration for eliminating regional imbalance due to extrinsic regional factors. Theoretical work is utilized to establish the large sample properties of the proposed methods. The methods are compared through extensive simulation studies and illustrated using a real MRCT on acute coronary syndromes.

1.3.3 Biomarker-Guided Adaptive and Enrichment Design

Biomarker-guided designs are increasingly used to evaluate personalized treatments based on patient's biomarker status in Phase II and III clinical trials. With adaptive enrichment, these designs can improve the efficiency of assessing the treatment effect in biomarker-positive patients by increasing their proportion in the randomized trial. While time-to-event outcomes are often used as the primary endpoint to measure treatment effects for a new therapy in severe diseases like cancer and cardiovascular diseases, there is limited research on biomarker-guided adaptive enrichment trials with time-to-event endpoints. Such trials almost always utilize HRs for statistically measuring treatment effects under the PH assumption. In this study, we propose a

novel two-stage biomarker-guided adaptive RMST design with threshold detection and patient enrichment, and our design strategy does no require the PH assumption. We develop sophisticated methods for identifying the optimal biomarker threshold, treatment effect estimators in the biomarker-positive subgroup, and approaches for type I error rate, power analysis, and sample size calculation. An extensive simulation study is presented to evaluate the performance of the proposed design.

2

Topic 1: Network Meta-Analysis of Time-to-Event Endpoints with Individual Participant Data Using Restricted Mean Survival Time Regression

2.1 Introduction

2.1.1 Individual Participant Data Network Meta-Analysis

As an extension of traditional pairwise meta-analysis (DerSimonian and Laird, 1986), network meta-analysis (NMA) evaluates more than two treatments simultaneously by integrating findings from multiple trials, particularly randomized controlled trials (RCTs), and accounting for heterogeneity across the trials (Lu and Ades, 2004; Salanti et al., 2008; Mills et al., 2013; White, 2015; Hong et al., 2016). With limited evidence about head-to-head comparisons in RCTs, NMA can leverage information about both "direct" and "indirect" comparisons of treatment, providing comprehensive and integrative inferences about the comparative effectiveness of treatments (Saramago et al., 2014; Lu and Ades, 2004). NMA often combines aggregate data (e.g., summary statistics), denoted by AD-NMA, but thanks to recent advances in data sharing and increasing access to individual participant data (IPD),

statistical methods for NMA using IPD, denoted by IPD-NMA, have been actively developed (Donegan et al., 2013; Hong et al., 2015; Debray et al., 2015). Using IPD in NMA allows researchers to inform personalized treatment decisions by assessing treatment effect moderation and subgroup effects using individual-level covariates (Lyman and Kuderer, 2005; Riley et al., 2010; Donegan et al., 2012; Hong et al., 2015).

2.1.2 Approaches in IPD-NMA

There are two approaches for conducting IPD-NMA (Simmonds et al., 2005): two-stage and one-stage approaches. The two-stage approach first estimates treatment effects within each study, and then combines the study-level effect estimates across studies using an AD-NMA model (Debray et al., 2013). On the other hand, the one-stage approach combines the IPD of all studies included in the NMA simultaneously as a single step by using a hierarchical model (Bowden et al., 2011). The differences between these two methods have been extensively discussed through empirical, theoretical, or simulation studies by Stewart et al. (2012); Debray et al. (2013); Burke et al. (2017); Hua et al. (2017), and Kontopantelis (2018). The benefits of using the one-stage methods include: 1) they use an exact likelihood specification instead of assuming the asymptotic normality of study estimates that are usually used in the two-stage methods; 2) they can account for within-study parameter correlation more efficiently, especially in the presence of missing outcomes; 3) the one-stage methods provide higher power and lower bias when estimating treatment-by-covariate interactions to assess effect moderation; 4) the one-stage methods can provide reliable results with limited numbers of studies and study participants or events per study are available. In contrast, the two-stage methods have fewer computational burdens compared to the one-stage methods, and the two-stage methods can be used when a

few studies provide IPD and the remaining studies provide aggregate data.

2.1.3 IPD-NMA with time-to-event outcome

When analyzing a time-to-event outcome in NMA, IPD has many advantages (Smith et al., 2005). First, we can apply the same outcome model and assess model assumptions consistently across trials. Second, a common metric quantifying treatment effects concerning time-to-event outcomes (e.g., HR or RMST) can be estimated directly in all available trials in IPD-NMA, regardless of what was reported in publications. Third, the individual-level covariates enable us to assess treatment effect moderation accurately compared to aggregate covariates. Many statistical methods for time-to-event outcomes typically rely on HRs, usually estimated by Cox PH models (Cox, 1972). NMA with time-to-event outcomes is not an exception, and IPD-NMA models based on Cox regression have been proposed by Smith et al. (2005) and Crowther et al. (2012). However, Cox regression models rely heavily on the PH assumption and could result in misleading HR estimates and an inappropriate summary of treatment effect when the assumption is violated (Lin and Wei, 1989).

In contrast, RMST has a more straightforward clinical interpretation than HRs and provides a valid measure for treatment effects under any distributions of time-to-event outcomes with no model assumptions (Perego et al., 2020). Another advantage of RMST is its ability to estimate treatment effects at different time points, avoiding the conceptual dilemma of averaging HRs over time, especially when survival curves between two groups cross each other, which is an indication of PH assumption violation. Leveraging these advantages, RMST methods have been recently incorporated into NMA with IPD (Wei et al., 2015; Lueza et al., 2016; Weir et al., 2021; Tang and Trinquart, 2022). A two-stage IPD-NMA model is commonly used where the first

stage obtains RMST estimates per treatment arm in each study, and then the second stage conducts a traditional study-level NMA using the RMST estimates. However, this model does not properly adjust for covariates in the first stage because RMSTs are estimated by a simple integration of a survival curve (i.e., Equation (1.1)) in the first stage. Moreover, the existing methods use a contrast-based parameterization that directly models relative treatment effects, such as an RMST difference between two treatments. As an alternative, an arm-based parameterization can directly model the responses observed in each of the treatment arms (Hawkins et al., 2016), and it enables direct modeling of RMSTs rather than RMST differences in NMA, which provides straightforward parameter interpretations (White et al., 2019). By estimating RMSTs, we can calculate the RMST differences for all pairwise comparisons of treatments.

To the best of our knowledge, no formal model has been developed for one-stage IPD-NMA using RMST. Moreover, if covariates are properly used to estimate treatment effect in the first stage, a two-stage model would still be able to adjust the covariates when estimating treatment effect. As such, in this paper, we propose advanced two-stage and one-stage IPD-NMA methods for time-to-event outcomes using the arm-based parameterization to estimate RMST. Our methods fully utilize individual-level covariates to adjust RMST by covariates and estimate treatment-by-covariate interaction effects.

2.1.4 Chapter Outlines

The remainder of this chapter is structured as follows. Chapter 2.2 introduces our proposed RMST IPD-NMA methods. Chapter 2.3 provides the results from an extensive simulation study. In Chapter 2.4, we illustrate the proposed methods using a real NMA example about treatments for patients with atrial fibrillation (Carnicelli

et al., 2021, 2022). We conclude the paper with discussions and limitations of the proposed methods in Chapter 2.5.

2.2 Methods

We propose two RMST IPD-NMA models under (1) two-stage and (2) one-stage approaches. Both models estimate RMST adjusted by covariates using the inverse probability of censoring weighted regression method (Tian et al., 2014). The models are developed under the consistency assumption, where direct and indirect comparisons of treatments are the same (Higgins et al., 2012).

2.2.1 Two-Stage RMST IPD-NMA Model

The first stage of the two-stage RMST IPD-NMA model analyzes individual studies separately to estimate RMSTs and treatment-by-covariate interaction effects by fitting an inverse probability of censoring weighted RMST regression as introduced in Chapter 1.2.2. Suppose T_{ij} is the time-to-event outcome and C_{ij} is the censoring time, where i and j index participants and studies such as $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. The truncated time-to-event outcome is $Y_{ij} = \min(T_{ij}, t^*)$. Suppose we evaluate K treatments and $Z_{ij} \in \{1, \ldots, K\}$ is the treatment indicator. We consider P covariates, where the p^{th} covariate is X_p, and we let $\mathbf{X}_{ij}$ be a $P \times 1$ covariate vector specific for Patient i in Study j. Let $\mu_{ij} = E(Y_{ij}|\mathbf{X}_{ij}, Z_{ij})$ be the conditional RMST, then the RMST regression model for the j^{th} study in the first stage is written as follows:

$$g(\mu_{ij}) = \sum_{k=1}^{K} \alpha_{jk} I[Z_{ij} = k] + \sum_{k=1}^{K} \mathbf{X}_{ij} \boldsymbol{\beta}_{jk} I[Z_{ij} = k]. \tag{2.1}$$

Typical choices of the link function $g(\cdot)$ can be log link or identity link. We will use the log link function (i.e., $g(\cdot) = \log(\cdot)$) throughout this chapter. Note that

our model under arm-based parameterization does not need a separate intercept or main covariate effect parameters in the model (Hong et al., 2015, 2016). Here, α_{jk} is the log-RMST up to t^* for Treatment k and $\mathbf{X}_{ij} = 0$ in the j^{th} study. We define $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots, \beta_{jkP})^T$, where β_{jkp} is the study-specific treatment-by-covariate interaction effect for $p = 1, \ldots, P$. The resulting estimates in the j^{th} study are denoted as $(\hat{\alpha}_{j1}, \ldots, \hat{\alpha}_{jK}, \hat{\boldsymbol{\beta}}_{j1}^T, \ldots, \hat{\boldsymbol{\beta}}_{jK}^T)$ and their associated study-specific variance-covariance matrix $\hat{\Sigma}_j$, where $\hat{\boldsymbol{\beta}}_{jk} = (\hat{\beta}_{jk1}, \ldots, \hat{\beta}_{jkP})^T$.

In the second stage, the estimated log-RMSTs, $(\hat{\alpha}_{j1}, \ldots, \hat{\alpha}_{jK})$, and treatment-by-covariate interactions, $(\hat{\boldsymbol{\beta}}_{j1}^T, \ldots, \hat{\boldsymbol{\beta}}_{jK}^T)$, are analyzed using a multivariate meta-analysis model with restricted maximum likelihood estimation (Chen et al., 2012; Schwarzer et al., 2015). The multivariate meta-analysis model is written as follows:

$$\begin{aligned} \hat{\alpha}_{jk} &= \alpha_k + a_{jk} + \epsilon_{jk}, \\ \hat{\beta}_{jkp} &= \beta_{kp} + b_{jkp} + \eta_{jkp}, \end{aligned} \tag{2.2}$$

for $j = 1, \ldots, J$, $k = 1, \ldots, K$, and $p = 1, \ldots, P$. Here, α_k is the fixed (mean) main treatment effect, interpreted by the log-RMST up to t^* of Treatment k for $\mathbf{X}_{ij} = 0$, and β_{kp} is the fixed (mean) treatment-by-covariate interaction effect between Treatment k and X_p. If X_p is a binary covariate, β_{kp} is the difference in log-RMST for Treatment k between $X_p = 1$ and $X_p = 0$. If X_p is a continuous covariate, β_{kp} is the change in log-RMST for Treatment k per unit change in X_p. Both a_{jk} and b_{jkp} are associated with random effects α_k and β_{kp}, respectively. Let $\mathbf{b}_{jk} = (b_{jk1}, \ldots, b_{jkP})^T$, we assume that the random effects parameter vector, $\mathbf{r}_j = (a_{j1}, \ldots, a_{jK}, \mathbf{b}_{j1}^T, \ldots, \mathbf{b}_{jK}^T)$, independently follows a multivariate normal distribution, namely $MVN(0, \mathbf{R}_2)$. Here $\mathbf{R}_2$ is an unstructured covariance-variance matrix of the random effects that describes the between-study heterogeneity for each parameter in Equations (2.2) and the diagonal element of $\mathbf{R}_2$ is the variance component of the between-study heterogene-

ity corresponding to each parameter. In addition, ϵ_{jk} and η_{jkp} are the associated residuals for α_k and β_{kp}, respectively, which are assumed to be independent. Let $\boldsymbol{\eta}_{jk} = (\eta_{jk1}, \ldots, \eta_{jkP})$, then the residual vector $(\epsilon_{j1}, \ldots, \epsilon_{jK}, \boldsymbol{\eta}_{j1}^T, \ldots, \boldsymbol{\eta}_{jK}^T)$ is assumed to follow a multivariate normal distribution with mean 0 and covariance-variance matrix $\hat{\Sigma}_j$, i.e. $MVN(0, \hat{\Sigma}_j)$.

We use the restricted maximum likelihood method (Jackson et al., 2011) to fit the model in Equations (2.2), implemented using R package `mvmeta` (version 1.0.3) (Gasparrini et al., 2012). Alternatively, one can also use a multivariate version of the DerSimonian-Laird method of moments estimator (Jackson et al., 2013) or Bayesian approach (Wei and Higgins, 2013) to fit this multivariate model with random effects.

2.2.2 One-Stage RMST IPD-NMA Model

In the one-stage model, IPD from all studies included in NMA are modeled simultaneously. The IPCW RMST regression model is extended to combine multiple studies and account for between-study heterogeneity. We use the penalized quasi-likelihood (PQL) method (Breslow and Clayton, 1993) to fit the one-stage model with random effects as our truncated event time outcome does not follow a normal distribution. Followed by the same notations as in the two-stage RMST IPD-NMA model, given the study level random effects, Y_{ij} are assumed to be conditionally independent with the mean $E(Y_{ij}|\mathbf{r}_j) = \mu_{ij}^{(\mathbf{r}_j)}$ and the variance $\text{Var}(Y_{ij}|\mathbf{r}_j) = \phi v(\mu_{ij}^{(\mathbf{r}_j)})$. Here, $\mu_{ij}^{(\mathbf{r}_j)}$ is related to the value of random effects $\mathbf{r}_j = (a_{j1}, \ldots, a_{jK}, \mathbf{b}_{j1}^T, \ldots, \mathbf{b}_{jK}^T)$, $v(\cdot)$ is a variance function as introduced in Chapter 1.2.2, and ϕ is the dispersion parameter. The one-stage RMST IPD-NMA model is written as follows:

$$g(\mu_{ij}^{(\mathbf{r}_j)}) = \sum_{k=1}^{K} \alpha_{jk} I[Z_{ij} = k] + \sum_{k=1}^{K} \mathbf{X}_{ij} \boldsymbol{\beta}_{jk} I[Z_{ij} = k]. \tag{2.3}$$

For $k = 1, \ldots, K$, we define $\alpha_{jk} = \alpha_k + a_{jk}$ and $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots, \beta_{jkP})^T$, where $\beta_{jkp} = \beta_{kp} + b_{jkp}$ $(p = 1, \ldots, P)$. Here, α_k, a_{jk}, β_{kp}, and b_{jkp} have the same interpretations as in Equation (2.2). Let $\mathbf{b}_{jk} = (b_{jk1}, \ldots, b_{jkP})$, we assume the random effects parameter vector, $\mathbf{r}_j$, independently following a multivariate normal distribution with mean 0 and an unstructured covariance-variance matrix $\mathbf{R}_1$, namely $MVN(0, \mathbf{R}_1)$. Here, $\mathbf{R}_1$ is the covariance matrix of the random effects in Equation (2.3), and each diagonal element of $\mathbf{R}_1$ measures the variance of the between-study heterogeneity. The PQL method is implemented using the `glmmPQL` function available under R package `MASS` (version 7.3-58.3) (Venables and Ripley, 2002). We show the algorithm for estimating the parameters in Equation (2.3) in Appendix B.

In both two-stage and one-stage models, including too many covariates can result in numerical difficulties when estimating random effects for all coefficients. In such cases, there are several ways to simplify the models. First, if we expect no interaction effect between covariate X_p and the treatments, we can replace β_{jkp} with β_{jp} in Equations (2.1) and (2.3). Second, random effects can be removed for some parameters if little or no between-study variability exists. Third, covariance matrices for random effects, $\mathbf{R}_2$ and $\mathbf{R}_1$, can be assumed to be structured, such as diagonal or blocked diagonal, rather than unstructured.

2.2.3 Existing RMST IPD-NMA Methods

We consider two existing two-stage IPD-NMA proposed by Weir et al. (2021) using a frequentist framework (denoted by nonparametric frequentist [NPF] two-stage model) and Tang and Trinquart (2022) using a Bayesian framework (denoted by nonparametric Bayesian [NPB] two-stage model). We modify these models by employing arm-based parameterization to compare their results with those under our methods. The first stage of the two models estimates the study-specific RMSTs for

a treatment using Equation (1.1). For the NPF model, the second stage combines the study-specific RMSTs by using a standard AD-NMA approach. The modified NPF model under the arm-based parameterization is fitted using the multivariate meta-analysis model (Chen et al., 2012; Schwarzer et al., 2015) yielding a restricted maximum likelihood estimator. For the modified NPB model, we adopt a Bayesian arm-based model proposed by Hong et al. (2015, 2016) to combine the study-specific RMSTs. Note that the existing models are not designed to study effect moderators because the first stage of these models uses a simple integration of the marginal survival function.

2.3 Simulation Study

2.3.1 Aims and Performance Measures

In this simulation study, we evaluate the finite sample performance of the proposed two-stage and one-stage RMST IPD-NMA models under various scenarios, and compare the proposed methods with the existing methods in terms of bias, mean squared error (MSE) and coverage probability of the RMST estimators. We use 1,000 Monte Carlo replications. The coverage probability is defined as the proportion of replications of which 95% confidence intervals (or credible intervals when fitting the NPB model) contain the true value out of the total number of replications.

2.3.2 Data-Generating Mechanisms

We simulate data for a network meta-analysis of randomized trials to assess three treatments: Treatment A ($k = 1$), Treatment B ($k = 2$) and Treatment C ($k = 3$). Our data-generating process is built based on the approaches proposed by Wang and Schaubel (2018) and Zhong and Schaubel (2022). Let nt denote the number of trials and n_j denote the number of patients in the j^{th} trial. For $j = 1, \ldots, nt$ and

$i = 1, \ldots, n_j$, the time-to-event outcome T_{ij} is generated using the following model:

$$\log(T_{ij}) = \sum_{k=1}^{3} \alpha_{jk}^* I[Z_{ij} = k] + X_{ij} \sum_{k=1}^{3} \beta_k^* I[Z_{ij} = k] + \epsilon_{ij} \sum_{k=1}^{3} \sigma_k, \tag{2.4}$$

where $X_{ij} \sim Bern(0.5)$, a binary covariate. Here, we sample α_{jk}^* from $N(\alpha_k^*, \tau^2)$, where τ measures heterogeneity across studies. We consider three values of $\tau = 0.1, 0.3, 0.5$, indicating low, moderate, and high between-study heterogeneity. Note that Equation (2.4) is an accelerated failure time model with random effects, where α_k^* is the fixed treatment effect for $X_{ij} = 0$, β_k^* is the fixed treatment-by-covariate interaction effect, σ_k is a scale parameter. We assume the residual $\epsilon_{ij} \sim N(0, 1)$. The prespecified values of α_k^*, β_k^*, and σ_k are provided in Table B.1 in Appendix B.

Next, we assume the censoring time C_{ij} is independent of T_{ij}, which is generated from an exponential distribution $Exp(0.15)$. We set a fixed truncation time for RMST at $t^* = 4$. Figure B.1 in Appendix B displays the simulated survival curves in the two subgroups, $X = 0$ and $X = 1$, for three treatments.

2.3.3 Simulation Scenarios

In this simulation, we vary one of the three factors while fixing the others. The three factors include (1) the total sample size of NMA data, (2) the number of trials included in an NMA, and (3) the network of treatments. The following scenarios determine the simulated NMA data structures:

Scenario 1: Vary $n_j = 200, 500, \text{or } 1000$ for all j while fixing $nt = 20$. All trials have three treatment arms.

Scenario 2: Vary $nt = 10, 20, \text{or } 30$ while fixing $n_j = 500$ for all j. All trials have three treatment arms.

Scenario 3: Vary treatment networks (shown in Table 2.1) while fixing $nt = 20$. Let the sample size in each trial be sampled from a discrete uniform distribution, $n_j \sim Unif(300, 700)$.

Table 2.1: Treatment networks in Scenario 3 from the simulation study. The upper table summarizes the number of trials with each treatment combination. The lower table summarizes the total number of trials that contain certain treatments.

	Number of Trials		
Treatments	Network 1	Network 2	Network 3
A,B,C	14	8	0
A,B	2	4	7
A,C	2	4	7
B,C	2	4	6
	Total Number of Trials Studying Individual Treatment		
Treatment	Network 1	Network 2	Network 3
A	18	16	14
B	18	16	13
C	18	16	13

Note that we assume an equal allocation of treatments in each trial. In Scenarios 1 and 2, we evaluate the bias, MSE, and coverage probability of RMSTs when varying nt and n_j, respectively, while fixing the other parameters. Scenario 3 considers realistic treatment networks where no trials compare all three treatments simultaneously (i.e., no three-arm trials). In Scenario 3, the three networks result in different numbers of trials included in NMA. For example, 18 trials study Treatment A in Network 1, while 14 trials study Treatment A in Network 3 (see the bottom of Table 2.1). Network 1 contains more data (i.e., information) about Treatment A than Network 3. Scenario 3 is designed as the number of trials studying each treatment decreases from Network 1 to Network 3.

2.3.4 Methods and Estimands

In three scenarios, we fit the proposed two-stage and one-stage RMST IPD-NMA models with the *log* link function. Additionally, we compare the proposed methods with the two existing two-stage methods, NPF and NPB.

The estimands are the log-RMSTs for each treatment and subgroup. The true log-RMSTs are calculated as 0.687, 1.070, and 0.877 for Treatment A, B, and C in subgroup $X = 0$, and 0.859, 1.186, and 1.056 for Treatment A, B, and C in subgroup $X = 1$. We discuss the true estimand calculation in Appendix B.

2.3.5 Results

Figure 2.1 presents the bias of log-RMST of three treatments for the subgroup with $x = 0$ under each scenario. Compared with the existing two-stage methods, our proposed models provide no or smaller bias across all scenarios. The most significant difference in bias between the existing and proposed methods is observed when $n_j = 200$ in Scenario 1. Across all scenarios, bias tends to increase as the level of between-study heterogeneity increases. That is, NMA data containing high heterogeneity tend to yield biased estimates. When considering the same level of heterogeneity, the bias from the existing approaches decreases as the sample size in each trial increases. However, the sample size, number of trials, and treatment network do not affect bias in our proposed methods under the same level of heterogeneity.

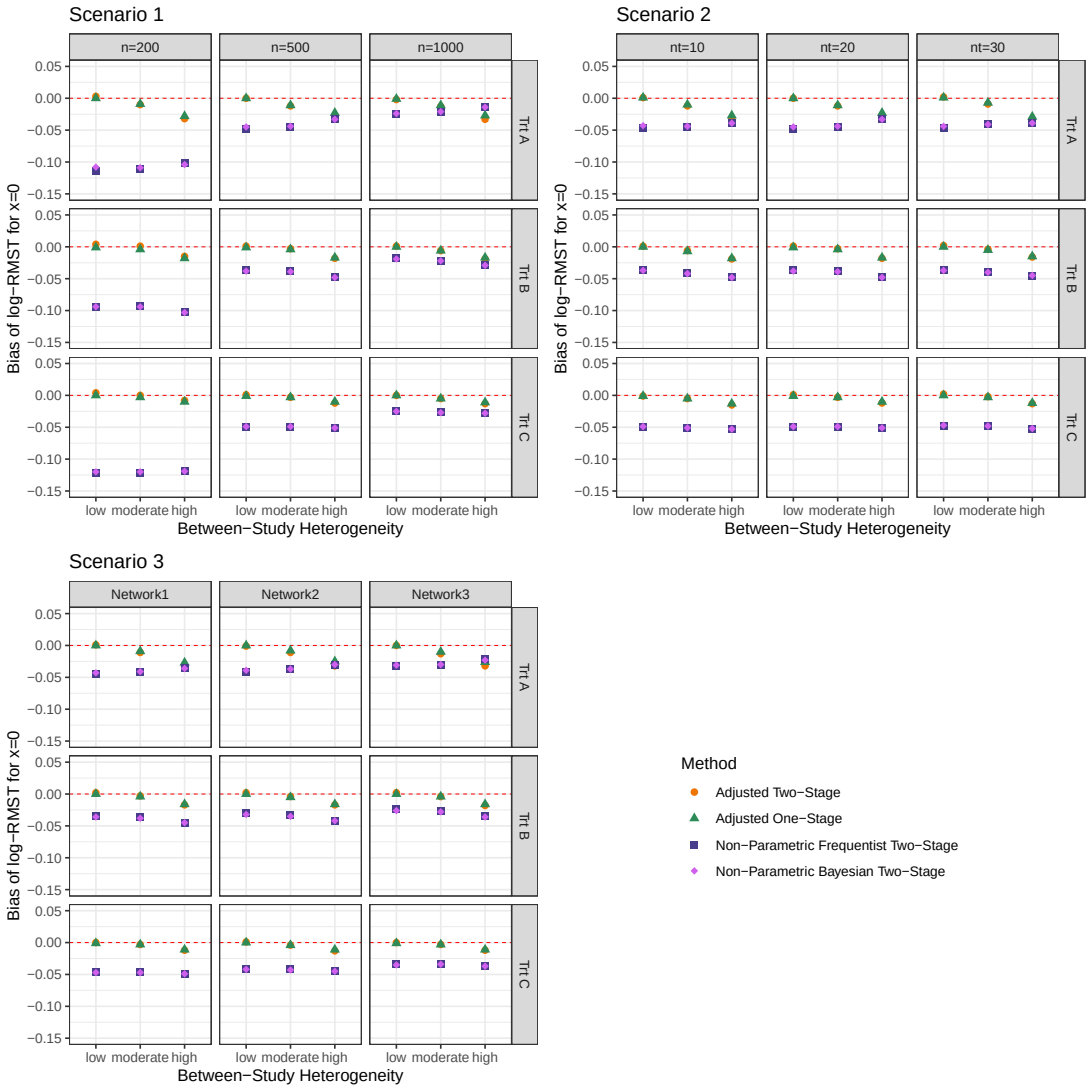

FIGURE 2.1: Simulation results of bias of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 0$ under three scenarios.

Figure 2.2 displays the MSE of log-RMST for the subgroup with $x = 0$. Overall, our proposed methods show MSEs no larger than those under the existing methods. In Scenario 1, the existing models provide much larger MSE than the proposed models when $n_j = 200$. In Scenarios 2 and 3, the MSE difference between the proposed and existing models is slight. Furthermore, MSEs tend to increase as the level of between-study heterogeneity increases. In addition, fixing the level of between-study

heterogeneity, all models yield decreasing MSEs when n_j increases in Scenario 1, nt increases in Scenario 2, and the network setting moves from Network 3 to Network 1 in Scenario 3. These results are expected because they show that MSEs get smaller when more data (e.g., large sample size, many trials) are used in NMA.

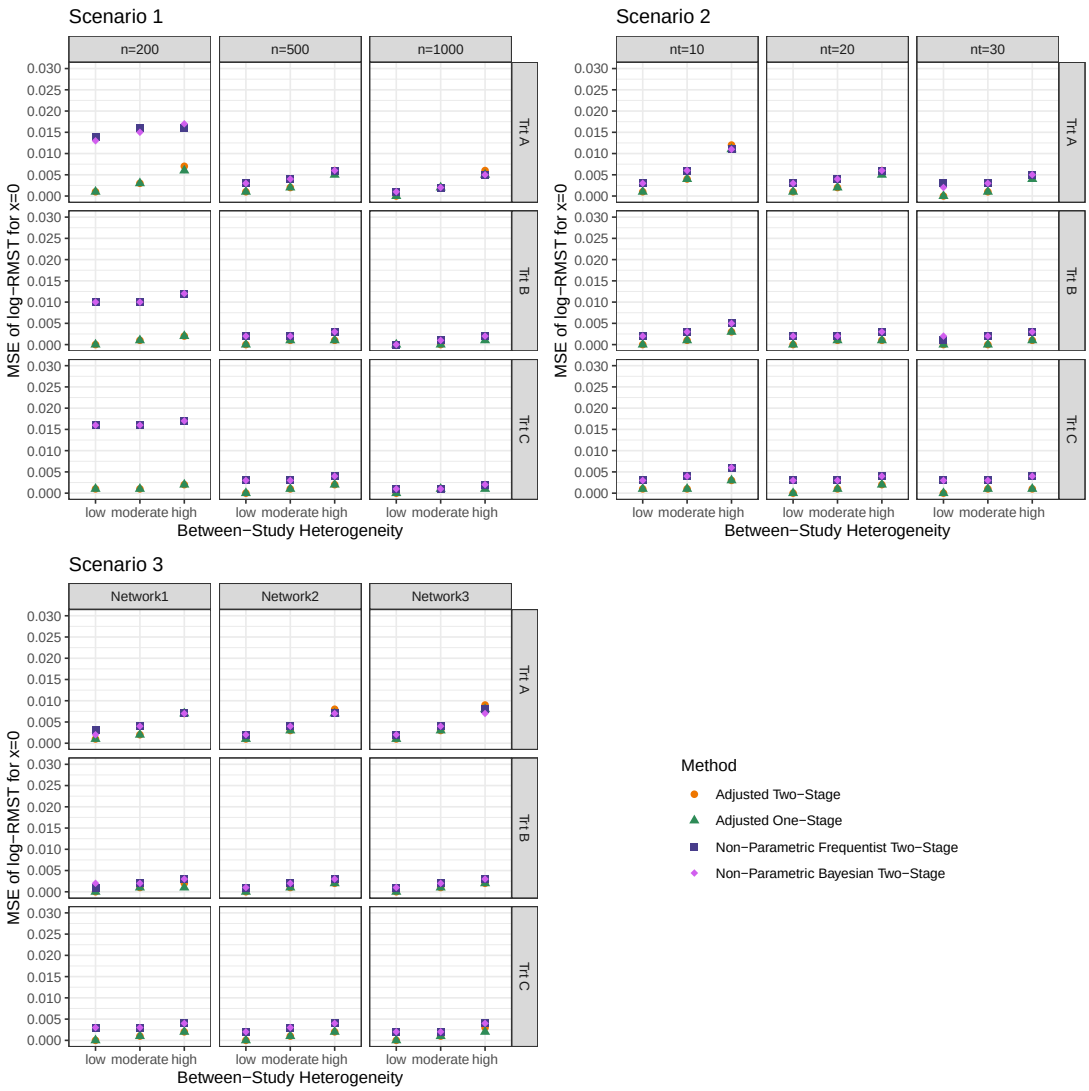

FIGURE 2.2: Simulation results of mean squared error (MSE) of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 0$ under three scenarios.

Figure 2.3 presents the coverage probability of log-RMST for the subgroup with $x = 0$. The existing methods provide coverage probabilities a lot lower than 95%, especially when between-study heterogeneity is low, and n_j or nt is small. In contrast, our proposed two-stage model yields coverage probabilities close to a nominal probability of 95% in most scenarios. However, our one-stage model shows a slightly lower coverage than the two-stage model with a coverage probability below 90% in Scenario 2 when $nt = 10$ and in Scenario 3 under Network 3. This result shows that the standard errors of log-RMST might be underestimated in our one-stage model due to a biased estimate of between-study variability (simulation results not shown). Fixing the level of between-study heterogeneity, the changes in n_j, nt, and treatment networks have minimal effect on the coverage probabilities in our proposed methods.

In summary, the proposed RMST IPD-NMA models outperform existing models with smaller biases, lower MSE, and coverage probabilities close to 95% across all scenarios. Notably, the proposed models perform exceptionally well when sample sizes are small. The simulation results for subgroup $x = 1$ provide similar results to those for $x = 0$, shown in Figures B.2 to B.4 in Appendix B.

2.4 Data Analysis

We illustrate the proposed two-stage and one-stage models using a real NMA example about treatments for patients with atrial fibrillation. The existing methods, NPF and NPB, are not applied because the data include continuous covariates that cannot be adequately handled properly in the existing methods.

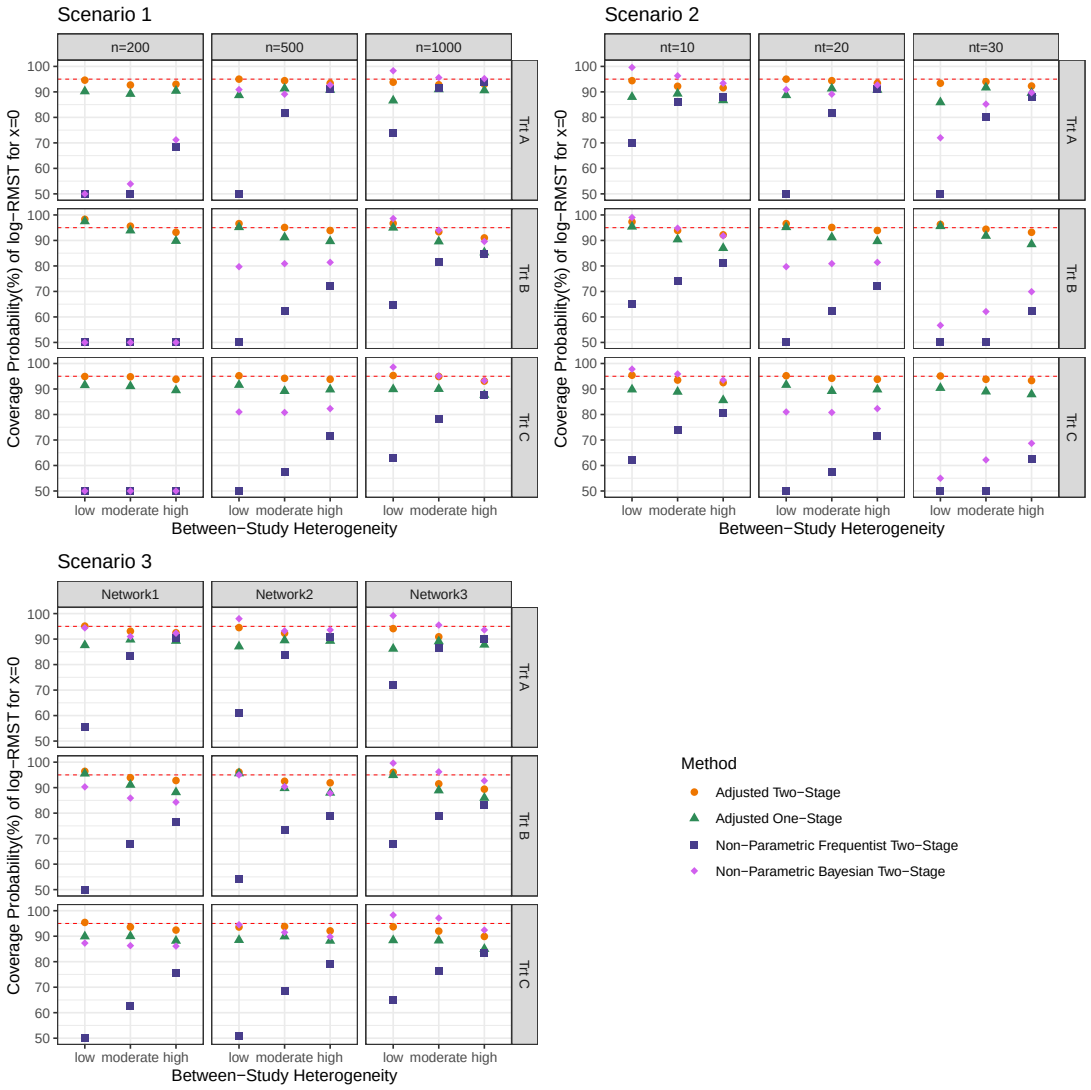

FIGURE 2.3: Simulation results of coverage probabilities of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 0$ under three scenarios. The coverage probabilities below 50% are truncated to 50% in the plots.

2.4.1 COMBINE-AF Data

The COMBINE-AF (A COllaboration between Multiple institutions to Better Investigate Non-vitamin K antagonist oral anticoagulant usE in Atrial Fibrillation) data contain IPD from four phase III randomized trials comparing direct oral anticoagulants (DOACs) with warfarin to prevent stroke for patients with atrial fibrillation (Carnicelli et al., 2021, 2022). Among the four trials, two are three-arm trials

comparing two dose levels of DOACs, such as standard-dose DOAC (SD-DOAC) and low-dose DOAC (LD-DOAC), with warfarin, and the other two compared SD-DOAC with warfarin. The network of treatments and detailed information on individual trials can be found elsewhere (Carnicelli et al., 2021, 2022).

In this analysis, we consider two outcomes: time to bleeding and time to all-cause mortality (death) as safety and efficacy profiles, respectively. The time-to-event outcomes are in the unit of month, and we use 32 months, the point at which less than 10% of patients remained at risk across all studies (Carnicelli et al., 2022), as the truncation time for RMST. We fit our proposed two-stage and one-stage models adjusted by two baseline covariates: sex as a binary variable and age as a continuous variable. The treatment-by-covariate interactions are included. We use males as the reference group and center the age to 70.55, the average age of the four trials.

2.4.2 Results

Figures 2.4 and 2.5 display the estimated RMSTs over age (from 50 to 90) by treatment and sex groups for the event of bleeding and all-cause mortality, respectively. The coefficient estimates of the fixed effect and the estimates of the standard deviation of random effects are reported in Tables B.3 and B.4 in Appendix B. The proposed two-stage and one-stage models yield similar point estimates.

In Figure 2.4, the estimated RMSTs for the bleeding outcome demonstrate a consistent decrease in older patients, and this pattern is observed across both female and male patients and all three treatments. According to both models, all treatments have significant interaction effects with age (p-value < 0.01), while the treatment effects do not significantly differ by sex (see Table B.3 for details). LD-DOAC is preferable for avoiding bleeding among three treatments across the age range of 50 to 90, regardless of the sex. However, we observe that the relative efficacy of SD-

DOAC compared to warfarin is smaller in senior patients.

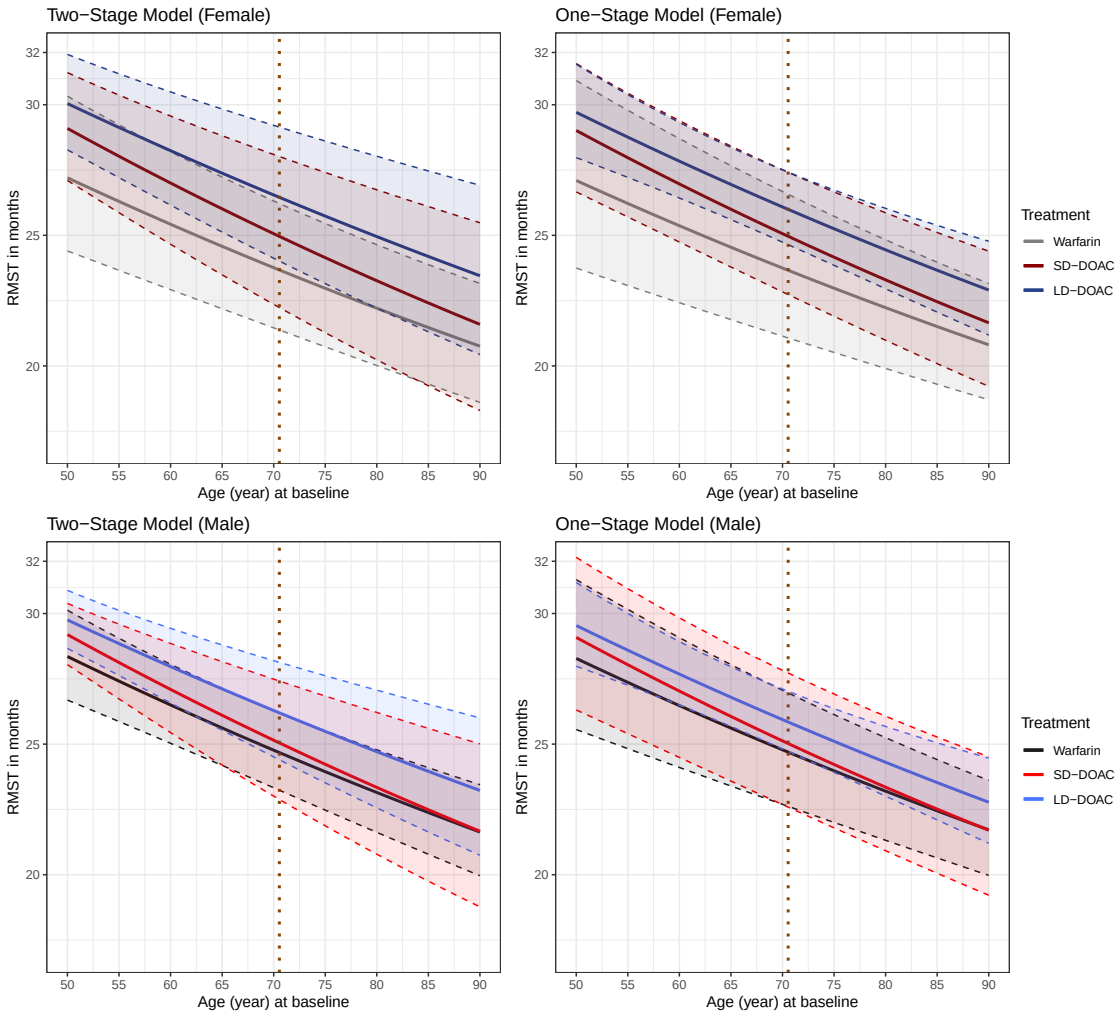

FIGURE 2.4: Estimated RMSTs (in months) for the bleeding outcome over age (from 50 years to 90 years) by treatment and sex groups. The results for female and male patients are shown in the first and second rows, respectively. The results from the two-stage and one-stage models are displayed in the first and second columns, respectively.

Under the two-stage model, given patients aged 70.55, the estimated average bleeding-free survival times at 32 months are 26.21 months for male patients and 26.47 months for female patients randomized to LD-DOAC. The one-stage model yields results of 25.84 months for males and 25.99 months for females. Meanwhile, the two-stage

model suggests a more significant benefit of LD-DOAC vs. warfarin among patients at 70.55 years old. Compared to the patients randomized to warfarin, those randomized to LD-DOAC show a significant enhancement in average bleeding-free survival times at 32 months, with a relative increase of 6% (95% CI: [4%, 9%]) for males and 12% (95% CI: [9%, 15%]) for females.

Figure 2.5 shows a similar declining pattern in the estimated RMSTs for all-cause mortality over age, commonly observed across sex and treatment groups. However, the estimated RMSTs are closely aligned among the three treatments. Both two-stage and one-stage models reveal significant interaction effects of age with all three treatments (p-value < 0.01). The two-stage model indicates that SD-DOAC and LD-DOAC have significantly different effects by sex. In contrast, the one-stage model shows that warfarin and SD-DOAC have significantly different effects by sex (see Table B.3 for details). Under both models, LD-DOAC is preferable in mitigating the risk of death for patients whose ages are between 40 and 90, regardless of sex. Additionally, SD-DOAC shows a more favorable impact on reducing all-cause mortality compared to warfarin in older patients.

Under the two-stage model, given patients aged 70.55, the estimated average survival times at 32 months, are 30.45 months for male patients and 30.85 months for female patients randomized to LD-DOAC. The one-stage model yields results of 30.48 months for males and 30.75 months for females. The relative RMST differences among treatments are all less than 1% according to both models.

We assess the between-study variability through the variance of random effects for the main treatment effects and treatment-by-covariate interactions (see Table B.3 for details). For the bleeding outcome, the variability seems moderate for the main treatment effects and treatment-by-sex interactions, while it is notably low for treatment-

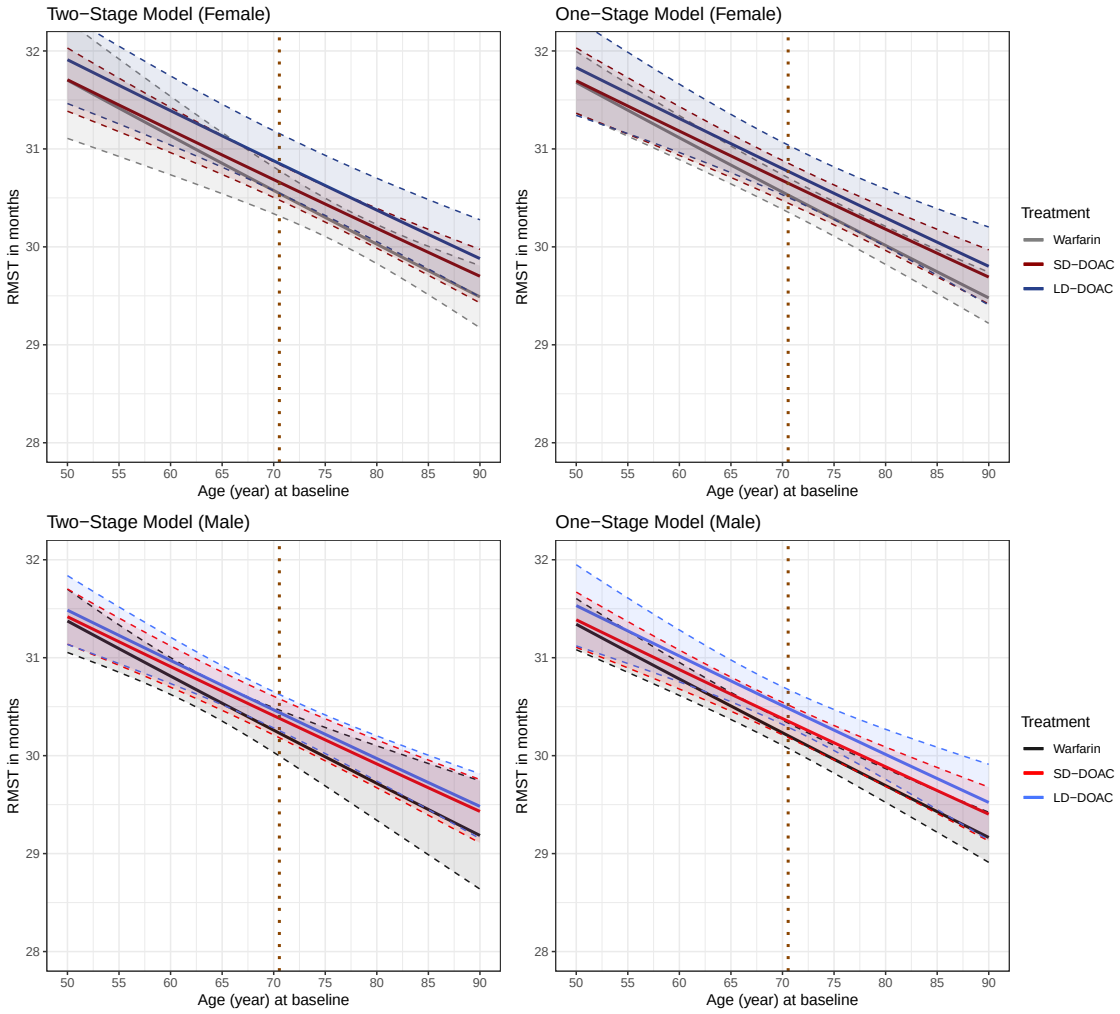

FIGURE 2.5: Estimated RMSTs (months) for the all-cause mortality outcome over age (from 50 years to 90 years) by treatment and sex groups. The results for female and male patients are shown in the first and second rows, respectively. The results from the two-stage and one-stage models are displayed in the first and second columns, respectively.

by-age interactions. For the death outcome, the between-study variability for all parameters is minimal, indicating the consistency of the results across trials.

2.5 Discussion

We developed two-stage and one-stage IPD-NMA methods that estimate RMST for time-to-event outcomes. These models used individual participant-level covariates and estimated RMST and treatment-by-covariate interaction effects using inverse probability of censoring weighted RMST regression. Our simulation study showed that the proposed two-stage and one-stage models provided unbiased RMST estimates with smaller MSE and proper coverage probability than the existing nonparametric two-stage approaches. In addition, we demonstrated the proposed methods using a real data example.

In the simulation study, we observed finite biases in our proposed one-stage and two-stage methods, mainly when between-study heterogeneity is high. For the two-stage approach, our findings align with those in several previous research (Chen et al., 2012; Jackson et al., 2013), where considerable biases and low coverage probability of coefficients were observed in multivariate meta-analysis under high between-study variability. For the one-stage approach, these biases arose due to the biased PQL estimators of either the fixed effect or the standard deviations of random effects (Lin and Breslow, 1996; Breslow, 2004; Lin, 2007; Jiang and Nguyen, 2007; Nugent et al., 2019). The PQL algorithm involves the Laplace approximation for the integrated quasi-likelihood (Breslow and Clayton, 1993), and this approach may lead to less accurate estimations when the number of clusters (e.g., the number of studies) is small and the variance component (e.g., the between-study heterogeneity) is large (Bellamy et al., 2005; Jang and Lim, 2009). However, it is essential to note that the PQL approach offers computational advantages because it can be applied to the RMST regression using the quasi-likelihood approach. It tends to perform well, and these biases can be ignored when the variances of random effects are small, and the outcomes become closer to a normal distribution (Breslow, 2004; Bellamy et al., 2005; Jiang

and Nguyen, 2007). To further enhance the estimation in our one-stage method, the adaptation of alternative approximation methods, such as Gauss-Hermite quadrature can be considered (Liu and Pierce, 1994; Pinheiro and Bates, 1995). However, this is beyond the scope of our work and remains as future research.

We used the COMBINE-AF as an illustrative example and considered selected outcomes to demonstrate the proposed methods described in the manuscript. As such, it is essential to interpret the results cautiously and refrain from employing them directly for clinical decision-making. Clinical decisions should be made by carefully considering other important outcomes and relevant medical information. A comprehensive discussion on clinical decision-making based on the COMBINE-AF can be found in Carnicelli et al. (2022).

In our data analysis, both two-stage and one-stage models produced similar coefficient estimates, but they provided slightly different estimates of the standard deviation of the between-study heterogeneity. This difference could be attributed to the varying estimation methods used in the two approaches and the different likelihood functions employed in each method. In specific, the two-stage model assumes normality for all parameters derived from the first stage, while the one-stage model uses quasi-likelihood functions.

In this study, all NMA models were fitted under the "consistency" assumption that assumes direct and indirect comparisons provide the same effects. In NMA, "inconsistency" is often defined as a statistical disagreement between direct and indirect comparisons (Higgins et al., 2012), and it could lead to biased results if not handled properly (White et al., 2012; Freeman et al., 2019). However, the treatment network of our data example needs more information to assess consistency. For time-to-event outcomes, violations of the PH assumptions could result in inconsistency with biased

indirect treatment comparisons (Jansen, 2011). However, formal methods have yet to be proposed to analyze NMA with time-to-event in the existence of inconsistency. This topic requires further research.

Software

The relevant R code for the methodology and simulation study is available at `https://github.com/kimihua1995/NMA_IPD_RMST`.

Data Availability Statement

The data that support the findings of this study are available to members of the COMBINE-AF executive committee. Restrictions apply to the availability of these data, which were used under license for this study. Individual investigators may reach out directly to the COMBINE-AF executive committee for collaboration.

3

Topic 2: Inference of Treatment Effect and Its Regional Modifiers Using Restricted Mean Survival Time in Multi-Regional Clinical Trials

3.1 Introduction

3.1.1 Multi-Regional Clinical Trials

With a growing demand for region-specific evidence of drug effects and a need for pooling patient data from multiple regions, the pharmaceutical industry has globalized the research and development to provide high-quality medical products around the world (Quan et al., 2013). Multi-regional clinical trials (MRCTs), in which patients are enrolled from multiple geographically separated regions, have become a common practice in recent years since they could accelerate the availability of critical medical products to needed patients worldwide and also help increase the developmental efficiency in regional clinical research (Chen et al., 2010). In 2016, the International Council for Harmonisation E17 (Food and Drug Administration (FDA), 2016) was issued to promote MRCT data acceptance by multiple regulatory

agencies.

In MRCT, there is a keen interest in estimating region-specific average treatment effects and assessing the heterogeneity of average treatment effects across regions. When treatment effects from multiple regions are consistent, combining data from these regions becomes feasible. Such a combination enhances the power to detect the overall treatment effect and extends the generalizability of trial findings. In such cases, the regional regulatory agency could expedite drug approval by bridging data from its region with data from other regions. Nevertheless, in the presence of inconsistency in treatment effects and safety profiles, mainly if the differences are substantial or opposite, a regional regulatory agency may choose to disapprove the drug in its market.

3.1.2 Essential and Inessential Traits in MRCT

In MRCT, the treatment effect inconsistency between regions can stem from various factors that can be classified into two categories: *exogenous (inessential)* imbalance and *intrinsic (essential)* imbalance of regional traits. Essential traits are intrinsic to a region, including a region's ethnic and racial constituents, genetic variants, and medical practice due to administrative, cultural, and historical reasons. These factors are typically important to distinguish regions but may be difficult to quantify and often not captured by the MRCTs.

On the other hand, inessential traits are exogenous to a region's identity, including patient characteristics (e.g., gender and age) and clinical factors (e.g., baseline performance status, tumor size, and comorbidity). The imbalances in inessential traits between regions are inevitable due to the recruitment of patients from those who "conveniently" come to the clinics or centers that conduct the trial in a specific region. The recruitment in each region cannot be considered a simple random sample

of the reference population. As the region-specific treatment effects are estimated over the population represented by the sample of recruited patients from each region, differences in treatment effects between regions may be partially attributed to disparities in the inessential traits of these region-specific samples.

With the obligation to protect the patients of their regions from ineffective or toxic treatments, regional regulatory agencies are usually interested in the inconsistency caused by essential traits (e.g., race and genetic variants). However, assessing the treatment effect consistency across regions may be biased if imbalanced inessential traits modify the treatment effect (i.e., the effect modifiers). Therefore, there is a need to eliminate any apparent difference introduced by inessential traits, especially the effect modifiers, in a particular sample to ensure a comparable evaluation of the treatment effect. Throughout this study, we will refer to inessential traits, which are effect modifiers as variables that modify the region-specific average treatment effect (which will be defined in Chapter 3.2). We will interchangeably use the terms "inessential traits" and "covariates" in this study.

3.1.3 Motivation Example: PLATO Trial

This study is motivated by the PLATO (PLATlet inhibition and patient Outcomes) trial (Wallentin et al., 2009), which compared the relative efficacy and safety of standard antiplatelet therapy, clopidogrel, versus a novel antiplatelet inhibitor, ticagrelor, for patients with acute coronary syndromes. The trial randomized 18,624 patients in 43 countries among four predefined geographic regions: i) Asia and Australia, ii) Central and South America, iii) Europe, Middle East, and Africa, and iv) North America. The primary endpoint was the time to a composite of cardiovascular death, myocardial infarction, or stroke. The analysis led by Carroll and Fleming (2013) showed a significant interaction between treatment and region. Their subsequent subgroup

analysis indicated that ticagrelor was less effective than clopidogrel, though not statistically significant when limiting participants to the United States (US). However, the results with participants from the rest of the world (non-US) showed the opposite results. The unbalanced distribution of aspirin dose usage between US and non-US mainly explained this substantial heterogeneity of region-specific treatment effects. In addition, another study showed that ticagrelor was associated with lower risk in primary outcome compared with clopidogrel for patients taking low-dose aspirin (Mahaffey et al., 2011).

In this study, our objective is to evaluate the region-specific treatment effect in the PLATO trial after the distributions of maintenance aspirin dose usage and other potential effect modifiers are balanced across regions. However, the existing methods in MRCT that rely on fixed or random effects hierarchical models (Chen et al., 2010; Tsong et al., 2012; Quan et al., 2013, 2014) have ignored the incomparability of patients' baseline covariates from different regions. These methods may lead to biased estimates of region-specific treatment effects and inaccurate assessment of treatment effect consistency across regions. To address this limitation, we propose advanced approaches that can correctly estimate the region-specific treatments and evaluate their consistency. These proposed methods can eliminate disparities arising from inessential traits, providing a more robust and unbiased assessment of region-specific treatment effects in MRCTs.

3.1.4 Generalization and Transportation Approaches in MRCT

To eliminate the disparities arising from inessential traits, we propose to "transport" or "generalize" the treatment effect from each region to a uniform target distribution of these covariates. By doing so, we can ascribe any variations in the transported or generalized region-specific treatment effects to the essential traits. The

differences between transportation and generalization problems have been discussed elsewhere (Lee et al., 2021; Josey et al., 2021b). *Generalization* refers to extending the findings obtained from a nested population to a larger target population. This concept is applicable when one wants to generalize treatment effects observed in each region to a broader, common superpopulation. On the other hand, *Transportation* is a term used when the regional and target populations are drawn from the same superpopulation. In this situation, it is used to apply results from a randomized control trial (RCT) conducted in one region to another where both regions are from the same superpopulation. In summary, whether to use the terminology of *Generalization* or *Transportation* depends on the selected target population related to our research questions.

Many transportation and generalization methods rely on inverse probability of sampling weighting (IPSW) approaches (Stuart et al., 2011; Kern et al., 2016; Westreich et al., 2017; Dahabreh et al., 2019, 2020). These IPSW-based approaches rely on correct specification of the sampling score model and sufficient overlap between study participants and target population (Degtiar and Rose, 2023). However, they become unstable due to model misspecification, extreme weights, or sparse covariates (Robins et al., 2007; Ben-Michael et al., 2021). Furthermore, IPSW-based approaches are ineffective in cases where individual patient data from the target population is not accessible (Chattopadhyay et al., 2022).

To address these issues from IPSW, recent research embraced the calibration weighting (CW) method to transport or generalize treatment effects from RCTs to the target population (Josey et al., 2021b, 2022; Lee et al., 2021). The CW approach is similar to the idea of entropy balancing proposed by Hainmueller (2012). It has been initially studied in estimating causal effects using observation data (Hainmueller, 2012; Zhao and Percival, 2016; Wang et al., 2019a; Lee et al., 2021). The calibration

weights are estimated by solving a (convex) entropy maximization problem under constraints of an exact balance of covariate moments (Zhao and Percival, 2016) so that the covariate distributions of the samples across groups will empirically match the common target population. Estimating the calibration weights does not require fitting a sampling score and an outcome model. In addition, this method is flexible as it can be applied when only the sample moments of the covariates in the target population are available (Josey et al., 2021b).

3.1.5 Study Objectives

In this study, we propose both IPSW and CW-based methods to balance the covariate distributions of the samples across different regions in an MRCT against a common target population that carries clinical relevance and interpretable validity. Although some limitations exist in the IPSW-based approaches, they have not been extensively discussed in the context of MRCTs. We contribute by proposing the IPSW-based estimators for assessing average region-specific treatment effects in MRCTs, thus expanding the methodological repertoire in this domain.

Our focus extends to proposing the CW-based treatment effect estimators from multiple regions, which enable us to consistently estimate the average region-specific treatment effects concerning the target population. Furthermore, we assess the consistency of average treatment effects across multiple regions and estimate a global average treatment effect if there is no evidence of heterogeneity. This approach can eventually enhance the interpretability of MRCTs.

While previous studies have primarily concentrated on binary outcomes, limited attention has been given to time-to-event outcomes. To fill this methodological and practical gap, we propose transportation and generalization methods for analyzing MRCTs with time-to-event endpoints. We incorporate RMSTs into our proposed

methods, leveraging the advantages of RMST in analyzing time-to-event data in clinical studies.

3.1.6 Chapter Outlines

The remainder of this chapter is structured as follows. Chapter 3.2 establishes the basic settings and assumptions for the problem. Chapter 3.3 introduces the two weighting methods in MRCT: the inverse probability of sampling weighting and calibration weighting. In Chapter 3.4, we propose four weighted estimators for region-specific average RMST difference and their large sample properties and introduce a region-specific consistency test of treatment effects in MRCTs. In Chapter 3.5, we show an extensive simulation study and assess the finite sample properties of the proposed estimators. In Chapter 3.6, we illustrate our methods on the PLATO trial. Chapter 3.7 concludes with a discussion of various issues.

3.2 Basic Setup

3.2.1 Notations and MRCT Data Structure

Figure 3.1 displays the MRCT data structure. Suppose there are M geographically separate regions ($M \geqslant 2$) with R being the region indicator ($R = 1, \ldots, M$). Let N_r denote the size of the patient population with a certain disease in Region r ($R = r$) under review of RCT enrollment, which is not necessarily known. Hypothetically, we assume the patient population from different regions share a common distribution of inessential traits, such as gender and age, while the distribution of some essential traits, such as race, can distinguish them. Suppose $\mathbf{X}$ is a p-dimensional vector of baseline covariates of interest, and F denotes the common distribution of $\mathbf{X}$ shared by the patient population across all regions (i.e., $X \sim F(\mathbf{X})$).

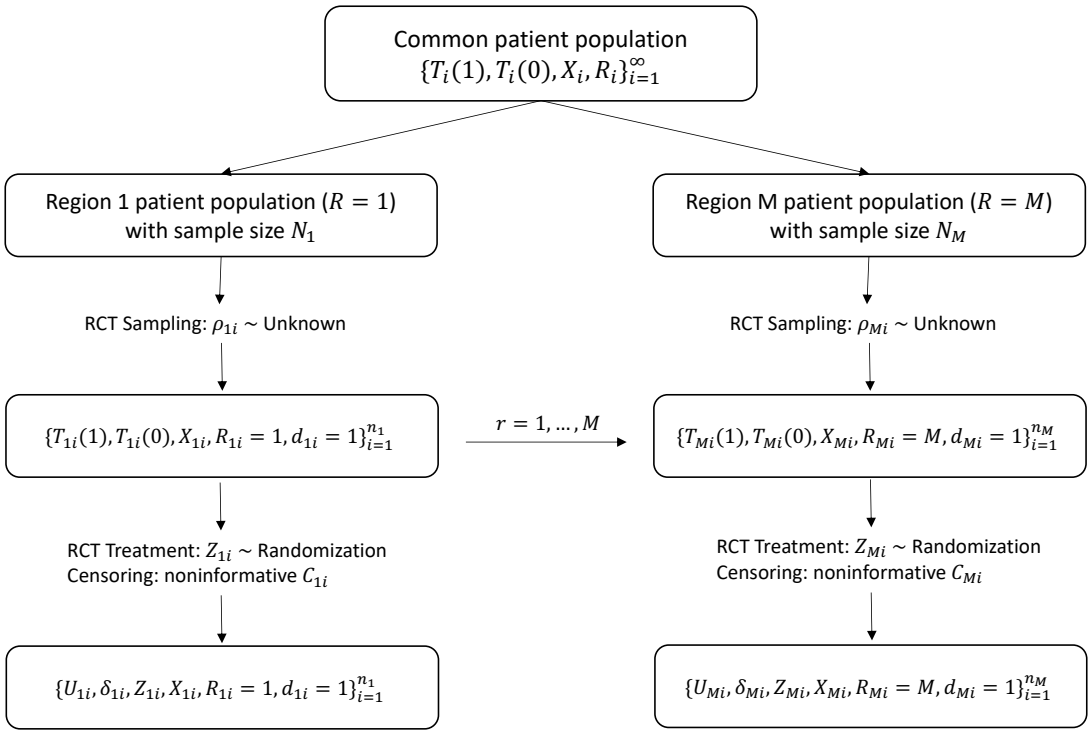

FIGURE 3.1: Data structure of multi-regional clinical trials.

We first consider a single RCT in Region r, comparing the efficacy of two treatments. Out of N_r individuals in Region r, suppose n_r patients are enrolled in the RCT. We indicate the selection using d, where $d = 1$ if an individual is enrolled in the RCT and $d = 0$ otherwise. In general, we can only observe the data of the enrolled patients (i.e., $d = 1$). Suppose $\rho_r(\mathbf{X}) = P(d = 1|\mathbf{X}, R = r)$ is an unknown sampling score for RCT participation in Region r given the baseline covariates $\mathbf{X}$. We denote the distribution of $\mathbf{X}$ for RCT participants as $F_r(\mathbf{X}) = F(\mathbf{X}|d = 1, R = r)$. Let Z be the indicator of the treatment assignment, where $Z = 1$ for the treatment group and $Z = 0$ for the control group. Let $\pi_r(\mathbf{X}) = P(Z = 1|\mathbf{X}, R = r, d = 1)$ be the treatment propensity score. In RCTs, $\pi_r(\mathbf{X}) = \pi_r$ can be commonly assumed to be known by design.

Following the potential outcomes framework in Rubin (1974), let $T(1)$ and $T(0)$

denote the potential time-to-event under treatment and control, respectively. By assuming the stability of potential outcomes, the time-to-event is $T = T(1)Z + T(0)(1-Z)$. Let C denote the censoring time. In the presence of right censoring, T is not observed for all patients. Instead, we observe $U = \min(T, C)$ and the censoring indicator $\delta = I[T \leqslant C]$, where $I[\cdot]$ is the indicator function. In summary, let i index the enrolled patient and r index the region, the observed data in Region r is $\{U_{ri}, \delta_{ri}, \mathbf{X}_{ri}, Z_{ri}, R_{ri} = r, d_{ri} = 1 | i = 1, \ldots, n_r\}$, where $r = 1, \ldots, M$.

3.2.2 Estimands and Assumptions

In Region r, we define the treatment-specific conditional survival function as $S_{rz}(t|\mathbf{X}) = E\{I[T(z) \geqslant t]|\mathbf{X}, R = r, d = 1\}$ for $z \in \{0, 1\}$. For a pre-specified truncation time t^*, the treatment-specific conditional RMST up to t^* is defined as $\mu_{rz}(t^*|\mathbf{X}) = \int_0^{t^*} S_{rz}(t|\mathbf{X})dt$. The conditional treatment effect is defined by the RMST difference such as $\Delta_r(t^*|\mathbf{X}) = \mu_{r1}(t^*|\mathbf{X}) - \mu_{r0}(t^*|\mathbf{X})$.

We seek to characterize the region-specific conditional RMST difference $\Delta_r(t^*|\mathbf{X})$ to a target population. Let F^* denote the distribution of $\mathbf{X}$ in the target population (i.e., $\mathbf{X} \sim F^*(\mathbf{X})$). Our estimand is a region-specific average RMST difference if the distribution of covariates $\mathbf{X}$ from RCT participants in Region r is the same as in the target population, which is written as:

$$\Delta_r(t^*) = \mu_{r1}(t^*) - \mu_{r0}(t^*), \tag{3.1}$$

where $\mu_{rz}(t^*) = E_{F*}[\mu_{rz}(t^*|\mathbf{X})]$ is the region-specific average RMST for $z \in \{0, 1\}$. The expectation is taken concerning the distribution of $\mathbf{X}$ in the target population. To identify the estimand from the observed data in MRCT, we make the following assumptions:

Assumption 1 (Ignorability and positivity of trial treatment assignment)
(i) $\{T(1), T(0)\} \perp Z|(d = 1, \mathbf{X}, R)$*, and (ii)* $0 < \pi_r(\mathbf{X}) < 1$ *with probability 1.*

Assumption 2 (Conditional noninformative censoring)
$\{T(1), T(0)\} \perp C|(d = 1, Z, \mathbf{X}, R)$*, which implies* $T \perp C|(d = 1, Z, \mathbf{X}, R)$*.*

Assumption 3 (Covariate overlap with target population)
The target distribution of $\mathbf{X}$*,* F^**, is absolutely continuous concerning the distribution of* $\mathbf{X}$ *for RCT participants in each region,* F_r*. That is, for any set of* $\mathbf{X}$*,* $A_{\mathbf{X}}$*, if* $A_{\mathbf{X}}$ *has zero probability in* F_r*, then it also has zero probability in* F^**.*

Assumption 4 (Conditional exchangeability of survival function and positivity of RCT participation)
(i) $S(t|\mathbf{X}, Z, R, d = 1) = S(t|\mathbf{X}, Z, R)$*, and (ii)* $\rho_r(\mathbf{X}) > 0$ *with probability 1.*

Assumption 4 also implies the conditional exchangeability of RMST, i.e., $\mu_{rz}(t^*|\mathbf{X}) = \int_0^{t^*} S(t|\mathbf{X}, Z = z, R = r, d = 1)dt = \int_0^{t^*} S(t|\mathbf{X}, Z = z, R = r)dt$. Under Assumptions 1 to 4, $\mu_{rz}(t^*)$ can be identified by $E[\mu_{rz}(t^*|\mathbf{X})\frac{dF^*}{dF_r}(\mathbf{X})]$, where $\frac{dF^*}{dF_r}(\mathbf{X})$ denotes the Radon-Nikodym derivative of the distribution F^* relative to the distribution of F_r. For example, if all covariates in $\mathbf{X}$ are discrete, $\frac{dF^*}{dF_r}(\mathbf{X})$ is the ratio of the probability mass functions of $\mathbf{X} = \mathbf{x}$ in both distributions, and if all covariates in $\mathbf{X}$ are continuous, $\frac{dF^*}{dF_r}(\mathbf{X})$ is the ratio of the probability density functions.

3.3 Weighting Methods in MRCT

3.3.1 The Inverse Probability of Sampling Weights

Following the concept of the balancing weight discussed in Li et al. (2018), to characterize the treatment effect from each region to the target population, we can weigh the distribution of F_r to the target population F^* by using the following weight

functions:

$$\gamma_{ri} = \frac{dF^*(\mathbf{X}_{ri})}{dF_r(\mathbf{X}_{ri})} = \frac{P_{F^*}(\mathbf{X} = \mathbf{X}_{ri})}{P_{F_r}(\mathbf{X} = \mathbf{X}_{ri})}, \tag{3.2}$$

where P_{F^*} and P_{F_r} denote the joint probability of $\mathbf{X} = \mathbf{X}_{ri}$ under the distribution of F^* and F_r, respectively. We can use the Kernel smoothing method (Scott, 2015) to estimate the density function for F_r and F^* if the distributions are unknown.

Alternatively, as the density functions may be difficult to estimate when $\mathbf{X}$ is high-dimensional (Westreich et al., 2017), we can use Bayes' rules to rearrange the γ_{ri} from Equation (3.2) into an inverse probability of sampling weight (IPSW) as follows:

$$\gamma_{ri} = \frac{P(d'_{ri} = 1|\mathbf{X}_{ri})}{P(d_{ri} = 1|\mathbf{X}_{ri}, R_{ri} = r)} \frac{P(d_{ri} = 1|R_{ri} = r)}{P(d'_{ri} = 1)} \propto \frac{P(d'_{ri} = 1|\mathbf{X}_{ri})}{P(d_{ri} = 1|\mathbf{X}_{ri}, R_{ri} = r)}, \tag{3.3}$$

where $d'_{ri} \in \{0, 1\}$ denotes the inclusion in the target population. As a result, the weight γ_{ri} from Equation (3.3) is proportional to the inverse ratio of a patient's sampling score for RCT participation in Region r as opposed to the probability of being in the target population, conditional on the covariates $\mathbf{X}_{ri}$. However, this approach requires prior knowledge of the sampling score of the target population, $P(d'_{ri} = 1|\mathbf{X}_{ri})$.

According to the MRCT data structure shown in Figure 3.1, we can only access the data from the RCT participation population in each region (i.e., $d = 1$), and the data from other unenrolled patients are unavailable. Therefore, we may not directly estimate the sampling score $\rho_r(\mathbf{X}) = P(d = 1|\mathbf{X}, R = r)$ in each region only with the data of $d = 1$. Alternatively, we estimate the sampling score by using a region-specific propensity score, $P(R = r|\mathbf{X}, d = 1)$, given the accessible data.

Many methods can be considered when estimating $P(R = r|\mathbf{X}, d = 1)$ such as 1) multinomial logistic regression model, 2) random forests (Lee et al., 2010), 3) gradient

boosted models (McCaffrey et al., 2013; Burgette et al., 2021), and 4) vector-based kernel weighting methods (Garrido et al., 2021). This study uses the gradient boosted model to estimate the region-specific propensity score. This method is implemented using the R package `twang` (version 1.4-9.5) (Cefalu et al., 2021). However, using a misspecified sampling score model may lead to biased results when implementing the IPSW. Consequently, to establish the large sample properties of the proposed estimators based on IPSW (to be introduced in Chapter 3.4), we assume the sampling score $\rho_r(\mathbf{X})$ is correctly specified.

3.3.2 The Calibration Weights

In contrast, the calibration weights (CW) can be estimated without fitting the sampling score. We utilize the Entropy Balancing method proposed by Hainmueller (2012) to estimate the calibration weights. Let p_{ri} denote the weights, where i and r index the patient and region such as $i = 1, \ldots, n_r$ and $r = 1, \ldots, M$. We estimate p_{ri} by solving the following optimization problem:

$$\min \left\{ \sum_{r=1}^{M} \sum_{i=1}^{n_r} p_{ri} log(p_{ri}) \right\}, \tag{3.4}$$

with constraints:

$$\sum_{i=1}^{n_r} p_{ri} \mathbf{g}(\mathbf{X}_{ri}) = \tilde{\mathbf{g}}, r = 1, \ldots, M. \tag{3.5}$$

$$\sum_{i=1}^{n_r} p_{ri} = 1, r = 1, \ldots, M. \tag{3.6}$$

The objective function in Equation (3.4) is the negative entropy of the CWs. Minimizing this function keeps the CW's empirical distribution close to the uniform, which in turn minimizes the variability due to heterogeneous weights (Lee et al.,

2021).

In Equation (3.5), we suppose $\mathbf{g}(\cdot)$ contains L covariate functions to be calibrated. Let $\mathbf{g}(\mathbf{X}) = [g_1(\mathbf{X}), \ldots, g_L(\mathbf{X})]$, each $g_l(\cdot)$ is a function of the covariates that can be any transformation of $\mathbf{X}$, but not necessarily the polynomial function (Zhao and Percival, 2016). The researcher imposes these functions to equalize the moments of the covariates between each region and the target population. A typical balance constraint contains the first k^{th} moments of the calibrated covariates (Hainmueller, 2012; Josey et al., 2021a). For example, when $k = 2$, $\mathbf{g}(\mathbf{X}) = [\mathbf{X}, \mathbf{X}^2]$, which forces the first and second moment of $\mathbf{X}$ to equal each region and the target population. In general, lower order k (e.g., $k \leqslant 2$) is often sufficient to empirically match the covariate distributions in RCT from every region to the target population (Signorovitch et al., 2010; Wang et al., 2019a). We let $\tilde{\mathbf{g}} = [\tilde{g}_1, \ldots, \tilde{g}_L]$ be the associated moment estimates of $\mathbf{g}(\mathbf{X})$ from any researcher-defined target population. Each $\tilde{g}_l$ can be calculated using individual data from the target population, when available or by setting it as a specified covariate sample moment of the target population.

Equation (3.6) implies that the calibration weights p_{ri} in each region sum to a normalization constant of one, which guarantees that p_{ri} are valid density functions in each region (Hainmueller, 2012; Zhao and Percival, 2016).

We calculate $\hat{p}_{ri}$ in (3.4) by using the Lagrange multiplier (De la Fuente, 2000).

$$\hat{p}_{ri} = \frac{\exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}},$$

where $\boldsymbol{\lambda}_r$ solves $\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}\{\mathbf{g}(\mathbf{X}_{ri}) - \tilde{\mathbf{g}}\} = 0$, for $r = 1, \ldots, M$. We provide the derivation in Appendix C. Additionally, to establish the large sample properties of the proposed estimators based on calibration weights (to be introduced in Chapter 3.4), we assume that for any uncalibrated covariates, denoted as $\mathbf{W}$, the sampling

score is conditionally exchangeable, i.e., $P(d = 1|\mathbf{X}, \mathbf{W}, R = r) = P(d = 1|\mathbf{X}, R = r) = \rho_r(\mathbf{X})$.

3.4 Weighted Estimators for Region-Specific RMST Difference

We propose four weighted estimators for the region-specific treatment effect in MRCT, i.e., the region-specific RMST difference $\Delta_r(t^*)$: 1) Weighted Kaplan-Meier (KM) estimator, 2) Weighted G-formula (GF) estimator, 3) Weighted Hajek (HJ) estimator, and 4) Weighted Augmented (AG) estimator. It is important to note that each proposed estimator has the flexibility to employ both weighting methods as introduced in Chapter 3.3. Throughout this chapter, we denote $\hat{\xi}_{ri}$ as the estimated weight function to balance the covariate distributions from each region to the target population, which can be either IPSW $\hat{\gamma}_{ri}$ or CW $\hat{p}_{ri}$. The proofs of all theoretical results are given in Appendix C. In the remainder of this paper, we denote "CW-adjusted estimators" and "IPSW-adjusted estimators" as the proposed weighted estimators using $\hat{p}_{ri}$ and $\hat{\gamma}_{ri}$, respectively.

3.4.1 Weighted Kaplan-Meier RMST Difference

The first estimator is derived based on a weighted RMST estimator proposed by Conner et al. (2019) in observational studies. They first estimated the propensity score, the probability of being in a particular treatment group. Then they integrated an inverse propensity score-adjusted KM estimator proposed by Xie and Liu (2005) to reduce confounding effects in an observational study. Based on their work, we propose a weighted KM estimator for the region-specific average RMST difference in MRCT.

For $i = 1, \ldots, n_r$, denote $N_{rzi}(t) = I[U_{ri} \leqslant t; \delta_{ri} = 1; Z_{ri} = z]$ as the individual treatment-specific counting process and $Y_{rzi}(t) = I[U_{ri} \geqslant t; Z_{ri} = z]$ as the individual

treatment-specific at risk process. Then for treatment $z \in \{0, 1\}$, the weighted KM RMST estimator for $\mu_{rz}(t^*)$ is written as:

$$\tilde{\mu}_{rz}(t^*) = \int_0^{t^*} \tilde{S}_{rz}(t)dt = \int_0^{t^*} \prod_{u \leqslant t} \left\{ 1 - \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} \right\} dt, \tag{3.7}$$

where $\tilde{S}_{rz}(t)$ is a weighted KM estimator for the region-specific average survival function. We define $\tilde{N}_{rz}(u) = \sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{rzi}(\mathbf{X}_{ri}) N_{rzi}(u)$ as the weighted counting process and $\tilde{Y}_{rz}(u) = \sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{rzi}(\mathbf{X}_{ri}) Y_{rzi}(u)$ as the weighted at-risk process. Here, $q_{rzi}(\mathbf{X}_{ri}) = \frac{I(Z_{ri}=z)}{\pi_{ri}(\mathbf{X}_{ri})^z (1-\pi_{ri}(\mathbf{X}_{ri}))^{1-z}}$ is an inverse of the individual treatment-specific propensity score, where $\pi_{ri}(\mathbf{X}_{ri})$ is the treatment propensity score that can be either estimated or assumed as constant.

The variance of $\tilde{\mu}_{rz}(t^*)$ is derived as follows:

$$\text{Var}[\tilde{\mu}_{rz}(t^*)] = \int_0^{t^*} \left\{ \int_u^{t^*} \tilde{S}_{rz}(t)dt \right\}^2 \frac{d\tilde{N}_{rz}(u)}{\tilde{W}_{rz}(u)(\tilde{Y}_{rz}(u) - \Delta\tilde{N}_{rz}(u))}, \tag{3.8}$$

where $\tilde{W}_{rz}(u) = \frac{\tilde{Y}_{rz}^2(u)}{\sum_{i=1}^{n_r} [\hat{\xi}_{ri} q_{rzi}(\mathbf{X}_{ri})]^2 Y_{rzi}(u)}$ and $\Delta\tilde{N}_{rz}(u) = \tilde{N}_{rz}(u) - \tilde{N}_{rz}(u-)$. If we assume that the treatment propensity score $\pi_{ri}(\mathbf{X}_{ri})$ is a constant in the RCT from each region, the term $q_{rzi}(\mathbf{X}_{ri})$ can be omitted when calculating $\tilde{\mu}_{rz}(t^*)$ and its variance.

The weighted KM estimator for $\Delta_r(t^*)$ is then defined as:

$$\hat{\Delta}_r^{KM}(t^*) = \tilde{\mu}_{r1}(t^*) - \tilde{\mu}_{r0}(t^*), \tag{3.9}$$

and we have the following theoretical properties for $\hat{\Delta}_r^{KM}(t^*)$:

Theorem 1

For any fixed $0 < t^* < \infty$*, as* $n_r \to \infty$*,*

$$\left|\hat{\Delta}_r^{KM}(t^*) - \Delta_r(t^*)\right| \xrightarrow{p} 0,$$

$$\sqrt{n_r}\left\{\hat{\Delta}_r^{KM}(t^*) - \Delta_r(t^*)\right\} \xrightarrow{d} N\left(0, \sigma^2_{r,KM}(t^*)\right).$$

Here, by $\xrightarrow{p}$ *and* $\xrightarrow{d}$ *we mean "converges in probability" and "converges in distribution", respectively. Let* $\Lambda_{rz}(t) = E_{\mathbf{X}}[-log\{S(t|\mathbf{X}, Z = z, R = r, d = 1)\}]$ *denote the cumulative incidence function, the asymptotic variance* $\sigma^2_{r,KM}(t^*) = n_r \sum_{z=0,1} \int_0^{t^*} \{\int_u^{t^*} S_{rz}(t)dt\}^2 \frac{d\Lambda_{rz}(u)}{\tilde{W}_{rz}(u)}$*, can be estimated by:*

$$\hat{\sigma}^2_{r,KM}(t^*) = n_r \sum_{z=0,1} \int_0^{t^*} \left\{\int_u^{t^*} \tilde{S}_{rz}(t)dt\right\}^2 \frac{d\tilde{N}_{rz}(u)}{\tilde{W}_{rz}(u)(\tilde{Y}_{rz}(u) - \Delta\tilde{N}_{rz}(u))}.$$

3.4.2 Weighted G-Formula RMST Difference

The second estimator employs the G-computation technique (Robins, 1986; Robins and Hernan, 2008; Naimi et al., 2017). This estimator is a direct regression estimator, and its outcome model can be identified by the IPCW RMST regression.

The IPCW RMST regression models with identity links are first fitted in each region. Here, we rewrite the Equation (1.3) by treatment-by-covariate interactions as follows:

$$\mu_r(t^*|\mathbf{X}_r, Z_r) = \beta_{r0} + \beta_{r1}Z_r + \boldsymbol{\beta}_{r2}\mathbf{g}(\mathbf{X}_r^T) + \boldsymbol{\beta}_{r3}Z_r\mathbf{g}(\mathbf{X}_r^T), \tag{3.10}$$

where $\mathbf{g}(\cdot)$ is the same function mentioned in Equation (3.5). Then, we define the outcome models for $Z_r \in \{0, 1\}$ in Region r as $m_{r0}(\mathbf{X}_r) = \hat{\beta}_{r0} + \hat{\boldsymbol{\beta}}_{r2}\mathbf{g}(\mathbf{X}_r^T)$ and $m_{r1}(\mathbf{X}_r) = \hat{\beta}_{r0} + \hat{\beta}_{r1} + \hat{\boldsymbol{\beta}}_{r2}\mathbf{g}(\mathbf{X}_r^T) + \hat{\boldsymbol{\beta}}_{r3}\mathbf{g}(\mathbf{X}_r^T)$. The weighted G-formula estimator for $\Delta_r(t^*)$ is written as:

$$\hat{\Delta}_r^{GF}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\{m_{r1}(\mathbf{X}_{ri}) - m_{r0}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}}. \tag{3.11}$$

By assuming that the RMST regression model in Equation (3.10) is not misspecified. We have the following theoretical properties for $\hat{\Delta}_r^{GF}(t^*)$:

Theorem 2

For any fixed $0 < t^* < \infty$*, as* $n_r \to \infty$*,*

$$\left|\hat{\Delta}_r^{GF}(t^*) - \Delta_r(t^*)\right| \xrightarrow{p} 0,$$

$$\sqrt{n_r}\left\{\hat{\Delta}_r^{GF}(t^*) - \Delta_r(t^*)\right\} \xrightarrow{d} N\left(0, \sigma^2_{r,GF}(t^*)\right).$$

Here, the asymptotic variance $\sigma^2_{r,GF}(t^*)$ *can be estimated using the Delta method (Dowd et al., 2014):*

$$\hat{\sigma}^2_{r,GF}(t^*) = n_r J_r^T \Sigma_r J_r,$$

where Σ_r *is the variance-covariance matrix of the parameter vector* $[\hat{\beta}_{r0}, \hat{\beta}_{r1}, \hat{\boldsymbol{\beta}}_{r2}, \hat{\boldsymbol{\beta}}_{r3}]^T$ *and*

$$J_r = \frac{\partial \hat{\Delta}_r^{GF}(t^*)}{\partial [\hat{\beta}_{r0}, \hat{\beta}_{r1}, \hat{\boldsymbol{\beta}}_{r2}, \hat{\boldsymbol{\beta}}_{r3}]^T}.$$

3.4.3 Weighted Hajek RMST Difference

The third estimator is proposed based on an alternative RMST estimation in Equation (1.2). For the i^{th} patient in Region r, let $Y_{ri} = \min(T_{ri}, t^*)$ denote the truncated time-to-event, and $w_{ri} = \delta^*_{ri}/\hat{G}(Y_{ri})$ denote the inverse probability of censoring weight. The weighted Hajek estimator for the average RMST difference in Region r is defined as:

$$\hat{\Delta}_r^{HJ}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r1i}(\mathbf{X}_{ri}) w_{ri} Y_{ri}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r1i}(\mathbf{X}_{ri}) w_{ri}} - \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r0i}(\mathbf{X}_{ri}) w_{ri} Y_{ri}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r0i}(\mathbf{X}_{ri}) w_{ri}}. \tag{3.12}$$

We have the following theoretical properties for $\hat{\Delta}_r^{HJ}(t^*)$:

Theorem 3

For any fixed $0 < t^* < \infty$*, as* $n_r \to \infty$*,*

$$\begin{aligned} \left|\hat{\Delta}_r^{HJ}(t^*) - \Delta_r(t^*)\right| &\xrightarrow{p} 0, \\ \sqrt{n_r}\left\{\hat{\Delta}_r^{HJ}(t^*) - \Delta_r(t^*)\right\} &\xrightarrow{d} N\left(0, \sigma^2_{r,HJ}(t^*)\right). \end{aligned}$$

This theorem can be proven by the M-estimator theory (Stefanski and Boos, 2002), and we provide the derivation of the asymptotic variance $\sigma^2_{r,HJ}(t^*)$ in Appendix C.

3.4.4 Weighted Augmented RMST Difference

The weighted Augmented RMST estimator combines the weighted G-formula and Hajek estimators:

$$\begin{aligned} \hat{\Delta}_r^{AG}(t^*) &= \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r1i}(\mathbf{X}_{ri}) w_{ri}\{Y_{ri} - m_{r1}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r1i}(\mathbf{X}_{ri}) w_{ri}} \\ &- \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r0i}(\mathbf{X}_{ri}) w_{ri}\{Y_{ri} - m_{r0}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri} q_{r0i}(\mathbf{X}_{ri}) w_{ri}} \\ &+ \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\{m_{r1}(\mathbf{X}_{ri}) - m_{r0}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}}. \end{aligned} \tag{3.13}$$

Based on the semiparametric theory (Tsiatis, 2006), the weighted Augmented estimator $\hat{\Delta}_r^{AG}(t^*)$ is doubly robust and does not require the correct specification of the outcome models (shown in Appendix C). The theoretical properties for $\hat{\Delta}_r^{AG}(t^*)$ is as follows:

Theorem 4

For any fixed $0 < t^* < \infty$*, as* $n_r \to \infty$*,*

$$\begin{aligned} \left|\hat{\Delta}_r^{AG}(t^*) - \Delta_r(t^*)\right| &\xrightarrow{p} 0, \\ \sqrt{n_r}\left\{\hat{\Delta}_r^{AG}(t^*) - \Delta_r(t^*)\right\} &\xrightarrow{d} N\left(0, \sigma^2_{r,AG}(t^*)\right). \end{aligned}$$

This theorem is also proven by the M-estimator theory, and we provide the derivation of the asymptotic variance $\sigma^2_{r,AG}(t^*)$ in Appendix C.

3.4.5 Regional Consistency Test

A Wald-type test can evaluate the consistency of treatment effects across regions. Given t^*, the null hypothesis of the treatment effect consistency is $H_0 : \Delta_1(t^*) = \cdots = \Delta_M(t^*)$. Regional consistency of treatment effects can be obtained if H_0 is not rejected. We define a χ^2 test statistics $U(t^*)$ as follows:

$$\tilde{U}(t^*) = \{\mathbf{E}\tilde{\mathbf{\Delta}}(t^*)\}^T\{\mathbf{E}\tilde{\mathbf{V}}(t^*)\mathbf{E}^T\}^{-1}\mathbf{E}\tilde{\mathbf{\Delta}}(t^*),$$

where $\mathbf{E}$ is an $(M-1)$-by-(M) contrast matrix $\begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & 0 & \dots & 1 \end{pmatrix}$,

$\tilde{\mathbf{\Delta}}(t^*) = [\tilde{\Delta}_1(t^*), \dots, \tilde{\Delta}_M(t^*)]^T$, and $\tilde{\mathbf{V}}(t^*)$ is a diagonal matrix of $\text{Var}[\tilde{\Delta}_r(t^*)]$, for $r = 1, \dots, M$. Here, $\tilde{\Delta}_r(t^*)$ can be each of the estimators among $\hat{\Delta}^{KM}_r(t^*)$, $\hat{\Delta}^{GF}_r(t^*)$, $\hat{\Delta}^{HJ}_r(t^*)$, and $\hat{\Delta}^{AG}_r(t^*)$. We assume the treatment effects are independent across regions. Under H_0, $\tilde{U}(t^*)$ follows a χ^2 distribution with $M-1$ degrees of freedom.

When the treatment effects are consistent across regions, we can combine the treatment effect estimates from each region and estimate a global treatment effect. The RMST difference for the global treatment effect is estimated by using an inverse variance weighted estimator (Sinha et al., 2011):

$$\tilde{\Delta}_G(t^*) = \frac{\sum_{r=1}^M \tilde{\Delta}_r(t^*)/\text{Var}[\tilde{\Delta}_r(t^*)]}{\sum_{r=1}^M 1/\text{Var}[\tilde{\Delta}_r(t^*)]}.$$

The variance of $\tilde{\Delta}_G(t^*)$ is $\sum_{r=1}^M \frac{1}{\text{Var}[\tilde{\Delta}_r(t^*)]}/\{\sum_{r=1}^M \frac{1}{\text{Var}[\tilde{\Delta}_r(t^*)]}\}^2$.

3.5 Simulation Study

3.5.1 Aims and Performance Measures

In this simulation study, we evaluate the finite sample performance of the proposed estimators for region-specific average RMST differences under various data-generating scenarios. The bias and variance of each estimator are evaluated under 1,000 Monte Carlo replications.

3.5.2 Data-Generating Mechanism

We simulate MRCTs conducted in three regions ($M = 3$) and suppose the numbers of enrolled participants in each region are $n_1 = 400$, $n_2 = 500$, and $n_3 = 600$. We consider two covariates $\mathbf{X} = [X_1, X_2]$ from a common population, where $X_1 \sim Unif(0, 1)$ and $X_2 \sim N(1, 1)$, and assume independence between X_1 and X_2. Suppose the distribution of $\mathbf{X}$ in the target population is the same as the common population. We consider two types of sampling score models, the log-linear model, and the logistic model:

$$\log\{\rho_r(\mathbf{X}_{ri})\} = \eta_{r0} + \eta_{r1}X_{1ri} + \eta_{r2}X_{2ri}, \text{ for } r = 1, 2, 3. \tag{3.14}$$

$$\text{logit}\{\rho_r(\mathbf{X}_{ri})\} = \eta^*_{r0} + \eta^*_{r1}X^*_{1ri} + \eta^*_{r2}X^*_{2ri}, \text{ for } r = 1, 2, 3. \tag{3.15}$$

In Equation (3.14), we assume the covariates have linear associations with the log of sampling score. In Equation (3.15), we assume the covariates have non-linear associations with the logistic of sampling score by letting $X^*_{1ri} = X_{1ri} \times X_{2ri}$ and $X^*_{2ri} = \exp\{X_{2ri}/10\}$. In each sampling score model, we consider two settings according to the level of similarity between enrolled patients in each region and the target population, which is quantified by the absolute standardized mean difference (SMD) (Austin, 2011) of covariates $\mathbf{X}$. As a result, we have four scenarios for the sampling as follows:

Scenario 1: Log-linear sampling with moderate SMDs.

Scenario 2: Log-linear sampling with large SMDs.

Scenario 3: Logistic-nonlinear sampling with moderate SMDs.

Scenario 4: Logistic-nonlinear sampling with large SMDs.

These simulation scenarios evaluate the robustness of the proposed estimators under different sampling score models and similarities between each region and the target population. The actual values of the parameters η_r's in Equation (3.14) and η_r^*'s in Equation (3.15), and the absolute SMDs from each scenario are given in Tables C.1 and C.2 in Appendix C.

The following settings are equivalent in all four scenarios. We assume the treatment propensity score $\pi_r(\mathbf{X}) = 0.5$ in all three regions, indicating the participants are randomized to two treatment groups with an allocation ratio of 1:1. The event time T_{rzi} is assumed to be generated from the following hazard function:

$$\begin{aligned} h(t|Z_{ri}, X_{1ri}, X_{2ri}, R_{ri}) &= \lambda(t|Z_{ri})\exp\{0.3I[R_{ri}=2] + 0.5I[R_{ri}=3] \\ &- X_{1ri} + 0.5X_{2ri} + 0.3Z_{ri}I[R_{ri}=2] + 0.5Z_{ri}I[R_{ri}=3] \\ &- Z_{ri}X_{1ri} - 0.5Z_{ri}X_{2ri} - 0.6I[R_{ri}=2]X_{1ri} + 0.3I[R_{ri}=2]X_{2ri} \\ &- I[R_{ri}=3]X_{1ri} + 0.5I[R_{ri}=3]X_{2ri}\}. \end{aligned} \tag{3.16}$$

We use different baseline hazard functions for each treatment group to simulate data under the non-PH assumption. We set the baseline hazard function for the experimental treatment group to be $\lambda(t|Z_{ri}=1) = 0.15t^{-0.7}$ and for the control group to be $\lambda(t|Z_{ri}=0) = 0.5$. The censoring time follows an exponential distribution, i.e., $C \sim \exp(0.15)$, which is common in all regions. We set a fixed truncation time for RMST at $t^* = 4$.

3.5.3 Estimands

The estimands are the region-specific average RMST differences concerning the target population. We present its derivation according to Equation (3.16) in Appendix C. The average RMST difference is 1.71 in Region 1, 1.51 in Region 2, and 1.15 in Region 3.

3.5.4 Methods

First, we consider a Naive estimator using the difference of standard unadjusted RMST (Equation (1.1)) between two treatment groups. The Naive estimator is calculated in each region. Then, we compare the proposed IPSW-adjusted and CW-adjusted estimators for the region-specific RMSTs. We consider two weighting methods using IPSW and one weighting method using CW, and each weighting method is applied to six weighted estimators. Overall, we have 18 weighted estimators and a Naive estimator.

In the weighted G-Formula estimator and weighted Augmented estimator, we fit two IPCW RMST regression models in each region:

$$\begin{aligned}
\mu_r(t^*|\mathbf{X}_r, Z_r) &= \beta_{r0} + \beta_{r1}Z_r + \beta_{r2}X_{r1} + \beta_{r3}X_{r2} + \beta_{r4}Z_rX_{r1} + \beta_{r5}Z_rX_{r2}, \\
\mu_r^{mis}(t^*|\mathbf{X}_r, Z_r) &= \beta'_{r0} + \beta'_{r1}Z_r + \beta'_{r2}X_{r1} + \beta'_{r3}Z_rX_{r1}.
\end{aligned}$$

We assume the first RMST regression model is correctly specified while the second RMST regression model is misspecified as it does not include the confounding effect of X_2. Accordingly, there are six estimators for each weighting method:

1. Weighted Kaplan-Meier estimator (KM).

2. Weighted G-formula estimator using correctly specified outcome models (GF).

3. Weighted G-formula estimator using mis-specified outcome models (GFmis).

4. Weighted Hajek estimator (HJ).

5. Weighted Augmented estimator using correctly specified outcome models (AG).

6. Weighted Augmented estimator using mis-specified outcome models (AGmis).

For the IPSW-adjusted estimators, with the known distribution of $\mathbf{X}$ in the target population, we derive the weight γ_{ri} as follows:

$$\gamma_{ri} = \frac{P_{F^*}(\mathbf{X} = \mathbf{X}_{ri})}{P_{F_r}(\mathbf{X} = \mathbf{X}_{ri})} \propto \frac{P_{F^*}(\mathbf{X} = \mathbf{X}_{ri})}{P(d_{ri} = 1 | X_{ri}, R_{ri} = r) P_{F^*}(\mathbf{X} = \mathbf{X}_{ri})} = \frac{1}{\rho_r(\mathbf{X}_{ri})}.$$

We consider two ways to calculate the sampling scores $\rho_r(\mathbf{X}) = P(d = 1 | X, R = r)$. The first uses the true sampling score as known by the data-generating mechanism. The second is estimated by the gradient boosted models (McCaffrey et al., 2013; Burgette et al., 2021) as introduced in Chapter 3.3.2.

For the CW-adjusted estimators, we consider the constraint function as $\mathbf{g}(\mathbf{X}) = [X_1, X_2, X_1^2, X_2^2]$, which makes constraints on the first and second moments of X_1 and X_2. This constraint function can empirically match the covariate distribution from each region to the target population.

3.5.5 Results

Figure 3.2 displays the results with box plots of the estimated RMST differences in Region 1 across four scenarios for different sampling score modeling. The first sub-figure shows the results from the IPSW-adjusted estimators under the true sampling score. The second sub-figure shows the results from the IPSW-adjusted estimators under the estimated sampling score. The third sub-figure shows the results from the CW-adjusted estimators. The Naive estimator is included in all panels for reference.

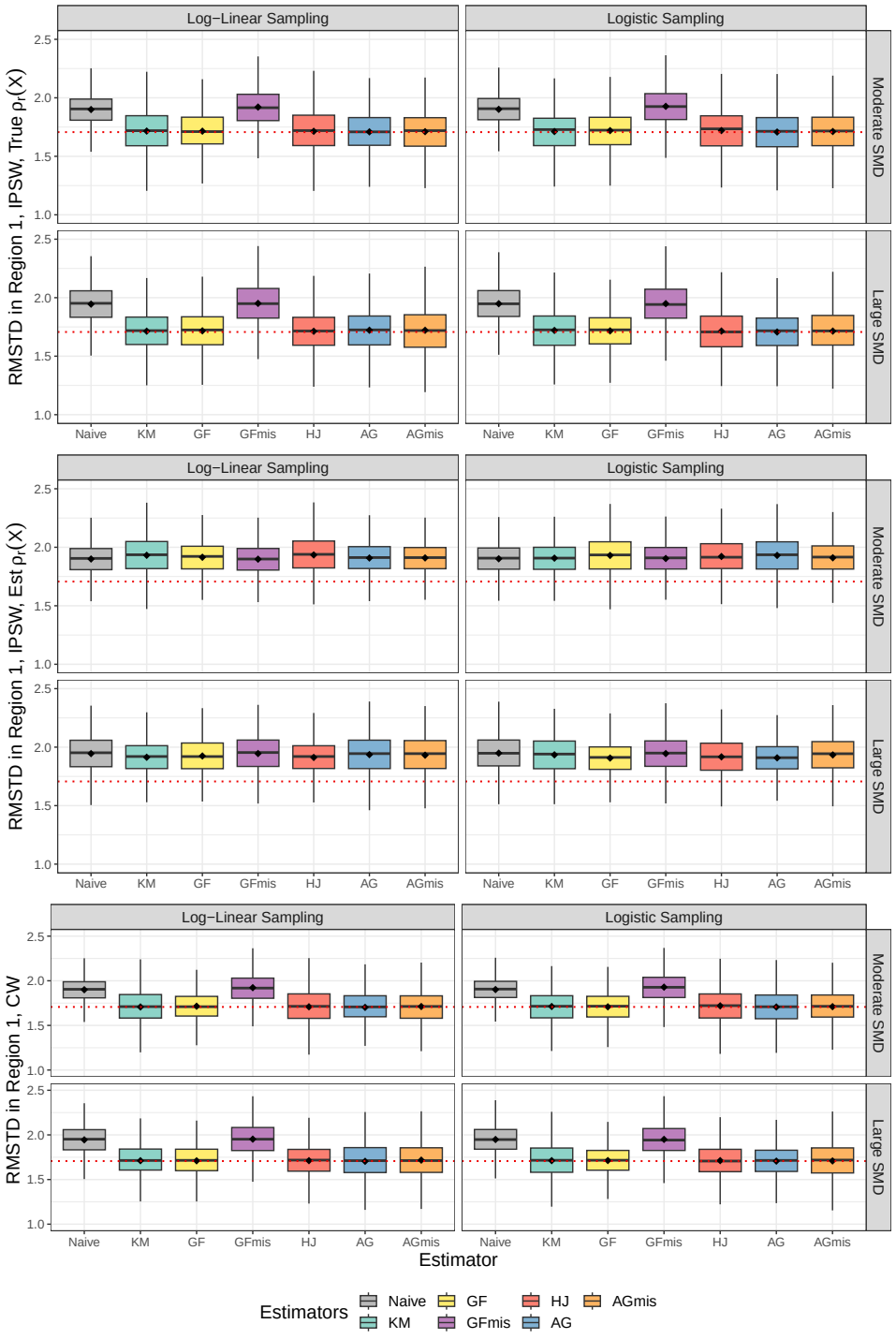

FIGURE 3.2: Boxplots of estimated average RMST difference (RMSTD) in Region 1 under four sampling scenarios in simulation study. Upper panel: IPSW-adjusted estimators with true sampling score; Middle panel: IPSW-adjusted estimators with estimated sampling score; Bottom panel: CW-adjusted estimators.

The Naive estimator and all IPSW-adjusted estimators with estimated sampling scores fail to generalize the treatment effects from each region to the target population under all scenarios. These results largely stem from the incorrect estimated sampling score from the gradient boosted model. In contrast, the IPSW-adjusted estimators with the true sampling scores and CW-adjusted estimators yield unbiased estimation across all scenarios, except for the weighted G-Formula estimator when the outcome models are misspecified. However, the weighted Augmented estimator demonstrates robustness under the outcome model misspecifications. The choice of sampling score models and similarity levels between each region and the target population do not impact bias, but larger SMDs would increase the variance of the estimators. The CW method yields a smaller variance within the same weighted estimator than the IPSW method. Among the five weighted estimators (excluding GFmis), when the outcome model is correctly specified, the weighted G-formula and weighted Augmented estimators exhibit similar and smaller variance than other estimators.

Figures C.1 and C.2 show the boxplots of estimated RMST differences in Region 2 and 3 respectively, which present similar results as in Region 1.

3.6 Data Analysis

We illustrate our proposed methods using the PLATO dataset under two scenarios: 1) a two-region analysis comparing treatment effects in US vs. non-US regions, and 2) a four-region analysis comparing treatment effects across four predefined geographic regions. The treatment effect is the RMST difference at the truncation time of $t^* = 360$ days. The results of the four-region analysis are presented in Appendix C.

3.6.1 Variable Selection

Previous effect modifier analyses suggested that the maintenance aspirin dose is the only variable that explained the statistically significant regional interaction effect with a p-value below 0.05 (Mahaffey et al., 2011; Carroll and Fleming, 2013). We conduct a univariable effect modifier analysis with the baseline characteristics listed in Mahaffey et al. (2011) to select more covariates for the weighting methods. We fit two univariable models in the pooled dataset for each candidate variable: 1) IPCW RMST regression model (Tian et al., 2014) with truncation time $t^* = 360$ days; 2) Cox regression model (Cox, 1972). Each model includes a treatment indicator, a single covariate, and a treatment-by-covariate interaction. A covariate will be selected if the p-value of its interaction with treatment is less than 0.2 in either of the two models. We choose the significance level of 0.2 to enhance the power to detect significant effect modifiers.

As a result, seven variables are selected as potential effect modifiers, including weight ($\geqslant 80$ kg), hypertension, percutaneous coronary intervention, smoking (ever), peripheral arterial disease, Killip classification (Level I vs. Level II-IV), and maintenance aspirin dose. The first six variables are binary, and we categorize maintenance aspirin dose into three levels (Mahaffey et al., 2011): 1) $\leqslant 100$ mg, 2) $(100, 300)$ mg, and 3) $\geqslant 300$ mg, and we let the first level be the reference group. Overall, we have eight binary covariates included in the weighting methods.

3.6.2 Methods

As we do not have data on the underlying super patient population for PLATO in the entire world, we define the target population as the patient population represented by a simple mixture of the enrolled patients from all regions for the MRCT. As such, we estimate the region-specific average RMST differences by using the proposed

weighted estimators concerning the patient population in the pooled dataset. The eight selected covariates are used to calculate the IPSW and CW. We assume that the treatment-specific propensity score $\pi_{ri}(\mathbf{X}_{ri})$ is constant based on the allocation rate from the randomized trial in each region. The outcome models are fitted by including all selected variables and their interaction with the treatment.

The Naive estimators use the standard unadjusted RMST differences in each region. The IPSW is derived based on Equation (3.3) as follows:

$$\gamma_{ri} \propto \frac{P(d'_{ri} = 1|\mathbf{X}_{ri})}{P(d_{ri} = 1|\mathbf{X}_{ri}, R_{ri} = r)} = \frac{\sum_r P(d_{ri} = 1|\mathbf{X}_{ri}, R_{ri} = r)P(R_{ri} = r|\mathbf{X}_{ri})}{P(d_{ri} = 1|\mathbf{X}_{ri}, R_{ri} = r)}.$$

As we do not have information on the patient population not enrolled in the RCT from each region, we approximate $P(R = r|X)$ by the proportion of the sample size, $n_r/\sum_r n_r$, for each patient. In the CW-adjusted estimators, we make constraints on the first moment as all covariates are binary.

The truncation time, $t^* = 360$ days, is shorter than the longest follow-up time among all regions. We use a weighted absolute SMD to evaluate the covariate similarity between each region and the target population before and after applying the weighting methods. Its mathematical definition is provided in Appendix C. When all weights are equal to 1 (i.e., unweighted), the weighted absolute SMD is equivalent to the standard unadjusted absolute SMD (Austin, 2011). Additionally, we conduct a region-specific consistency test of treatment effects for all estimators.

3.6.3 Results

Figure 3.3 and Table 3.1 present the estimated average RMST differences and the associated 95% confidence intervals (CI) comparing ticagrelor and clopidogrel for the primary outcome in the two-region analysis. The Naive RMST differences are

-4.0 days (95% CI: -13.6, 5.6) in US and 5.4 days (95% CI: 2.6, 8.1) in non-US regions, indicating the clopidogrel is more effective than ticagrelor in US, though not statistically significant, while ticagrelor is significantly more effective than clopidogrel in non-US. The consistency test for the Naive estimator shows a strong treatment effect heterogeneity between US and non-US, though not significant ($p = 0.07$).

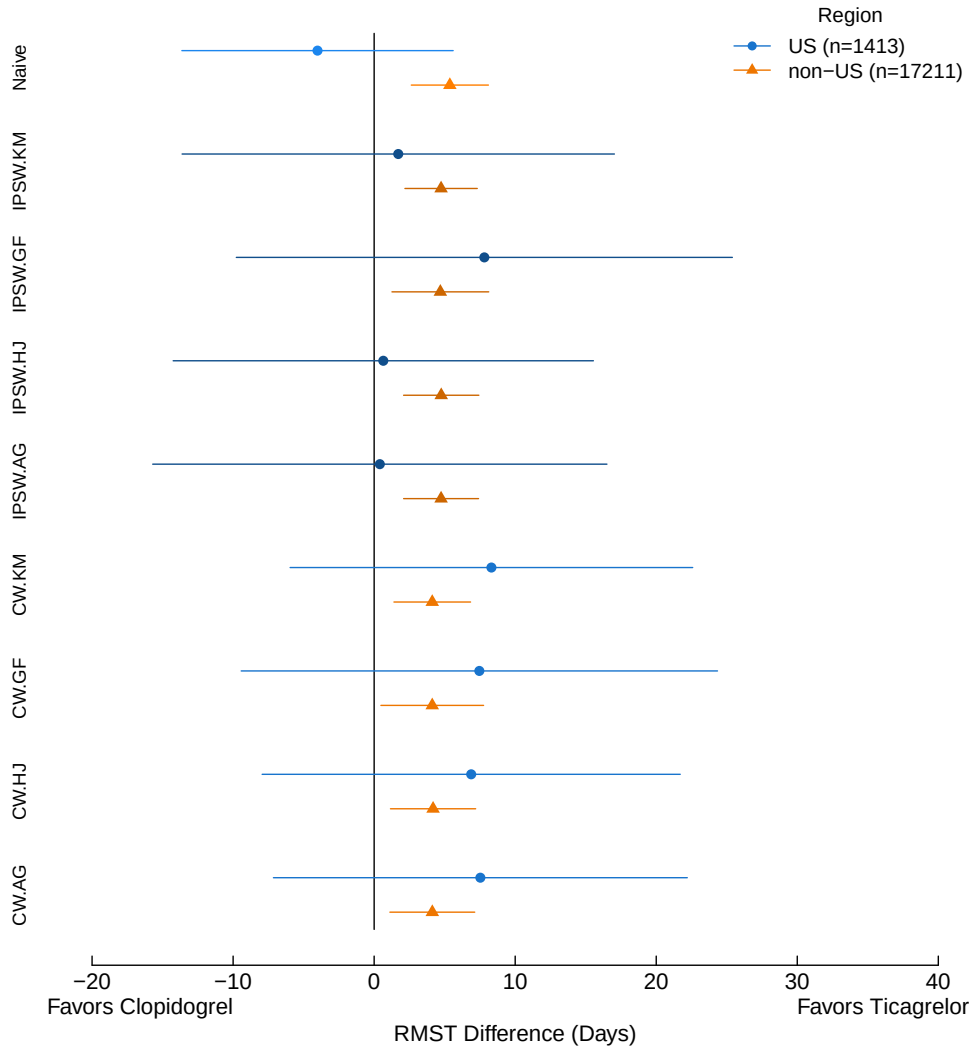

FIGURE 3.3: Forest plot of estimated average RMST differences with 95% CIs at $t^* = 360$ days from US and non-US in the two-region analysis for PLATO trial.

In the non-US region, the estimated RMST differences are very similar across four

estimators within the same weighting method. The IPSW-adjusted estimators yield slightly higher RMST differences than the CW-adjusted estimators (e.g., 4.7 vs. 4.2 days). Both weighting methods indicate that the ticagrelor is significantly more effective than clopidogrel in the non-US region. Under both weighting approaches, the weighted G-formula estimators have the largest variance, potentially due to misspecification of the fitted RMST regression model. Such misspecification could arise from the RMST regression model not including all confounders or a non-linear association between the selected variables and the outcome.

In the US region, the estimated RMST differences vary between the two weighting methods. Among the CW-adjusted estimators, the average RMST differences are close across four estimators, with the highest value in the weighted Augmented estimator at 9.1 days (95% CI: -4.4, 22.5). These results indicate that ticagrelor is more effective than clopidogrel for the patients in the US after having the selected variable distributions, notably the maintenance aspirin dose, be resembled in the target population. However, the average RMST difference in the US from the IPSW-adjusted Kaplan-Meier, Hajek, and Augmented estimators are much lower than the corresponding results from the CW-adjusted estimators. For example, the average RMST difference from the IPSW-adjusted Augmented estimator is 2.2 days (95% CI: -12.5, 17) in the US. The IPSW-adjusted G-formula estimator yields similar a result to the CW-adjusted G-formula estimator in the US region.

The consistency tests for all CW-adjusted and IPSW-adjusted estimators reveal no regional treatment effect heterogeneity (see p-values in Table 3.1). Consequently, we estimate a global RMST difference by combining the patients from both regions, indicating that ticagrelor is significantly more globally effective than clopidogrel. For example, the global RMST differences from the CW-adjusted and IPSW-adjusted Augmented estimator are 4.4 days (95% CI: 1.5, 7.3) and 4.7 days (95% CI: 2.0, 7.3),

respectively.

Table 3.1: Estimated average RMST differences with 95% CIs at $t^* = 360$ days from US and non-US in the two-region analysis, and the results from the regional consistency test with p-values and global RMST difference.

Method	RMSTD in US	RMSTD in non-US	P-value	Global RMSTD
Naive	-4.0 (-13.6, 5.6)	5.4 (2.6, 8.1)	0.07	4.7 (2.0, 7.3)
IPSW.KM	2.3 (-12.0, 16.5)	4.7 (2.2, 7.3)	0.74	4.6 (2.1, 7.2)
IPSW.GF	8.7 (-8.5, 25.9)	4.7 (1.3, 8.2)	0.66	4.9 (1.5, 8.2)
IPSW.HJ	1.3 (-12.5, 15.0)	4.7 (2.1, 7.4)	0.63	4.6 (2.0, 7.2)
IPSW.AG	2.2 (-12.5, 17.0)	4.7 (2.1, 7.4)	0.74	4.7 (2.0, 7.3)
CW.KM	7.4 (-6.0, 20.8)	4.1 (1.4, 6.9)	0.64	4.3 (1.6, 6.9)
CW.GF	8.3 (-8.3, 24.9)	4.1 (0.5, 7.8)	0.63	4.3 (0.8, 7.9)
CW.HJ	6.4 (-6.9, 19.7)	4.2 (1.2, 7.2)	0.75	4.3 (1.4, 7.2)
CW.AG	9.1 (-4.4, 22.5)	4.2 (1.2, 7.2)	0.48	4.4 (1.5, 7.3)

Figure 3.4 presents the weighted absolute SMDs of the eight selected covariates in the two-region analysis, comparing US and non-US regions to the target population. Since the non-US population dominates the target population, the absolute SMDs under the Naive estimator between non-US and the target population are close to 0 for all variables, except for the high aspirin dose group ($\geqslant$ 300 mg). However, the absolute SMDs reveal a notable imbalance of the selected covariates between the US and the target population. In contrast, the absolute SMDs under the CW method are 0 for all variables when comparing the US with the target population. The IPSW method shows good balances for all variables between the US and the target population except for the high aspirin dose group ($\geqslant$ 300 mg).

3.7 Discussion

We proposed the calibration weighting (CW) and inverse probability of sampling weighting (IPSW) adjusted estimators to transport or generalize the region-specific

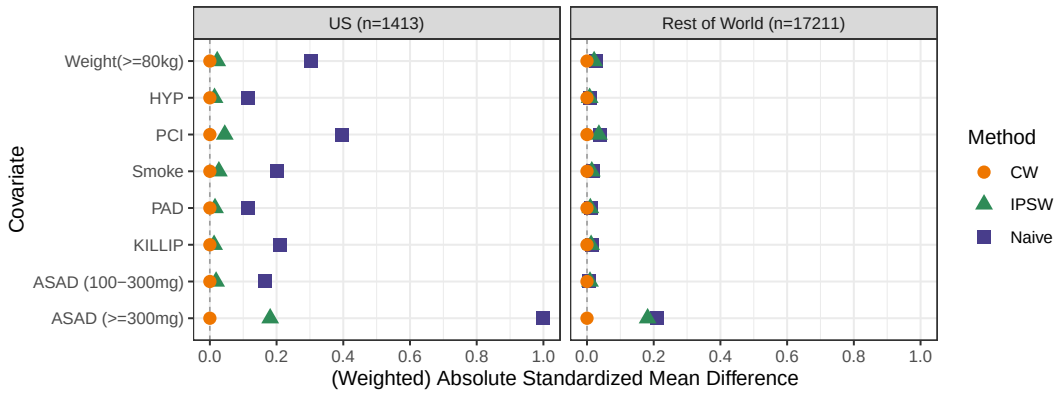

FIGURE 3.4: Weighted absolute standardized mean differences of eight covariates comparing US and non-US regions to the target population in the two-region analysis for PLATO trial. HYP: hypertension, PCI: percutaneous coronary intervention, PAD: peripheral arterial disease, KILLIP: Killip classification (Level I vs. Level II-IV), ASAD: aspirin dose.

treatment effect to a target population in the MRCT. These methods eliminate the disparities arising from the inessential traits. The large sample properties for the CW-adjusted and IPSW-adjusted estimators for the region-specific treatment effects were established. Furthermore, we developed a Wald-type test for the regional consistency test of treatment effect and provided the global treatment effect estimator when the consistency held. Our simulation study demonstrated that the CW-adjusted and IPSW-adjusted estimators with the true sampling scores consistently yielded unbiased estimations across all scenarios. Furthermore, the CW-adjusted estimators exhibited smaller variances compared to the IPSW-adjusted estimators. We highlighted that the IPSW-adjusted estimators are sensitive to the specification of the sampling score models, while the CW-adjusted estimators are more robust. Among the four proposed weighted estimators, the weighted Augmented estimator shows a smaller variance and higher robustness than other estimators.

In the PLATO analysis, our approaches strengthened the previous work that the apparent differences in ticagrelor effects between the US and non-US environment

were explained by the difference in the maintenance aspirin doses and a few other factors. After implementing the CW and IPSW approaches, no significant interaction existed between region and ticagrelor effects. We need to incorporate the difference in these inessential traits to correctly evaluate the region-specific treatment effect, mainly caused by essential traits such as race or genetic variants. We could use other reference populations than the simple mixture of enrolled patients across all regions based on different research questions. For example, if one is interested in knowing how much difference the treatment effect of US patients is from non-US patients, we could transport the treatment effects from US patients to the non-US patient population. Suppose there are no more differences in the treatment effects of these two regions after eliminating the difference from inessential traits. In that case, the US drug regulatory agency may be convinced that the apparent difference before transportation can be attributed to the imbalanced inessential traits (e.g., aspirin dose usage) between the two regions. If the region difference still exists, it may be attributed to other hidden region-specific factors (essential traits), such as racial and genetic differences between the two regions.

There are several avenues for future research to enhance our proposed method. First, when unobserved confounders have associations with both treatment effect and sampling scores, further research is required to reduce the bias caused by such factors. Second, we selected the effect modifiers in the PLATO analysis by fitting univariable Cox regression and RMST regression models. A more robust criteria model needs to be further investigated for high dimensional covariates to refine our approach. Third, our proposed methods could be extended to the MRCT design by incorporating power analysis, sample size calculation, and type I error control.

Our proposed method allows rigorous assessment of region-specific treatment effects against the target population and the treatment effect heterogeneity across regions by

eliminating the potential effect modifications of imbalanced baseline characteristics across regions in an MRCT. While our discussion of this work is in the context of MRCT analysis, it is worthy to note that this method can be applied to assessing treatment effect heterogeneity across patient subgroups for which the effect modifying covariates are not balanced across subgroups in randomized clinical trials.

Software

The relevant R code for the methodology and simulation study is available on `https://github.com/kimihua1995/CW_MRCT_RMST`.

Data Availability Statement

The PLATO data is available to members of the PLATO executive committee. Without the permission of the third parties and to avoid unintended leakage of patient privacy, we elect to not share the data. Individual investigators may reach out directly to the PLATO executive committee for collaboration.

4

Topic 3: Biomarker-Guided Adaptive Design with Threshold Detection and Patient Enrichment of Restricted Mean Survival Time

4.1 Introduction

4.1.1 Biomarker-Guided Designs with Adaptive Enrichment

Due to advancements in targeted therapies and precision medicine over the last two decades, there is a growing trend in the use of biomarker-guided designs for assessing personalized treatments in Phase II and Phase III clinical trials, especially in the field of oncology (Simon, 2014; Antoniou et al., 2016, 2017). Biomarkers associated with clinical outcomes, such as genomic, proteomic, or other biological characteristics, can be categorized as either prognostic or predictive (Antoniou et al., 2016). Prognostic biomarkers offer insights into the potential progression of a disease regardless of any specific treatment, whereas predictive biomarkers provide information about the patient's outcome given a specific treatment. As such, predictive biomarkers are often used in biomarker-guided trials to develop classifiers to identify appropriate patients as either excellent or poor candidates for clinical decisions to optimize therapy selec-

tions (Bailey et al., 2014; Chen et al., 2014; Lin and He, 2015; Landeck et al., 2016).

Recently, the designs with adaptive enrichment have become increasingly attractive in the context of biomarker-guided therapies as they provide additional flexibility during the trial (Fedorov and Liu, 2007; Wang et al., 2009; Simon and Simon, 2013; Antoniou et al., 2016). In this study, we consider the setting in which a single continuous biomarker is available at baseline. In general, a fixed threshold of the biomarker is pre-specified that dichotomies patients into "biomarker-positive" and "biomarker-negative" subgroups, and it is assumed that the biomarker-positive patients will benefit more from the new therapy over the active control. The biomarker-guided adaptive enrichment designs are typically conducted in a two-stage manner (Frieri et al., 2023; Stallard, 2023). In the first stage, they start with randomizing patients from the whole population. At the end of this stage, the accumulating data are used to decide whether to restrict accrual in the biomarker-positive subgroup in the second stage. Such adaptive enrichment could avoid further including patients who do not benefit from an intervention and may experience some negative effects (Wang et al., 2018; Lai et al., 2019; Thall, 2021). Existing literature demonstrates that such designs improve the efficiency of clinical trials by reducing the number of patients needed for randomization, particularly when the therapy of interest is known to be effective only in biomarker-positive patients and the biomarker assessment is highly accurate (Simon and Maitournam, 2004; Maitournam and Simon, 2005; Wang et al., 2019b).

However, the optimal cutoff for classifying the patients into positive and negative subgroups is not available and challenging to be pre-specified when the confirmatory randomized phase III trial is designed (Simon and Simon, 2013). To address this challenge, more flexible adaptive enrichment designs have been proposed by re-

searchers (Freidlin and Simon, 2005; Freidlin and Korn, 2010; Simon and Simon, 2013, 2018; Wang et al., 2020). While the trial is ongoing, they use the observed data to identify the subgroup of patients most likely to benefit from a treatment (i.e., biomarker-positive subgroup) based on the biomarker values and certain utility functions. These designs allow us to simultaneously identify the optimal biomarker threshold, define the biomarker-positive subgroup, and estimate and test the treatment effect in this subgroup. These approaches provide much of the efficiency of the enrichment strategy without the need to select a subset in advance, and the meaningful treatment effect estimates without being diluted by biomarker-negative patients who receive no or fewer benefits (Wason, 2023).

4.1.2 Study Motivations

Within the existing literature on biomarker-adaptive enrichment designs, there has been limited attention on the time-to-event outcomes (Mehta et al., 2014; Diao et al., 2018; Park et al., 2022; Johnston et al., 2022). For the studies with time-to-event outcomes, their designs were developed based on HR, which is commonly estimated by the Cox PH model (Cox, 1972) or requires testing treatment effect with log-rank test (Bland and Altman, 2004). However, violations of the PH assumption would render HR a potentially misleading and inappropriate summary of treatment effect (Lin and Wei, 1989). They may also considerably impact the statistical power of the log-rank test to detect the treatment difference (Mukhopadhyay et al., 2022). This study emphasizes a compelling alternative method based on the RMST. Although many RMST-based clinical trial design methods have recently been proposed (Trinquart et al., 2016; Weir and Trinquart, 2018; Luo et al., 2019; Trinquart et al., 2019; Eaton et al., 2020; Lu and Tian, 2021; Mao, 2023), limited attention has been given to the biomarker-adaptive enrichment design. We address this methodological and prac-

tical gap by proposing a two-stage biomarker-guided adaptive RMST design with threshold detection and patient enrichment.

Our proposed design is motivated by JAVELIN Lung 200 (see introduction in Chapter 1.1.1). The exploratory analyses in this trial revealed that avelumab was associated with improved OS and PFS compared to docetaxel in patients with higher PD-L1 expression at $\geqslant 50\%$ and $\geqslant 80\%$ cutoffs. However, these analyses were post hoc, and the biomarker thresholds were pre-specified and thus possibly not optimal. Moreover, the PH assumption was violated for both OS and PFS, so the characterization of the treatment effect based on the HR and its confidence interval estimated from the Cox PH model was not valid, and the hypothesis testing based on the suboptimal log-rank test was less powerful. Consequently, there is a need for an RMST-based design to identify an optimal threshold and define the biomarker-positive subgroup while the trial is ongoing. Enrichment is enabled to increase the accrual of patients from the selected subgroup and enhance the design's efficiency and power.

4.1.3 Chapter Outlines

The remainder of this chapter is structured as follows. The fundamental setups and design procedures are introduced in Chapter 4.2. Chapter 4.3 proposes two methods for identifying the biomarker threshold. In Chapter 4.4, we propose five estimators for treatment effect in the biomarker-positive subgroup. Chapter 4.5 outlines the approaches for type I error rate control, power analysis, and sample size calculation. Chapter 4.6 presents a numerical example of re-designing the JAVELIN Lung 200 study using simulated data. An extensive simulation study and an assessment of the finite sample properties of the proposed design and methods are provided in Chapter 4.7. The paper concludes by discussing various limitations and future works

in Chapter 4.8.

4.2 Design

4.2.1 Notations and Assumptions

We consider a two-stage enrichment design where patients are recruited and randomized between experimental and control treatments. Let Z be the indicator of the treatment assignment, where $Z = 1$ for the experimental and $Z = 0$ for the control. Let $X \in [0, 1]$ represent a single continuous biomarker of interest and is assumed to follow a standard uniform distribution, i.e., $X \sim Unif[0, 1]$. If X is a continuous marker with other distributions, the proposed design is still applicable when the values of X can be scaled to $[0, 1]$ through percentile transformations.

Following the potential outcomes framework (Rubin, 1974), let $T(1)$ and $T(0)$ be the potential time-to-event under treatment and control, respectively. By assuming the consistency of potential outcomes, the time-to-event is $T = T(1)Z + T(0)(1-Z)$. Let C be the censoring time, and we assume conditionally independent censoring given treatment Z and biomarker X, formally written as $C \perp T|Z, X$. In the presence of right censoring, T is not observed for all patients. Instead, we observe $U = \min(T, C)$ and the censoring indicator $\delta = I[T \leqslant C]$, where $I[\cdot]$ is the indicator function.

We denote $h_1(t|X)$ and $h_0(t|X)$ as the hazard functions for patients with $Z = 1$ and $Z = 0$, respectively. By assuming both $h_1(t|X)$ and $h_0(t|X)$ follow piecewise exponential models, the hazard functions are written as follows:

$$h_z(t|X) \quad = \quad \lambda_z(t) \times \exp\{\gamma_z X\}, \text{ for } z = 0, 1 \tag{4.1}$$

where $\lambda_z(t)$ is the baseline hazard and γ_z is the log HR which represents a constant association between biomarker and hazard. Consider J_z time intervals for the control and experimental treatments with change points $0 = \tau_{z,0} < \tau_{z,1} < \cdots < \tau_{z,J_z} = \infty$,

respectively. The j^{th} interval is defined as $[\tau_{z,j-1}, \tau_{z,j})$. We assume the baseline hazard is constant within each interval, so that:

$$\lambda_z(t) = \lambda_{z,j}, \text{ for } t \in [\tau_{z,j-1}, \tau_{z,j}) \tag{4.2}$$

Note that our proposed design is versatile and applicable to various hazard functions. The hazard functions in Equations (4.1) and (4.2) are employed here for illustrative purposes, and alternative hazard models can be considered based on the requirement of a specific design.

4.2.2 Optimal Biomarker Threshold

Several methods have been proposed for selecting the biomarker threshold and identifying biomarker-positive subgroups (Su et al., 2013; Lai et al., 2014; Renfro et al., 2014; Zhang et al., 2017; Lai et al., 2019; Wang et al., 2020; Zhao and LeBlanc, 2020). For example, the subgroup can be selected by either maximizing the product of its sample size and effect size (Zhao and LeBlanc, 2020) or maximizing the test statistic of treatment effect difference between two subgroups (Su et al., 2013; Renfro et al., 2014). Other approaches are based on maximizing the power of the test of treatment effect to select a subgroup (Lai et al., 2014; Zhang et al., 2017). However, defining the best subgroup using these methods involves a trade-off between the size and treatment effect in that subgroup (Wang et al., 2020). For example, there may be better choices than choosing a smaller subgroup with a higher average treatment effect. Furthermore, these approaches typically necessitate pre-specifying a set of candidate thresholds or employing a greedy search to identify the optimal one.

To address these issues, we adopt the methods introduced by Wang et al. (2020) to identify the optimal threshold by the intersection point of the biomarker response curves of two treatment groups (i.e., the RMST curves as functions of biomarker

value). This method parametrically models the biomarker response curves without mandating the pre-specification or greedy search for thresholds. Another advantage of using the intersection point is that it ensures the benefit of the experimental treatment for the biomarker-positive subgroup including all patients with positive treatment effects.

In the context of RMST, the biomarker-positive subgroup comprises patients whose biomarker values are within the range that satisfies the conditional RMST differences between experimental and control treatments are all positive, given X. Conversely, the biomarker-negative group comprises the range of X where the conditional RMST differences are all negative. We refer to the threshold between biomarker-positive and negative groups as the biomarker cutpoint, denoted as c_{t^*} concerning the truncation time t^*. Without loss of generality, we assume the biomarker value is positively associated with the RMST difference between experimental and control treatments. As such, the biomarker-positive and negative groups are defined as $X \in (c_{t^*}, 1]$ and $X \in [0, c_{t^*})$, respectively.

We denote $\mu_z(t^*|X)$ as the conditional RMST function for $z \in \{0, 1\}$, which can be written as:

$$\mu_z(t^*|X) = \int_0^{t^*} S_z(t|X)dt = \int_0^{t^*} \exp\left\{-\int_0^t h_z(u|X)du\right\} dt,$$

where $S_z(t|X)$ is the conditional survival function. The biomarker cutpoint, $X = c_{t^*}$, can be identified by solving the following equation concerning X:

$$\mu_1(t^*|X) - \mu_0(t^*|X) = 0. \tag{4.3}$$

Note that c_{t^*} may fall outside the range of $[0, 1]$. In such instances, the biomarker cutpoint will be truncated at 0 or 1, which is formally written as $c_{t^*} = min\{max\{c_{t^*}, 0\}, 1\}$.

4.2.3 Two-Stage Adaptive Enrichment Design

Figure 4.1 visualizes the diagram of the proposed design. Stage I is conducted from the calendar time 0 to t_1, and accrues n_1 patients per treatment arm regardless of the biomarker value. At the end of Stage I (at t_1), a biomarker cutpoint, denoted as $\hat{c}_0$, is estimated concerning the RMST difference at a pre-specified truncation time t^* related to a specific research question. Based on the result, two enrichment criteria can be introduced for Stage II: 1) accrue patients regardless of the biomarker value (i.e., no enrichment) in Stage II if $\hat{c}_0$ is not identified (e.g., all biomarker values are positive or negative to patients); 2) enrich and restrict the accrual in Stage II to patients in the biomarker-positive subgroup. Note that the trial will not be terminated in the first condition as the data are not mature in the early stage, and the treatment effect is not tested.

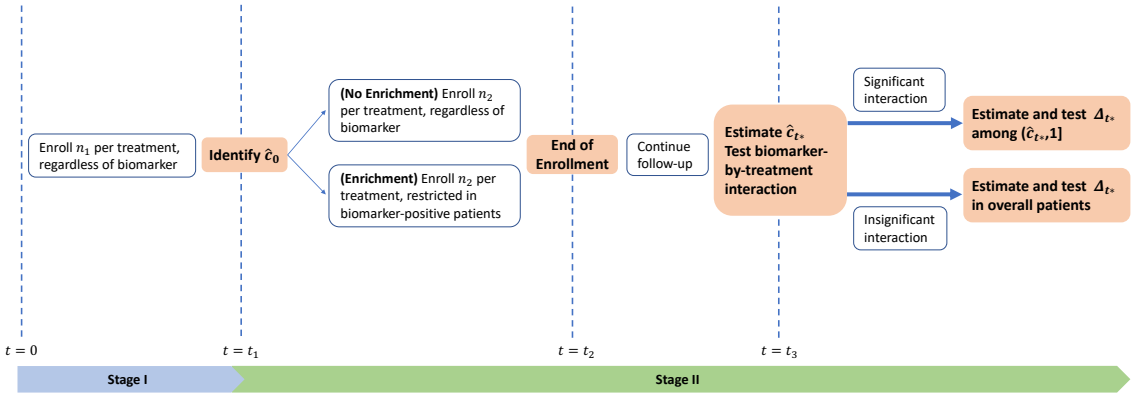

FIGURE 4.1: Diagram of the two-stage adaptive and enrichment design.

In Stage II, n_2 patients per treatment arm are accrued, and the enrollment continues from t_1 to t_2, resulting in a total of $n = 2n_1 + 2n_2$. All patients are continuously followed up until t_3, and we assume $t_3 > t^*$. After completing the follow-up, we estimate a biomarker cutpoint, denoted as $\hat{c}_{t*}$. Additionally, we test whether there exists a statistically significant positive relationship between the biomarker and treatment effect. This hypothesis test, denoted as Hypothesis Zero, can be written as follows:

- Null hypothesis H_{00}: There is no positive biomarker-by-treatment interaction for RMST.
- Alternative hypothesis H_{a0}: There is a positive biomarker-by-treatment interaction for RMST.

Based on the testing result, two mutually exclusive hypotheses for treatment effects are considered.

First, if H_{00} is rejected, we estimate and test the treatment effect (i.e., RMST difference) within the biomarker-positive subgroup $(\hat{c}_{t*}, 1]$. This hypothesis test, denoted as Hypothesis One, is written as:

- Null hypothesis H_{01}: The RMST difference in the positive group = 0.
- Alternative hypothesis H_{a1}: The RMST difference in the positive group > 0.

Second, if H_{00} is not rejected, we estimate and test the treatment effect in the overall population. This hypothesis test, denoted as Hypothesis Two, is written as:

- Null hypothesis H_{02}: The RMST difference in the overall population = 0.
- Alternative hypothesis H_{a2}: The RMST difference in the overall population > 0.

In Chapters 4.3 to 4.5, we will further describe the methods for identifying biomarker cutpoint, testing biomarker-by-treatment interaction, and estimating the corresponding treatment effect.

4.3 Biomarker Cutpoint Estimation

We explore two methods for identifying the biomarker cutpoint: 1) the prediction method and 2) the RMST regression method. In the proposed design, we employ

the first approach in Stage I and the second approach in Stage II.

4.3.1 Prediction Method

The prediction method involves two key steps. First, it fits a hazard function, $h_z(t|X)$, for each treatment based on the observed data. Second, it integrates the fitted hazard functions into Equation (4.3) and numerically estimates the biomarker cutpoint by solving this equation. This approach necessitates certain parametric assumptions on the survival curves or hazard functions within each treatment group. Examples of such assumptions include exponential, piecewise exponential, Weibull, log-normal, and others. In this context, we focus on the piecewise exponential models, specified in Equations (4.1) and (4.2).

For the i^{th} patient and j^{th} time interval, the Equations (4.1) and (4.2) can be jointly written as follows:

$$\log(\lambda_{z_i,ij}) = \log(\lambda_{z_i,j}) + \gamma_{z_i} X_i, \text{ for } t \in [\tau_{z_i,j-1}, \tau_{z_i,j}), \tag{4.4}$$

where $z_i \in \{0, 1\}$ is the treatment indicator and $\lambda_{z_i,j}$ is the baseline hazard in the j^{th} interval. We use a Poisson log-linear regression method (Holford, 1980; Laird and Olivier, 1981) to fit the model in Equation (4.4) and estimate the parameters $\hat{\lambda}_{z,j}$ and $\hat{\gamma}_z$ for $z \in \{0, 1\}$. The details are presented in Appendix D.

If the number of intervals J_z and the change points $\tau_{z,0}, \ldots, \tau_{z,J_z}$ are known, fitting the model in Equation (4.4) is straightforward. However, without such prior information, the challenge lies in identifying the number of change points and their locations. In such instances, we adapt the approach proposed by Goodman et al. (2011) to determine J_z and estimate $\tau_{z,1}, \ldots, \tau_{z,J_z-1}$. They proposed a sequential testing approach for determining the number of change points in the piecewise hazard function. Given a fixed number of intervals $J_z \geqslant 2$, we at first compute the

maximum likelihood estimates for both the piecewise hazards $\alpha_{z,j}$ and change points $\tau_{z,j}$ for $j = 1, \ldots, J_z - 1$. Then, the process begins by testing the null hypothesis that there are no change points against the alternative hypothesis with one change point. If the null hypothesis is rejected, we test the subsequent test involving the null hypothesis indicating one change point against the alternative with two change points in the model. This procedure continues until a null hypothesis cannot be rejected. If the overall significance level is α^*, they suggested using $\alpha^*(k) = \frac{\alpha^*}{2^{k-1}}$ for the k^{th} test to control the type I error rate of this test (Goodman et al., 2011).

The strength of this prediction method lies in its ability to predict a biomarker cutpoint concerning a future truncation time t^*, which is particularly valuable in survival adaptive design where an early enrichment is sought to accrue more patients benefiting from the new therapy. In the context of our design, this approach is particularly advantageous in Stage I when the longest follow-up time t_1 is shorter than t^*. However, one limitation is its dependence on a parametric assumption on the survival curves. Moreover, under piecewise exponential models, it assumes no change point in the hazard functions between t_1 and t^*. As such, when a sufficient number of patients are being followed up beyond t^*, an alternative method (e.g., RMST regression method) can be used to estimate the biomarker cutpoint without the parametric model assumptions.

4.3.2 RMST Regression Method

The RMST regression method adapts the IPCW RMST regression (see Chapter 1.2.3) to estimate the biomarker cutpoint. For a specific truncation time t^*, let $Y = \min(T, t^*)$ denote the truncated time-to-event outcome at t^*, the RMST up to t^* given the biomarker and treatment is defined by $\mu(t^*|X, Z) = E(Y|X, Z)$. The

RMST regression model is written as follows:

$$g(\mu(t^*|X,Z)) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX, \tag{4.5}$$

where $g(\cdot)$ is the link function. We use the identity link function throughout this study, but the log link can also be used. The model parameters are denoted as $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$.

Under the identity link, the conditional RMSTs given X under the experimental and control treatments can be written as follows:

$$\begin{aligned} \hat{\mu}(t^*|X,1) &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X + \hat{\beta}_3 X, \\ \hat{\mu}(t^*|X,0) &= \hat{\beta}_0 + \hat{\beta}_2 X. \end{aligned} \tag{4.6}$$

Consequently, the conditional RMST difference comparing the experimental and control is $\hat{\mu}(t^*|X,1) - \hat{\mu}(t^*|X,0) = \hat{\beta}_1 + \hat{\beta}_3 X$. Under the assumption that the treatment effect is positively related to the biomarker value, i.e., $\beta_3 > 0$, the cutpoint $\hat{c}_{t^*} = -\hat{\beta}_1/\hat{\beta}_3$ and the biomarker-positive group is defined as $X \in [-\hat{\beta}_1/\hat{\beta}_3, 1]$.

The advantage of this approach is that it does not require parametric assumptions on the hazard functions. However, it is essential to ensure that the truncation time t^* does not surpass the maximum of the observed event time (i.e., $U = \min(T, C)$) within each treatment group. At the end of Stage II of the proposed design, we utilize this method to estimate the biomarker cutpoint as the final analysis time $t_3 > t^*$.

4.4 Treatment Effect Estimators

We propose five methods to estimate the treatment effect, i.e., the RMST difference in the biomarker-positive subgroup, at the end of Stage II: 1) Naive unadjusted estimator, 2) Calibration weighted Kaplan-Meier estimator, 3) Calibration weighted

G-formula estimator, 4) Calibration weighted Hajek estimator, and 5) Calibration weighted augmented estimator.

Let c_{t*} denote the true biomarker cutpoint, the treatment effect estimand, denoted as $\Delta_{t*}^{(P)}$, is defined by the marginal RMST difference between the experimental and control treatments among the true biomarker-positive subgroup $(c_{t*}, 1]$. It can be formally written as:

$$\Delta_{t*}^{(P)} = \int_{c_{t*}}^{1} \{\mu_1(t^*|x) - \mu_0(t^*|x)\}\, dF_+(x), \tag{4.7}$$

where $\mu_1(t^*|x)$ and $\mu_0(t^*|x)$ are the conditional RMST for the experiment and control treatments, respectively, and $F_+(x)$ is the cumulative distribution function of X within the true biomarker-positive subgroup. Note that $\Delta_{t*}^{(P)} = \Delta_{t*}^{(O)}$ when $c_{t*} = 0$ in Equation (4.7).

4.4.1 Naive Unadjusted Estimator

The naive unadjusted method employs Equation (1.1) to estimate Δ_{t*}. Assume at the end of Stage II, there are a total of n_+ patients from both stages in the estimated biomarker-positive subgroup $(\hat{c}_{t*}, 1]$. For $i = 1, ..., n_+$ and $z = 0, 1$, let $N_{zi}(t) = I[U_i \leqslant t; \delta_i = 1; Z_i = z]$ denote the individual treatment-specific counting process and $Y_{zi}(t) = I[U_i \geqslant t; Z_i = z]$ denote the individual treatment-specific at-risk process. The treatment-specific KM estimator for the survival function is $\hat{S}_z(t) = \prod_{u \leqslant t} \left\{1 - \frac{dN_z(u)}{Y_z(u)}\right\}$, where $N_z(u) = \sum_{i=1}^{n_+} N_{zi}(u)$ and $Y_z(u) = \sum_{i=1}^{n_+} Y_{zi}(u)$. The naive unadjusted RMST difference estimator is:

$$\hat{\Delta}_1(t^*) = \int_0^{t^*} \left\{\hat{S}_1(t) - \hat{S}_0(t)\right\} dt. \tag{4.8}$$

The estimator $\hat{\Delta}_1(t^*)$ is consistent with $\Delta_{t^*}^{(P)}$ when the estimated biomarker cutpoint $\hat{c}_{t^*}$ is unbiased to c_{t^*} and the biomarker distribution of n_+ patients aligns with $F_+(x)$. According to Zhao et al. (2016); Tian et al. (2018), as the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_1(t^*) - \Delta_{t^*}^{(P)}\} \xrightarrow{d} N\left(0, \sigma_1^2(t^*)\right)$, where $\sigma_1^2(t^*)$ is estimated by:

$$\hat{\sigma}_1^2(t^*) = \int_0^{t^*} \left\{ \int_t^{t^*} \hat{S}_1(u)du \right\}^2 \frac{d\hat{\Lambda}_1(t)}{Y_1(t)/n_+} + \int_0^{t^*} \left\{ \int_t^{t^*} \hat{S}_0(u)du \right\}^2 \frac{d\hat{\Lambda}_0(t)}{Y_0(t)/n_+}.$$

Here, $\hat{\Lambda}_z(t) = -\log\{\hat{S}_z(t)\}$ is the Nelson-Aalen estimator for the cumulative hazard function for $z \in \{0, 1\}$.

However, since the proposed design uses a different biomarker cutpoint $\hat{c}_0$ estimated at the end of Stage I for enrichment, the distribution of the included biomarker-positive patients may not align with the distribution for $X \in (\hat{c}_{t^*}, 1]$. Specifically, when $\hat{c}_0 \leqslant \hat{c}_{t^*}$, the biomarker value of the enriched patients in Stage II would cover the range of $X \in (\hat{c}_{t^*}, 1]$. In such a situation, the distribution of n_+ patients matches with the true distribution for $X \in (\hat{c}_{t^*}, 1]$. When $\hat{c}_0 > \hat{c}_{t^*}$, the biomarker values of the enriched patients represent only a subset of $(\hat{c}_{t^*}, 1]$. In such a situation, the distribution of these patients' biomarkers would be truncated to $X \in (\hat{c}_0, 1]$. This could lead to an overestimation of $\hat{\Delta}_1(t^*)$ for $\Delta_{t^*}^{(P)}$ under the assumption that the biomarker is positively related to the treatment effect. A weighting approach is needed to balance the distribution of the included n_+ biomarker-positive patients against the true distribution of $F_+(x)$.

4.4.2 Calibration Weighted Estimators

In our proposed design, we use the calibration weighting (CW) method, as introduced in Chapter 3.3.2, to balance the distributions and characterize the RMST difference from the included positive patients to the true biomarker-positive subgroup with

distribution $F_+(X)$. As patients in Stage I are accrued regardless of the biomarker value, we define the target population as the biomarker-positive subgroup in Stage I. Suppose that among the overall n_+ patients in the biomarker-positive subgroup, the first $\tilde{n}_+$ patients are accrued from Stage I. For each patient $i = 1, \ldots, n_+$, the CW p_i is calculated through an optimization problem using the negative entropy objective function as follows:

$$\min \left\{ \sum_{r=1}^{n_+} p_i log(p_i) \right\},$$

with constraints:

$$\begin{aligned} \sum_{i=1}^{n_+} p_i \mathbf{g}(X_i) &= \tilde{\mathbf{g}} \\ \sum_{i=1}^{n_+} p_i &= 1. \end{aligned}$$

Here, we let $\mathbf{g}(X_i) = [X_i, X_i^2]^T$ and $\tilde{\mathbf{g}} = \sum_{i=1}^{\tilde{n}_+} [X_i, X_i^2]^T / \tilde{n}_+$. As such, the first constraint equalizes the first and second moments of X between the overall biomarker-positive patients and the target population. It ensures that the weighted biomarker distribution among the positive patients in the proposed design empirically matches the distribution of X from the target distribution. The second constraint implies that p_i's sum to a normalization constant of one, which guarantees a valid density function (Hainmueller, 2012; Zhao and Percival, 2016).

The CW p_i can be solved by using Lagrange Multiplier (De la Fuente, 2000).

$$\hat{p}_i = \frac{\exp\{\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{X}_i)\}}{\sum_{i=1}^{n_+} \exp\{\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{X}_i)\}}, \tag{4.9}$$

where $\boldsymbol{\lambda}$ solves $\sum_{i=1}^{n_+} \exp\{\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{X}_i)\}\{\mathbf{g}(\mathbf{X}_i) - \tilde{\mathbf{g}}\} = 0$.

Next, we present four CW-adjusted estimators of $\Delta_{t*}^{(P)}$. The structure and large

sample properties of these estimators are similar to those proposed in Chapter 3.4.

Calibration Weighted Kaplan-Meier Estimator

For $z \in \{0,1\}$, let $\tilde{N}_z(t) = \sum_{i=1}^{n_+} \hat{p}_i N_{zi}(t)$ denote the CW treatment-specific counting process and $\tilde{Y}_z(t) = \sum_{i=1}^{n_+} \hat{p}_i Y_{zi}(t)$ denote the CW treatment-specific at-risk process, the CW Kaplan-Meier estimator is written as:

$$\hat{\Delta}_2(t^*) = \tilde{\mu}_1(t^*) - \tilde{\mu}_0(t^*), \tag{4.10}$$

where $\tilde{\mu}_z(t^*)$ is the treatment-specific CW RMST, defined as:

$$\tilde{\mu}_z(t^*) = \int_0^{t^*} \tilde{S}_z(t)dt = \int_0^{t^*} \prod_{u \leqslant t} \left\{ 1 - \frac{d\tilde{N}_z(u)}{\tilde{Y}_z(u)} \right\} dt.$$

According to Theorem 1 in Chapter 3.4.1, $\hat{\Delta}_2(t^*)$ is consistent with $\Delta_{t^*}^{(P)}$, and as the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_2(t^*) - \Delta_{t^*}^{(P)}\} \xrightarrow{d} N\left(0, \sigma_2^2(t^*)\right)$, where $\sigma_2^2(t^*)$ is estimated by:

$$\hat{\sigma}_2^2(t^*) \;=\; n_+ \sum_{z=0,1} \int_0^{t^*} \left\{ \int_u^{t^*} \tilde{S}_z(t)dt \right\}^2 \frac{d\tilde{N}_z(u)}{\tilde{W}_z(u)(\tilde{Y}_z(u) - \Delta\tilde{N}_z(u))},$$

where $\tilde{W}_z(u) = \tilde{Y}_z^2(u) / \sum_{i=1}^{n_+} \hat{p}_i^2 Y_{zi}(u)$ and $\Delta\tilde{N}_z(u) = \tilde{N}_z(u) - \tilde{N}_z(u-)$.

Calibration Weighted G-Formula Estimator

The CW G-Formula estimator combines the IPCW RMST regression and the CW. The outcome models in this estimator are identified by the conditional RMSTs defined in Equations (4.6). Here, we let the outcome models be $m_1(X) = \hat{\mu}(t^*|X,1)$ and $m_0(X) = \hat{\mu}(t^*|X,0)$. The CW G-formula estimator is written as:

$$\hat{\Delta}_3(t^*) = \frac{\sum_{i=1}^{n_+} \hat{p}_i\{m_1(x_i) - m_0(x_i)\}}{\sum_{i=1}^{n_+} \hat{p}_i} = \sum_{i=1}^{n_+} \hat{p}_i(\hat{\beta}_1 + \hat{\beta}_3 x_i) \tag{4.11}$$

According to Theorem 2 in Chapter 3.4.2, when the outcome models are not misspecified, $\hat{\Delta}_3(t^*)$ is consistent with $\Delta_{t^*}^{(P)}$. As the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_3(t^*) - \Delta_{t^*}^{(P)}\} \xrightarrow{d} N\left(0, \sigma_3^2(t^*)\right)$, where $\sigma_3^2(t^*)$ can be estimated using the Delta method (Dowd et al., 2014):

$$\hat{\sigma}_3^2(t^*) = n_+ J_{\hat{\beta}}^T \Sigma_{\hat{\beta}} J_{\hat{\beta}},$$

where $\Sigma_{\hat{\beta}} = \begin{pmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_3) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_3) & \text{Var}(\hat{\beta}_3) \end{pmatrix}$ is the variance-covariance matrix of the parameter vector $[\hat{\beta}_1, \hat{\beta}_3]^T$ and

$$J_{\hat{\beta}} = \sum_{i=1}^{n_+} \frac{\partial \hat{p}_i(\hat{\beta}_1 + \hat{\beta}_3 x_i)}{\partial [\hat{\beta}_1, \hat{\beta}_3]^T} = [1, \sum_{i=1}^{n_+} \hat{p}_i x_i]^T.$$

Calibration Weighted Hajek Estimator

By combining the IPCW estimator of RMST, $\hat{\mu}_{IPCW}(t^*)$ as defined in Equation (1.2), and the CW, the CW Hajek estimator is written as:

$$\hat{\Delta}_4(t^*) = \frac{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i Y_i}{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i} - \frac{\sum_{i=1}^{n_+} \hat{p}_i (1 - Z_i) w_i Y_i}{\sum_{i=1}^{n_+} \hat{p}_i (1 - Z_i) w_i}, \tag{4.12}$$

where $Y_i = \min(T_i, t^*)$ is the truncated time-to-event outcome and $w_i = \delta_i^* / \hat{G}(Y_i)$ is the IPCW for the i^{th} patient. According to Theorem 3 in Chapter 3.4.3, $\hat{\Delta}_4(t^*)$ is consistent with $\Delta_{t^*}^{(P)}$. As $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_4(t^*) - \Delta_{t^*}^{(P)}\} \xrightarrow{d} N\left(0, \sigma_4^2(t^*)\right)$. The derivation of the asymptotic variance $\hat{\sigma}_4^2(t^*)$ is provided in Appendix D.

Calibration Weighted Augmented Estimator

The CW augmented estimator combines the CW Hajek and CW G-formula estimators:

$$\begin{aligned}\hat{\Delta}_5(t^*) &= \frac{\sum_{i=1}^{n_+}\hat{p}_i Z_i w_i\{Y_i - m_1(x_i)\}}{\sum_{i=1}^{n_+}\hat{p}_i Z_i w_i} - \frac{\sum_{i=1}^{n_+}\hat{p}_i (1-Z_i) w_i\{Y_i - m_0(x_i)\}}{\sum_{i=1}^{n_+}\hat{p}_i (1-Z_i) w_i} \\ &+ \frac{\sum_{i=1}^{n_+}\hat{p}_i\{m_1(x_i) - m_0(x_i)\}}{\sum_{i=1}^{n_+}\hat{p}_i}. \end{aligned} \tag{4.13}$$

Based on the semiparametric theory (Tsiatis, 2006), this estimator is doubly robust and does not require the correct specification of the outcome models. According to Theorem 4 in Chapter 3.4.4, $\hat{\Delta}_5(t^*)$ is consistent with $\Delta_{t^*}^{(P)}$. As the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_5(t^*) - \Delta_{t^*}^{(P)}\} \xrightarrow{d} N\left(0, \sigma_5^2(t^*)\right)$, and we derive the asymptotic variance $\hat{\sigma}_5^2(t^*)$ in Appendix D.

4.5 Type I Error Rate, Power, and Sample Size

4.5.1 Type I Error Rate and Critical Values

In this study, we focus on the approaches that control the type I error rate under the global null setting such that the hazard functions for the experimental and control treatments are identical, i.e., $h_0(t|X) = h_1(t|X)$ for any $X \in [0,1]$. As outlined in Chapter 4.2.3, our proposed design incorporates one null hypothesis H_{00} for testing positive biomarker-by-treatment interaction and two null hypotheses H_{01} and H_{02} for testing treatment effect. We discuss the type I error rate controls on these two aspects.

First, we assign α_0 as the significant level for testing H_{00}. Based on the RMST regression (Equation (4.5)) fitted at the end of Stage II, we evaluate whether the biomarker exhibits a significant positive interaction with the treatment by testing if $\beta_3 > 0$. As such, the Hypothesis Zero can be rewritten as:

- H_{00}: $\beta_3 = 0$.
- H_{a0}: $\beta_3 > 0$.

The test statistic is defined as $Z_{\beta_3} = \sqrt{n}\hat{\beta}_3/\hat{\sigma}_{\beta_3}$, where $\hat{\sigma}_{\beta_3}/\sqrt{n}$ is the standard error of $\hat{\beta}_3$. Under the null hypothesis H_{00}, $Z_{\beta_3} \sim N(0,1)$. To control the Type I error rate at α_0, the critical value for rejecting H_{00} is $q_0 = \Phi^{-1}(1-\alpha_0)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of standard normal distribution.

Second, we define the family-wise type I error rate α for testing the treatment effect as the probability of rejecting H_{01} or H_{02} under the global null. We allocate α_1 and α_2 to test H_{01} and H_{02}, respectively. Since H_{01} and H_{02} are mutually exclusive, we have $\alpha = \alpha_1 + \alpha_2$. We denote $Z_l^{(P)}$ and $Z_l^{(O)}$ as the test statistics when evaluating the treatment effect in the biomarker-positive and overall patients, respectively. According to the five treatment effect estimators proposed in Chapter 4.4, the test statistics are defined as $Z_l^{(P)} = \sqrt{n_+}\hat{\Delta}_l^{(P)}(t^*)/\hat{\sigma}_l^{(P)}(t^*)$ and $Z_l^{(O)} = \sqrt{n}\hat{\Delta}_l^{(O)}(t^*)/\hat{\sigma}_l^{(O)}(t^*)$, for $l = 1, ..., 5$. Here, $\hat{\Delta}_l^{(P)}(t^*)$ is the treatment effect estimator for the biomarker-positive subgroup, and $\hat{\Delta}_l^{(O)}(t^*)$ is the treatment effect estimator for the overall patients which can be estimated by letting the estimated biomarker cutpoint $\hat{c}_{t^*} = 0$. As such, α_1 and α_2 are calculated as follows:

$$\begin{aligned}
\alpha_1 &= P(H_{01} \text{ is rejected} \mid \text{global null}) \\
&= P(H_{01} \text{ is rejected} \mid H_{00} \text{ is rejected}) \times P(H_{00} \text{ is rejected} \mid \text{global null}) \\
&= P(Z_l^{(P)} > q | H_{00} \text{ is rejected}) \times P(Z_{\beta_3} > q_0 | \text{ global null}) \\
&= \pi_1 \times \alpha_0, \qquad (4.14)
\end{aligned}$$

$$
\begin{aligned}
\alpha_2 &= P(H_{02} \text{ is rejected} \mid \text{global null}) \\
&= P(H_{02} \text{ is rejected} \mid H_{00} \text{ is not rejected}) \times P(H_{00} \text{ is not rejected} \mid \text{ global null}) \\
&= P(Z_l^{(O)} > q \mid H_{00} \text{ is not rejected}) \times P(Z_{\beta_3} \leqslant q_0 \mid \text{ global null}) \\
&= \pi_2 \times (1 - \alpha_0). \qquad (4.15)
\end{aligned}
$$

Here, q is the critical value for testing the conditional treatment effects in both biomarker-positive and overall patients. We assume a significance level of $\tilde{\alpha}$ for testing the conditional treatment effects, such that $q = \Phi^{-1}(1 - \tilde{\alpha})$. Here, π_1 is the conditional error rate for H_{01} when H_{00} is rejected and π_2 is the conditional error rate for H_{02} when H_{00} is not rejected. When H_{00} is not rejected, $Z_l^{(O)} \sim N(0, 1)$, thus we have $\pi_2 = \tilde{\alpha}$. When H_{00} is rejected, a biomarker is estimated from the prediction method as introduced in Chapter 4.3.1, and we calculate π_1 through Monte Carlo method. Taking JAVELIN Lung 200 as an example, π_1 can be approximated by bootstrapping the data from patients with docetaxel (the standard care) for both treatment groups. As a result, the family-wise type I error rate can be controlled at α by adjusting the critical values q and q_0.

4.5.2 Global Power

To estimate the power of our design, we need to consider a specific alternative setting with hazard functions $h_1(t|X)$ and $h_0(t|X)$. We assume H_{a1} as an alternative setting such that a positive biomarker-by-treatment interaction exists and the RMST difference in the positive group is greater than 0. As under H_{a1}, the estimated biomarker cutpoint $\hat{c}_{t*}$ is identifiable and unbiased to the true cutpoint c_{t*} (shown in simulation study), the test statistic $Z_l^{(P)}$ is calculated for the true biomarker-positive subgroup $(c_{t*}, 1]$.

Once the critical values q and q_0 are determined in Equations (4.14) and (4.15), the

global power can be calculated as:

$$\begin{aligned} \text{Global Power} &= P(Z_l^{(P)} > q | H_{00} \text{ is rejected}, H_{a1}) \times P(Z_{\beta_3} > q_0 | H_{a1}) \\ &+ P(Z_l^{(O)} > q | H_{00} \text{ is not rejected}, H_{a1}) \times P(Z_{\beta_3} \leqslant q_0 | H_{a1}), \end{aligned} \tag{4.16}$$

where $q = \Phi^{-1}(1-\tilde{\alpha})$ and $q_0 = \Phi^{-1}(1-\alpha_0)$.

4.5.3 Sample Size Calculation

Given the global power and the critical values q and q_0, we calculate the sample size using Equation (4.16) and the asymptotic distribution of the test statistics as detailed in Chapter 4.4. Let n_+ and n represent the required sample size in the biomarker-positive subgroup and the overall patients, respectively. Note that $n_+ = f(n, c_{t*})$ is a function of n and c_{t*} based on the enrichment strategy and accrual rate in the proposed design. For example, in the all-comer design with no enrichment, $f(n, c_{t*}) = n(1-c_{t*})$.

Under H_{a1}, we define the true RMST difference in the biomarker-positive subgroup and the overall patients as $\Delta_{t*}^{(P)}$ and $\Delta_{t*}^{(O)}$, respectively. Then we have:

$$\frac{\hat{\Delta}_l^{(P)}(t^*) - \Delta_{t*}^{(P)}}{\hat{\sigma}_l^{(P)}(t^*)/\sqrt{n_+}} \sim N(0,1) \text{ and } \frac{\hat{\Delta}_l^{(O)}(t^*) - \Delta_{t*}^{(O)}}{\hat{\sigma}_l^{(O)}(t^*)/\sqrt{n}} \sim N(0,1), \text{ for } l = 1, \ldots, 5.$$

To test the positive biomarker-by-treatment interaction, we have $\frac{\hat{\beta}_3 - \beta_3}{\hat{\sigma}_{\beta_3}/\sqrt{n}} \sim N(0,1)$ under H_{a1}. Therefore, $P(Z_{\beta_3} > q_0 | H_{a1}) = 1 - \Phi(q_0 - \frac{\beta_3}{\hat{\sigma}_{\beta_3}/\sqrt{n}})$. Let $\eta = P(Z_{\beta_3} > q_0 | H_{a1})$, the Equation (4.16) can be written as:

$$\begin{aligned} \text{Global Power} &= \left[1 - \Phi(q - \frac{\Delta_{t*}^{(P)}}{\hat{\sigma}_l^{(P)}(t^*)/\sqrt{f(n, c_{t*})}})\right] \times \eta \\ &+ \left[1 - \Phi(q - \frac{\Delta_{t*}^{(O)}}{\hat{\sigma}_l^{(O)}(t^*)/\sqrt{n})}\right] \times (1-\eta) \end{aligned} \tag{4.17}$$

By solving this equation, we can calculate the required sample size, n_+, in the biomarker-positive subgroup to achieve the specific global power in our proposed design. The values of $\Delta_{t*}^{(P)}$ and $\Delta_{t*}^{(O)}$ can be theoretically derived from the hazard functions $h_1(t|X)$ and $h_0(t|X)$. Additionally, both $\sigma_l^{(P)}(t^*)$ and $\sigma_l^{(O)}(t^*)$ can be estimated by Monte Carlo method as described in Lu and Tian (2021). Take $\sigma_l^{(O)}(t^*)$ for example, one may first simulate a data set of a large sample size M under H_{a1}, and then calculate the centered RMST estimates,

$$I = \sqrt{M}\left\{\hat{\Delta}_l^{(O)}(t^*) - \Delta_{t*}^{(O)}\right\}.$$

After repeating this process a large number of B times, $\hat{\sigma}_l^{(O)}(t^*)$ can be approximated by $\sqrt{\sum_{b=1}^{B} I_b^2/B}$.

4.6 A Numerical Example

In this chapter, we present a numerical example of re-designing the JAVELIN Lung 200 study using our proposed methodology. Given the unavailability of data from the original trial, we use the simulated data to mimic the settings in JAVELIN Lung 200. We select progression-free survival (PFS) as the endpoint of interest in this RMST-based design, particularly due to notable PH violations observed in the treatment effect of avelumab, as discussed in Chapter 1.1.1. We denote $Z = 1$ to represent the experimental treatment of avelumab and $Z = 0$ for the standard care of docetaxel, the randomization ratio is one to one. The hazard functions for each treatment are assumed as follows:

$$\begin{aligned} h_0(t|X) &= 2.5\log(2), \\ h_1(t|X) &= \begin{cases} 6\log(2) \times \exp\{-0.8X\} & \text{for } t \leqslant 1/6 \\ 2\log(2) \times \exp\{-0.8X\} & \text{for } t > 1/6. \end{cases} \end{aligned} \tag{4.18}$$

Here, time is measured in years. We designate the PD-L1 expression on tumor cells as the biomarker X. Similar to in JAVELIN Lung 200, we focus on the patients with PD-L1-positive tumors, defined by PD-L1 expression $\geqslant 1\%$, in this design. As such, we assume the distribution of X follows $Unif(0.01, 1)$. As docetaxel is not a biomarker-related therapy, the hazard function for $Z = 0$ is supposed to follow an exponential model and is unrelated to the biomarker. The hazard function for avelumab is assumed to be piecewise exponential, and according to the Supplemental Figure 3 in Barlesi et al. (2018), we posit a single change point at 2 months (i.e., $t = 1/6$). Figure 4.2 displays the survival curves derived from the simulated data among patients with PD-L1 expression $\geqslant 1\%$ and $\geqslant 50\%$. These curves closely resemble those depicted in Supplemental Figure 3 of Barlesi et al. (2018), where the PH assumption is clearly violated.

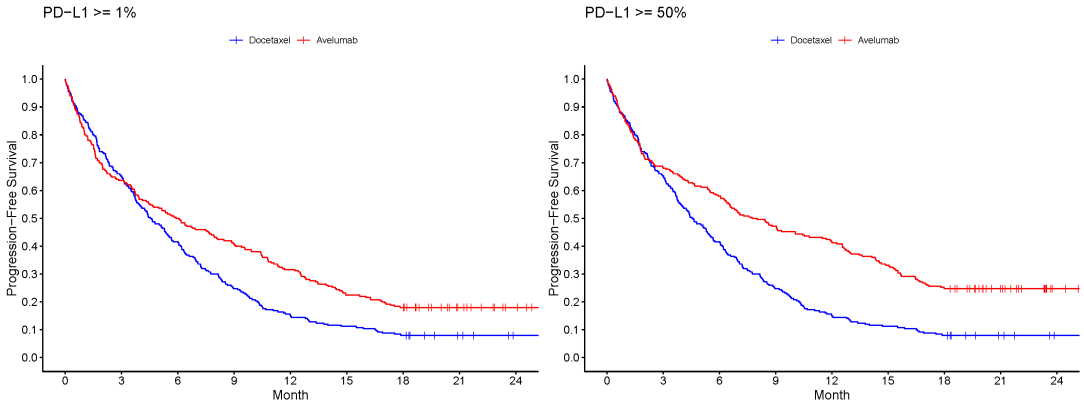

FIGURE 4.2: Simulated RMST curves of progression-free survival in patients with PD-L1 $\geqslant 1\%$ (left panel) and PD-L1 $\geqslant 50\%$ (right panel).

In the two-stage enrichment design, we assume patients are uniformly enrolled from time 0 to $t_1 = 0.5$ year in Stage I. The enrichment occurs at the end of Stage I at $t_1 = 0.5$. Subsequently, in Stage II, patients are uniformly enrolled from t_1 to $t_2 = 1$ year, with restriction in the biomarker-positive subgroup. We assume equivalent accrual rates across both stages. All patients will be followed up until $t_3 = 2.5$ years.

We assume that the annual dropout rate is 5% with a median dropout time of 2 years in both arms following an exponential distribution. As such, the lost to follow-up time L_i follows an exponential distribution with a rate $-\log(1-0.05)/2$.

We use the RMST difference between avelumab and docetaxel up to $t^* = 1.5$ years as the measure of treatment effect. Based on Equations (4.3) and (4.18), the optimal biomarker cutpoint is calculated as $c_{t*} = 29.6\%$. Consequently, the marginal RMST difference among the biomarker-positive subgroup $(c_{t*}, 1]$ is $\Delta_{t*}^{(P)} = 0.137$ years and the marginal RMST difference in the overall patients is $\Delta_{t*}^{(O)} = 0.082$ years.

We use Equation (4.17) for calculating the global power across a range of sample size n under each proposed estimator. Given that the accrual time and rates in two stages are equivalent, the sample size n_+ in the biomarker-positive subgroup in this equation is determined as $\frac{n}{2}(1 + \frac{1-c_{t*}}{0.99})$.

Regarding the CW-adjusted estimators, it is essential to note that during the design for sample size calculation, we use the true biomarker cutpoint for enrichment and determining the biomarker-positive subgroup, i.e., $\hat{c}_0 = \hat{c}_{t*} = c_{t*}$. Consequently, the distribution of the included biomarker-positive patients aligns with the true distribution of $F_+(x)$. In such a situation, utilizing the calibration weighting method to balance distributions becomes unnecessary, and we can set $\hat{p}_i = 1$ in each CW-adjusted estimator. By doing so, the CW Kaplan-Meier estimator degenerates to the naive unadjusted estimator.

We plan to control the family-wise type I error rate at 2.5%, and we allocate a significance level of 2.5% for testing positive biomarker-by-treatment interaction (i.e., $\alpha_0 = 2.5\%$ for testing H_{00}). Next, we use Monte Carlo method to determine the significance level of $\tilde{\alpha}$ for testing the conditional treatment effects (i.e., H_{10} and H_{20}). We conduct a simulation under the global null setting where $h_1(t|X) = h_0(t|X) = 2.5\log(2)$ with

a total sample size $n = 10,000$ across a range of $\tilde{\alpha} \in \{1.5\%, 1.6\%, ..., 2.5\%\}$. Based on the testing results from 10,000 simulated datasets, the family-wise type I error rate can be controlled at 2.5% with $\tilde{\alpha} = 2.3\%$. As an illustration, this simulation utilizes the naive estimator, but other estimators can also be employed. However, the results might be slightly different due to the different variances of each estimator. As a result, the critical values used in Equation (4.17) are $q_0 = \Phi^{-1}(0.975)$ and $q = \Phi^{-1}(0.977)$. Additionally, the asymptotic variance $\hat{\sigma}^2_{\beta_3}$, $[\hat{\sigma_l}^{(P)}(t^*)]^2$, and $[\hat{\sigma_l}^{(O)}(t^*)]^2$ in Equation (4.17) are estimated by Monte Carlo method as introduced in Chapter 4.5.3 with a total sample size $n = 10,000$ and 10,000 replicates.

Figure 4.3 presents the curves of global power across a range of total sample size from the proposed enrichment design, using the Kaplan-Meier estimator ($\hat{\Delta}_2(t^*)$), G-Formula estimator ($\hat{\Delta}_3(t^*)$), Hajek estimator ($\hat{\Delta}_4(t^*)$), and Augmented estimator ($\hat{\Delta}_5(t^*)$). Additionally, we compare the result from an all-comer design with no enrichment using the naive unadjusted estimator. Notably, different estimators employed in the enrichment design yield very similar results. Compared to the all-comer design, the enrichment design requires fewer sample sizes to achieve the same global power. For example, to attain 90% of the global power, the total sample size required in the all-comer design is $n = 940$, while in the enrichment design with the G-Formula estimator, it reduces to $n = 835$. Furthermore, within the enrichment design, the G-Formula estimator demands the smallest sample size to achieve the same global power, while the Hajek estimator requires the largest sample size.

4.7 Simulation Studies

4.7.1 Aims

This simulation study aims to assess the properties of the proposed biomarker cut-point estimating methods and treatment effect estimators within the framework of

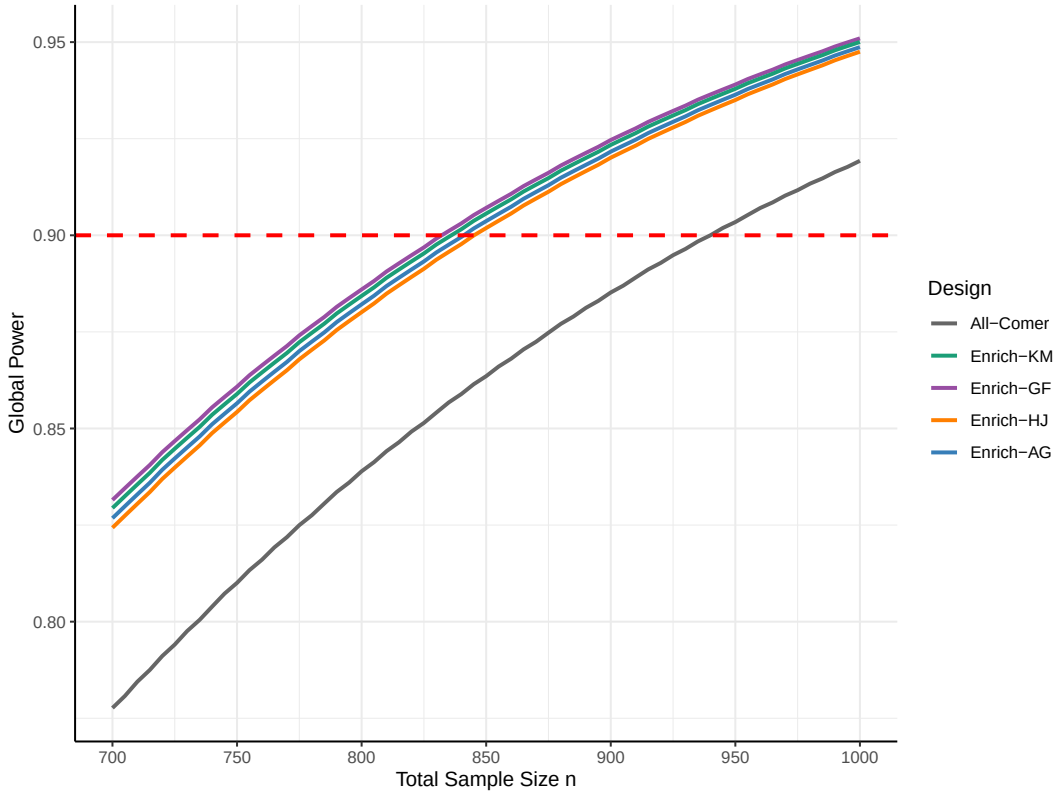

FIGURE 4.3: Curves of the global power by total sample size required in the all-comer design using naive unadjusted estimator and enrichment design using Kaplan-Meier (KM), G-Formula (GF), Hajek (HJ), and Augmented (AG) estimators.

our proposed two-stage enrichment design. Additionally, we seek to evaluate the efficacy of the proposed design. Three primary questions are addressed:

1. Can the proposed methods accurately estimate the biomarker cutpoint in the proposed design?

2. Are the proposed treatment effect estimators consistent with the true marginal RMST difference within the biomarker-positive subgroup in the proposed design?

3. Does the proposed design effectively control the type I error rate while maintaining sufficient power? Additionally, what advantages does this design offer compared to the all-comer design?

4.7.2 Data-Generating Mechanisms

We simulate 10,000 datasets under two designs: 1) the proposed two-stage adaptive enrichment design and 2) the all-comer design with no enrichment.

The two-stage adaptive enrichment design follows the procedure outlined in Chapter 4.2.3. In Stage I, n_1 patients are accrued for each of the experimental and control treatments. The accrual in the first stage is regardless of patient's biomarker value. We assume a uniform enrollment from calendar time 0 to $t_1 = 1$ year. For each patient i, the enrolled time $E_i \sim Unif(0, t_1)$. In Stage II, an additional n_2 patients are accrued per treatment. The accrual in Stage II is determined based on the enrichment strategy established at the end of Stage I. A uniform enrollment is assumed from t_1 to $t_2 = 2$ years. All patients will be followed up until $t_3 = 4$ years.

The all-comer design accrues and randomizes a total of $n = 2 \times (n_1 + n_2)$ patients with one to one allocation regardless of their biomarker values. To follow a similar recruitment scheme as in the enrichment design, we assume that the first $2n_1$ patients are uniformly enrolled from time 0 to t_1 and the other $2n_2$ patients are uniformly enrolled from t_1 to t_2. All patients will be followed up until t_3.

In both designs, the event time T_i for $Z \in \{0, 1\}$ is generated from the following hazard functions:

$$\begin{aligned} h_0(t|X) &= 0.9, \\ h_1(t|X) &= \begin{cases} 0.9 \times \exp\{0.9(1-X)\} & \text{for } t \leqslant 0.25 \\ 0.45 \times \exp\{0.9(1-X)\} & \text{for } t > 0.25, \end{cases} \end{aligned} \tag{4.19}$$

Here, the time is measured in years. These hazard functions represent the scenario where the experimental treatment exhibits a delayed treatment effect after 3 months.

We assume the lost to follow-up time L_i follows an exponential distribution, specifically $L_i \sim \exp(0.12)$. For an analysis conducted at time $t' \in \{t_1, t_3\}$, the censoring time for the i^{th} patient is defined as $C_i = \min(L_i, t' - E_i)$. In Figure 4.4, we present the true RMST curves by the biomarker values and their intersection point between the two treatments.

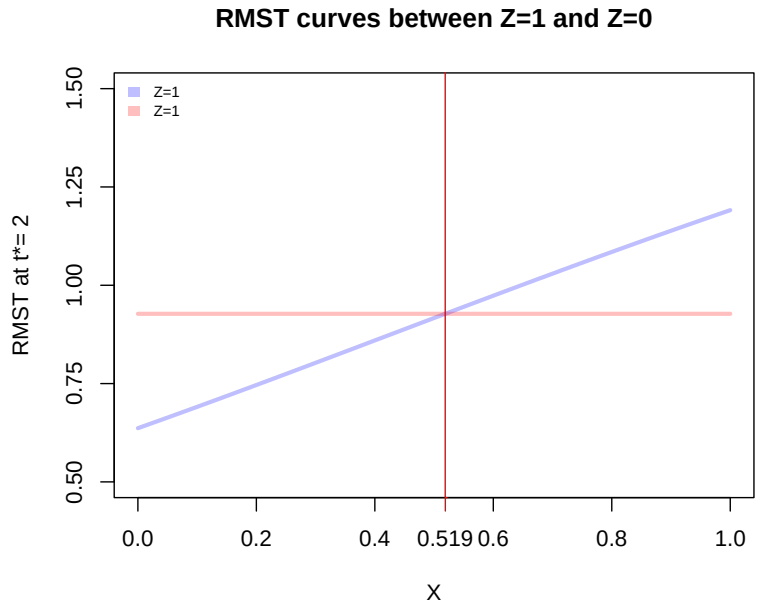

FIGURE 4.4: RMST curves by the biomarker values for $Z = 1$ and $Z = 0$ in simulation study.

We use the RMST difference between experimental and control arms up to $t^* = 2$ years to measure the treatment effect. Based on Equations (4.3) and (4.19), the true biomarker cutpoint is $c_{t*} = 0.519$, and the true marginal RMST differences among the biomarker-positive subgroup $(c_{t*}, 1]$ and overall patients are $\Delta_{t*}^{(P)} = 0.134$ and $\Delta_{t*}^{(O)} = -0.012$, respectively.

We follow the same procedure as in the numerical example in Chapter 4.6 to calculate the sample size. Assuming equal accrual rates in Stages I and II, and the critical

values $q = q_0 = \Phi^{-1}(0.975)$, the total sample size required to achieve a global power of 80% in an all-comer design using the naive unadjusted estimator is $n = 2020$. As such, the sample size per stage and treatment arm is $n_1 = n_2 = 505$.

4.7.3 Methods and Performance Measures

The methods and performance measures to assess operating characteristics are summarized according to the three aforementioned primary questions. We present four design scenarios as follows:

Design Scenario 1: All-comer design without enrichment.

Design Scenario 2: Enrichment design using Prediction Method A.

Design Scenario 3: Enrichment design using Prediction Method B.

Design Scenario 4: Enrichment design using Prediction Method C.

We will introduce the three prediction methods below.

Biomarker Cutpoint Estimation

At the end of Stage I, we consider three different prediction methods to estimate the cutpoint according to the prior information on the piecewise exponential hazard functions:

Prediction Method A: The number of change points in the piecewise exponential models and their locations are known.

Prediction Method B: The number of change points in the piecewise exponential models is known, but the locations are unknown.

Prediction Method C: Both the number of change points and their locations in the piecewise exponential models are unknown. The number of change points is tested under the significance level of 0.05.

We assess the bias of the estimated cutpoint, $\hat{c}_0$, in each prediction method and the accuracy of the detected number and locations of change points in the piecewise exponential models in Prediction Methods B and C.

At the end of Stage II, we use the RMST regression method to re-estimate the biomarker cutpoint. We assess the bias of the estimated cutpoint, $\hat{c}_{t*}$, under all four design scenarios.

Treatment Effect Estimation

We compare five treatment effect estimators, $\hat{\Delta}_l(t^*)$ for $l = 1, \ldots, 5$, for the marginal RMST difference in the biomarker-positive subgroup under four design scenarios. Each estimator is evaluated using both the estimated biomarker cutpoint $\hat{c}_{t*}$ and the true fixed biomarker cutpoint $c_{t*} = 0.519$ to determine the biomarker-positive subgroup.

When estimating the marginal RMST difference using the estimated biomarker cutpoint $\hat{c}_{t*}$, we assess the bias and coverage probabilities of each estimator. Coverage probability is defined as the proportion of replications in which 95% confidence intervals contain the true value out of the total number of replications. When estimating the marginal RMST difference using the true biomarker cutpoint c_{t*}, we evaluate the bias, coverage probability, estimated standard errors ($E[\hat{\sigma}_l(t^*)]$) and the standard deviations ($\sqrt{Var[\hat{\Delta}_l(t^*)]}$) of each estimator. This helps verify the correctness of the proposed variance of each estimator (i.e., $\hat{\sigma}_l^2(t^*)$ for $l = 1, \ldots, 5$).

Operating Characteristics

We compare the operating characteristics in terms of global power, type I error rates and included true negative patients of the two-stage enrichment and all-comer designs under four design scenarios using five treatment effect estimators.

First, in the power analysis, we specify the significance levels for testing positive biomarker-by-treatment interaction and conditional treatment effect as $\alpha_0 = \tilde{\alpha} = 2.5\%$ (see definitions in Chapter 4.5.1), so that the associated critical values are $q = q_0 = \Phi^{-1}(0.975)$. We compare the global power and number of included true negative patients in each design scenario and estimator, where the true negative patients are those whose biomarker values are within the range $[0, c_{t*})$.

Second, the family-wise type I error rate is evaluated under the global null setting, and we also assess the type I error rate for testing positive biomarker-by-treatment interaction. We use the significance levels of $\alpha_0 \in \{1.5\%, 2.0\%, 2.5\%, 3.5\%\}$ for testing the positive biomarker-by-treatment interaction, and $\tilde{\alpha} \in \{2.0\%, 2.3\%, 2.5\%\}$ for testing the conditional treatment effects. As a result, we present a total of 12 combinations of nominal significance levels. Under the global null setting, the Prediction Methods A and B are equivalent since there are no change points in both treatments. As such, we only present results of type I error rate under Design Scenarios 1, 2, and 4.

4.7.4 Results

Biomarker Cutpoint Estimation

Table 4.1 summarizes the predicted biomarker cutpoint, $\hat{c}_0$, in Stage I from the enrichment design using three prediction methods. The Prediction Method A yields the most accurate prediction (bias = 0.001) since the number of change points and

their locations in the piecewise exponential models are known. For the Prediction Method B, with a known number of change points, the average estimated change point from the hazard function for $Z = 1$ is 0.267, which is close to the true (0.25). The estimated biomarker cutpoint from this method is $\hat{c}_0 = 0.506$ (bias = -0.013). For Prediction Method C, where both change points and their locations are unknown, the estimated cutpoint is $\hat{c}_0 = 0.509$ (bias = -0.010). For $Z = 0$, the frequency of correctly identifying the number of change points (i.e., $k = 0$) in its hazard function is 79.3%, while for $Z = 1$, 87.3% of replications can correctly identify the number of change points (i.e., $k = 1$).

Table 4.1: Estimated biomarker cutpoint $\hat{c}_0$ in Stage I in the two-stage enrichment design by using three prediction methods in simulation study. S.D.: standard deviation.

Prediction Method	$\hat{c}_0$	Bias of $\hat{c}_0$	S.D. of $\hat{c}_0$
A	0.520	0.001	0.130
B	0.506	-0.013	0.132
C	0.509	-0.010	0.163

Table 4.2: Estimated biomarker cutpoint $\hat{c}_{t*}$ in Stage II using RMST regression methods under four simulation design scenarios. S.D.: standard deviation.

Design Scenario	$\hat{c}_{t*}$	Bias of $\hat{c}_{t*}$	S.D. of $\hat{c}_{t*}$
1	0.522	0.003	0.062
2	0.519	0.000	0.067
3	0.520	0.001	0.066
4	0.519	0.000	0.067

Table 4.2 presents the estimated biomarker cutpoint, $\hat{c}_{t*}$, in Stage II from the RMST regression method in four design scenarios. The RMST regression method yields

unbiased cutpoint estimates in all design scenarios regardless of the choice of the prediction method in Stage I.

Treatment Effect Estimation

Table 4.3 displays the estimated RMST differences in the biomarker-positive subgroup and their coverage probabilities under four design scenarios using the estimated biomarker cutpoint $\hat{c}_{t*}$. In the all-comer design, all five estimators are unbiased with coverage probabilities over 98%. In the enrichment designs, different prediction methods applied in Stage I do not alter the treatment effect estimations, as Design Scenarios 2-4 give similar results. Overall, the naive unadjusted estimator tends to overestimate the treatment effect in the enrichment designs since the biomarker distribution of the included positive patients may not align with the true distribution of the positive patients (see discussion in Chapter 4.4.1). The four CW-adjusted estimators show unbiased results because the calibration weighting method balances the distribution of the included biomarker-positive patients to the true distribution of $F_+(x)$. The associated coverage probabilities are around 98%.

Table 4.4 presents the estimated RMST differences under four design scenarios using the fixed true biomarker cutpoint. Similar to the findings in Table 4.3, the naive unadjusted estimator is unbiased in the all-comer design but is biased in the enrichment designs. In contrast, the four CW-adjusted estimators show unbiased results under all design scenarios. The estimated standard error of each estimator is close to its standard deviation under all design scenarios, with the associated coverage probabilities approximately 95%. These results confirm the correctness of the variance estimators proposed in Chapter 4.4. The CW G-formula estimator, $\hat{\Delta}_3(t^*)$, demonstrates the smallest variance among the four CW-adjusted estimators.

Table 4.3: Estimated marginal RMST differences in the estimated biomarker-positive subgroup under four simulation design scenarios, using the estimated biomarker cut-point $\hat{c}_{t^*}$. RMSTD: marginal RMST difference in the estimated biomarker-positive subgroup, C.P.: coverage probability. $\hat{\Delta}_1(t^*)$: Naive unadjusted estimator; $\hat{\Delta}_2(t^*)$: CW Kaplan-Meier estimator; $\hat{\Delta}_3(t^*)$: CW G-formula estimator; $\hat{\Delta}_4(t^*)$: CW Hajek estimator; $\hat{\Delta}_5(t^*)$: CW augmented estimator.

Design Scenario	Estimator	Est. RMSTD	Bias of RMSTD	C.P. of RMSTD
1	$\hat{\Delta}_1(t^*)$	0.135	0.001	98.5%
(All-Comer)	$\hat{\Delta}_2(t^*)$	0.134	0.000	98.4%
	$\hat{\Delta}_3(t^*)$	0.134	0.000	99.0%
	$\hat{\Delta}_4(t^*)$	0.134	0.000	98.8%
	$\hat{\Delta}_5(t^*)$	0.135	0.001	98.8%
2	$\hat{\Delta}_1(t^*)$	0.144	0.010	96.5%
(Enrichment)	$\hat{\Delta}_2(t^*)$	0.134	0.000	97.8%
	$\hat{\Delta}_3(t^*)$	0.135	0.001	98.5%
	$\hat{\Delta}_4(t^*)$	0.135	0.001	98.2%
	$\hat{\Delta}_5(t^*)$	0.135	0.001	98.1%
3	$\hat{\Delta}_1(t^*)$	0.143	0.009	96.8%
(Enrichment)	$\hat{\Delta}_2(t^*)$	0.135	0.001	97.8%
	$\hat{\Delta}_3(t^*)$	0.135	0.001	98.5%
	$\hat{\Delta}_4(t^*)$	0.135	0.001	98.1%
	$\hat{\Delta}_5(t^*)$	0.135	0.001	98.1%
4	$\hat{\Delta}_1(t^*)$	0.145	0.010	95.4%
(Enrichment)	$\hat{\Delta}_2(t^*)$	0.135	0.001	97.5%
	$\hat{\Delta}_3(t^*)$	0.135	0.001	98.2%
	$\hat{\Delta}_4(t^*)$	0.135	0.001	97.8%
	$\hat{\Delta}_5(t^*)$	0.135	0.001	97.8%

Table 4.4: Estimated marginal RMST differences in the true biomarker-positive subgroup under four design scenarios, using the true biomarker-cutpoint c_{t*}. RMSTD: marginal RMST difference in the true biomarker-positive subgroup, S.E. of RMSTD: estimated standard error of RMSTD, S.D. of RMSTD: standard deviation of estimated RMSTD's, C.P. of RMSTD: coverage probability of RMSTD.

Design Scenario	Estimator	Est. RMSTD	Bias of RMSTD	S.E. of RMSTD	S.D. of RMSTD	C.P. of RMSTD
1 (All-Comer)	$\hat{\Delta}_1(t^*)$	0.135	0.001	0.048	0.048	94.7%
	$\hat{\Delta}_2(t^*)$	0.134	-0.000	0.050	0.051	94.7%
	$\hat{\Delta}_3(t^*)$	0.134	-0.000	0.044	0.045	94.6%
	$\hat{\Delta}_4(t^*)$	0.134	-0.000	0.052	0.051	95.8%
	$\hat{\Delta}_5(t^*)$	0.134	-0.000	0.052	0.051	95.7%
2 (Enrichment)	$\hat{\Delta}_1(t^*)$	0.144	0.010	0.040	0.040	94.5%
	$\hat{\Delta}_2(t^*)$	0.133	-0.001	0.044	0.046	94.8%
	$\hat{\Delta}_3(t^*)$	0.133	-0.001	0.038	0.039	94.7%
	$\hat{\Delta}_4(t^*)$	0.133	-0.001	0.046	0.046	95.7%
	$\hat{\Delta}_5(t^*)$	0.133	-0.001	0.046	0.046	95.6%
3 (Enrichment)	$\hat{\Delta}_1(t^*)$	0.143	0.009	0.040	0.040	94.6%
	$\hat{\Delta}_2(t^*)$	0.133	-0.001	0.043	0.045	94.7%
	$\hat{\Delta}_3(t^*)$	0.133	-0.001	0.038	0.039	94.7%
	$\hat{\Delta}_4(t^*)$	0.134	-0.000	0.045	0.046	95.5%
	$\hat{\Delta}_5(t^*)$	0.134	-0.000	0.045	0.046	95.4%
4 (Enrichment)	$\hat{\Delta}_1(t^*)$	0.145	0.011	0.041	0.042	93.5%
	$\hat{\Delta}_2(t^*)$	0.134	-0.000	0.045	0.048	94.6%
	$\hat{\Delta}_3(t^*)$	0.134	-0.000	0.038	0.040	94.4%
	$\hat{\Delta}_4(t^*)$	0.134	-0.000	0.047	0.048	95.3%
	$\hat{\Delta}_5(t^*)$	0.134	-0.000	0.047	0.048	95.2%

Operating Characteristics

Table 4.5 summarizes the results from the power analysis. We first compare the average numbers of included negative patients in each design scenario. The all-comer design includes 1048 (51.9%) true negative patients on average. However, the enrichment designs include significantly fewer negative patients. When less prior information on piecewise exponential hazard functions is used in the prediction method

in Stage I, the average number of true negative patients increases from 628 (Design Scenario 2) to 653 (Design Scenario 4).

Table 4.5: Results from power analysis.

Design Scenario	Avg. Number of True Negative Patients	Power of Using $\hat{\Delta}_1(t^*)$	$\hat{\Delta}_2(t^*)$	$\hat{\Delta}_3(t^*)$	$\hat{\Delta}_4(t^*)$	$\hat{\Delta}_5(t^*)$
1	1048	83.7%	80.0%	89.8%	77.4%	77.7%
2	628	95.8%	90.2%	96.7%	88.9%	88.8%
3	640	95.6%	90.6%	96.5%	89.1%	89.1%
4	653	95.1%	88.5%	96.3%	86.8%	86.8%

Subsequently, we compare the global power across five estimators. Within each design scenario, the testing based on the CW G-formula estimator, $\hat{\Delta}_3(t^*)$, has the largest global power due to its smallest variance, as outlined in Web Table 1. Notably, the high global powers from the naive unadjusted estimator, $\hat{\Delta}_1(t^*)$, in the enrichment design stem from its overestimation of the RMST difference (see Table 4.3). The global powers from the other three estimators are similar but lower than those from the CW G-formula estimator due to their larger variance (see Table 4.4). When considering the same estimator, the global power in the all-comer design is smaller than in the enrichment designs. Within the enrichment designs, the global powers are similar between using the first and second prediction methods (Design Scenario 2 and 3), and they are slightly higher than using the third prediction method (Design Scenario 4).

Table 4.6 presents the type I error rate for testing the positive biomarker-by-treatment interaction using four nominal significance levels of α_0 under Design Scenarios 1, 2, and 4. As outlined in Chapter 4.5.1, under the global null, $P(Z_{\beta_3} > q_0) = \alpha_0$, where $q_0 = \Phi^{-1}(1 - \alpha_0)$. Consequently, the type I error rate should align with the nominal

significance level (i.e., α_0). Table 4.6 shows that type I errors rates are effectively controlled in all design scenarios.

Table 4.6: Type I error rate for testing positive biomarker-by-treatment interaction. α_0: significance level for testing Hypothesis Zero.

α_0	Design Scenario 1	Design Scenario 2	Design Scenario 4
0.015	0.016	0.016	0.016
0.020	0.021	0.022	0.021
0.025	0.026	0.027	0.027
0.035	0.037	0.037	0.037

Table 4.7 displays the family-wise type I error rates using 12 combinations of nominal significance levels of $\tilde{\alpha}$ and α_0 for each estimator under Design Scenarios 1, 2, and 4. First, under the same design scenario and treatment effect estimator, the family-wise type I error rate increases with higher α_0 or $\tilde{\alpha}$. Notably, the family-wise type I error rate is more sensitive to changes in $\tilde{\alpha}$ than α_0. Second, under the same significance levels and treatment effect estimator, the enrichment designs have slightly lower family-wise type I error rates than the all-comer design, and different prediction methods have minimal effect on the results. Finally, under the same design scenario and significance levels, the family-wise type I error rates are similar among the first three estimators. The CW Hajek and augmented estimators ($\hat{\Delta}_4(t^*)$ and $\hat{\Delta}_5(t^*)$) provide lower family-wise type I errors than other estimators. However, this is attributed to slightly overestimating the variance in these two estimators under the global null setting (simulation results not shown). According to Table 4.7, if we want to control the family-wise type I error under 0.025 in our proposed design using the CW G-formula estimator ($\hat{\Delta}_3(t^*)$), one option is to set $\alpha_0 = 0.015$ and $\tilde{\alpha} = 0.023$.

Table 4.7: Family-wise type I error rate under each design scenario. α_0: Significance level for testing positive biomarker-by-treatment interaction; $\tilde{\alpha}$: Significance level for testing conditional treatment effect.

Design Scenario	α_0	$\tilde{\alpha}$	$\hat{\Delta}_1(t^*)$	$\hat{\Delta}_2(t^*)$	$\hat{\Delta}_3(t^*)$	$\hat{\Delta}_4(t^*)$	$\hat{\Delta}_5(t^*)$
1	0.015	0.025	0.029	0.029	0.028	0.024	0.024
(All-Comer)		0.023	0.027	0.027	0.026	0.021	0.022
		0.020	0.024	0.024	0.022	0.019	0.019
	0.020	0.025	0.029	0.030	0.028	0.024	0.024
		0.023	0.027	0.028	0.026	0.022	0.022
		0.020	0.024	0.024	0.023	0.019	0.019
	0.025	0.025	0.030	0.031	0.030	0.025	0.025
		0.023	0.028	0.028	0.028	0.022	0.023
		0.020	0.025	0.025	0.024	0.019	0.019
	0.035	0.025	0.031	0.031	0.030	0.025	0.025
		0.023	0.029	0.029	0.028	0.023	0.023
		0.020	0.025	0.025	0.024	0.020	0.020
2	0.015	0.025	0.029	0.029	0.027	0.023	0.023
(Enrichment)		0.023	0.027	0.026	0.025	0.021	0.021
		0.020	0.024	0.023	0.022	0.018	0.018
	0.020	0.025	0.029	0.029	0.028	0.023	0.024
		0.023	0.027	0.027	0.026	0.021	0.022
		0.020	0.024	0.023	0.022	0.019	0.019
	0.025	0.025	0.030	0.030	0.029	0.024	0.024
		0.023	0.028	0.028	0.027	0.022	0.022
		0.020	0.025	0.024	0.023	0.019	0.019
	0.035	0.025	0.031	0.031	0.029	0.025	0.025
		0.023	0.029	0.028	0.027	0.022	0.023
		0.020	0.025	0.025	0.024	0.020	0.020
4	0.015	0.025	0.029	0.028	0.027	0.023	0.023
(Enrichment)		0.023	0.027	0.026	0.025	0.021	0.021
		0.020	0.023	0.023	0.022	0.018	0.018
	0.020	0.025	0.029	0.029	0.028	0.023	0.024
		0.023	0.027	0.027	0.026	0.021	0.022
		0.020	0.024	0.023	0.022	0.019	0.019
	0.025	0.025	0.030	0.030	0.029	0.024	0.024
		0.023	0.028	0.028	0.027	0.022	0.022
		0.020	0.025	0.024	0.023	0.019	0.019
	0.035	0.025	0.031	0.031	0.030	0.025	0.025
		0.023	0.028	0.028	0.028	0.022	0.023
		0.020	0.025	0.025	0.024	0.020	0.020

4.8 Discussion

We presented a two-stage adaptive RMST design incorporating biomarker threshold detection and patient enrichment. Our approach involves two methods to identify the biomarker cutpoint and five estimators of the RMST difference in the biomarker-positive subgroup. Furthermore, we defined the family-wise type I error rate and global power in the context of our design and proposed a method for controlling the family-wise type I error rate under the global null by numerically determining the critical values for test statistics and a complementary method for calculating the sample size. We re-designed the JAVELIN Lung 200 using our design scheme and illustrated the sample size calculation process. Our findings demonstrated that the proposed enrichment design offers substantial reductions in required sample size compared to the all-comer design, while maintaining equivalent global power. Another important strength of the proposed design is its independence from needing a pre-specification of the biomarker-positive threshold, particularly for a continuous biomarker. Therefore, adopting the new design to oncology trials could avoid the potential failures of betting on an inaccurate biomarker cutpoint for defining the biomarker-positive groups at the time of trial design. This is especially crucial when investigators lack sufficient data to determine the optimal biomarker cutpoint. Our flexible design allows adaptively identifying the threshold while simultaneously determining the biomarker-positive subgroup and estimating and testing the treatment effect in the "selected" positive group. Furthermore, this RMST design applies to designs with time-to-event data when non-proportional hazards are expected, such as the delayed effect in the experimental treatment.

Our theoretical work and extensive simulation study demonstrated that the proposed design and related methods addressed three key questions with satisfaction. First, in Stage I, the prediction methods can effectively identify the biomarker cutpoint

for enrichment, and the accuracy increases as more prior information on the piecewise hazard functions becomes available. In Stage II, the RMST regression method provides an unbiased estimation of the biomarker cutpoint to define the biomarker-positive subgroup. Second, the calibration weighting method successfully balances the distribution of the included biomarker-positive patients with the true distribution of positive biomarkers. Consequently, the CW-adjusted estimators are all unbiased in the enrichment designs. The CW G-formula estimator demonstrates the smallest variance with the largest statistical power among all estimators. Third, compared to the all-comer design with an equivalent overall sample size, our proposed enrichment design features significantly more biomarker-positive patients on average and achieves higher global power while retaining a well-controlled family-wise type I error rate.

When outlining the proposed design for time-to-event outcomes, we primarily focused on utilizing the piecewise exponential model for hazard functions. This model is flexible as we allow an arbitrary number of change points, enabling us to approximate a wide range of event distributions. Notably, our proposed RMST regression method for estimating the biomarker cutpoint and the associated estimators for RMST difference does not necessitate the piecewise exponential hazard assumption. However, it is essential to highlight that this assumption becomes necessary when employing the prediction method to identify the biomarker cutpoint. In this method, we adopted the approach outlined in Goodman et al. (2011) to detect the number and location of the change points in the piecewise exponential hazard model. It is important to note that this approach assumes that the change points only affect the parameters of the baseline hazard function in Equation (4.1), that is, γ_z remains constant over time. In situations where γ_z varies over time, alternative methods can be employed, as proposed by He et al. (2013) and Zhang et al. (2014). Additionally, it is crucial to

underscore that our proposed design applies to other types of hazard functions, provided there is sufficient prior information on the parametric assumptions involved.

We focused on a practical scenario in the clinical trial with time-to-event endpoints, where the enrichment is implemented early to reduce the inclusion of biomarker-negative patients. Consequently, the treatment effects are not tested at the end of Stage I as the follow-up time in this stage may not be sufficiently long. However, an additional statistical test can be introduced at the end of Stage I if the follow-up time exceeds the truncation time t^* or if the RMST difference test uses a different truncation time earlier than t^*. Furthermore, an extension of our design could involve a more complicated group-sequential RMST design (Luo et al., 2019; Lu and Tian, 2021). This extension allows for incorporating $k(k \geqslant 1)$ additional interim analysis before the final analysis at the end of the design. At each interim analysis, enrichment can be simultaneously conducted with decisions on early termination for futility or efficacy. While our proposed design primarily considers a single truncation time, it is essential to note that challenges may arise when multiple truncation times for RMSTs are used to define treatment effects. The challenges include merged multiplicity testing due to different treatment effect estimands, and the difficulties when identifying the biomarker cutpoint for enrichment under various truncation times. This topic remains for further research. There are several other avenues for future research to enhance and generalize our proposed design. First, an in-depth exploration of how the design's performance is impacted by the length of follow-up time in both stages or the ratio of enrollment sizes between stages could provide valuable insights. Second, investigating methods for determining optimal or minimax designs would contribute to refining our approach. Third, an extension of our proposed design from a single biomarker to multiple biomarkers with variable

selection and biomarker-based subgroup boundary selection, or even expanding to high-dimensional biomarkers using machine learning algorithms to characterize patient subgroups, presents an opportunity for developing the advanced designs and methods in biomarker-guided clinical trials.

Software

The relevant R code for the methodology and simulation study is available at `https://github.com/kimihua1995/RMST_DESIGN`.

5

Conclusion

In this dissertation, we proposed novel RMST-based methodologies and demonstrated their advantages in three research areas in clinical trials with time-to-event outcomes: 1) individual participant-level network meta-analysis; 2) inference in multi-regional clinical trials; and 3) biomarker-guided adaptive and enrichment design.

In the first topic, we developed two-stage and one-stage RMST-based IPD-NMA methods, enabling the incorporation of individual covariates and allowing the assessment of treatment effect moderations. Through extensive simulations, our proposed methods demonstrated reduced bias in RMST estimates with smaller MSE and proper coverage probabilities than the existing approaches. Utilizing the COMBINE-AF dataset, we used our advanced methods to illustrate the assessment of treatment-by-covariate interactions and the evaluation of between-trial heterogeneity. Future research will focus on improving the approximation method involved in the one-stage model.

In the second topic, we proposed CW and IPSW-adjusted methods to transport or

generalize the treatment effect from each region to the target population in MRCT. Our methods enable the accurate assessment of treatment effect heterogeneity across regions, which is caused by the essential traits, through eliminating imbalances in the inessential traits across regions. We proposed four weighted estimators for the region-specific RMST difference and established their theoretical properties. Simulation studies demonstrated that within the context of MRCT, the CW method was more robust than the IPSW method, and the weighted Augmented estimator displayed smaller variance and higher robustness than other estimators. In the analysis of the PLATO trial, we illustrated that ticagrelor proved more effective than clopidogrel among US patients after achieving balance in aspirin usage and other effect modifiers with the non-US patients. Future research will focus on the methods for variable selection, adjustment for unobserved confounders, and extension to MRCT study designs.

In the third topic, we proposed a two-stage biomarker-guided adaptive enrichment RMST design, allowing for identifying an optimal biomarker cutpoint during an ongoing trial without pre-specifying the biomarker thresholds and simultaneously estimating and testing the RMST difference in the selected biomarker-positive group. Additionally, we developed approaches for calculating family-wise type I error rate, global power, and sample size in this design. Simulation studies demonstrated the effectiveness of our methods in accurately identifying biomarker cutpoints, unbiasedly estimating treatment effects, and achieving higher global power, fewer biomarker-negative enrollments, and a well-controlled type I error rate compared to the all-comer design. Moreover, we re-designed the JAVELIN Lung 200 using our design scheme and illustrated the sample size calculation process. Future research can focus on evaluating the proposed design's performance about follow-up time and enrollment ratio between stages, exploring optimal or minimax designs, and extending from

a single biomarker to multiple biomarkers for advanced methods in variable and cutpoint selection.

Finally, it is essential to recognize that while the RMST offers a more straightforward interpretation with time-to-event outcomes and entails fewer model assumptions than HRs, HRs and Cox models remain widely accepted among researchers or clinicians in biomedical studies. We acknowledge the limitations of RMST (Kloecker et al., 2020), including challenges in selecting the truncation time t^*, handling complex correlations between different truncation times, and difficulties in comparing studies of heterogeneous designs and in inferring the long-term effects of treatments using trials of short duration. However, the development of RMST-based methodologies holds significance, mainly serving as a valid approach in situations of non-proportional hazards or as a meaningful alternative or sensitivity analysis method alongside the HR-based approaches.

Appendix A

Appendix for Chapter 1

Appendix A is organized as follows. Appendix A.1 shows the proof of the large sample properties of $\hat{\boldsymbol{\beta}}$ in the IPCW RMST regression. Appendix A.2 briefly introduces an alternative RMST regression method based on pseudo observations.

A.1 Large Sample Properties of $\hat{\boldsymbol{\beta}}$ in IPCW RMST Regression

The large sample properties of the parameters $\hat{\boldsymbol{\beta}}$ in the IPCW RMST regression can be derived from the M-estimator theory (Stefanski and Boos, 2002). Assume that $\hat{\boldsymbol{\beta}}$ is a unique root from the following weighted quasi-score estimating equation:

$$U_n(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} w_i\{Y_i - g^{-1}(\tilde{X}_i^T\hat{\boldsymbol{\beta}})\}\tilde{X}_i^T = 0.$$

Let $U(\boldsymbol{\beta}) = E[w_i\{Y_i - g^{-1}(\tilde{X}_i^T\hat{\boldsymbol{\beta}})\}\tilde{X}_i^T]$ and $\boldsymbol{\beta}$ be the unique root of $U(\boldsymbol{\beta}) = 0$. It follows that $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ in probability.

Applying a first-order Taylor expansion of $U_n(\hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}$, we have

$$0 = U_n(\hat{\boldsymbol{\beta}}) \approx U_n(\boldsymbol{\beta}) + \left(\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

It follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = -\left(n^{-1}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \hat{\boldsymbol{\beta}}}\right)^{-1} \frac{1}{\sqrt{n}} U_n(\boldsymbol{\beta}).$$

And we have

$$\begin{aligned} -\frac{1}{n}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \hat{\boldsymbol{\beta}}} &= \frac{1}{n}\sum_{i=1}^{n} w_i g'^{-1}(\tilde{X}_i\hat{\boldsymbol{\beta}})\tilde{X}_i\tilde{X}_i^T \\ &\xrightarrow{p} E\left[g'^{-1}(\tilde{X}_i\hat{\boldsymbol{\beta}})\tilde{X}_i\tilde{X}_i^T\right] \equiv A_0 \end{aligned}$$

As such,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = A_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(Y_i - g^{-1}(\tilde{X}_i^T\hat{\boldsymbol{\beta}}))\tilde{X}_i^T.$$

The influence function of $\hat{\boldsymbol{\beta}}$ is $A_0^{-1}w_i\{Y_i - g^{-1}(\tilde{X}_i^T\hat{\boldsymbol{\beta}})\}\tilde{X}_i^T$. Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotic normal with mean (vector) zero, and the covariance matrix $A_0^{-1}B_0A_0^{-1}$, where

$$B_0 = E\left[\{Y_i - g^{-1}(\tilde{X}_i^T\hat{\boldsymbol{\beta}})\}^2\tilde{X}_i^T\tilde{X}_i\right].$$

A.2 Pseudo-Value RMST Regression

Andersen et al. (2004) proposed a RMST regression model using pseudo observations. Begin with the dataset $\{(U_i, \delta_i, \tilde{\mathbf{X}}_i), i = 1, \ldots, n\}$, the pseudo value of RMST for each patient is derived by the jackknife approach as follows:

1. Let the parameter of interest be $\mu(t^*) = E[\min(T, t^*)]$.

2. For $i = 1, \ldots, n$, define the conditional expectation as $\mu_i(t^*) = E(\min(T_i, t^*)|\tilde{\mathbf{X}}_i)$.

3. The i^{th} pseudo observation of RMST is defined as

$$\begin{aligned} \hat{\mu}_i(t^*) &= n \cdot \hat{\mu}(t^*) - (n-1) \cdot \hat{\mu}^{-i}(t^*) \\ &= n \int_0^{t^*} \hat{S}(t)dt - (n-1) \int_0^{t^*} \hat{S}^{-i}(t)dt, \end{aligned}$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$ among all n patients while $\hat{S}^{-i}(t)$ is the "leave-one-out" Kaplan-Meier estimator by excluding the i^{th} patient. Since the Kaplan-Meier estimator is approximately unbiased for the survival function in large samples, both $\hat{\mu}(t^*)$ and $\hat{\mu}^{-i}(t^*)$ are approximately unbiased to $\mu(t^*)$. Thus we have $E(\hat{\mu}_i(t^*)) = \mu(t^*)$ for $i = 1, \ldots, n$. Then the pseudo RMST regression model is written as follows:

$$g(\mu_i(t^*)) = g(E[\hat{\mu}_i(t^*)|\tilde{\mathbf{X}}]) = \tilde{\mathbf{X}}^T\boldsymbol{\gamma}, \tag{A.1}$$

where $g(\cdot)$ is the link function (e.g., log link or identity link). The regression coefficients $\boldsymbol{\gamma}$ can be estimated from a generalized estimation equation (Andersen and Pohar Perme, 2010) as follows:

$$U_n(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \left\{ \frac{\partial g^{-1}(\tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}} V_i^{-1}[\hat{\mu}_i(t^*) - g^{-1}(\tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\gamma}})] \right\} = 0,$$

where V_i^{-1} is the working covariance matrix for $\hat{\mu}_i(t^*)$ (Liang and Zeger, 1986). The derivation of the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is similar to $\hat{\boldsymbol{\beta}}$ from the IPCW RMST regression.

Appendix B

Appendix for Chapter 2

Appendix B is organized as follows. Appendix B.1 briefly shows the algorithm for fitting the one-stage RMST IPD-NMA model (Equation 2.3). Appendix B.2 provides the simulation settings, including the prespecified values of parameters in Equation 2.4 (Table B.1) and the discussion about calculating the true estimands. Figure B.1 shows the survival curves used in the simulation study. Appendix B.3 presents the additional simulation results of bias, mean squared error, and coverage probability of RMSTs for the subgroup with $x = 1$ under three scenarios in Figures B.2 to B.4, respectively. In Appendix B.4, Tables B.3 and B.4 present additional data analysis results with the outcomes of bleeding and all-cause mortality under our proposed two-stage and one-stage methods, respectively.

B.1 PQL Estimation for the One-Stage RMST IPD-NMA Model

Consider the total sample size of $N = \sum_{j=1}^{J} n_j$, Equation (2.3) can be written in a matrix form as follows:

$$\begin{aligned} g(\boldsymbol{\mu}^{(\mathbf{r})}) &= (\mathbf{A}, \mathbf{X}_1, \ldots, \mathbf{X}_P)(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_P^T)^T + \mathbf{Z}_{A1}\mathbf{a}_1' + \cdots + \mathbf{Z}_{AK}\mathbf{a}_K' \\ &+ \sum_{l=1}^{P}[\mathbf{Z}_{X1p}\mathbf{b}_{1p}' + \cdots + \mathbf{Z}_{XKp}\mathbf{b}_{Kp}'] \\ &= \tilde{\mathbf{X}}\boldsymbol{\gamma} + \sum_{k=1}^{K}\mathbf{Z}_{Ak}\mathbf{a}_k' + \sum_{l=1}^{P}\sum_{k=1}^{K}\mathbf{Z}_{Xkp}\mathbf{b}_{kp}' \\ &= \tilde{\mathbf{X}}\boldsymbol{\gamma} + \tilde{\mathbf{Z}}\mathbf{r} \end{aligned} \tag{B.1}$$

In Equation (B.1), $\boldsymbol{\mu}^{(\mathbf{r})} = (\mu_i^{(\mathbf{r})})_{i=1,\ldots,N}$ is a $N \times 1$ vector, where $\mu_i^{(\mathbf{r})} = E(Y_i|\mathbf{r})$ is the conditional expectation of the response Y_i. The conditional variance for Y_i is $\text{Var}(Y_i|\mathbf{r}) = \phi v(\mu_i^{(\mathbf{r})})$. $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ and $\boldsymbol{\beta}_p = (\beta_{1p}, \ldots, \beta_{Kp})^T$ are the parameter vectors for the fixed effect. $\mathbf{A}$ is an $N \times K$ design matrix for the treatments and $\mathbf{X}_p$ is an $N \times K$ design matrix for the treatment-by-the p^{th} covariate, for $p = 1, \ldots, P$. We let $\tilde{\mathbf{X}}$ be partitioned as $(\mathbf{A}, \mathbf{X}_1, \ldots, \mathbf{X}_P)$ and $\boldsymbol{\gamma}$ be partitioned as $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T)^T$. $\mathbf{Z}_{Ak}$ is a $N \times J$ design matrix for the random treatment effects $\mathbf{a}_k' = (a_{1k}, \ldots, a_{Jk})^T$, and $\mathbf{Z}_{Xkp}$ is the $N \times J$ design matrix for the random treatment-by-the p^{th} covariate effects $\mathbf{b}_{kp}' = (b_{1kp}, \ldots, b_{Jkp})^T$. The elements of the matrices $\mathbf{Z}_{Ak}$ and $\mathbf{Z}_{Xkp}$ are 0 or 1. We let $\tilde{\mathbf{Z}}$ be partitioned as $(\mathbf{Z}_{A1}, \ldots, \mathbf{Z}_{AK}, \mathbf{Z}_{X11}, \ldots, \mathbf{Z}_{XK1}, \ldots, \mathbf{Z}_{X1P} \ldots, \mathbf{Z}_{XKP})$ and $\mathbf{r}$ be partitioned as $(\mathbf{a}_1', \ldots, \mathbf{a}_K', \ldots, \mathbf{b}_{11}', \ldots, \mathbf{b}_{K1}', \ldots, \mathbf{b}_{1P}', \ldots, \mathbf{b}_{KP}')^T$. We assume the variance-covariance matrix of the random effects vector $\mathbf{r}$ is $\mathbf{D}$ and we refer to the parameters contained in $\mathbf{D}$ as $\boldsymbol{\tau}$. Note that $\boldsymbol{\tau}$ is a composite of parameters that are involved in $\mathbf{R}_1$, which is defined in Chapter 2.2.2.

In the proposed one-stage model, we use the penalized quasi-likelihood (PQL) method (Breslow and Clayton, 1993) to estimate the parameters $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T)^T$ for

the fixed effect and $\boldsymbol{\tau}$ for the random effects. The objective function for estimating the parameters $(\boldsymbol{\gamma}, \boldsymbol{\tau})$ is an integrated quasi-likelihood given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\tau}) \propto |\mathbf{R}_1|^{-\frac{1}{2}} \int exp\left[-\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i, \mu_i^{(\mathbf{r})}) - \frac{1}{2}\mathbf{r}^T \mathbf{D}^{-1}\mathbf{r} \right] d\mathbf{r}, \tag{B.2}$$

where

$$d_i(y, \mu) = -2w_i \int_y^{\mu} \frac{y-u}{v(u)} du.$$

Since the integral from Equation (B.2) cannot be evaluated as a closed form, the Laplace's method is applied for the integral approximation. As such, maximizing Equation (B.2) is approximated to maximize

$$\text{PQL}(\boldsymbol{\gamma}, \mathbf{r}) = -\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i, \mu_i^{(\mathbf{r})}) - \frac{1}{2}\mathbf{r}^T \mathbf{D}^{-1}\mathbf{r}, \tag{B.3}$$

which is called the *penalized quasi-likelihood (PQL)*. Here, the components of $\mathbf{r}$ serve as the predictors of the random effects, and $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\tau})$ and $\mathbf{r} = \mathbf{r}(\boldsymbol{\tau})$ are functions of $\boldsymbol{\tau}$. For fixing $\boldsymbol{\tau}$, differentiation of Equation (B.3) concerning $\boldsymbol{\gamma}$ and $\mathbf{r}$ leads to the following score functions:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i^{(\mathbf{r})})\tilde{X}_i}{\phi w_i v(\mu_i^{(\mathbf{r})}) g'(\mu_i^{(\mathbf{r})})} = 0, \tag{B.4}$$

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i^{(\mathbf{r})})\tilde{Z}_i}{\phi w_i v(\mu_i^{(\mathbf{r})}) g'(\mu_i^{(\mathbf{r})})} = \mathbf{D}^{-1}\mathbf{r}, \tag{B.5}$$

where $\tilde{X}_i$ and $\tilde{Z}_i$ are denote as the i^{th} row in $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$.

Breslow and Clayton (1993) used the Fisher scoring method to solve Equations (B.4) and (B.5) for fixed $\boldsymbol{\tau}$ as an iterated weighted least squares problem involving a working dependent variable and a weight matrix that are updated at each iteration.

After the maximizers, $\hat{\boldsymbol{\gamma}}(\boldsymbol{\tau})$ and $\hat{\mathbf{r}}(\boldsymbol{\tau})$, are calculated in Equation (B.3) for fixed $\boldsymbol{\tau}$, they employed further approximation to derive the standard REML estimation equation for $\boldsymbol{\tau}$. We briefly summarize the algorithms as follows:

Step 1: Given $\boldsymbol{\tau}$ and $\mathbf{r}$, the fixed effect $\boldsymbol{\gamma}$ is solved from

$$(\tilde{\mathbf{X}}^T\mathbf{V}^{-1}\tilde{\mathbf{X}})\boldsymbol{\gamma} = \tilde{\mathbf{X}}^T\mathbf{V}^{-1}\mathbf{Y}^*, \tag{B.6}$$

where $\mathbf{Y}^*$ is a $N \times 1$ working vector having the component $g(\mu_i^{(\mathbf{r})}) + (y_i - \mu_i^{(\mathbf{r})})g'(\mu_i^{(\mathbf{r})})$, $\mathbf{V} = \mathbf{Q}^{-1} + \tilde{\mathbf{Z}}\mathbf{D}\tilde{\mathbf{Z}}^T$, and $\mathbf{Q}$ is a $N \times N$ diagonal matrix with diagonal terms $\{\phi w_i v(\mu_i^{(\mathbf{r})})[g'(\mu_i^{(\mathbf{r})})]^2\}^{-1}$.

Step 2: For $\hat{\boldsymbol{\gamma}}$ calculated from Equation (B.6), the random effects $\mathbf{r}$ can be estimated by:

$$\hat{\mathbf{r}} = \mathbf{D}\tilde{\mathbf{Z}}^T\mathbf{V}^{-1}(\mathbf{Y}^* - \tilde{\mathbf{X}}\hat{\boldsymbol{\gamma}}). \tag{B.7}$$

Step 3: Let τ_i be the component of $\boldsymbol{\tau}$, τ_i can be obtained from the following estimating equation:

$$-\frac{1}{2}\left[(\mathbf{Y}^* - \tilde{\mathbf{X}}\hat{\boldsymbol{\gamma}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\tau_i}\mathbf{V}^{-1}(\mathbf{Y}^* - \tilde{\mathbf{X}}\hat{\boldsymbol{\gamma}}) - \mathrm{tr}(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\tau_i})\right] = 0, \tag{B.8}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\mathbf{V}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{V}^{-1}$.

Step 4: Update $\mathbf{Y}^*$ at the end of each iteration. The PQL estimators are established once convergence is achieved.

B.2 Simulation Settings

Table B.1: Parameter setup used in Equation (2.4)

	$k=1$	$k=2$	$k=3$
α_k^*	0.5	1.5	1
β_k^*	0.3	0.5	0.7
σ_k	1	1.5	2

We use the Monto-Carlo method to calculate the true values of the estimands, i.e., the log-RMST for each treatment and subgroup. By doing this, we fit an RMST regression model on a simulated dataset with a single study with a size of 10 million samples, sampled from Equation (2.4) with $nt = 1$ and $\tau = 0$. The RMST regression model is written as follows:

$$\log\left(E[T_i \wedge t^* | X_i, Z_i]\right) = \sum_{k=1}^{3} \alpha_k I[Z_i = k] + X_i \sum_{k=1}^{3} \beta_k I[Z_i = k], \tag{B.9}$$

The true values of the parameters in Equation (B.9) are shown in Table B.2.

Table B.2: True value of the parameters in Equation (B.9)

	$k=1$	$k=2$	$k=3$
α_k	0.687	1.070	0.877
β_k	0.171	0.116	0.179

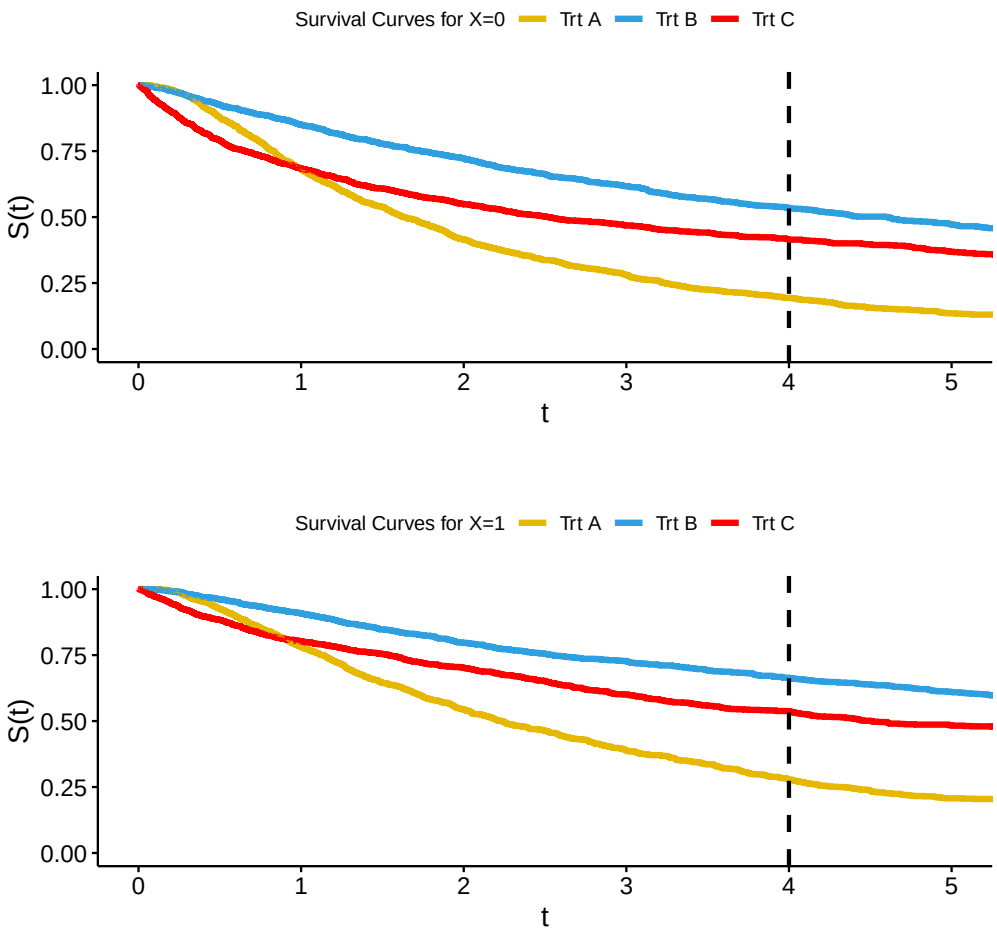

FIGURE B.1: Survival curves for three treatments by subgroups $x = 0$ and $x = 1$, which are considered in the simulation study. These curves are generated from a simulated dataset under Equation (2.4) assuming $nt = 1$ and $\tau = 0$ (i.e., fixed effect model) with a sample size of 10,000.

B.3 Additional Simulation Study Results

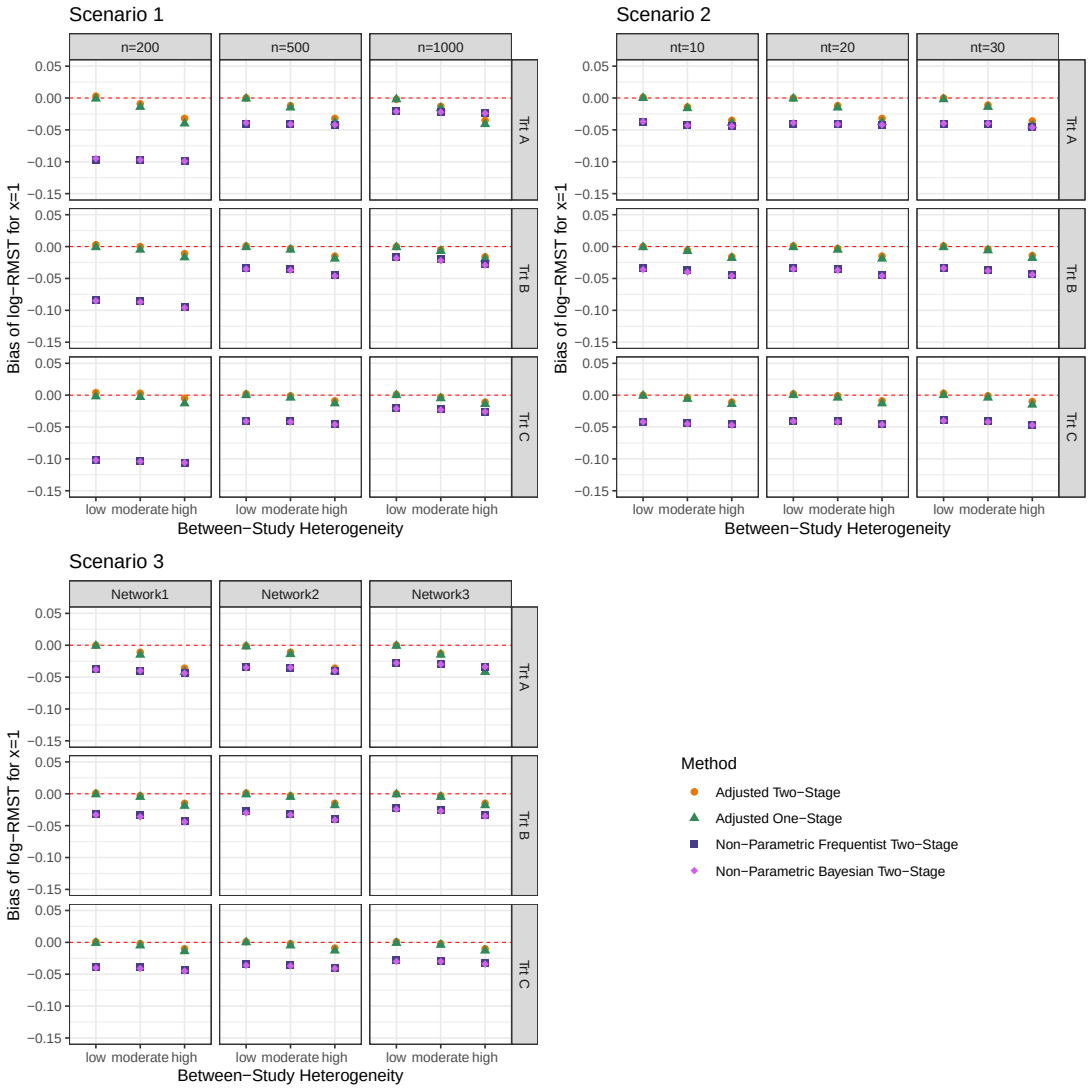

FIGURE B.2: Simulation results of bias of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 1$ under three scenarios.

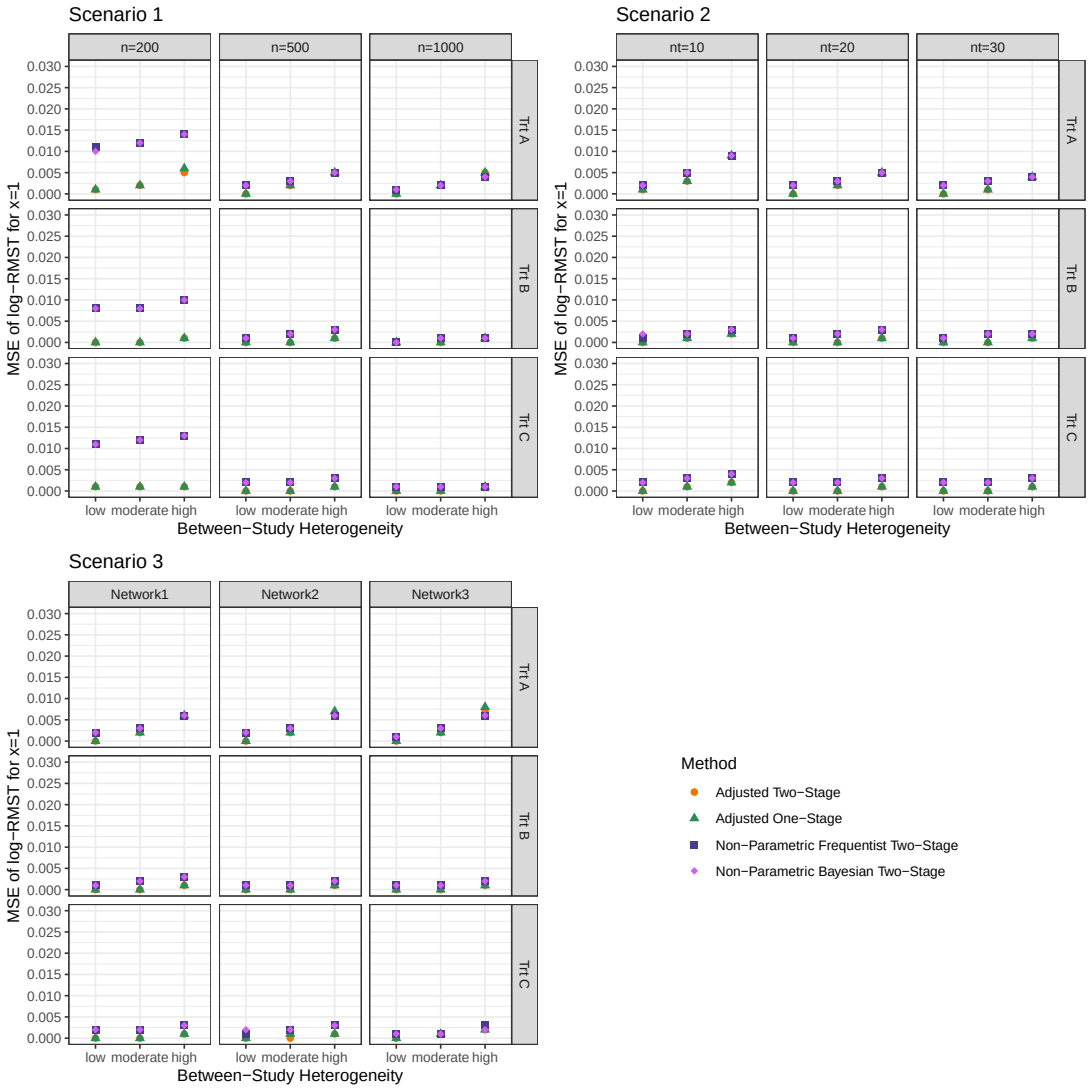

FIGURE B.3: Simulation results of mean squared error (MSE) of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 1$ under three scenarios.

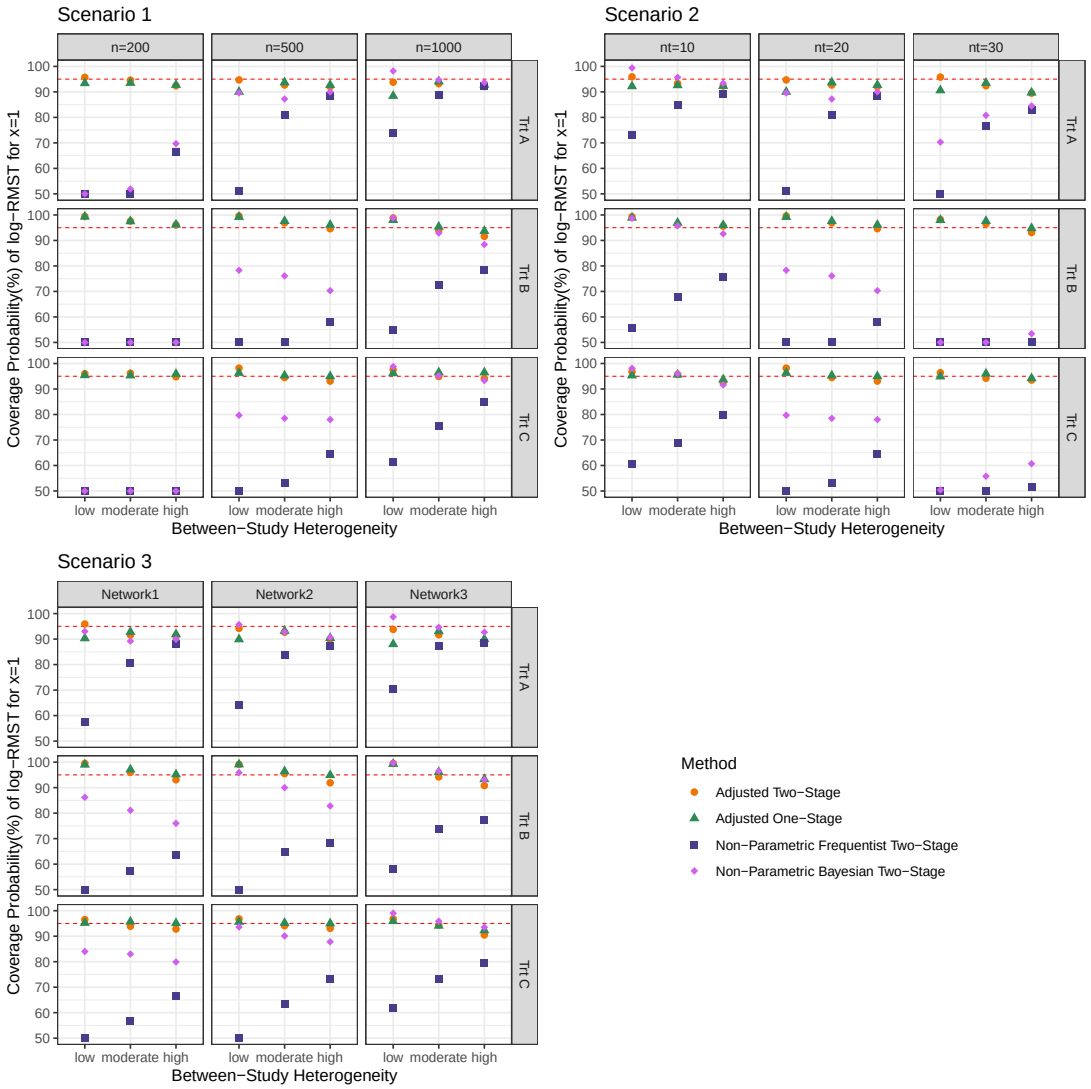

FIGURE B.4: Simulation results of coverage probabilities of log-RMST of three treatments (A, B, and C) for the subgroup with $x = 1$ under three scenarios. The coverage probabilities below 50% are truncated to 50% in the plots.

B.4 Additional Data Analysis Results

Table B.3: Data analysis results for the event of bleeding using the proposed two-stage and one-stage RMST IPD-NMA models with a log link function. The coefficient estimates for log RMSTs of treatments (first three rows in each table) and treatment-by-covariate interactions (denoted as treatment:covariate) are presented along with the 95% confidence intervals (CI), Wald test p-values (P-Val), and the estimates of standard deviations of random effects (τ) as a measure of between-study heterogeneity. W: warfarin; SD: standard-dose DOAC; LD: low-dose DOAC.

Any Bleeding						
	Two-Stage Model			One-Stage Model		
Variable	Coef (95% CI)	P-Val	τ	Coef (95% CI)	P-Val	τ
Warfarin	3.206	<.01	0.061	3.207	<.01	0.089
	(3.145, 3.267)			(3.119, 3.295)		
SD-DOAC	3.221	<.01	0.092	3.220	<.01	0.104
	(3.131, 3.311)			(3.118, 3.322)		
LD-DOAC	3.266	<.01	0.071	3.252	<.01	0.034
	(3.195, 3.337)			(3.207, 3.297)		
W:Sex	-0.041	0.11	0.050	-0.043	0.06	0.043
	(-0.092, 0.010)			(-0.088, 0.002)		
SD:Sex	-0.003	0.92	0.059	-0.002	0.94	0.049
	(-0.064, 0.058)			(-0.053, 0.049)		
LD:Sex	0.010	0.56	0.026	0.006	0.71	0.015
	(-0.023, 0.043)			(-0.025, 0.037)		
W:Age	-0.007	<.01	0.002	-0.007	<.01	0.001
	(-0.009, -0.005)			(-0.009, -0.005)		
SD:Age	-0.007	<.01	0.003	-0.007	<.01	0.002
	(-0.009, -0.005)			(-0.009, -0.005)		
LD:Age	-0.006	<.01	0.002	-0.007	<.01	0.001
	(-0.008, -0.004)			(-0.009, -0.005)		

Table B.4: Data analysis results for all-cause mortality using the proposed two-stage and one-stage RMST IPD-NMA models with a log link function. The coefficient estimates for log RMSTs of treatments (first three rows in each table) and treatment-by-covariate interactions (denoted as treatment:covariate) are presented along with the 95% confidence intervals (CI), Wald test p-values (P-Val), and the estimates of standard deviations of random effects (τ) as a measure of between-study heterogeneity. W: warfarin; SD: standard-dose DOAC; LD: low-dose DOAC.

All-Cause Mortality						
	Two-Stage Model			One-Stage Model		
Variable	Coef (95% CI)	P-Val	τ	Coef (95% CI)	P-Val	τ
Warfarin	3.409	<.01	0.006	3.408	<.01	0.000
	(3.401, 3.417)			(3.404, 3.412)		
SD-DOAC	3.414	<.01	0.005	3.413	<.01	0.003
	(3.408, 3.420)			(3.407, 3.419)		
LD-DOAC	3.416	<.01	0.003	3.417	<.01	0.000
	(3.410, 3.422)			(3.411, 3.423)		
W:Sex	0.010	0.15	0.011	0.011	0.01	0.000
	(-0.004, 0.024)			(0.003, 0.019)		
SD:Sex	0.009	0.02	0.004	0.010	0.01	0.000
	(0.001, 0.017)			(0.002, 0.018)		
LD:Sex	0.013	0.06	0.009	0.009	0.07	0.001
	(-0.001, 0.027)			(-0.001, 0.019)		
W:Age	-0.002	<.01	0.001	-0.002	<.01	0.000
	(-0.002, -0.002)			(-0.002, -0.002)		
SD:Age	-0.002	<.01	0.000	-0.002	<.01	0.000
	(-0.002, -0.002)			(-0.002, -0.002)		
LD:Age	-0.002	<.01	0.000	-0.002	<.01	0.000
	(-0.002, -0.002)			(-0.002, -0.002)		

Appendix C

Appendix for Chapter 3

Appendix C is organized as follows. Appendix C.1 shows the derivation of the calibration weight $\hat{p}_{ri}$ from the convex optimization problem using the Lagrange multiplier. In Appendix C.2 to C.4, we provide the proofs of Theorems 1 to 4 for the proposed weighted estimators of region-specific RMST difference. The definition of the weighted absolute standardized mean difference is presented in Appendix C.5. Appendix C.6 includes the additional simulation set-ups and results. We show the results of the four-region analysis from the PLATO trial in Appendix C.7.

C.1 Derivation of Calibration Weights

The objective function from the convex problem as discussed in Chapter 3.3.2 is

$$L(\boldsymbol{\lambda}) = \sum_{r=1}^{M}\sum_{i=1}^{n_r} p_{ri} log(p_{ri}) - \sum_{r=1}^{M} \boldsymbol{\lambda}_r^T \left\{ \sum_{i=1}^{n_r} p_{ri}\mathbf{g}(\mathbf{X}_{ri}) - \tilde{\mathbf{g}} \right\} - \sum_{r=1}^{M} \lambda_{r0} \left\{ \sum_{i=1}^{n_r} p_{ri} - 1 \right\}.$$

To minimize $L(\boldsymbol{\lambda})$, for $i = 1, \ldots, n_r$, let $\frac{\partial L}{\partial p_{ri}} = 1 + log(p_{ri}) - \boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri}) - \lambda_{r0} = 0$, we get $p_{ri} = \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}/\exp\{1 - \lambda_{r0}\}$. Since $\sum_{i=1}^{n_r} p_{ri} = 1$, then $\exp\{1 - \lambda_{r0}\} =$

$\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}$. Therefore,

$$\hat{p}_{ri} = \frac{\exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}}.$$

Since $\sum_{i=1}^{n_r} p_{ri}\mathbf{g}(\mathbf{X}_{ri}) = \tilde{\mathbf{g}}$, after plugging in the $\hat{p}_{ri}$, we have

$$\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}\{\mathbf{g}(\mathbf{X}_{ri}) - \tilde{\mathbf{g}}\} = 0.$$

C.2 Proofs of Theorem 1

We will first show the proof for Theorem 1 concerning the calibration weights $\hat{p}_{ri}$ (i.e., the CW-adjusted Kaplan-Meier estimator). The proofs for the inverse probability of sampling weights, $\hat{\gamma}_{ri}$, are similar, and we will provide some discussions on the key steps.

C.2.1 Notations

For $i = 1, \ldots, n_r$, we let the calibration weighted counting process and at risk process for Treatment z in Region r be

$$\begin{aligned} \tilde{N}_{rz}(t) &= \sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{rzi} N_{rzi}(t), \\ \tilde{Y}_{rz}(t) &= \sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{rzi} Y_{rzi}(t), \end{aligned}$$

where $N_{rzi}(t) = I[Y_{ri} \leqslant t; \delta_{ri} = 1; Z_{ri} = z]$ and $Y_{rzi}(t) = I[Y_{ri} \geqslant t; Z_{ri} = z]$. We denote $\hat{q}_{rzi} = \frac{I(Z_{ri}=z)}{\hat{\pi}_{ri}(\mathbf{X}_{ri})^z(1-\hat{\pi}_{ri}(\mathbf{X}_{ri}))^{1-z}}$ as the estimated inverse individual treatment-specific propensity score for $q_{rzi}(\mathbf{X}_{ri})$. Suppose the process $M_{rzi}(t) = N_{rzi}(t) -$

$\int_0^t Y_{rzi}(u)d\Lambda_{rz}(u)$, where $\Lambda_{rz}(t) = E_{\mathbf{X}}[-\log\{S(t|\mathbf{X}, Z = z, R = r, d = 1)\}]$ is the cumulative incidence function. $M_{rzi}(t)$ is a martingale with its derivation $dM_{rzi}(t) = dN_{rzi}(t) - Y_{rzi}(u)d\Lambda_{rz}(u)$. We define the entropy weighted process $\tilde{M}_{rz}(t) = \tilde{N}_{rz}(t) - \int_0^t \tilde{Y}_{rz}(u)d\Lambda_{rz}(u)$, with its derivation:

$$\begin{aligned} d\tilde{M}_{rz}(t) &= d\tilde{N}_{rz}(t) - \tilde{Y}_{rz}(t)d\Lambda_{rz}(t) \\ &= \sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{rzi}[dN_{rzi}(t) - Y_{rzi}(t)d\Lambda_{rz}(t)] \\ &= \sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{rzi}dM_{rzi}(t). \end{aligned}$$

For $r = 1, \ldots, M$ and $z = 0, 1$, let $\omega_{rz}(t) = P(Y_{ri} \geqslant t|Z_{ri} = z)$ and $\bar{Y}_{rz}(t) = \sum_{i=1}^{n_r} \hat{q}_{rzi}Y_{rzi}(t)$. We assume that for any fixed $0 < t^* < \infty$, as $n_r \to \infty$, $\sup_{0 \leqslant t \leqslant t^*} |\frac{\bar{Y}_{rz}(t)}{\sum_{i=1}^{n_r} I[Z_{ri}=z]} - \omega_{rz}(t)| \xrightarrow{p} 0$. By $\xrightarrow{p}$ we mean "converges in probability".

C.2.2 Lemma 1

Lemma 1. *Assume the sampling score of RCT participation in each region is proportional to a log-linear model with respect to* $\mathbf{g}(\mathbf{X})$*, that is,* $\rho_r(\mathbf{X}) \propto exp\{\boldsymbol{\eta}_r^T\mathbf{g}(\mathbf{X})\}$*, the estimated calibration weights satisfy* $\hat{p}_{ri} - (N_r\rho_r(\mathbf{X}_i))^{-1} \xrightarrow{p} 0$*, as* $n_r \to \infty$*.*

Proof. Let $d_{ri} = 1$ for RCT participants and $d_{ri} = 0$ for non-participants in region r, then $E[d_r/\rho_r(\mathbf{X})] = 1$. Let $\boldsymbol{\mu}_{g0} = E[\mathbf{g}(\mathbf{X})]$, by using the M-estimator theory, we write the objective function, $L(\boldsymbol{\lambda})$, proposed in Appendix C.1 as the following estimating equations:

$$\frac{1}{N_r}\sum_{i=1}^{N_r} \xi_1(\mathbf{X}_i, d_{ri}; \boldsymbol{\mu}_g) = \frac{1}{N_r}\sum_{i=1}^{N_r} d_{ri}\{\mathbf{g}(\mathbf{X}_i) - \boldsymbol{\mu}_g\} = 0. \quad \text{(C.1)}$$

$$\frac{1}{N_r}\sum_{i=1}^{N_r} \xi_2(\mathbf{X}_i, d_{ri}; \boldsymbol{\mu}_g) = \frac{1}{N_r}\sum_{i=1}^{N_r} d_{ri}\exp\{\boldsymbol{\lambda}_r^T\mathbf{g}(\mathbf{X}_i)\}\{\mathbf{g}(\mathbf{X}_i) - \boldsymbol{\mu}_g\} = 0. \quad \text{(C.2)}$$

Notice that $\boldsymbol{\mu}_{g0}$ is the solution to $E[\xi_1(\mathbf{X}, d_r; \boldsymbol{\mu}_g)] = 0$. Taking expectation on the left hand side of Equation C.2 with $\boldsymbol{\mu}_g = \boldsymbol{\mu}_{g0}$ and under the assumption of $\rho_r(\mathbf{X}) \propto \exp\{\boldsymbol{\eta}_r^T \mathbf{g}(\mathbf{X})\}$ leads to

$$\begin{aligned}
E[\xi_2(\mathbf{X}, d_r; \boldsymbol{\mu}_{g0})] &= E\left[d_r \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X})\}\{\mathbf{g}(\mathbf{X}) - \boldsymbol{\mu}_{g0}\}\right] \\
&= E\left[E[d_r \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X})\}\{\mathbf{g}(\mathbf{X}) - \boldsymbol{\mu}_{g0}\}|\mathbf{X}]\right] \\
&= E\left[\exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X})\}\{\mathbf{g}(\mathbf{X}) - \boldsymbol{\mu}_{g0}\} Pr(A = 1|\mathbf{X}, R = r)\right] \\
&\propto E\left[\exp\{(\boldsymbol{\lambda}_r + \boldsymbol{\eta}_r)^T \mathbf{g}(\mathbf{X})\}\{\mathbf{g}(\mathbf{X}) - E[\mathbf{g}(\mathbf{X})]\}\right].
\end{aligned}$$

For the above expectation to be zero, one needs $\exp\{(\boldsymbol{\lambda}_r + \boldsymbol{\eta}_r)^T \mathbf{g}(\mathbf{X})\}$ to be a constant, that is, $\boldsymbol{\eta}_r = -\boldsymbol{\lambda}_r$. Therefore,

$$\begin{aligned}
\hat{p}_{ri}[N_r \rho_r(\mathbf{X}_{ri})] &= \frac{\exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\}}[N_r \rho_r(\mathbf{X}_{ri})] \\
&= \frac{N_r}{\sum_{i=1}^{N_r} \exp\{\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{X}_{ri})\} d_{ri}} \\
&= N_r \Big/ \sum_{i=1}^{N_r} \frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})} \\
&\xrightarrow{p} 1.
\end{aligned}$$

In Lemma 1, we made a log-linear assumption on the sampling score to show the validity of the CW-adjusted estimators in the following Appendix. Under this assumption, we showed that there is a direct correspondence between the calibration weights $\hat{p}_{ri}$ and the estimated sampling score, that is, $\hat{p}_{ri} - (N_r \rho_r(\mathbf{X}_i))^{-1} \xrightarrow{p} 0$. The model assumption of the sampling score is related to the objective function we use in the optimization problem, i.e., the entropy function. We can also assume that the sampling score follows a logistic regression model with $\sum_{r=1}^{M} \sum_{i=1}^{n_r} (p_{ri} - 1) log(p_{ri} - 1)$ as the objective function (Josey et al., 2021a). And it can be shown that the log-linear model is close to the logistic model when the fraction of RCT participants to the

reference population, i.e., n_r/N_r, is small (Lee et al., 2021). Note that the log-linear sampling score assumption simplifies the proofs of the large sample properties of the CW-adjusted estimators, but calculating the calibration weights does not require the estimation of the unknown sampling scores. To validate its robustness, we showed in the simulation study that the CW-adjusted estimators were unbiased under both the log-linear and the logistic sampling score model. A less strict assumption remains in future research.

C.2.3 Large Sample Properties of CW-Adjusted Survival Function

Before we show the large sample properties of the CW-adjusted KM estimator of the region-specific average RMST, we first show the large sample properties of the CW-adjusted KM estimator of the region-specific average survival function, $\tilde{S}_{rz}$, where

$$\tilde{S}_{rz}(t) = \prod_{u \leqslant t} \left\{ 1 - \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} \right\}$$

Let $S_{rz}(t) = E_{F*}[S_{rz}(t|\mathbf{X})]$ denote the region-specific average survival function, we have the following theorem for $\tilde{S}_{rz}$:

Theorem 1*. *For any fixed* $0 < t^* < \infty$, *as* $n_r \to \infty$,

$$\sup_{0 \leqslant t \leqslant t^*} \left| \tilde{S}_{rz}(t) - S_{rz}(t) \right| \xrightarrow{p} 0,$$

$$\sqrt{n_r} \left\{ \tilde{S}_{rz}(t) - S_{rz}(t) \right\} \xrightarrow{d} N\left(0, S_{rz}^2(t)\sigma_{rz}^2(t)\right),$$

where $\sigma_{rz}^2(t) = n_{rz} \int_0^t \frac{d\Lambda_{rz}(u)}{\tilde{W}_{rz}(u)}$ *and* $\Lambda_{rz}(t) = E_{\mathbf{X}}[-log\{S(t|\mathbf{X}, Z = z, R = r, d = 1)\}]$ *is the cumulative incidence function. Let* $\hat{\sigma}_{rz}^2(t) = \int_0^t \frac{n_r d\tilde{N}_{rz}(u)}{\tilde{W}_{rz}(u)(\tilde{Y}_{rz}(u) - \Delta\tilde{N}_{rz}(u))}$, *as* $n_r \to \infty$,

$$\sup_{0 \leqslant t \leqslant t^*} |\hat{\sigma}_{rz}^2(t) - \sigma_{rz}^2(t)| \xrightarrow{p} 0.$$

Proof of the consistency in Theorem 1*.

If $S_{rz}(t) > 0$, by the formulas for integration by parts (referred by the proof of Theorem 3.2.3 in Fleming and Harrington (1991))

$$\begin{aligned}
\frac{\tilde{S}_{rz}(t)}{S_{rz}(t)} &= 1 - \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \left\{ \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} - d\Lambda_{rz}(u) \right\}. \\
\frac{S_{rz}(t) - \tilde{S}_{rz}(t)}{S_{rz}(t)} &= \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \left\{ \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} - d\Lambda_{rz}(u) \right\} \\
&= \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \left\{ \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} - \frac{\tilde{Y}_{rz}(u)}{\tilde{Y}_{rz}(u)} I[\tilde{Y}_{rz}(u) > 0] d\Lambda_{rz}(u) \right. \\
&- \left. I[\tilde{Y}_{rz}(u) = 0] d\Lambda_{rz}(u) \right\} \\
&= \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d\tilde{M}_{rz}(u) - B_{rz}(t),
\end{aligned}$$

where $B_{rz}(t) = \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} I[\tilde{Y}_{rz}(u) = 0] d\Lambda_{rz}(u)$. Let $\tau = \inf\{u : \tilde{Y}_{rz}(u) = 0\}$ and $\tilde{B}_{rz}(t) = B_{rz}(t) S_{rz}(t)$. Then $\forall u > \tau$, $\tilde{Y}_{rz}(u) = 0$,

$$\begin{aligned}
\tilde{B}_{rz}(t) &= S_{rz}(t) I[\tau < t] \int_\tau^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} d\Lambda_{rz}(u) \\
&= S_{rz}(t) I[\tau < t] \int_\tau^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{-dS_{rz}(u)}{S_{rz}(u-)} \\
&= S_{rz}(t) I[\tau < t] \tilde{S}_{rz}(\tau) \int_\tau^t d\frac{1}{S_{rz}(u)} \\
&= I[\tau < t] \frac{\tilde{S}_{rz}(\tau)\{S_{rz}(\tau) - S_{rz}(t)\}}{S_{rz}(\tau)}.
\end{aligned}$$

As $n_r \to \infty$,

$$\begin{aligned} E[\tilde{B}_{rz}(t)] &\leqslant E\left[I[\tau < t]\{1 - \frac{S_{rz}(t)}{S_{rz}(\tau)}\}\right] \\ &\leqslant E\left[I[\tau < t]\{1 - S_{rz}(t)\}\right] \\ &= \{1 - S_{rz}(t)\}P(\tilde{Y}_{rz}(t) = 0) \\ &= \{1 - S_{rz}(t)\}\{1 - \omega_{rz}(t)\}^{\sum_{i=1}^{n_r} I[Z_{ri}=z]} \\ &\to 0. \end{aligned}$$

This implies that as $n_r \to \infty$, $B_{rz}(t) \xrightarrow{p} 0$. Therefore, for any fixed $0 < t^* < \infty$,

$$P\left(\frac{S_{rz}(t) - \tilde{S}_{rz}(t)}{S_{rz}(t)} = V_{rz}(t), t \in [0, t^*]\right) \to 1,$$

where $V_{rz}(t) = \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u)>0]}{\tilde{Y}_{rz}(u)} d\tilde{M}_{rz}(u)$.

Based on Lemma 1,

$$\begin{aligned} V_{rz}(t) &= \sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d_{ri}\hat{p}_{ri}\hat{q}_{rzi} dM_{rzi}(u) \\ &\xrightarrow{p} \frac{1}{N_r} \sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} \frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})} \hat{q}_{rzi} dM_{rzi}(u) \\ &\xrightarrow{p} \frac{1}{N_r} \sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\bar{Y}_{rz}(u) > 0]}{\omega_{rz}(u)} \hat{q}_{rzi} \frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})} dM_{rzi}(u) \\ &\xrightarrow{p} E\left[\int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\bar{Y}_{rz}(u) > 0]}{\omega_{rz}(u)} dM_{rzi}(u)\right], \end{aligned}$$

where $\bar{Y}_{rz}(u) = \sum_{i=1}^{n_r} \hat{q}_{rzi} Y_{rzi}(u)$. The second "$\xrightarrow{p}$" is derived as the following:

1. Since $\hat{p}_{ri} > 0$, $I[\tilde{Y}_{rz}(u) > 0] = I[\sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{rzi} Y_{rzi}(u) > 0] =$ $I[\sum_{i=1}^{n_r} \hat{q}_{rzi} Y_{rzi}(u) > 0] = I[\bar{Y}_{rz}(u) > 0]$.

2. $\tilde{Y}_{rz}(u) = \sum_{i=1}^{N_r} d_{ri}\hat{p}_{ri}\hat{q}_{rzi}I[Y_{rzi} \geqslant u] \xrightarrow{p} \frac{1}{N_r}\sum_{i=1}^{N_r}\frac{d_{ri}}{\rho_r(X_i)}\hat{q}_{rzi}I[Y_{rzi} \geqslant u] \xrightarrow{p} P(Y_{rzi} \geqslant u) = \omega_{rz}(u)$.

We let

$$\begin{aligned} V^*_{rz}(t) &= \frac{1}{n_r}\sum_{i=1}^{n_r}\int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)}\frac{I[\bar{Y}_{rz}(u)>0]}{\omega_{rz}(u)}\hat{q}_{rzi}dM_{rzi}(u) \\ &\xrightarrow{p} E\left[\int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)}\frac{I[\bar{Y}_{rz}(u)>0]}{\omega_{rz}(u)}dM_{rzi}(u)\right], \end{aligned}$$

then $\sup_{0\leqslant t\leqslant t^*}|V_{rz}(t) - V^*_{rz}(t)| \xrightarrow{p} 0$. Therefore, to show consistency of $\tilde{S}_{tz}(t)$, it is sufficient to show $\sup_{0\leqslant t\leqslant t^*}|V^*_{rz}(t)|^2 \xrightarrow{p} 0$. Since the process $M_{rzi}(u)$ is a martingale and the process $\frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)}\frac{I[\bar{Y}_{rz}(u)>0]}{\omega_{rz}(u)}\hat{q}_{rzi}$ is predictable and bounded, then by Corollary 3.4.1 in Fleming and Harrington (1991) and the assumption of $\sup_{0\leqslant t\leqslant t^*}|\frac{\bar{Y}_{rz}(t)}{\sum_{i=1}^{n_r} I[Z_{ri}=z]} - \omega_{rz}(t)| \xrightarrow{p} 0$, $\forall \epsilon, \eta > 0$,

$$\begin{aligned} & P\{\sup_{0\leqslant t\leqslant t^*}|V^*_{rz}(t)|^2 \geqslant \epsilon\} \\ \leqslant\ & \frac{\eta}{\epsilon} + P\{\sum_{i=1}^{n_r}\int_0^{t^*}\frac{\tilde{S}^2_{rz}(u-)}{S^2_{rz}(u)}\frac{I[\bar{Y}_{rz}(u)>0]}{n_r^2\omega^2_{rz}(u)}\hat{q}^2_{rzi}Y_{rzi}(u)d\Lambda_{rz}(u) \geqslant \eta\} \\ \leqslant\ & \frac{\eta}{\epsilon} + P\{\int_0^{t^*}\frac{\tilde{S}^2_{rz}(u-)}{S^2_{rz}(u)}\frac{I[\bar{Y}_{rz}(u)>0]}{n_r^2\omega^2_{rz}(u)}\bar{Y}_{rz}(u)d\Lambda_{rz}(u) \geqslant \eta\} \\ \leqslant\ & \frac{\eta}{\epsilon} + P\{\frac{\Lambda_{rz}(t^*)\sum_{i=1}^{n_r} I[Z_{ri}=z]}{S^2_{rz}(t^*)n_r^2\omega_{rz}(u)} \geqslant \eta\}. \end{aligned}$$

By letting $\eta \to 0$, the right-hand side of the above converge to 0 as $n_r \to \infty$. Therefore, as $n_r \to \infty$, $\sup_{0\leqslant t\leqslant t^*}|V_{rz}(t)| \xrightarrow{p} 0$, thus $\sup_{0\leqslant t\leqslant t^*}|\tilde{S}_{rz}(t) - S_{rz}(t)| \xrightarrow{p} 0$.

Proof of the asymptotic normality in Theorem 1*.

We first show a corollary. Let $\tilde{\Lambda}_{rz}(t) = \int_0^t \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)}$, for any fixed $0 < t^* < \infty$, as $n_r \to \infty$,

$$\sup_{0 \leqslant t \leqslant t^*} |\tilde{\Lambda}_{rz}(t) - \Lambda_{rz}(t)| \xrightarrow{p} 0.$$

Proof of this corollary. For any $t \in [0, t^*]$,

$$\begin{aligned}
\left|\tilde{\Lambda}_{rz}(t) - \Lambda_{rz}(t)\right| &= \left|\int_0^t \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} - \int_0^t d\Lambda_{rz}(u)\right| \\
&\leqslant \left|\int_0^t \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)} - \int_0^t \frac{\tilde{Y}_{rz}(u)}{\tilde{Y}_{rz}(u)} I[\tilde{Y}_{rz}(u) > 0] d\Lambda_{rz}(u)\right| \\
&+ \left|\int_0^t I[\tilde{Y}_{rz}(u) = 0] d\Lambda_{rz}(u)\right| \\
&\leqslant \left|\int_0^t \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d\tilde{M}_{rz}(u)\right| + I[\tilde{Y}_{rz}(t) = 0]\Lambda_{rz}(t).
\end{aligned}$$

Note that $I[\tilde{Y}_{rz}(t) = 0]\Lambda_{rz}(t) \xrightarrow{p} 0$ as $n_r \to \infty$. We let $Q_{rz}(t) = \int_0^t \frac{I[\tilde{Y}_{rz}(u)>0]}{\tilde{Y}_{rz}(u)} d\tilde{M}_{rz}(u)$.

$$\begin{aligned}
Q_{rz}(t) &= \sum_{i=1}^{N_r} \int_0^t \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d_{ri}\hat{p}_{ri}\hat{q}_{rzi} dM_{rzi}(u) \\
&\xrightarrow{p} \frac{1}{N_r} \sum_{i=1}^{N_r} \int_0^t \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} \frac{d_{ri}}{\rho(\mathbf{X}_i)} d_{ri}\hat{p}_{ri}\hat{q}_{rzi} dM_{rzi}(u) \\
&\xrightarrow{p} E\left[\int_0^t \frac{I[\bar{Y}_{rz}(u) > 0]}{\omega_{rz}(u)} dM_{rzi}(u)\right].
\end{aligned}$$

We let $Q^*_{rz}(t) = \frac{1}{n_r} \sum_{i=1}^{n_r} \int_0^t \frac{I[\bar{Y}_{rz}(u)>0]}{\omega_{rz}(u)} \hat{q}_{rzi} dM_{rzi}(u)$, then $\sup_{0 \leqslant t \leqslant t^*} |Q_{rz}(t) - Q^*_{rz}(t)| \xrightarrow{p} 0$.

By Corollary 3.4.1 in Fleming and Harrington (1991), $\forall \epsilon, \eta > 0$,

$$
\begin{aligned}
P\{\sup_{0\leqslant t\leqslant t^*}|Q^*_{rz}(t)|^2 \geqslant \epsilon\} &\leqslant \frac{\eta}{\epsilon} + P\{\int_0^{t^*} \frac{I[\bar{Y}_{rz}(u)>0]}{n_r^2\omega_{rz}^2(u)}\bar{Y}_{rz}(u)d\Lambda_{rz}(u) \geqslant \eta\} \\
&\leqslant \frac{\eta}{\epsilon} + P\{\frac{\Lambda_{rz}(t^*)\sum_{i=1}^{n_r} I[Z_{ri}=z]}{n_r^2\omega_{rz}(u)} \geqslant \eta\}.
\end{aligned}
$$

By letting $\eta \to 0$, the right-hand side of the above converges to 0 as $n_r \to \infty$. Therefore, as $n_r \to \infty$, $\sup_{0\leqslant t\leqslant t^*}|Q_{rz}(t)| \xrightarrow{p} 0$, thus $\sup_{0\leqslant t\leqslant t^*}|\tilde{\Lambda}_{rz}(t) - \Lambda_{rz}(t)| \xrightarrow{p} 0$.

Now back to the proof of the asymptotic normality, as shown above,

$$
\sqrt{n_r}\left\{S_{rz}(t) - \tilde{S}_{rz}(t)\right\} \xrightarrow{p} S_{rz}(t)\sqrt{n_r}\sum_{i=1}^{N_r}\int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)}\frac{I[\tilde{Y}_{rz}(u)>0]}{\tilde{Y}_{rz}(u)}d_{ri}\hat{p}_{ri}\hat{q}_{rzi}dM_{rzi}(u).
$$

We define $H_{rzi}(t) = \sqrt{n_r}\frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)}\frac{I[\tilde{Y}_{rz}(u)>0]}{\tilde{Y}_{rz}(u)}d_{ri}\hat{p}_{ri}\hat{q}_{rzi}$, $U_{rz}(t) = \sum_{i=1}^{N_r}\int_0^t H_{rzi}(u)dM_{rzi}(u)$ and $U_{rz,\epsilon}(t) = \sum_{i=1}^{N_r}\int_0^t H_{rzi}(u)I[|H_{rzi}(u)| \geqslant \epsilon]dM_{rzi}(u)$. Let $< \cdot,\cdot >$ denote predictable variation process of martingale, then as $n_r \to \infty$,

$$
\begin{aligned}
< U_{rz}, U_{rz} > (t) &= n_r\sum_{i=1}^{N_r}\int_0^t H^2_{rzi}(u)Y_{rzi}(u)d\Lambda_{rz}(u) \\
&= n_r\sum_{i=1}^{n_r}\int_0^t \frac{\tilde{S}^2_{rz}(u-)}{S^2_{rz}(u)}\frac{I[\tilde{Y}_{rz}(u)>0]}{\tilde{Y}^2_{rz}(u)}\hat{p}^2_{ri}[\hat{q}_{rzi}]^2Y_{rzi}(u)d\Lambda_{rz}(u) \\
&\xrightarrow{p} n_r\int_0^t \frac{\sum_{i=1}^{n_r}[\hat{p}_{ri}I[Z_{ri}=z]/\pi^z_{Zi}(1-\pi_{Zi})^{1-z}]^2Y_{rzi}(u)}{\tilde{Y}^2_{rz}(u)}d\Lambda_{rz}(u) \\
&= n_r\int_0^t \frac{d\Lambda_{rz}(u)}{\tilde{W}_{rz}(u)} = \sigma^2_{rz}(t), \\
< U_{rz,\epsilon}, U_{rz,\epsilon} > (t) &= n_r\sum_{i=1}^{N_r}\int_0^t H^2_{rzi}(u)I[|H_{rzi}(u)| \geqslant \epsilon]Y_{rzi}(u)d\Lambda_{rz}(u) \xrightarrow{p} 0.
\end{aligned}
$$

Therefore by the martingale central limit theorem (see Theorem 5.1.1 in Fleming and Harrington (1991)), $\sqrt{n_r}\left\{\tilde{S}_{rz}(t) - S_{rz}(t)\right\} \xrightarrow{d} N\left(0, S^2_{rz}(t)\sigma^2_{rz}(t)\right)$. Let $\tilde{\Lambda}_{rz}(t) =$

$\int_0^t \frac{d\tilde{N}_{rz}(u)}{\tilde{Y}_{rz}(u)}$, as $n_r \to \infty$,

$$\hat{\sigma}^2_{rz}(t) = \int_0^t \frac{n_r d\tilde{N}_{rz}(u)}{\tilde{W}_{rz}(u)(\tilde{Y}_{rz}(u) - \Delta\tilde{N}_{rz}(u))} \xrightarrow{p} n_r \int_0^t \frac{d\tilde{\Lambda}_{rz}(u)}{\tilde{W}_{rz}(u)}.$$

By the corollary showed above, we have $\sup_{0 \leqslant t \leqslant t^*} |\hat{\sigma}^2_{rz}(t) - \sigma^2_{rz}(t)| \xrightarrow{p} 0$.

C.2.4 Proof of Theorem 1

By the consistency of $\tilde{S}_{rz}(t)$ from Theorem 1*, we can obtain $\sup_{0 \leqslant t \leqslant t^*} |\tilde{S}_{rz}(t) - S_{rz}(t)| \xrightarrow{p} 0$. Then for any fixed $0 < t^* < \infty$,

$$\begin{aligned} |\tilde{\mu}_{rz}(t^*) - \mu_{rz}(t^*)| &= \left| \int_0^{t^*} \tilde{S}_{rz}(t)dt - \int_0^{t^*} S_{rz}(t)dt \right| \\ &\leqslant t^* \sup_{0 \leqslant t \leqslant t^*} |\tilde{S}_{rz}(t) - S_{rz}(t)| \\ &\xrightarrow{p} 0. \end{aligned}$$

According to the limiting distribution of $\tilde{S}_{rz}(t)$ derived in Theorem 1*, and using the asymptotic theory for functional of $\tilde{S}_{rz}(t)$ (see Theorem 2.1 and Theorem 3.1 in Gill (1983)), we have

$$\sqrt{n_r}\,\{\tilde{\mu}_{rz}(t^*) - \mu_{rz}(t^*)\} \xrightarrow{d} N(0, \tau^2_{rz}(t^*)),$$

where $\tau^2_{rz}(t^*) = n_r \int_0^{t^*} \{\int_u^{t^*} S_{rz}(t)dt\}^2 \frac{d\Lambda_{rz}(u)}{\tilde{W}_{rz}(u)}$, and by the consistency of $\tilde{\Lambda}_{rz}(t)$, we have

$$|\hat{\tau}^2_{rz}(t^*) - \tau^2_{rz}(t^*)| \xrightarrow{p} 0,$$

where $\hat{\tau}^2_{rz}(t^*) = n_r \int_0^{t^*} \{\int_u^{t^*} \tilde{S}_{rz}(t)dt\}^2 \frac{d\tilde{N}_{rz}(u)}{\tilde{W}_{rz}(u)(\tilde{Y}_{rz}(u) - \Delta\tilde{N}_{rz}(u))}$.

C.2.5 Theorem 1 for IPSW

The large sample properties in Theorem 1, are also valid for the IPSW-adjusted estimators. Most of the proofs are similar as shown for the CW-adjusted estimator, here we provide the distinctions when showing the consistency of $\tilde{S}_{rz}(t)$. Lemma 1 is not required for the proof, but we make another assumption as follows:

Assumption 5

Let d'_{ri} denote the inclusion in the target population, we assume that the marginal probability of being included in the target population is a constant, i.e., $E[P(d'_{ri}|X_{ri})] = C_r$. *In addition, the conditional probability of being included in the target population, $P(d'_{ri}|X_{ri})$, is independent with the observed time Y_{ri}.*

In Appendix C.2.3, we have shown that for any fixed $0 < t^* < \infty$,

$$P\left(\frac{S_{rz}(t) - \tilde{S}_{rz}(t)}{S_{rz}(t)} = V_{rz}(t), t \in [0, t^*]\right) \to 1,$$

where

$$\begin{aligned}
V_{rz}(t) &= \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d\tilde{M}_{rz}(u) \\
&= \sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)} d_{ri}\gamma_{ri}\hat{q}_{rzi} dM_{rzi}(u) \\
&= \frac{1}{N_r}\sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\tilde{Y}_{rz}(u) > 0]}{\tilde{Y}_{rz}(u)/N_r} \frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})} P(d'_{ri} = 1|\mathbf{X}_{ri})\hat{q}_{rzi} dM_{rzi}(u) \\
&\xrightarrow{p} \frac{1}{N_r}\sum_{i=1}^{N_r} \int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\bar{Y}_{rz}(u) > 0]}{\omega_{rz}(u)C_r} \hat{q}_{rzi} \frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})} P(d'_{ri} = 1|\mathbf{X}_{ri}) dM_{rzi}(u) \\
&\xrightarrow{p} E\left[\int_0^t \frac{\tilde{S}_{rz}(u-)}{S_{rz}(u)} \frac{I[\bar{Y}_{rz}(u) > 0]}{\omega_{rz}(u)} dM_{rzi}(u)\right],
\end{aligned}$$

where $\bar{Y}_{rz}(u) = \sum_{i=1}^{n_r} \hat{q}_{rzi} Y_{rzi}(u)$. And by Assumption 5, we have $\tilde{Y}_{rz}(u)/N_r = \frac{1}{N_r}\sum_{i=1}^{N_r} d_{ri}\gamma_{ri}\hat{q}_{rzi} I[Y_{rzi} \geqslant u] = \frac{1}{N_r}\sum_{i=1}^{N_r} \frac{d_{ri}}{\rho_r(X_i)} P(d'_{ri} = 1|\mathbf{X}_{ri})\hat{q}_{rzi} I[Y_{rzi} \geqslant u] \xrightarrow{p}$ $P(Y_{rzi} \geqslant u)C_r = \omega_{rz}(u)C_r$. The remaining of the proofs are similar as for CW-adjusted estimator in Theorem 1.

C.3 Proofs of Theorem 2

We show the consistency of the weighted G-formula estimators by using CW $\hat{p}_{ri}$ and IPSW $\hat{\gamma}_{ri}$. The asymptotic distribution of the CW-adjusted and IPSW-adjusted G-formula estimators can be derived by the M-estimator theory (Stefanski and Boos, 2002) and the Delta method (Dowd et al., 2014). Let the outcome models for $z_r \in \{0, 1\}$ in Region r be $m_{r0}(\mathbf{X}_r) = \hat{\beta}_{r0} + \hat{\boldsymbol{\beta}}_{r2}\mathbf{g}(\mathbf{X}_r^T)$ and $m_{r1}(\mathbf{X}_r) = \hat{\beta}_{r0} + \hat{\beta}_{r1} + \hat{\boldsymbol{\beta}}_{r2}\mathbf{g}(\mathbf{X}_r^T) + \hat{\boldsymbol{\beta}}_{r3}\mathbf{g}(\mathbf{X}_r^T)$. Here, we assume that the outcome models are not misspecified. We re-write $\hat{\Delta}_r^{GF}(t^*)$ as follows:

$$\hat{\Delta}_r^{GF}(t^*) = \hat{\mu}_{r1}^{GF}(t^*) - \hat{\mu}_{r0}^{GF}(t^*),$$

where $\hat{\mu}_{r1}^{GF}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}}$ and $\hat{\mu}_{r0}^{GF}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri} m_{r0}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}}$.

C.3.1 Consistency of CW-adjusted G-formula estimator

$$\begin{aligned}
\hat{\mu}_{r1}^{GF}(t^*) &= \frac{\sum_{i=1}^{n_r} \hat{p}_{ri} m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{p}_{ri}} \\
&= \hat{\beta}_{r0} + \hat{\beta}_{r1} + \hat{\boldsymbol{\beta}}_{r2}\tilde{\mathbf{g}}^T + \hat{\boldsymbol{\beta}}_{r3}\tilde{\mathbf{g}}^T,
\end{aligned}$$

where $\tilde{\mathbf{g}}$ is the sample moment estimates of $\mathbf{g}(\mathbf{X})$ from the target population. The equation on the second line is obtained by the constraint function in Equation (3.5).

Therefore,

$$\begin{aligned} E_{F*}[\hat{\mu}_{r1}^{GF}(t^*)] &\xrightarrow{p} E_{F*}[\beta_{r0} + \beta_{r1} + \boldsymbol{\beta}_{r2}\mathbf{g}(\mathbf{X}^T) + \boldsymbol{\beta}_{r3}\mathbf{g}(\mathbf{X}^T)] \\ &= E_{F*}[m_{r1}(\mathbf{X})] \\ &= \mu_{r1}(t^*). \end{aligned}$$

Here, the expectation is taken with respect on the distribution in the target population. Similarly, we have

$$\begin{aligned} E_{F*}[\hat{\mu}_{r1}^{GF}(t^*)] &\xrightarrow{p} E_{F*}[\beta_{r0} + \boldsymbol{\beta}_{r2}\mathbf{g}(\mathbf{X}^T)] \\ &= E_{F*}[m_{r0}(\mathbf{X})] \\ &= \mu_{r0}(t^*). \end{aligned}$$

Then,

$$E_{F*}[\hat{\Delta}_r^{GF}(t^*)] \xrightarrow{p} \mu_{r1}(t^*) - \mu_{r0}(t^*) = \Delta_r(t^*).$$

C.3.2 Consistency of IPSW-adjusted G-formula estimator

We use the M-estimator theory to show the consistency of the IPSW-adjusted G-formula estimator. Let $\boldsymbol{\theta}_r^{GF} = [\theta_{r0}^{GF}, \theta_{r1}^{GF}]^T$ as the collection of parameters to be estimated. Then $\hat{\Delta}_r^{GF}(t^*) = \hat{\theta}_{r1}^{GF} - \hat{\theta}_{r0}^{GF}$ jointly solves the estimation equations as follows:

$$\sum_{i=1}^{n_r} \boldsymbol{\Phi}_{ri}^{GF}(\boldsymbol{\theta}_r^{GF}) = \sum_{i=1}^{n_r} \begin{pmatrix} \hat{\gamma}_{ri}\{m_{r1}(\mathbf{X}_{ri}) - \theta_{r1}^{GF}\} \\ \hat{\gamma}_{ri}\{m_{r0}(\mathbf{X}_{ri}) - \theta_{r0}^{GF}\} \end{pmatrix} = 0.$$

Taking expectation of $\mathbf{\Phi}_{ri}^{GF}(\boldsymbol{\theta}_r^{GF})$ with respect on the distribution of F_r, we have

$$\begin{aligned}
E_{F_r}\mathbf{\Phi}_{ri}^{GF}(\boldsymbol{\theta}_r^{GF}) &= \begin{pmatrix} E_{F_r}[\hat{\gamma}_{ri}\{m_{r1}(\mathbf{X}_{ri}) - \theta_{r1}^{GF}\}] \\ E_{F_r}[\hat{\gamma}_{ri}\{m_{r0}(\mathbf{X}_{ri}) - \theta_{r0}^{GF}\}] \end{pmatrix} \\
&= \begin{pmatrix} E_{F_r}[\frac{dF^*(\mathbf{X}_{ri})}{dF_r(\mathbf{X}_{ri})}\{m_{r1}(\mathbf{X}_{ri}) - \theta_{r1}^{GF}\}] \\ E_{F_r}[\frac{dF^*(\mathbf{X}_{ri})}{dF_r(\mathbf{X}_{ri})}\{m_{r0}(\mathbf{X}_{ri}) - \theta_{r0}^{GF}\}] \end{pmatrix} \\
&= \begin{pmatrix} E_{F^*}[m_{r1}(\mathbf{X}_{ri}) - \theta_{r1}^{GF}] \\ E_{F^*}[m_{r0}(\mathbf{X}_{ri}) - \theta_{r0}^{GF}] \end{pmatrix} \\
&= \begin{pmatrix} E_{F^*}[m_{r1}(\mathbf{X}_{ri})] - \theta_{r1}^{GF} \\ E_{F^*}[m_{r0}(\mathbf{X}_{ri})] - \theta_{r0}^{GF} \end{pmatrix}.
\end{aligned}$$

By solving $E_{F_r}\mathbf{\Phi}_{ri}^{GF}(\boldsymbol{\theta}_r^{GF}) = 0$, we have $\theta_{r1}^{GF} = E_{F^*}[m_{r1}(\mathbf{X}_{ri})]$ and $\theta_{r0}^{GF} = E_{F^*}[m_{r0}(\mathbf{X}_{ri})]$. As such, $\hat{\Delta}_r^{GF}(t^*)$ is consistent with $\theta_{r1}^{GF} - \theta_{r0}^{GF} = E_{F^*}[m_{r1}(\mathbf{X}_{ri})] - E_{F^*}[m_{r0}(\mathbf{X}_{ri})] = \Delta_r(t^*)$.

C.4 Proofs of Theorems 3 and 4

The large sample properties for the weighted Hajek and weighted Augmented estimators are proofed by the M-estimator theory.

C.4.1 Theorem 3

We re-write the weighted Hajek estimator, $\hat{\Delta}_r^{HJ}(t^*)$, as follows:

$$\hat{\Delta}_r^{HJ}(t^*) = \hat{\mu}_{r1}^{HJ}(t^*) - \hat{\mu}_{r0}^{HJ}(t^*),$$

where $\hat{\mu}_{r1}^{HJ}(t^*) = \frac{\sum_{i=1}^{n_r}\hat{\xi}_{ri}\hat{q}_{r1i}w_{ri}Y_{ri}}{\sum_{i=1}^{n_r}\hat{\xi}_{ri}\hat{q}_{r1i}w_{ri}}$ and $\hat{\mu}_{r0}^{HJ}(t^*) = \frac{\sum_{i=1}^{n_r}\hat{\xi}_{ri}\hat{q}_{r0i}w_{ri}Y_{ri}}{\sum_{i=1}^{n_r}\hat{\xi}_{ri}\hat{q}_{r0i}w_{ri}}$. Let $\boldsymbol{\theta}_r^{HJ} = [\theta_{r0}^{HJ}, \theta_{r1}^{HJ}]^T$ as the collection of parameters to be estimated. Then $\hat{\Delta}_r^{HJ}(t^*) = \hat{\theta}_{r1}^{HJ} - \hat{\theta}_{r0}^{HJ}$ jointly solves the estimation equations as follows:

$$\sum_{i=1}^{n_r}\mathbf{\Phi}_{ri}^{HJ}(\boldsymbol{\theta}_r^{HJ}) = \sum_{i=1}^{n_r}\begin{pmatrix} \hat{\xi}_{ri}\hat{q}_{r1i}w_{ri}\{Y_{ri} - \theta_{r1}^{HJ}\} \\ \hat{\xi}_{ri}\hat{q}_{r0i}w_{ri}\{Y_{ri} - \theta_{r0}^{HJ}\} \end{pmatrix} = 0.$$

First, for the calibration weight (i.e., $\hat{\xi}_{ri} = \hat{p}_{ri}$), by Lemma 1, we have:

$$\begin{aligned}
\sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{r1i}w_{ri}\{Y_{ri} - \theta_{r1}^{HJ}\} &\xrightarrow{p} \frac{1}{N_r}\sum_{i=1}^{N_r}\frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})}w_{ri}\{Y_{ri} - \theta_{r1}^{HJ}\} \\
&\xrightarrow{p} E[\frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})}\frac{\delta_i^*}{\hat{G}(Y_i)}\{Y_{ri} - \theta_{r1}^{HJ}\}] \\
&\xrightarrow{p} E[Y_{ri}] - \theta_{r1}^{HJ}.
\end{aligned}$$

By solving $E[Y_{ri}] - \theta_{r1}^{HJ} = 0$, we have $\hat{\theta}_{r1}^{HJ} = E[Y_{ri}] = \mu_{r1}(t^*)$. So that $\hat{\theta}_{r1}^{HJ}$ is consistent with $\mu_{r1}(t^*)$. Similarly, we can show $\hat{\theta}_{r0}^{HJ}$ is consistent with $\mu_{r0}(t^*)$. Therefore, $\hat{\Delta}_r^{HJ}(t^*)$ is consistent with $\Delta_r(t^*)$ under the calibration weight.

Second, for the inverse probability of sampling weight (i.e., $\hat{\xi_{ri}} = \hat{\gamma}_{ri}$), under the Assumption 5, we have:

$$\begin{aligned}
\sum_{i=1}^{n_r} \hat{\gamma}_{ri}\hat{q}_{r1i}w_{ri}\{Y_{ri} - \theta_{r1}^{HJ}\} &\xrightarrow{p} \frac{1}{N_r}\sum_{i=1}^{N_r}\frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})}P(d'_{ri} = 1|\mathbf{X}_{ri})w_{ri}\{Y_{ri} - \theta_{r1}^{HJ}\} \\
&\xrightarrow{p} E[\frac{d_{ri}}{\rho_r(\mathbf{X}_{ri})}\frac{\delta_i^*}{\hat{G}(Y_i)}C_r\{Y_{ri} - \theta_{r1}^{HJ}\}] \\
&\xrightarrow{p} E[Y_{ri}] - \theta_{r1}^{HJ}.
\end{aligned}$$

Therefore, $\hat{\Delta}_r^{HJ}(t^*)$ is consistent with $\Delta_r(t^*)$ under the inverse probability of sampling weight.

By the M-estimator theory, as $n_r \to \infty$, $\sqrt{n_r}(\hat{\boldsymbol{\theta}}_r^{HJ} - \boldsymbol{\theta}_r^{HJ}) \xrightarrow{d} N(0, n_r\Sigma_r^{HJ})$, where Σ_r^{HJ} is the sandwich variance estimator for $\hat{\boldsymbol{\theta}}_r^{HJ}$ as:

$$\Sigma_r^{HJ} = \left\{\sum_{i=1}^{n_r}\frac{\partial\boldsymbol{\Phi}_{ri}^{HJ}(\boldsymbol{\theta}_r^{HJ})}{(\partial\hat{\boldsymbol{\theta}}_r^{HJ})^T}\right\}^{-1}\left\{\sum_{i=1}^{n_r}\boldsymbol{\Phi}_{ri}^{HJ}(\hat{\boldsymbol{\theta}}_r^{HJ})\boldsymbol{\Phi}_{ri}^{HJ}(\hat{\boldsymbol{\theta}}_r^{HJ})^T\right\}\left\{\sum_{i=1}^{n_r}\frac{\partial\boldsymbol{\Phi}_{ri}^{HJ}(\boldsymbol{\theta}_r^{HJ})}{(\partial\hat{\boldsymbol{\theta}}_r^{HJ})^T}\right\}^{-1}.$$

Then by continuous mapping theorem, $\sqrt{n_r}\left\{\hat{\Delta}_r^{HJ}(t^*) - \Delta_r(t^*)\right\} \xrightarrow{d} N\left(0, \sigma_{r,HJ}^2(t^*)\right)$. The asymptotic variance $\sigma_{r,HJ}^2(t^*)$ can be estimated by $n_r[1,-1]\Sigma_r^{HJ}[1,-1]^T$.

C.4.2 Theorem 4

The proofs of Theorem 4 are similar in Theorem 3 based on the M-estimator theory, and the consistency of $\hat{\Delta}_r^{AG}(t^*)$ is ensured by the consistency of $\hat{\Delta}_r^{GF}(t^*)$ and $\hat{\Delta}_r^{HJ}(t^*)$. Here, we show that the weighted Augmented estimator is also consistent with $\Delta_r(t^*)$ when the outcome models are mis-specified.

We re-write the weighted Augmented estimator, $\hat{\Delta}_r^{AG}(t^*)$, as follows:

$$\hat{\Delta}_r^{AG}(t^*) = \hat{\mu}_{r1}^{AG}(t^*) - \hat{\mu}_{r0}^{AG}(t^*),$$

where $\hat{\mu}_{r1}^{AG}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r1i}w_{ri}\{Y_{ri} - m_{r1}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r1i}w_{ri}} + \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}}$ and

$\hat{\mu}_{r0}^{AG}(t^*) = \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r0i}w_{ri}\{Y_{ri} - m_{r0}(\mathbf{X}_{ri})\}}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r0i}w_{ri}} + \frac{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r0i}m_{r0}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\xi}_{ri}\hat{q}_{r0i}}$.

First, for the calibration weight (i.e., $\hat{\xi_{ri}} = \hat{p}_{ri}$), we have

$$\begin{aligned}
\hat{\mu}_{r1}^{AG}(t^*) &\xrightarrow{p} \frac{\sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{r1i}w_{ri}Y_{ri}}{\sum_{i=1}^{n_r} \hat{p}_{ri}\hat{q}_{r1i}w_{ri}} + \sum_{i=1}^{n_r} \hat{p}_{ri}m_{r1}(\mathbf{X}_{ri}) - \frac{\sum_{i=1}^{n_r} \hat{p}_{ri}w_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{p}_{ri}w_{ri}} \\
&= \hat{\mu}_{r1}^{HJ}(t^*) + \sum_{i=1}^{n_r} \left\{\hat{p}_{ri} - \frac{\hat{p}_{ri}w_{ri}}{\sum_{i=1}^{n_r} \hat{p}_{ri}w_{ri}}\right\} m_{r1}(\mathbf{X}_{ri}) \\
&= \hat{\mu}_{r1}^{HJ}(t^*) + \sum_{i=1}^{n_r} \hat{p}_{ri}\left\{1 - \frac{w_{ri}}{\sum_{i=1}^{n_r} \hat{p}_{ri}w_{ri}}\right\} m_{r1}(\mathbf{X}_{ri}).
\end{aligned}$$

Since $w_{ri}/\sum_{i=1}^{n_r} \hat{p}_{ri}w_{ri} \xrightarrow{p} 1$, then we have $\hat{\mu}_{r1}^{AG}(t^*) \xrightarrow{p} \hat{\mu}_{r1}^{HJ}(t^*)$. Similarly, we can show that $\hat{\mu}_{r0}^{AG}(t^*) \xrightarrow{p} \hat{\mu}_{r0}^{HJ}(t^*)$. Therefore, we see that $\hat{\Delta}_r^{AG}(t^*)$ is consistent with $\Delta_r(t^*)$

regardless of the outcome model specification under the CW.

Second, for the inverse probability of sampling weight (i.e., $\hat{\xi_{ri}} = \hat{\gamma}_{ri}$), we have

$$\hat{\mu}_{r1}^{AG}(t^*) \xrightarrow{p} \hat{\mu}_{r1}^{HJ}(t^*) - \frac{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}\hat{q}_{r1i}w_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}\hat{q}_{r1i}w_{ri}} + \frac{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}}.$$

By M-estimator theory (see proof in Appendix C.3.2), we can show that

$$\begin{aligned}
\frac{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}\hat{q}_{r1i}w_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}\hat{q}_{r1i}w_{ri}} &\xrightarrow{p} E_{F^*}(m_{r1}(\mathbf{X}_{ri})), \\
\frac{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}m_{r1}(\mathbf{X}_{ri})}{\sum_{i=1}^{n_r} \hat{\gamma}_{ri}} &\xrightarrow{p} E_{F^*}(m_{r1}(\mathbf{X}_{ri})).
\end{aligned}$$

As such, $\hat{\mu}_{r1}^{AG}(t^*) \xrightarrow{p} \hat{\mu}_{r1}^{HJ}(t^*)$, and similarly, $\hat{\mu}_{r0}^{AG}(t^*) \xrightarrow{p} \hat{\mu}_{r0}^{HJ}(t^*)$. Therefore, $\hat{\Delta}_r^{AG}(t^*)$ is consistent with $\Delta_r(t^*)$ regardless of the outcome model specification under the IPSW.

C.5 Weighted Standardized Mean Difference

We consider the weighted absolute standardized mean difference (SMD) between variable X between Region r_1 and r_2. We let i index the patient and suppose the sample sizes in two regions are n_1 and n_2. We let X_{1i} and X_{2i} denote the individual variables of X in the two regions. We let ξ_{1i} and ξ_{2i} denote the individual weight in the two regions.

1. If X is a continuous variable, the weighted absolute SMD is

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{s_1^2/2 + s_2^2/2}},$$

where $\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}\xi_{1i}}{\sum_{i=1}^{n_1} \xi_{1i}}$ and $\bar{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}\xi_{2i}}{\sum_{i=1}^{n_2} \xi_{2i}}$ are the weighted sample mean of X in each region. s_1^2 and s_2^2 are the weighted sample variance (Galassi et al.,

2002) of X in each region, where

$$s_1^2 = \frac{\sum_{i=1}^{n_1} \xi_{1i}}{(\sum_{i=1}^{n_1} \xi_{1i})^2 - \sum_{i=1}^{n_1} \xi_{1i}^2} \sum_{i=1}^{n_1} \xi_{1i}(X_{1i} - \bar{X}_1)^2,$$

and s_2^2 has the similar definition.

2. If X is a binary variable, the weighted absolute SMD is

$$d = \frac{|p_1 - p_2|}{\sqrt{p_1(1-p_1)/2 + p_2(1-p_2)/2}},$$

where $p_1 = \frac{\sum_{i=1}^{n_1} X_{1i}\xi_{1i}}{\sum_{i=1}^{n_1} \xi_{1i}}$ and $p_2 = \frac{\sum_{i=1}^{n_2} X_{2i}\xi_{2i}}{\sum_{i=1}^{n_2} \xi_{2i}}$ denote the weighted prevalence of the dichotomous variable X in each region.

Note that when the weight $\xi_i = 1$ for all individuals, the weighted absolute SMD degenerates to the standard absolute SMD.

C.6 Additional Simulation Set-Ups and Results

True Average Region-Specific RMST and RMST Difference

The average region-specific RMST at t^*, $\mu_{rz}(t^*)$, is averaged respect to the distribution of $\mathbf{X}$ from the target population represented by a common distribution F^* : $X_1 \sim Unif(0,1)$ and $X_2 \sim N(1,1)$. Its true value $\bar{\mu}_{rz}(t^*)$ is

$$\begin{aligned}
\bar{\mu}_{rz}(t^*) &= \int \mu_{rz}(t^*|\mathbf{x}) dF^*(\mathbf{x}) \\
&= \int \int_0^{t^*} S_{rz}(t|\mathbf{x}) dF^*(\mathbf{x}) \\
&= \int \int_0^{t^*} exp\{-\int_0^t h(u|z,r,\mathbf{x})\} dF^*(\mathbf{x}).
\end{aligned}$$

where $h(u|z,r,\mathbf{x}) = h(u|Z=z, R=r, X_1=x_1, X_2=x_2)$ is the hazard function of Equation (3.16). The true average region-specific RMST difference (i.e., treatment

effect) is $\bar{\Delta}_r(t^*) = \bar{\mu}_{r1}(t^*) - \bar{\mu}_{r0}(t^*)$.

Table C.1: Values of parameters in the sampling score models, Equations (3.14) and (3.15) under four sampling scenarios. Scenario 1: Log-linear sampling (Equation (3.14)) with moderate SMDs; Scenario 2: Log-linear sampling with large SMDs; Scenario 3: Logistic-nonlinear sampling (Equation (3.15)) with moderate SMDs; Scenario 4: Logistic-nonlinear sampling with large SMDs.

Equation (3.14)	Scenario 1			Scenario 2		
Region	η_{r0}	η_{r1}	η_{r2}	η_{r0}	η_{r1}	η_{r2}
$r = 1$	-5	0.8	0.30	-5	2.5	0.50
$r = 2$	-5	0.7	0.27	-5	2.3	0.55
$r = 3$	-5	0.6	0.25	-5	2.0	0.60
Equation (3.15)	Scenario 3			Scenario 4		
Region	η^*_{r0}	η^*_{r1}	η^*_{r2}	η^*_{r0}	η^*_{r1}	η^*_{r2}
$r = 1$	-3	0.6	-0.15	-2.3	3.0	-0.20
$r = 2$	-3	0.5	-0.10	-2.3	2.5	-0.15
$r = 3$	-3	0.4	-0.05	-2.3	2.0	-0.10

Table C.2: Absolute standardized mean differences of X_1 and X_2 between the enrolled patients in each region and the target population in simulation study. $X_1 \sim Unif(0, 1)$ and $X_2 \sim N(1, 1)$ in the target population. We generate a large data with sample size of 90,000, 30,000 individuals in each region, under four scenarios of the sampling models.

	Region	Scenario 1	Scenario 2	Scenario 3	Scenario 4
X_1	$r = 1$	0.235	0.692	0.208	0.517
	$r = 2$	0.207	0.656	0.164	0.497
	$r = 3$	0.161	0.564	0.126	0.481
X_2	$r = 1$	0.305	0.502	0.297	0.696
	$r = 2$	0.276	0.545	0.245	0.683
	$r = 3$	0.238	0.609	0.199	0.647

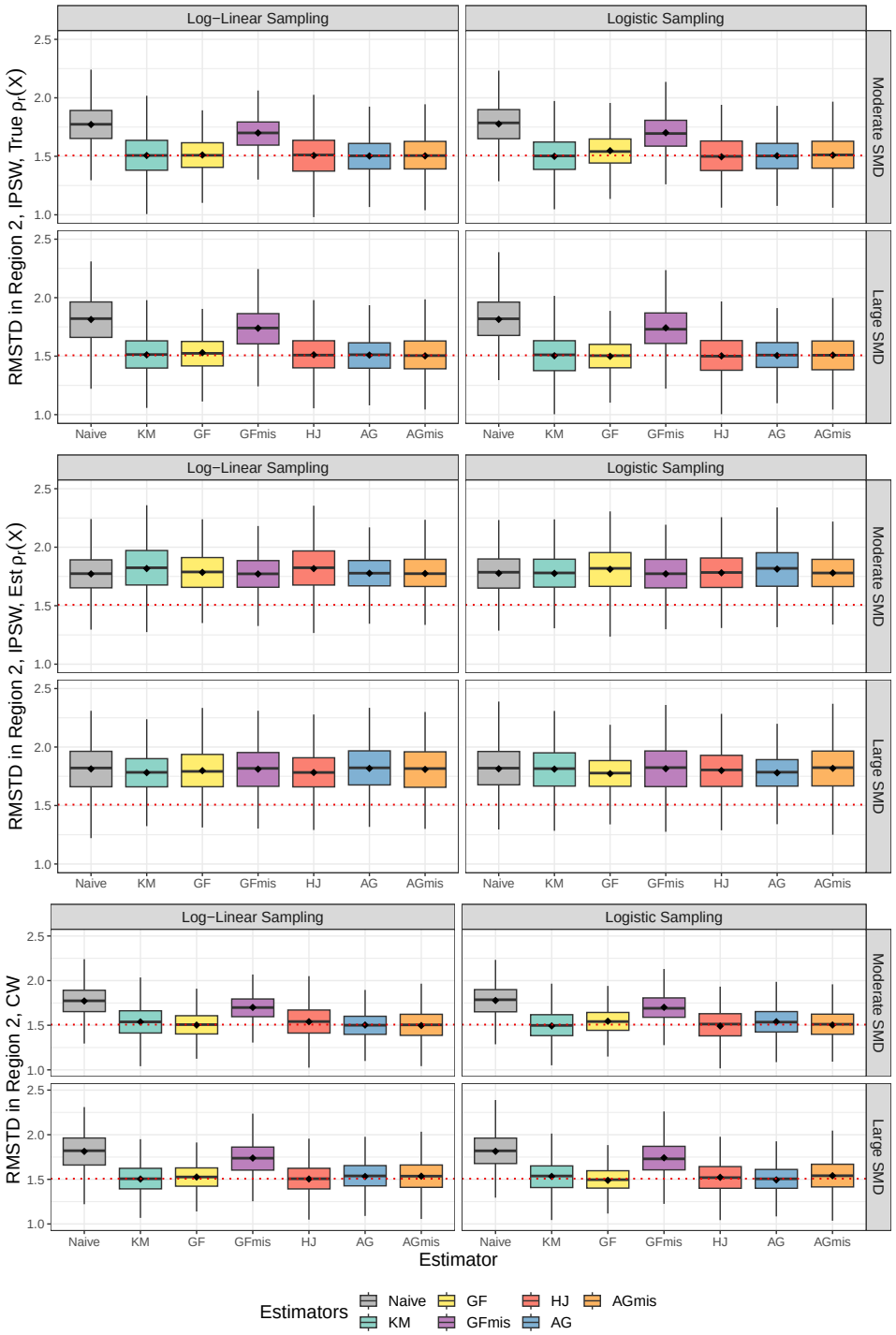

FIGURE C.1: Boxplots of estimated average RMST difference (RMSTD) in Region 2 under four sampling scenarios in simulation study. Upper panel: IPSW-adjusted estimators with true sampling score; Middle panel: IPSW-adjusted estimators with estimated sampling score; Bottom panel: CW-adjusted estimators.

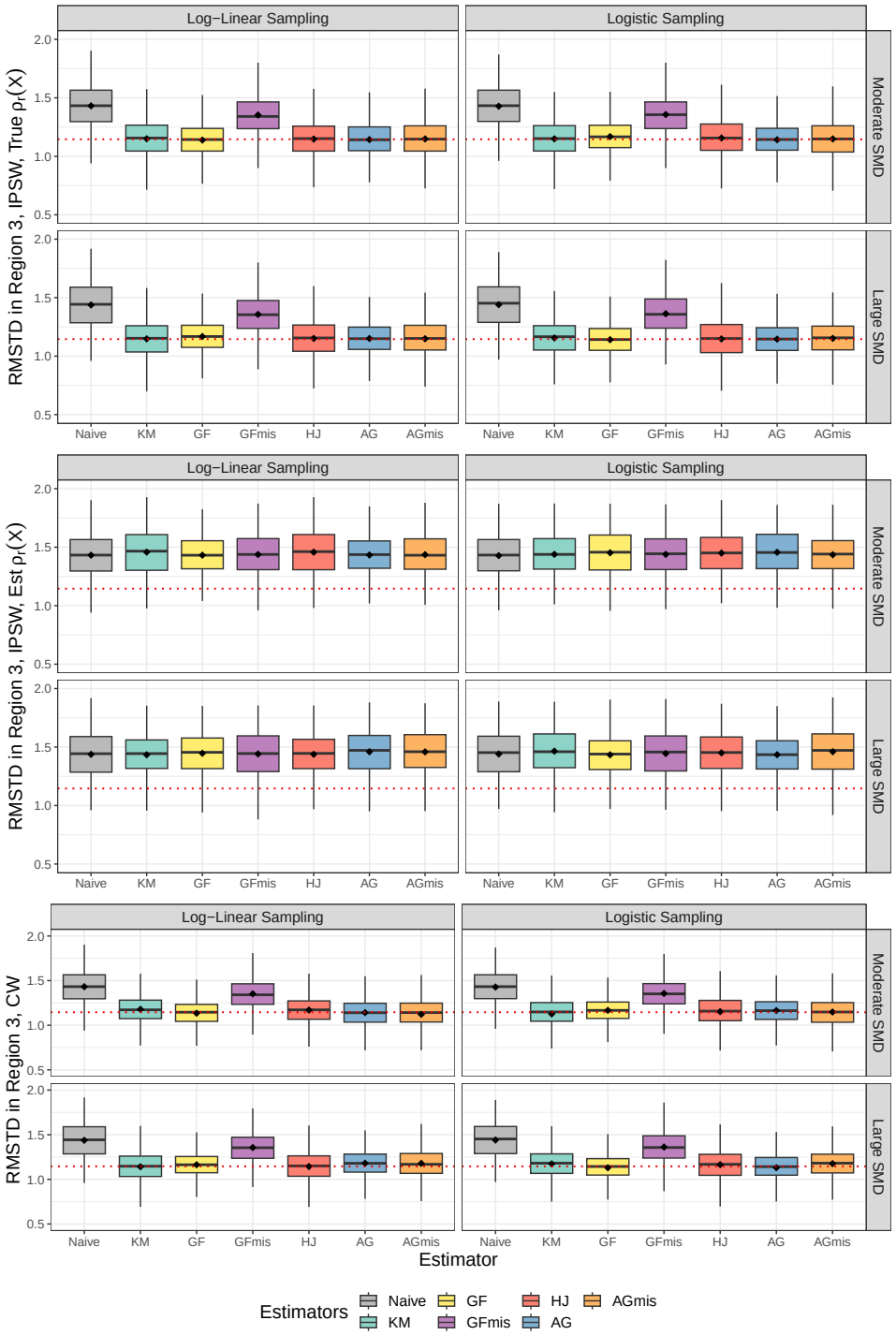

FIGURE C.2: Boxplots of estimated average RMST difference (RMSTD) in Region 3 under four sampling scenarios in simulation study. Upper panel: IPSW-adjusted estimators with true sampling score; Middle panel: IPSW-adjusted estimators with estimated sampling score; Bottom panel: CW-adjusted estimators.

C.7 Four-Region Analysis Results from PLATO Trial

Figure C.3 and Table C.3 present the estimated average RMST differences and the associated 95% CIs comparing ticagrelor and clopidogrel for the primary outcome among four regions. The unadjusted RMST differences are 6.6 days (95% CI: -3.1, 16.3) in Asia and Australia, 6.2 days (95% CI: -6.7, 19) in Central and South America, 5.5 days (95% CI: 2.6, 8.4) in Europe, Middle East, and Africa, and -4.1 days (95% CI: -12.5, 4.4) in North America. The unadjusted results indicate that ticagrelor is numerically, though not statistically significantly, more effective than clopidogrel in Asia and Australia, and Central and South America, and is significantly more effective in Europe, the Middle East, and Africa. However, it is numerically less effective than clopidogrel in North America. The consistency test for the Naive estimator shows that treatment effects are not heterogeneous across four regions ($p = 0.20$).

In Europe, Middle East, and Africa, the dominant region, the estimated RMST differences are very similar across four estimators within the same weighting method. The CW-adjusted estimators yield slightly higher RMST differences than the IPSW-adjusted estimators. Both weighting methods indicate that ticagrelor is significantly more effective than clopidogrel in this region. Under both weighting approaches, the weighted G-formula estimators have the largest variance. In North America, the average RMST differences from all weighted estimators are close to 0, indicating that ticagrelor has the same effect as clopidogrel in North America after balancing the distributions of the selected variables across four regions. However, the IPSW and CW methods yield opposite results in the other two regions. For example, the average RMST difference in Central and South America from all IPSW-adjusted estimators is positive, while it is negative from all CW-adjusted estimators. The results from Asia and Australia show a similar pattern. This discrepancy is potentially due to

misspecification of the estimated sampling score in the IPSW-adjusted estimators.

The consistency tests for most CW-adjusted and IPSW-adjusted estimators reveal no regional treatment effect heterogeneity (see p-values in Table C.3). However, the test using the CW-adjusted Augmented estimators demonstrates that the treatment effects are significantly heterogeneous across regions (p = 0.05). The global RMST differences from other estimators are summarized in Table C.3.

Table C.3: Estimated average RMST differences with 95% CIs at $t^* = 360$ days from 1) A&A: Asia and Australia, 2) C&SA: Central and South America, 3) U,ME&A: Europe, Middle East and Africa, and 4) NA: North America in the four-region analysis, and the results from the regional consistency test with p-values and global RMST difference.

Method	RMSTD (95% CI) in				P-Val	Global
	A&A	C&SA	U,ME&A	NA		RMSTD
Naive	6.6	6.2	5.5	-4.1	0.20	4.7
	(-3.1, 16.3)	(-6.7, 19)	(2.6, 8.4)	(-12.5, 4.4)		(2.0, 7.3)
IPSW.KM	6.6	6.1	5.2	0.3	0.83	5.0
	(-6.9, 20.1)	(-6, 18.2)	(2.4, 7.9)	(-10.1, 10.7)		(2.4, 7.5)
IPSW.GF	-2.9	2.7	5.8	1.6	0.76	5.2
	(-21.7, 16)	(-14.8, 20.2)	(2.2, 9.5)	(-12.5, 15.6)		(1.8, 8.6)
IPSW.HJ	5.9	4.5	5.2	-0.2	0.82	4.9
	(-6.1, 17.8)	(-7.2, 16.1)	(2.4, 8.1)	(-11.3, 10.8)		(2.3, 7.5)
IPSW.AG	2.0	2.6	5.8	0.7	0.74	5.2
	(-11.1, 15.1)	(-9.1, 14.2)	(3.0, 8.7)	(-10.4, 11.8)		(2.6, 7.8)
CW.KM	-5.0	-0.9	4.9	1.2	0.43	3.9
	(-18.9, 9.0)	(-13.6, 11.9)	(1.8, 7.9)	(-8.7, 11.0)		(1.1, 6.7)
CW.GF	-11.0	-3.0	6.6	1.6	0.24	5.1
	(-30.3, 8.2)	(-21.5, 15.5)	(2.4, 10.8)	(-12.0, 15.3)		(1.3, 8.9)
CW.HJ	-6.1	-1.3	5.3	0.8	0.33	4.0
	(-20.3, 8.0)	(-14.6, 12.1)	(1.8, 8.7)	(-9.5, 11.1)		(0.9, 7.1)
CW.AG	-10.8	-2.3	6.5	2.4	0.05	4.7
	(-23.9, 2.3)	(-15.9, 11.2)	(3.1, 10.0)	(-7.9, 12.7)		(1.6, 7.8)

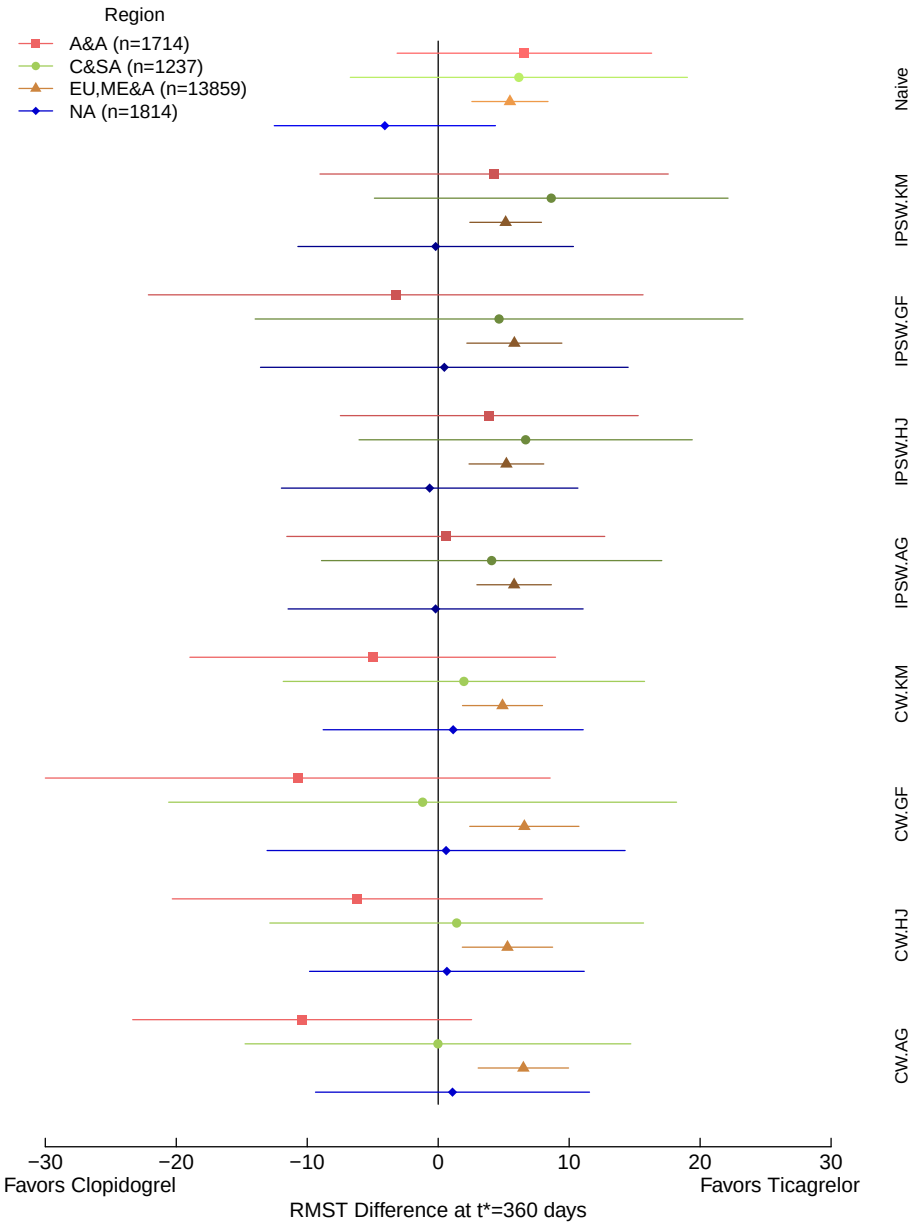

FIGURE C.3: Forest plot of estimated average RMST differences with 95% CIs at $t^* = 360$ days from 1) A&A: Asia and Australia, 2) C&SA: Central and South America, 3) EU,ME&A: Europe, Middle East and Africa, and 4) NA: North America in the four-region analysis for PLATO trial.

Figure C.4 presents the weighted absolute SMD of eight variables in the four-region analysis, comparing the population from each region to the target population. The unadjusted absolute SMDs between the dominant region, Europe, Middle East, and

Africa, and the target population are close to 0 for most variables except for the maintenance aspirin doses, which is a similar finding from the two-region analysis. However, notable imbalances of the covariates are observed from the unadjusted absolute SMDs between the other three regions and the target population. The CW absolute SMDs are always 0. However, the IPSW cannot balance the distributions of maintenance aspirin doses well.

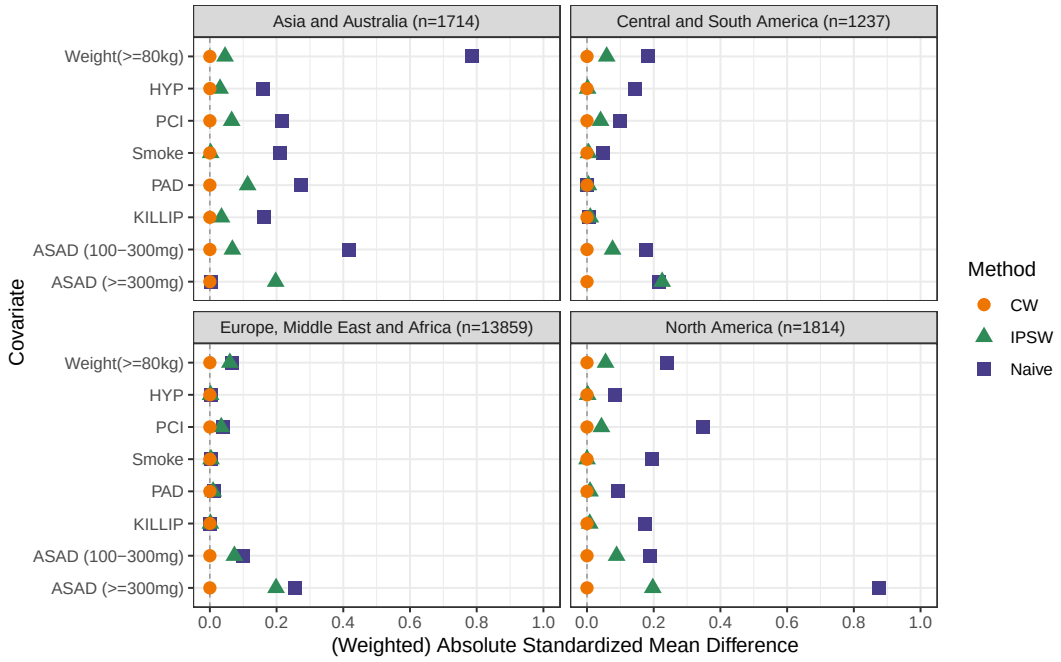

FIGURE C.4: Weighted absolute standardized mean differences of eight covariates comparing 1) Asia and Australia, 2) Central and South America, 3) Europe, Middle East and Africa, and 4) North America to the target population in the four-region analysis for PLATO trial. HYP: hypertension, PCI: percutaneous coronary intervention, PAD: peripheral arterial disease, KILLIP: Killip classification (Level I vs. Level II-IV), ASAD: aspirin dose.

Appendix D

Appendix for Chapter 4

Appendix D is organized as follows. In Appendix D.1, we show the details for fitting the piecewise exponential hazard models in the prediction method in Chapter 4.3.1. Appendix D.2 provides the derivation of the asymptotic variance of the CW Hajek and Augmented estimators.

D.1 Model Fitting in Prediction Method

For the i^{th} patient and j^{th} time interval, the piecewise exponential hazard function is written as follows:

$$\log(\lambda_{z_i,ij}) = \log(\lambda_{z_i,j}) + \gamma_{z_i}(1 - x_i), \text{ for } t \in [\tau_{z_i,j-1}, \tau_{z_i,j}), \tag{D.1}$$

where $z_i \in \{0, 1\}$ is the treatment indicator and $\lambda_{z_i,j}$ is the baseline hazard in the j^{th} interval.

For patient $i \in \{1, \ldots, n\}$, we first create some pseudo observations of U_i and δ_i in each time interval, where $U_i = \min(T_i, C_i)$ and $\delta_i = I[T_i \leqslant C_i]$. Let u_{ij} denote the event free time by the i^{th} patient in the j^{th} interval, $[\tau_{z_i,j-1}, \tau_{z_i,j})$. If this patient has

event or is censored after $\tau_{z_i,j}$, then $u_{ij} = \tau_{z_i,j} - \tau_{z_i,j-1}$. Otherwise, if this patient has an event or is censored in this interval, $u_{ij} = U_i - \tau_{z_i,j-1}$. In addition, let δ_{ij} denote the censoring indicator in the j^{th} time interval, such that $\delta_{ij} = 1$ if the i^{th} patients has event in this interval and $\delta_{ij} = 0$ otherwise. Then, the piecewise exponential model can be fitted to data by treating δ_{ij} as if they are independent Poisson observations with means $\mu_{ij} = u_{ij}\lambda_{z_i,ij}$. By replacing $\lambda_{z_i,ij} = \mu_{ij}/u_{ij}$, the Equation (D.1) can be written as follows:

$$\log(\mu_{ij}) = \log(u_{ij}) + \alpha_{z_i,j} + \gamma_{z_i}(1 - x_i), \tag{D.2}$$

where $\alpha_{z_i,j} = \log(\lambda_{z_i,j})$. As such, the initial piecewise exponential model can be reformulated and adapted using the Poisson log-linear regression, which incorporates pseudo observations of $\{\delta_{ij}, u_{ij}\}$ and employs $\log(u_{ij})$ as an offset. Then the parameters $\hat{\lambda}_{z,j}$ and $\hat{\gamma}_z$ can be estimated from this Poisson regression model.

D.2 Calibration Weighted Hajek and Augmented Estimators

D.2.1 CW Hajek Estimator

The CW Hajek estimator is defined as:

$$\hat{\Delta}_4(t^*) = \frac{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i Y_i}{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i} - \frac{\sum_{i=1}^{n_+} \hat{p}_i (1 - Z_i) w_i Y_i}{\sum_{i=1}^{n_+} \hat{p}_i (1 - Z_i) w_i}. \tag{D.3}$$

The large sample properties of this estimator can be derived by the M-estimator theory. Let $\boldsymbol{\theta} = [\theta_0, \theta_1]^T$ as the collection of parameters to be estimated. Then $\hat{\Delta}_4(t^*) = \hat{\theta}_1 - \hat{\theta}_0$ jointly solves the estimation equations as follows:

$$\sum_{i=1}^{n_+} \boldsymbol{\Phi}_i^H(\boldsymbol{\theta}) = \sum_{i=1}^{n_+} \begin{pmatrix} \hat{p}_i Z_i w_i (Y_i - \theta_1) \\ \hat{p}_i (1 - Z_i) w_i (Y_i - \theta_0) \end{pmatrix} = 0.$$

The sandwich variance estimator for $\hat{\boldsymbol{\theta}}$ is:

$$\Sigma_{\hat{\boldsymbol{\theta}}} = \left\{\sum_{i=1}^{n_+} \frac{\partial \boldsymbol{\Phi}_i^H(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T}\right\}^{-1} \left\{\sum_{i=1}^{n_+} \boldsymbol{\Phi}_i^H(\hat{\boldsymbol{\theta}})\boldsymbol{\Phi}_i^H(\hat{\boldsymbol{\theta}})^T\right\} \left\{\sum_{i=1}^{n_+} \frac{\partial \boldsymbol{\Phi}_i^H(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T}\right\}^{-1}.$$

By the M-estimator theory, $\hat{\boldsymbol{\theta}}$ is consistent to $\boldsymbol{\theta}$, and as the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\} \xrightarrow{d} N\left(0, n_+\Sigma_{\hat{\boldsymbol{\theta}}}\right)$. Then by continuous mapping theorem, $\sqrt{n_+}\{\hat{\Delta}_4(t^*) - \Delta_{t^*}\} \xrightarrow{d} N\left(0, \sigma_4^2(t^*)\right)$, where $\sigma_4^2(t^*)$ can be estimated by $n_+[1,-1]\Sigma_{\hat{\boldsymbol{\theta}}}[1,-1]^T = n_+[\mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_0) - 2\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_0)]$.

D.2.2 CW Augmented Estimator

The CW augmented estimator is defined as:

$$\begin{aligned}\hat{\Delta}_5(t^*) &= \frac{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i \{Y_i - m_1(x_i)\}}{\sum_{i=1}^{n_+} \hat{p}_i Z_i w_i} - \frac{\sum_{i=1}^{n_+} \hat{p}_i (1-Z_i) w_i \{Y_i - m_0(x_i)\}}{\sum_{i=1}^{n_+} \hat{p}_i (1-Z_i) w_i} \\ &+ \frac{\sum_{i=1}^{n_+} \hat{p}_i \{m_1(x_i) - m_0(x_i)\}}{\sum_{i=1}^{n_+} \hat{p}_i}. \end{aligned} \tag{D.4}$$

Its asymptotic distribution is also derived by the M-estimator theory. Let $\boldsymbol{\nu} = [\nu_0, \nu_1, \nu_2]^T$ denote as the collection of parameters to be estimated. Then $\hat{\Delta}_5(t^*) = \hat{\nu}_1 - \hat{\nu}_0 + \hat{\nu}_2$ jointly solves the following estimation equations:

$$\sum_{i=1}^{n_+} \boldsymbol{\Phi}_i^A(\boldsymbol{\nu}) = \sum_{i=1}^{n_+} \begin{pmatrix} \hat{p}_i Z_i w_i \{Y_i - m_1(x_i) - \nu_1\} \\ \hat{p}_i (1-Z_i) w_i \{Y_i - m_0(x_i) - \nu_0\} \\ \hat{p}_i \{m_1(x_i) - m_0(x_i) - \nu_2\} \end{pmatrix} = 0.$$

The sandwich variance estimator for $\hat{\boldsymbol{\nu}}$ is:

$$\Sigma_{\hat{\boldsymbol{\nu}}} = \left\{\sum_{i=1}^{n_+} \frac{\partial \boldsymbol{\Phi}_i^A(\hat{\boldsymbol{\nu}})}{\partial \boldsymbol{\nu}^T}\right\}^{-1} \left\{\sum_{i=1}^{n_+} \boldsymbol{\Phi}_i^A(\hat{\boldsymbol{\nu}})\boldsymbol{\Phi}_i^A(\hat{\boldsymbol{\nu}})^T\right\} \left\{\sum_{i=1}^{n_+} \frac{\partial \boldsymbol{\Phi}_i^A(\hat{\boldsymbol{\nu}})}{\partial \boldsymbol{\nu}^T}\right\}^{-1}.$$

As the sample size $n_+ \to \infty$, $\sqrt{n_+}\{\hat{\Delta}_5(t^*) - \Delta_{t^*}\} \xrightarrow{d} N\left(0, \sigma_5^2(t^*)\right)$, where $\sigma_5^2(t^*)$ can be estimated by $n_+[1,-1,1]\Sigma_{\hat{\boldsymbol{\nu}}}[1,-1,1]^T$.

Bibliography

Andersen, P., Hansen, M., and Klein, J. (2004), "Regression analysis of restricted mean survival time based on pseudo-observations," *Lifetime Data Analysis*, 10, 335–350.

Andersen, P. K. and Pohar Perme, M. (2010), "Pseudo-observations in survival analysis," *Statistical Methods in Medical Research*, 19, 71–99.

Antoniou, M., Jorgensen, A. L., and Kolamunnage-Dona, R. (2016), "Biomarker-guided adaptive trial designs in phase II and phase III: a methodological review," *PloS one*, 11, e0149803.

Antoniou, M., Kolamunnage-Dona, R., and Jorgensen, A. L. (2017), "Biomarker-guided non-adaptive trial designs in phase II and phase III: a methodological review," *Journal of Personalized Medicine*, 7, 1.

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

Austin, P. C. (2018), "Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model," *Journal of Statistical Computation and Simulation*, 88, 533–552.

Bailey, A. M., Mao, Y., Zeng, J., Holla, V., Johnson, A., Brusco, L., Chen, K., Mendelsohn, J., Routbort, M. J., Mills, G. B., et al. (2014), "Implementation of biomarker-driven cancer therapy: existing tools and remaining gaps," *Discovery medicine*, 17, 101.

Bang, H. and Tsiatis, A. A. (2000), "Estimating medical costs with censored data," *Biometrika*, 87, 329–343.

Barlesi, F., Vansteenkiste, J., Spigel, D., Ishii, H., Garassino, M., de Marinis, F., Özgüroğlu, M., Szczesna, A., Polychronis, A., Uslu, R., et al. (2018), "Avelumab versus docetaxel in patients with platinum-treated advanced non-small-cell lung cancer (JAVELIN Lung 200): an open-label, randomised, phase 3 study," *The Lancet Oncology*, 19, 1468–1479.

Bellamy, S. L., Li, Y., Lin, X., and Ryan, L. M. (2005), "Quantifying PQL bias in estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials," *Statistica Sinica*, pp. 1015–1032.

Bellera, C. A., MacGrogan, G., Debled, M., de Lara, C. T., Brouste, V., and Mathoulin, P. S. (2010), "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer," *BMC Medical Research Methodology*, 10, 1–12.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021), "The balancing act in causal inference," *arXiv preprint arXiv:2110.14831*.

Bland, J. M. and Altman, D. G. (2004), "The logrank test," *BMJ*, 328, 1073.

Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E., et al. (2015), "Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer," *New England Journal of Medicine*, 373, 1627–1639.

Bowden, J., Tierney, J. F., Simmonds, M., Copas, A. J., and Higgins, J. P. (2011), "Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model," *Research Synthesis Methods*, 2, 150–162.

Brahmer, J., Reckamp, K. L., Baas, P., Crinò, L., Eberhardt, W. E., Poddubskaya, E., Antonia, S., Pluzanski, A., Vokes, E. E., Holgado, E., et al. (2015), "Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer," *New England Journal of Medicine*, 373, 123–135.

Breslow, N. (2004), "Whither PQL?" in *Proceedings of the second seattle symposium in biostatistics: analysis of correlated data*, pp. 1–22, Springer.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Burgette, L., Griffin, B. A., and McCaffrey, D. (2021), "Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package," *R package. and Corporation*.

Burke, D., Ensor, J., and Riley, R. (2017), "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ," *Statistics in Medicine*, 36, 855–875.

Carnicelli, A., Hong, H., Giugliano, R. P., Connolly, S. J., Eikelboom, J., Patel, M. R., Wallentin, L., Morrow, D. A., Wojdyla, D., Hua, K., et al. (2021), "Individual Patient Data from the Pivotal Randomized Controlled Trials of Non-Vitamin

K Antagonist Oral Anticoagulants in Patients with Atrial Fibrillation (COMBINE AF): Design and Rationale: From the COMBINE AF (A Collaboration between Multiple institutions to Better Investigate Non-vitamin K antagonist oral anticoagulant use in Atrial Fibrillation) Investigators," *American Heart Journal*, 233, 48–58.

Carnicelli, A. P., Hong, H., Connolly, S. J., Eikelboom, J., Giugliano, R. P., Morrow, D. A., Patel, M. R., Wallentin, L., Alexander, J. H., Bahit, M. C., et al. (2022), "Direct oral anticoagulants versus warfarin in patients with atrial fibrillation: patient-level network meta-analyses of randomized clinical trials with interaction testing by age and sex," *Circulation*.

Carroll, K. and Fleming, T. (2013), "Statistical evaluation and analysis of regional interactions: the PLATO trial case study," *Statistics in Biopharmaceutical Research*, 5, 91–101.

Cefalu, M., Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2021), *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*, R package version 2.5.

Chattopadhyay, A., Cohn, E. R., and Zubizarreta, J. R. (2022), "One-step weighting to generalize and transport treatment effect estimates to a target population," *arXiv preprint arXiv:2203.08701*.

Chen, H., Manning, A., and Dupuis, J. (2012), "A method of moments estimator for random effect multivariate meta-analysis," *Biometrics*, 68, 1278–1284.

Chen, J., Quan, H., Binkowitz, B., Ouyang, S. P., Tanaka, Y., Li, G., Menjoge, S., Ibia, E., and Consistency Workstream of the PhRMA MRCT Key Issue Team (2010), "Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review," *Pharmaceutical Statistics*, 9, 242–253.

Chen, J. J., Lu, T.-P., Chen, D.-T., and Wang, S.-J. (2014), "Biomarker adaptive designs in clinical trials," *Translational Cancer Research*, 3, 279–92.

Conner, S., Sullivan, L. M., Benjamin, E. J., LaValley, M. P., Galea, S., and Trinquart, L. (2019), "Adjusted restricted mean survival times in observational studies," *Statistics in Medicine*, 38, 3832–3860.

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

Crowther, M. J., Riley, R. D., Staessen, J. A., Wang, J., Gueyffier, F., and Lambert, P. C. (2012), "Individual patient data meta-analysis of survival data using Poisson regression models," *BMC Medical Research Methodology*, 12, 1–14.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019), "Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals," *Biometrics*, 75, 685–694.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. (2020), "Extending inferences from a randomized trial to a new target population," *Statistics in Medicine*, 39, 1999–2014.

De la Fuente, A. (2000), *Mathematical methods and models for economists*, Cambridge University Press.

Debray, T. P., Moons, K. G., Abo-Zaid, G. M. A., Koffijberg, H., and Riley, R. D. (2013), "Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?" *PloS One*, 8, e60650.

Debray, T. P. A., Moons, K. G., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H., Reitsma, J. B., and Group, G. M. R. (2015), "Get real in individual participant data (IPD) meta-analysis: a review of the methodology," *Research Synthesis Methods*, 6, 293–309.

Degtiar, I. and Rose, S. (2023), "A review of generalizability and transportability," *Annual Review of Statistics and Its Application*, 10, 501–524.

DerSimonian, R. and Laird, N. (1986), "Meta-analysis in clinical trials," *Controlled Clinical Trials*, 7, 177–188.

Diao, G., Dong, J., Zeng, D., Ke, C., Rong, A., and Ibrahim, J. G. (2018), "Biomarker threshold adaptive designs for survival endpoints," *Journal of Biopharmaceutical Statistics*, 28, 1038–1054.

Donegan, S., Williamson, P., D'Alessandro, U., and Smith, C. T. (2012), "Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates," *Statistics in Medicine*, 31, 3840–3857.

Donegan, S., Williamson, P., D'Alessandro, U., Garner, P., and Smith, C. T. (2013), "Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials," *Statistics in Medicine*, 32, 914–930.

Dowd, B. E., Greene, W. H., and Norton, E. C. (2014), "Computation of standard errors," *Health services research*, 49, 731–750.

Eaton, A., Therneau, T., and Le-Rademacher, J. (2020), "Designing clinical trials with (restricted) mean survival time endpoint: practical considerations," *Clinical Trials*, 17, 285–294.

Fedorov, V. and Liu, T. (2007), "Enrichment design," *Wiley Encyclopedia of Clinical Trials*, pp. 1–8.

Fehrenbacher, L., Spira, A., Ballinger, M., Kowanetz, M., Vansteenkiste, J., Mazieres, J., Park, K., Smith, D., Artal-Cortes, A., Lewanski, C., et al. (2016), "Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial," *The Lancet*, 387, 1837–1846.

Finn, R., Qin, S., Ikeda, M., Galle, P. R., Ducreux, M., Kim, T.-Y., Kudo, M., Breder, V., Merle, P., Kaseb, A. O., et al. (2020), "Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma," *New England Journal of Medicine*, 382, 1894–1905.

Fleming, T. and Harrington, D. (1991), *Counting processes and survival analysis*, John Wiley & Sons.

Food and Drug Administration (FDA) (2016), "General Principles for Planning and Design of Multi-Regional Clinical Trials E17," .

Food and Drug Administration (FDA) (2019a), "FDA Briefing Document for Oncologic Drugs Advisory Committee Meeting," .

Food and Drug Administration (FDA) (2019b), "FDA CDRH Medical Devices Advisory Committee Circulatory System Panel Meeting," .

Food and Drug Administration (FDA) (2020), "Acute Myeloid Leukemia: Developing Drugs and Biological Products for Treatment Guidance for Industry," .

Freeman, S. C., Fisher, D., White, I. R., Auperin, A., and Carpenter, J. R. (2019), "Identifying inconsistency in network meta-analysis: Is the net heat plot a reliable method?" *Statistics in Medicine*, 38, 5547–5564.

Freidlin, B. and Korn, E. L. (2010), "Biomarker-adaptive clinical trial designs," *Pharmacogenomics*, 11, 1679–1682.

Freidlin, B. and Simon, R. (2005), "Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients," *Clinical Cancer Research*, 11, 7872–7878.

Frieri, R., Rosenberger, W. F., Flournoy, N., and Lin, Z. (2023), "Design considerations for two-stage enrichment clinical trials," *Biometrics*, 79, 2565–2576.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F., and Ulerich, R. (2002), *GNU scientific library*, Network Theory Limited Godalming.

Gardiner, J. (2021), "Restricted Mean Survival Time Estimation: Nonparametric and Regression Methods," *Journal of Statistical Theory and Practice*, 15, 1–15.

Garrido, M., Lum, J., and Pizer, S. (2021), "Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings," *Statistics in Medicine*, 40, 1204–1223.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2012), "Multivariate meta-analysis for non-linear and other multi-parameter associations," *Statistics in Medicine*, 31, 3821–3839.

Gill, R. (1983), "Large sample behaviour of the product-limit estimator on the whole line," *The Annals of Statistics*, pp. 49–58.

Goodman, M. S., Li, Y., and Tiwari, R. C. (2011), "Detecting multiple change points in piecewise constant hazard functions," *Journal of Applied Statistics*, 38, 2523–2532.

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

Gregson, J., Sharples, L., Stone, G. W., Burman, C.-F., Öhrn, F., and Pocock, S. (2019), "Nonproportional hazards for time-to-event outcomes in clinical trials: JACC review topic of the week," *Journal of the American College of Cardiology*, 74, 2102–2112.

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

Harrington, D. P. and Fleming, T. R. (1982), "A class of rank test procedures for censored survival data," *Biometrika*, 69, 553–566.

Hawkins, N., Scott, D. A., and Woods, B. (2016), "'Arm-based'parameterization for network meta-analysis," *Research Synthesis Methods*, 7, 306–313.

He, P., Fang, L., and Su, Z. (2013), "A sequential testing approach to detecting multiple change points in the proportional hazards model," *Statistics in Medicine*, 32, 1239–1245.

Higgins, J., Jackson, D., Barrett, J., Lu, G., Ades, A., and White, I. (2012), "Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies," *Research Synthesis Methods*, 3, 98–110.

Holford, T. R. (1980), "The analysis of rates and of survivorship using log-linear models," *Biometrics*, pp. 299–305.

Hong, H., Fu, H., Price, K. L., and Carlin, B. P. (2015), "Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment," *Statistics in Medicine*, 34, 2794–2819.

Hong, H., Chu, H., Zhang, J., and Carlin, B. P. (2016), "A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," *Research Synthesis Methods*, 7, 6–22.

Hua, H., Burke, D. L., Crowther, M. J., Ensor, J., Tudur Smith, C., and Riley, R. D. (2017), "One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information," *Statistics in Medicine*, 36, 772–789.

Hua, K., Wojdyla, D., Carnicelli, A., Granger, C., Wang, X., and Hwanhee, H. (2024), "Network meta-analysis with individual participant-level data of time-to-event outcomes using Cox regression," *Statistics in Medicine*.

Irwin, J. (1949), "The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice," *Epidemiology & Infection*, 47, 188–189.

Jackson, D., Riley, R., and White, I. R. (2011), "Multivariate meta-analysis: potential and promise," *Statistics in Medicine*, 30, 2481–2498.

Jackson, D., White, I., and Riley, R. (2013), "A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression," *Biometrical Journal*, 55, 231–245.

Jang, W. and Lim, J. (2009), "A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects," *Communications in Statistics-Simulation and Computation*, 38, 692–702.

Jansen, J. P. (2011), "Network meta-analysis of survival data with fractional polynomials," *BMC Medical Research Methodology*, 11, 1–14.

Jiang, J. and Nguyen, T. (2007), *Linear and generalized linear mixed models and their applications*, vol. 1, Springer.

Johnston, S. E., Lipkovich, I., Dmitrienko, A., and Zhao, Y. D. (2022), "A two-stage adaptive clinical trial design with data-driven subgroup identification at interim analysis," *Pharmaceutical Statistics*, 21, 1090–1108.

Josey, K., Juarez-Colunga, E., Yang, F., and Ghosh, D. (2021a), "A framework for covariate balance using Bregman distances," *Scandinavian Journal of Statistics*, 48, 790–816.

Josey, K. P., Berkowitz, S. A., Ghosh, D., and Raghavan, S. (2021b), "Transporting experimental results with entropy balancing," *Statistics in Medicine*, 40, 4310–4326.

Josey, K. P., Yang, F., Ghosh, D., and Raghavan, S. (2022), "A calibration approach to transportability and data-fusion with observational data," *Statistics in Medicine*, 41, 4511–4531.

Kantoff, P. W., Higano, C. S., Shore, N. D., Berger, E. R., Small, E. J., Penson, D. F., Redfern, C. H., Ferrari, A. C., Dreicer, R., Sims, R. B., et al. (2010), "Sipuleucel-T immunotherapy for castration-resistant prostate cancer," *New England Journal of Medicine*, 363, 411–422.

Karrison, T. G. (2016), "Versatile tests for comparing survival curves based on weighted log-rank statistics," *The Stata Journal*, 16, 678–690.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016), "Assessing methods for generalizing experimental impact estimates to target populations," *Journal of Research on Educational Effectiveness*, 9, 103–127.

Kim, D., Uno, H., and Wei, L. (2017), "Restricted mean survival time as a measure to interpret clinical trial results," *JAMA Cardiology*, 2, 1179–1180.

Kleinbaum, D. G. and Klein, M. (2012), "Evaluating the proportional hazards assumption," in *Survival Analysis*, pp. 161–200, Survival Analysis.

Kloecker, D. E., Davies, M. J., Khunti, K., and Zaccardi, F. (2020), "Uses and limitations of the restricted mean survival time: illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes," *Annals of Internal Medicine*, 172, 541–552.

Kontopantelis, E. (2018), "A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study," *Research Synthesis Methods*, 9, 417–430.

Lai, T. L., Lavori, P. W., and Liao, O. Y.-W. (2014), "Adaptive choice of patient subgroup for comparing two treatments," *Contemporary Clinical Trials*, 39, 191–200.

Lai, T. L., Lavori, P. W., and Tsang, K. W. (2019), "Adaptive enrichment designs for confirmatory trials," *Statistics in Medicine*, 38, 613–624.

Laird, N. and Olivier, D. (1981), "Covariance analysis of censored survival data using log-linear analysis techniques," *Journal of the American Statistical Association*, 76, 231–240.

Landeck, L., Kneip, C., Reischl, J., and Asadullah, K. (2016), "Biomarkers and personalized medicine: current status and further perspectives with special focus on dermatology," *Experimental Dermatology*, 25, 333–339.

Lee, B., Lessler, J., and Stuart, E. (2010), "Improving propensity score weighting using machine learning," *Statistics in Medicine*, 29, 337–346.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. (2021), "Improving trial generalizability using observational studies," *Biometrics*.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, 113, 390–400.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lin, D. and Wei, L. (1989), "The robust inference for the Cox proportional hazards model," *Journal of the American Statistical Association*, 84, 1074–1078.

Lin, D. Y., Wei, L.-J., and Ying, Z. (1993), "Checking the Cox model with cumulative sums of martingale-based residuals," *Biometrika*, 80, 557–572.

Lin, J.-A. and He, P. (2015), "Reinventing clinical trials: a review of innovative biomarker trial designs in cancer therapies," *British Medical Bulletin*, 114, 17–27.

Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., Leon, L. F., Liao, J. J., Liu, R., Luo, X., et al. (2020), "Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis," *Statistics in Biopharmaceutical Research*, 12, 187–198.

Lin, X. (2007), "Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models," *Lifetime Data Analysis*, 13, 533–544.

Lin, X. and Breslow, N. E. (1996), "Bias correction in generalized linear mixed models with multiple components of dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Liu, Q. and Pierce, D. A. (1994), "A note on Gauss—Hermite quadrature," *Biometrika*, 81, 624–629.

Lu, G. and Ades, A. E. (2004), "Combination of direct and indirect evidence in mixed treatment comparisons," *Statistics in Medicine*, 23, 3105–3124.

Lu, Y. and Tian, L. (2021), "Statistical considerations for sequential analysis of the restricted mean survival time for randomized clinical trials," *Statistics in Biopharmaceutical Research*, 13, 210–218.

Lueza, B., Mauguen, A., Pignon, J.-P., Rivero-Arias, O., Bonastre, J., and Group, M.-L. C. (2016), "Difference in restricted mean survival time for cost-effectiveness analysis using individual patient data meta-analysis: evidence from a case study," *PLoS One*, 11, e0150032.

Luo, X., Huang, B., and Quan, H. (2019), "Design and monitoring of survival trials based on restricted mean survival times," *Clinical Trials*, 16, 616–625.

Lyman, G. H. and Kuderer, N. M. (2005), "The strengths and limitations of meta-analyses based on aggregate data," *BMC Medical Research Methodology*, 5, 1–7.

Mahaffey, K., Wojdyla, D. M., Carroll, K., Becker, R. C., Storey, R. F., Angiolillo, D. J., Held, C., Cannon, C. P., James, S., Pieper, K. S., et al. (2011), "Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) trial," *Circulation*, 124, 544–554.

Maitournam, A. and Simon, R. (2005), "On the efficiency of targeted clinical trials," *Statistics in Medicine*, 24, 329–339.

Mao, L. (2023), "Study design for restricted mean time analysis of recurrent events and death," *Biometrics*.

McCaffrey, D., Griffin, B., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013), "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, 32, 3388–3414.

McCullagh, P. (2019), *Generalized linear models*, Routledge.

Mehta, C., Schäfer, H., Daniel, H., and Irle, S. (2014), "Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints," *Statistics in Medicine*, 33, 4515–4531.

Mills, E. J., Thorlund, K., and Ioannidis, J. P. (2013), "Demystifying trial networks and network meta-analysis," *BMJ*, 346.

Mukhopadhyay, P., Ye, J., Anderson, K. M., Roychoudhury, S., Rubin, E. H., Halabi, S., and Chappell, R. J. (2022), "Log-rank test vs MaxCombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in immuno-oncology trials: a systematic review and meta-analysis," *JAMA Oncology*.

Naimi, A. I., Cole, S. R., and Kennedy, E. H. (2017), "An introduction to g methods," *International Journal of Epidemiology*, 46, 756–762.

Ng'andu, N. H. (1997), "An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model," *Statistics in Medicine*, 16, 611–626.

Nugent, J., Doone, B., and Kleinman, K. (2019), "Bias induced by fitting GLMMs with dichotomous outcomes using penalized quasi-likelihood," *Unpublished*.

Park, Y., Liu, S., Thall, P. F., Yuan, Y., et al. (2022), "Bayesian group sequential enrichment designs based on adaptive regression of response and survival time on baseline biomarkers," *Biometrics*, 78, 60–71.

Pepe, M. S. and Fleming, T. R. (1989), "Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data," *Biometrics*, pp. 497–507.

Perego, C., Sbolli, M., Specchia, C., Fiuzat, M., McCaw, Z. R., Metra, M., Oriecuia, C., Peveri, G., Wei, L.-J., O'Connor, C. M., et al. (2020), "Utility of restricted mean survival time analysis for heart failure clinical trial evaluation and interpretation," *Heart Failure*, 8, 973–983.

Persson, I. and Khamis, H. J. (2007), "A comparison of graphical methods for assessing the proportional hazards Assumptions in the Cox model," *Journal of Statistics and Applications*, 2, 1.

Pinheiro, J. C. and Bates, D. M. (1995), "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of Computational and Graphical Statistics*, 4, 12–35.

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., and Zhao, P.-L. (2013), "Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials," *Statistics in Medicine*, 32, 1691–1706.

Quan, H., Mao, X., Chen, J., Shih, W. J., Ouyang, S. P., Zhang, J., Zhao, P.-L., and Binkowitz, B. (2014), "Multi-regional clinical trial design and consistency assessment of treatment effects," *Statistics in Medicine*, 33, 2191–2205.

Rajkumar, S. V., Jacobus, S., Callander, N. S., Fonseca, R., Vesole, D. H., Williams, M. E., Abonour, R., Siegel, D. S., Katz, M., and Greipp, P. R. (2010), "Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial," *The Lancet Oncology*, 11, 29–37.

Reck, M., Rodríguez-Abreu, D., Robinson, A. G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., et al. (2016), "Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer," *New England Journal of Medicine*, 375, 1823–1833.

Renfro, L. A., Coughlin, C. M., Grothey, A. M., and Sargent, D. J. (2014), "Adaptive randomized phase II design for biomarker threshold selection and independent evaluation," *Chinese Clinical Oncology*, 3.

Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010), "Meta-analysis of individual participant data: rationale, conduct, and reporting," *BMJ : British Medical Journal*, 340:C221.

Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., Von Pawel, J., Gadgeel, S. M., Hida, T., Kowalski, D. M., Dols, M. C., et al. (2017), "Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial," *The Lancet*, 389, 255–265.

Robert, C., Long, G. V., Brady, B., Dutriaux, C., Maio, M., Mortier, L., Hassel, J. C., Rutkowski, P., McNeil, C., Kalinka-Warzocha, E., et al. (2015), "Nivolumab in previously untreated melanoma without BRAF mutation," *New England Journal of Medicine*, 372, 320–330.

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

Robins, J. and Hernan, M. (2008), "Estimation of the causal effects of time-varying exposures," *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pp. 553–599.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable," *Statistical Science*, 22, 544–559.

Royston, P. and Parmar, M. K. (2013), "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, 1–15.

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

Salanti, G., Higgins, J. P., Ades, A., and Ioannidis, J. P. (2008), "Evaluation of networks of randomized trials," *Statistical Methods in Medical Research*, 17, 279–301.

Sankar, K., Ye, J. C., Li, Z., Zheng, L., Song, W., and Hu-Lieskovan, S. (2022), "The role of biomarkers in personalized immunotherapy," *Biomarker Research*, 10, 1–13.

Saramago, P., Chuang, L.-H., and Soares, M. O. (2014), "Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data," *BMC Medical Research Methodology*, 14, 1–11.

Schoenfeld, D. (1980), "Chi-squared goodness-of-fit tests for the proportional hazards regression model," *Biometrika*, 67, 145–153.

Schwarzer, G., Carpenter, J. R., Rücker, G., et al. (2015), *Meta-analysis with R*, vol. 4784, Springer.

Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

Shi, W.-J. and Zhao, W. (2020), "Biomarkers or factors for predicting the efficacy and adverse effects of immune checkpoint inhibitors in lung cancer: achievements and prospective," *Chinese Medical Journal*, 133, 2466–2475.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R., and Mulani, P. M. (2010), "Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept," *Pharmacoeconomics*, 28, 935–945.

Simmonds, M. C., Higginsa, J. P., Stewartb, L. A., Tierneyb, J. F., Clarke, M. J., and Thompson, S. G. (2005), "Meta-analysis of individual patient data from randomized trials: a review of methods used in practice," *Clinical Trials*, 2, 209–217.

Simon, N. and Simon, R. (2013), "Adaptive enrichment designs for clinical trials," *Biostatistics*, 14, 613–625.

Simon, N. and Simon, R. (2018), "Using Bayesian modeling in frequentist adaptive enrichment designs," *Biostatistics*, 19, 27–41.

Simon, R. (2014), "Biomarker based clinical trial design," *Chin Clin Oncol*, 3, 8.

Simon, R. and Maitournam, A. (2004), "Evaluating the efficiency of targeted designs for randomized clinical trials," *Clinical Cancer Research*, 10, 6759–6763.

Sinha, B., Hartung, J., and Knapp, G. (2011), *Statistical meta-analysis with applications*, John Wiley & Sons.

Smith, C. T., Williamson, P. R., and Marson, A. G. (2005), "Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes," *Statistics in Medicine*, 24, 1307–1319.

Stallard, N. (2023), "Adaptive enrichment designs with a continuous biomarker," *Biometrics*, 79, 9–19.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, 56, 29–38.

Stensrud, M. J. and Hernán, M. A. (2020), "Why test for proportional hazards?" *JAMA*, 323, 1401–1402.

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., and Stewart, L. A. (2012), "Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice," *PLoS One*.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011), "The use of propensity scores to assess the generalizability of results from randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369–386.

Su, M., Fang, L., and Su, Z. (2013), "A likelihood and resampling based approach to dichotomizing a continuous biomarker in medical research," *Journal of Biopharmaceutical Statistics*, 23, 637–647.

Tang, X. and Trinquart, L. (2022), "Bayesian multivariate network meta-analysis model for the difference in restricted mean survival times," *Statistics in Medicine*, 41, 595–611.

Thall, P. F. (2021), "Adaptive enrichment designs in clinical trials," *Annual Review of Statistics and its Application*, 8, 393–411.

Tian, L., Zhao, L., and Wei, L. (2014), "Predicting the restricted mean event time with the subject's baseline covariates in survival analysis," *Biostatistics*, 15, 222–233.

Tian, L., Fu, H., Ruberg, S. J., Uno, H., and Wei, L.-J. (2018), "Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations," *Biometrics*, 74, 694–702.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016), "Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials," *Journal of Clinical Oncology*, 34, 1813–1819.

Trinquart, L., Bill-Axelson, A., and Rider, J. R. (2019), "Restricted Mean Survival Times to Improve Communication of Evidence from Cancer Randomized Trials and Observational Studies." *European Urology*, 76, 137–139.

Tsiatis, A. A. (2006), "Semiparametric theory and missing data," .

Tsong, Y., Chang, W., Dong, X., and Tsou, H. (2012), "Assessment of regional treatment effect in a multiregional clinical trial," *Journal of Biopharmaceutical Statistics*, 22, 1019–1036.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., et al. (2014), "Moving beyond the hazard

ratio in quantifying the between-group difference in survival analysis," *Journal of Clinical Oncology*, 32, 2380.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer, New York, fourth edn., ISBN 0-387-95457-0.

Wallentin, L., Becker, R. C., Budaj, A., Cannon, C. P., Emanuelsson, H., Held, C., Horrow, J., Husted, S., James, S., Katus, H., et al. (2009), "Ticagrelor versus clopidogrel in patients with acute coronary syndromes," *New England Journal of Medicine*, 361, 1045–1057.

Wang, S.-J., James Hung, H., and O'Neill, R. T. (2009), "Adaptive patient enrichment designs in therapeutic trials," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51, 358–374.

Wang, T., Wang, X., George, S., and Zhou, H. (2020), "Design and analysis of biomarker-integrated clinical trials with adaptive threshold detection and flexible patient enrichment," *Journal of Biopharmaceutical Statistics*, 30, 1060–1076.

Wang, X. and Schaubel, D. E. (2018), "Modeling restricted mean survival time under general censoring mechanisms," *Lifetime Data Analysis*, 24, 176–199.

Wang, X., Zhou, J., Wang, T., and George, S. (2018), "On enrichment strategies for biomarker stratified clinical trials," *Journal of Biopharmaceutical Statistics*, 28, 292–308.

Wang, X., Bai, F., Pang, H., and George, S. (2019a), "Bias-adjusted Kaplan–Meier survival curves for marginal treatment effect in observational studies," *Journal of Biopharmaceutical Statistics*, 29, 592–605.

Wang, X., Cai, J., and George, S. (2019b), "Enriched Biomarker-Driven Trials," *Wiley StatsRef: Statistics Reference*.

Wason, J. M. (2023), "Discussion on "Adaptive enrichment designs with a continuous biomarker" by Nigel Stallard," *Biometrics*, 79, 23–25.

Wei, Y. and Higgins, J. P. (2013), "Bayesian multivariate meta-analysis with multiple outcomes," *Statistics in Medicine*, 32, 2911–2934.

Wei, Y., Royston, P., Tierney, J., and Parmar, M. (2015), "Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data," *Statistics in Medicine*, 34, 2881–2898.

Weir, I., Tian, L., and Trinquart, L. (2021), "Multivariate meta-analysis model for the difference in restricted mean survival times," *Biostatistics*, 22, 82–96.

Weir, I. R. and Trinquart, L. (2018), "Design of non-inferiority randomized trials using the difference in restricted mean survival times," *Clinical Trials*, 15, 499–508.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017), "Transportability of trial results using inverse odds of sampling weights," *American Journal of Epidemiology*, 186, 1010–1014.

White, I. R. (2015), "Network meta-analysis," *The Stata Journal*, 15, 951–985.

White, I. R., Barrett, J. K., Jackson, D., and Higgins, J. P. (2012), "Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression," *Research Synthesis Methods*, 3, 111–125.

White, I. R., Turner, R. M., Karahalios, A., and Salanti, G. (2019), "A comparison of arm-based and contrast-based models for network meta-analysis," *Statistics in Medicine*, 38, 5197–5213.

Xie, J. and Liu, C. (2005), "Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data," *Statistics in Medicine*, 24, 3089–3110.

Xue, Y. and Schifano, E. D. (2017), "Diagnostics for the Cox model," *Communications for Statistical Applications and Methods*, 24, 583–604.

Zhang, W., Qian, L., and Li, Y. (2014), "Semiparametric sequential testing for multiple change points in piecewise constant hazard functions with long-term survivors," *Communications in Statistics-Simulation and Computation*, 43, 1685–1699.

Zhang, Z., Li, M., Lin, M., Soon, G., Greene, T., and Shen, C. (2017), "Subgroup selection in adaptive signature designs of confirmatory clinical trials," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66, 345–361.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M. A., Solomon, S. D., Trippa, L., and Wei, L. (2016), "On the restricted mean survival time curve in survival analysis," *Biometrics*, 72, 215–221.

Zhao, Q. and Percival, D. (2016), "Entropy balancing is doubly robust," *Journal of Causal Inference*, 5, 20160010.

Zhao, Y.-Q. and LeBlanc, M. L. (2020), "Designing precision medicine trials to yield a greater population impact," *Biometrics*, 76, 643–653.

Zhong, Y. and Schaubel, D. E. (2022), "Restricted mean survival time as a function of restriction time," *Biometrics*, 78, 192–201.

Biography

Kaiyuan Hua received a PhD degree of Biostatistics in the Department of Biostatistics and Bioinformatics at Duke University School of Medicine, under supervision of Dr. Xiaofei Wang and Dr. Hwanhee Hong. He received a bachelor's degree of science in mathematics at Fudan University, Shanghai, China, and a master's dergree of Biostatistics at Duke University School of Medicine. His research interests include survival analysis, clinical trials, and meta-analysis. He will work at FDA as a mathematical statistician.