Bayesian Methods for snSMART Designs with External Controls and Dynamic Prediction of Landmark Survival Time in Cancer Clinical Trials

by

Sidi Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2024

Doctoral Committee:

Professor Kelley M. Kidwell, Chair
Professor Thomas M. Braun
Associate Professor Daniel L. Hertz
Professor Min Zhang

Sidi Wang

sidiwang@umich.edu

ORCID iD: 0000-0003-4838-0842

ACKNOWLEDGMENTS

In the quiet moments of reflection, as I pen down these acknowledgements, my heart overflows with profound gratitude. This journey, culminating in the pages of this dissertation, has been illuminated by the wisdom, support, and encouragement of many extraordinary souls.

Foremost in my heart is gratitude for Dr. Kelley Kidwell, whose guidance was the compass by which I navigated this journey. Dr. Kidwell, you have been more than an advisor; you have been a mentor, a supporter, and an inspiration. Your empathetic approach, combined with your academic acumen, has not only guided my academic pursuits but also shaped my personal ethos. Your ability to see the world from a student's perspective, your considerate nature, and your unyielding support have been the bedrock of my PhD experience. For this, and for the wonderful life lessons, I owe you a debt of gratitude that mere words cannot express.

To the esteemed members of my committee—Dr. Thomas Braun, Dr. Daniel Hertz, and Dr. Min Zhang—thank you for your invaluable insights and constructive feedback. Your contributions have been instrumental in shaping my academic narrative.

A special word of thanks to Dr. Satrajit Roychoudhury, whose expertise and guidance in the realm of Bayesian methods and research strategies have profoundly impacted my work. Your approach to research and your extensive knowledge have been a source of inspiration and learning.

A heartfelt thanks to Dan Barker, Michael Kleinsasser, and Dr. Roy Tamura, whose assistance with my projects was invaluable. And to my classmates and study groups, thank you for the shared struggles and triumphs, for the laughter and the late nights, and for all the moments in between that have enriched this journey.

The academic path is rarely walked alone, and I have been fortunate to be guided by many luminaries along the way. To Drs. Nicholas Henderson, Veera Baladandayuthapani, Jennifer Smith, Xiang Zhou, and many others, your encouragement, wisdom, and support have been pivotal at many crossroads in my life.

To Drs. Fang Fang and Holly Hartman, thank you for sharing the code and methodologies that laid the groundwork for my own research. Your generosity has been a beacon of collaborative spirit.

The administrative staff in the Biostatistics office deserves immense thanks for their behind-the-scenes work that made my journey smoother. Your reminders, guidance, and support have been indispensable.

And to my friends and family, especially Richard Higgins, Xuelin Gu, Di Wang, Kaiwei Lei, and my pillars of strength, Mom and Dad—your unwavering support, love, and belief in me have been my greatest source of motivation. You've nurtured my spirit and shaped my character, making me a better researcher and human being.

This dissertation represents not just my work, but the collective effort, support, and love of everyone mentioned here and many more. I'm deeply thankful for every one of you. This journey would have been much harder and lonelier without your support. Thank you from the bottom of my heart.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ACRONYMS

6MWD 6-minute walk distance

ANCOVA analysis of covariance

ANOVA analysis of variance

BLPM Bayesian longitudinal piecewise meta-analytic combined

BJSM Bayesian joint stage model

BMA Bayesian model averaging

CI credible interval

CINRG Cooperative International Neuromuscular Research Group

CR coverage rate

DMD Duchenne muscular dystrophy

DNHS Duchenne Natural History Study

ESS effective sample size

FDA US Food and Drug Administration

IPTW inverse probability treatment weighting

MAC meta-analytic combined

MAP meta-analytic predictive

MCMC Markov chain Monte Carlo

MMRM mixed-effects model of repeated measures

NSAA North Star Ambulatory Assessment

OLE open-label extension

OS overall survival

PD progressive disease

PFS progression-free survival

PPD posterior predictive distribution

PS propensity score

RCT randomized controlled trial

RECIST Response Evaluation Criteria in Solid Tumors

rMSE root mean squared error

snSMART small sample, sequential, multiple assignment, randomized trial

SMART sequential, multiple assignment, randomized trial

TTP time-to-progression

ABSTRACT

In Duchenne muscular dystrophy (DMD) and other rare diseases, recruiting patients into clinical trials is challenging. Additionally, assigning patients to long-term, multi-year placebo arms raises ethical and trial retention concerns. This poses a significant challenge to the traditional sequential drug development paradigm. In this dissertation, we present small sample, sequential, multiple assignment, randomized trial (snSMART) designs and methods that formally incorporate external control data under both the non-longitudinal and longitudinal settings.

After introducing the integration of snSMART with external control data in Chapter 1, Chapter 2 proposes an snSMART design that integrates dose selection and confirmatory assessment into a single trial. This multi-stage design evaluates the effects of multiple doses of a promising drug, rerandomizing patients to appropriate dose levels based on their stage 1 dose response. Our approach enhances the efficiency of treatment effect estimates by: (i) enriching the placebo arm with external control data, and (ii) utilizing data from all stages. We combine data from external controls and different stages using a robust meta-analytic combined (MAC) approach, accounting for various sources of heterogeneity and potential selection bias. Upon reanalyzing data from a DMD trial with our proposed method, MAC-snSMART, we observe that MAC-snSMART estimators offer improved efficiency over the original trial results. The robust MAC-snSMART method frequently provides more accurate estimators than traditional analytical methods. Overall, our proposed methodology provides a promising candidate for efficient drug development in DMD and other rare diseases.

In Chapter 3, we present Bayesian longitudinal piecewise meta-analytic combined (BLPM), a notable advancement on the robust MAC-snSMART method from Chapter 2. This enhancement introduces significant improvements to snSMART research by: (1) enabling longitudinal data analysis, (2) incorporating patient baseline characteristics, (3) utilizing multiple imputation for missing data, (4) reducing heterogeneity with propensity score (PS), and (5) managing stage-wise treatment effect non-exchangeability. These developments significantly increase the snSMART design's utility and efficiency in rare disease drug development. BLPM applies PS trimming, inverse probability treatment weighting (IPTW), and the MAC framework to navigate heterogeneity and cross-stage treatment effects. Our evaluations, through simulation studies and the reanalysis of a DMD trial, show that BLPM methods consistently achieve the lowest rMSE across tested scenarios, underscoring its potential to enhance rare disease drug development.

In Chapter 4, we propose a multivariate, joint modeling approach to assess the underlying dynamics of progression-free survival (PFS) components to forecast the death times of trial participants. Through Bayesian model averaging (BMA), our proposed method improves the accuracy of the overall survival (OS) forecast by combining joint models developed from each granular component of PFS. A case study of a renal cell carcinoma trial is conducted, and our method provides the most accurate predictions across all tested scenarios. The reliability of our proposed method is verified through extensive simulation studies, which include a scenario where OS is completely independent of PFS. Overall, the proposed methodology emerges as a promising candidate for reliable OS prediction in solid tumor oncology studies.

CHAPTER 1

Introduction

1.1 Background on snSMART

Conducting a randomized controlled trial (RCT), the most rigorous way of estimating the efficacy or effectiveness of treatment, can be difficult when the number of individuals affected is small. While useful in small samples, traditional methods such as crossover and N-of-1 trials have limitations that restrict their use and often encounter problems with recruitment and retention. In order to address these challenges, a new approach known as an snSMART (small sample, sequential, multiple assignment, randomized trial) has been developed in recent years.

The snSMART design is similar to a classic sequential, multiple assignment, randomized trial (SMART) design in that it first randomizes participants to one of several first-stage treatments and then conducts a second stage randomization based on the outcome of the first stage. However, the snSMART design differs in that it measures the same treatment outcome at the end of both stages and the time length of both stages is equal. Additionally, the goal of an snSMART is to identify the superior first-stage treatment using data from both stages, as opposed to a SMART's goal of identifying an effective personalized two-stage treatment sequence.

The snSMART design is particularly useful for disorders or diseases that affect a small group of people and remain stable over the duration of the trial. Additionally, the snSMART design allows for the sharing of data across stages, resulting in more precise estimates of first-stage treatment effects. However, the snSMART design will have a smaller sample size and be less flexible than a classic SMART design, requiring a different set of analytic methods.

Recent years have seen significant progress in the development of statistical methods for analyzing trial data from different snSMART designs: an snSMART with three active treatments (Wei et al. 2018, 2020, Chao, Trachtman, Gipson, Spino, Braun & Kidwell 2020), a group sequential snSMART with three active treatments (Chao, Braun, Tamura & Kidwell 2020), an snSMART with placebo and two dose levels of one treatment (Fang et al. 2021, 2023), and an snSMART with continuous outcomes (Hartman et al. 2021).

1.2 Background on External Control Data Integration

Integrating external control data in clinical trials is a topic of growing importance in the field of biostatistics. The ability to incorporate historical control data can improve the efficiency and power of clinical trials, allowing for smaller sample sizes and more precise estimation of treatment effects.

Rosenbaum & Rubin (1983) proposed using propensity score methods in observational studies to account for potential confounders. This approach has been extended to using historical control data in clinical trials (Viele et al. 2014, Pocock 1976). Propensity score methods are useful in adjusting for any observed covariates that may lead to bias in the treatment effect estimates.

Ibrahim & Chen (2000) proposed using power prior distributions for regression models, which can be useful for incorporating historical control data. Power priors allow for the incorporation of historical information in a Bayesian framework, which can help to improve the precision of treatment effect estimates.

Hobbs et al. (2011) developed hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. This approach can be useful for trials where the historical data may not be directly comparable to the current trial.

Spiegelhalter et al. (2004) discussed Bayesian approaches to clinical trials and health-care evaluation, which emphasizes the use of prior information to improve the precision of treatment effect estimates. Neuenschwander et al. (2010) proposed methods for summarizing historical information on controls in clinical trials. These methods can be used to extract relevant information from historical data to improve the power and precision of current trials.

Schmidli et al. (2014) proposed robust meta-analytic predictive (MAP) priors in clinical trials with historical control information. This approach combines historical data with current data in a meta-analytic framework to improve the precision of treatment effect estimates. Neuenschwander et al. (2016) discussed the use of co-data in clinical trials through MAC. This approach utilizes both historical and concurrent control data to improve the precision of treatment effect estimates.

Overall, there is a growing body of literature on integrating external control data in clinical trials. The use of propensity scores, power prior distributions, hierarchical models and meta-analytic approach are promising approaches for incorporating historical control data in a statistically rigorous manner. With the growing focus on efficiency and precision in clinical trials, the use of historical control data is likely to become increasingly important in the field of biostatistics. Thus far, there are no formal methods to incorporate external control data in the analysis of an snSMART design.

1.3 Summary of Objectives

None of the existing snSMART designs take into account external control data, and there is no established method for integrating control data from multiple sources in a robust manner. In Chapter 2, we aim to fill this gap by introducing a new snSMART design that incorporates external control data, allowing for a reduction in the number of participants on the placebo arm. In our proposed study design, eligible patients have a chance of 1:2:2 or 1:3:3 to be assigned to the placebo, low-dose or high-dose treatment group in the first stage of the trial. In the second stage, patients are either re-assigned or re-randomized to the same or a different dose depending on their initial treatment and the outcome of the first stage. For example, patients who received placebo in stage 1 will be re-randomized with an equal probability to either the low-dose or high-dose treatment group, regardless of their stage 1 response. Those who received the low-dose in stage 1 will continue on that dose if they responded positively or switch to the high-dose if they did not respond. Participants who first received high-dose and responded are re-randomized between low and high-dose, whereas those who did not respond to high-dose are taken off the trial in the second stage to discuss further treatment options with their physician. To utilize all external control data and information from both stages of the snSMART to enhance the accuracy of treatment effect estimations, we propose the MAC-snSMART method, which is a robust, flexible, hierarchical model to make inferences about stage 1. We simulated trials to compare results from the MAC-snSMART method to a traditional method where only stage 1 data is used for analysis. We compare the accuracy and efficacy of the treatment effect estimates. This work is published in *Biometrics Practice* (Wang et al. 2023).

In Chapter 3, we apply a trial design similar to that in Chapter 2, but with the addition of longitudinal data collected at each stage. Our goal remains the same: to efficiently estimate the stage 1 treatment effect by utilizing data from both stages of the trial and incorporating external data. This design is more versatile, accommodating a larger volume of longitudinal data from DMD natural history studies. To evaluate the treatment effect in our proposed study design, we employ a Bayesian longitudinal piecewise meta-analytic combined model (BLPM) and address between-study heterogeneities through PS trimming, IPTW weighting, and the MAC framework. BLPM is a statistical model designed for analyzing repeated measures data, accounting for correlations between measurements on the same subject. IPTW is a technique aimed at balancing covariate distributions across treatment groups, thus reducing bias from confounding variables. The MAC framework allows consideration of various sources of heterogeneity and potential selection bias when combining external control data with data from different trial stages. To assess our method's performance, we conducted both simulation studies and an example data analysis, comparing it to traditional analytic methods. These evaluations provide insights into our approach's strengths and limitations, highlighting conditions under which it excels.

The next chapter shifts focus from clinical trial design to improving the prediction of overall survival in solid tumor oncology studies. In oncology clinical trials, the choice of endpoint is a complex process. Two commonly used endpoints are PFS and OS. PFS is the length of time during and after treatment in which a patient lives without progressive disease (PD), while OS is the duration from the start of study treatment to the date of death due to any cause. OS remains the clinical gold standard for assessing patient benefit, however, powering a trial to show an OS benefit can be challenging. Factors such as longer duration of OS trials, patients crossing over to alternative treatment after progression, starting other anti-cancer therapy, or loss to follow-up can make it difficult. Additionally, OS data are often not mature enough to draw proper statistical inferences at the time of the primary analysis of PFS. Accurate prediction of OS can aid in resource allocation, future planning, and understanding the probability of success in oncology trials. It can also guide patient care and use of limited healthcare resources. In this project, we explore a model-based approach for forecasting the death times of trial participants using available, mature PFS data. We propose a multivariate joint modeling approach to assess the underlying dynamics of the PFS components (i.e., target lesion, non-target lesion, and new lesion) to predict OS. We build joint/marginal models based on OS and each component of PFS, and obtain real-time OS predictions based on each model. In total, four groups of intermediate predictions are generated. The final OS prediction is derived based on all four models simultaneously using BMA (Hoeting et al. 1999). This topic has gained interest in the statistical and clinical literature and has implications for both drug development and patient care. In addition, we enhance existing methods like the copula model and the multi-state model for predicting survival times of ongoing patients in trials in an unprecedented manner. We conducted extensive simulation studies and analyses of example data to compare the performance of the multivariate joint modeling approach with various existing models.

CHAPTER 2

Dynamic Enrichment of Bayesian Small Sample, Sequential, Multiple Assignment Randomized Trial (snSMART) Design Using Natural History Data: A Case Study from Duchenne Muscular Dystrophy

2.1 Introduction

DMD is a potentially deadly inherited genetic disease with a birth prevalence of 19.8 per 100,000 live male births (Crisafulli et al. 2020). Patients often progressively lose the ability to walk or function independently and die at a young age from lung or heart problems. There is an unmet need for effective treatments in this patient population. Given the limited number of patients affected by DMD, conducting separate dose-finding and randomized confirmatory trials is nearly impossible. Moreover, variability in disease progression, challenges in maintaining patient blindness to treatment, and ethical issues with long-term follow-up for placebo controls complicate trial design in DMD and other rare diseases (Muntoni et al. 2022). This paper aims to address some of the challenges in DMD and rare disease trial research and is motivated by the SPITFIRE trial (NCT03039686), a randomized, 2-phase, double-blind, placebo-controlled study (See Figure 2.2a) assessing the efficacy, safety, and tolerability of a new therapy (RO7239361) in 6-11 years old, ambulatory boys with DMD. We develop statistical methodology that improves the design and analysis of DMD trials using all relevant information in the trial and external control data (i.e., relevant patient information gathered from sources outside of the prospective study).

In the SPITFIRE trial, patients were equally randomized to one of two doses (low and high) of treatment or a placebo arm in stage 1. After completion of stage 1 (48 weeks after the study entry), patients in the placebo arm were randomized to either high or low dose in stage 2. The trial's primary analysis planned to use only stage 1 data which ignores critical information generated in stage 2 about the drug. Moreover, the SPITFIRE design did not allow patients who performed poorly with low dose in stage 1 to receive high dose in stage 2. Finally, the SPITFIRE trial did not formally incorporate external control data. Further innovations are necessary to consider a trial design where patients receive multiple stages of treatment and analytic methods where treatment effect estimates consider data from all stages and from external controls.

This paper proposes a new snSMART design with an efficient Bayesian analytic approach that synthesizes all relevant information and provides precise treatment effects for better decision-making about drugs in rare, life-threatening diseases like DMD. However, combining data from different sources poses unique challenges, including the choice of relevant external information, potential conflict between external and concurrent control, and "drift" in clinical outcomes between the two stages due to differences in the population. Lack of consideration of these factors may result in biased treatment effect estimates and increase false positive results. We modify existing snSMART designs and develop a model-based approach that enables the use of external control data and combines data across trial stages resulting in more precise estimates of the treatment effects. Given the small sample size, the proposed model must be parsimonious and robust when there is conflict between different data sources.

2.2 SPITFIRE Trial: A Motivating Example of Phase IIb/III Trial in DMD

We use the SPITFIRE trial setting and published results to demonstrate the utility of the proposed design and analytic approaches. Though SPITFIRE was stopped early for futility, it reflects the development paradigm for a new drug in many rare diseases. For example, a similar design was used for the ATMOSPHERE (NCT04704921) study. Note that, we have no intention to comment on or judge the clinical activity of the treatments involved or any decisions by sponsors or regulators associated with the trial or compound.

In the SPITFIRE trial, a total of 166 participants were randomized to receive one of two doses of RO7239361 or placebo (1:1:1) for 48 weeks at stage 1. After completion of stage 1, subjects entered an open-label stage where participants who originally received placebo were randomized into low or high dose groups, and all other participants stayed on their original treatment for up to 192 weeks. The primary and one of the secondary outcomes were the change (from baseline to week 48) of stage 1 North Star Ambulatory Assessment (NSAA) total score and 6-minute walk distance (6MWD) test.

The original SPITFIRE design is similar to an snSMART (Tamura et al. 2016) but lacks many fundamental features. SPITFIRE a) re-randomized only the participants in the control group in stage 2 and b) used only stage 1 data in the primary analysis. This differs from the snSMART

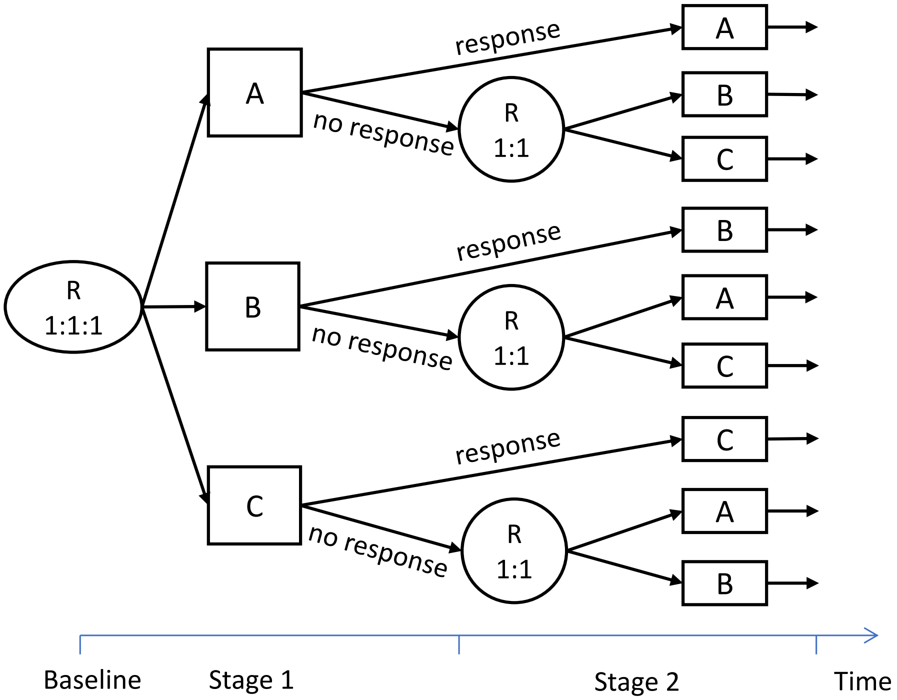


Figure 2.1: Study design of a standard snSMART. In the study, participants are randomly assigned to one of three treatment options, referred to as arms A, B, and C, in a 1:1:1 ratio. They are then monitored for a specified period of time. Stage 1 responders will continue to receive the same treatment in stage 2, while the non-responders will be re-randomized to one of the other two treatment options in stage 2.

design used in the ARAMIS study (NCT02939573; as shown in Figure 2.1; Tamura et al. (2016)), where data from both stages were used to estimate stage 1 treatment effects. There has been notable progress in snSMART methods (Wei et al. 2018, 2020, Hartman et al. 2021) that have been coded into an R package (Wang & Kidwell 2022). Other snSMART extensions include a group-sequential design that allows early stopping of an arm (Chao, Braun, Tamura & Kidwell 2020) and inclusion of multiple dose levels (Fang et al. 2021, 2023).

In recent years, there has been an increased interest to consider the use of external control data to expedite the development of a new drug in rare diseases with an unmet medical need. Including external control data allows more patients to receive potentially more effective treatments than a placebo aiding in recruitment and retention, a smaller sample size, and therefore faster development. However, heterogeneity between external control and concurrent trial data is often a limiting factor in real-life applications. Yet, many approaches have been proposed to leverage external data in clinical trial while addressing heterogeneity between data sources, different types of available data (i.e., individual patient data or aggregated data), and outcomes of interest. Bayesian approaches for borrowing external control information differ in terms of assumptions regarding the relevance and exchangeability of external data with current trial data (Wadsworth et al. 2018). Use of the power prior approach directly down-weights the external control using fixed weights (Ibrahim & Chen 2000). Dynamic methods adjust the informativeness of the prior based on measures of conflict between external control and the new trial data. Notable dynamic methods include normalized power prior (Duan 2005, Neuenschwander et al. 2009), commensurate priors (Hobbs et al. 2011), and robust meta-analytic-predictive priors (Spiegelhalter et al. 2004, Neuenschwander et al. 2010, Schmidli et al. 2014, Neuenschwander et al. 2016). In the context of basket trials, Ouma et al. (2022) explored Bayesian treatment effect borrowing and treatment response borrowing models that can be expanded to enable external control data borrowing. In addition to Bayesian methods, frequentist methods such as propensity score based matching (Rosenbaum & Rubin 1983), stratification, and inverse probability weighting (Lin et al. 2018) are widely used when aggregate level information and baseline covariates are available.

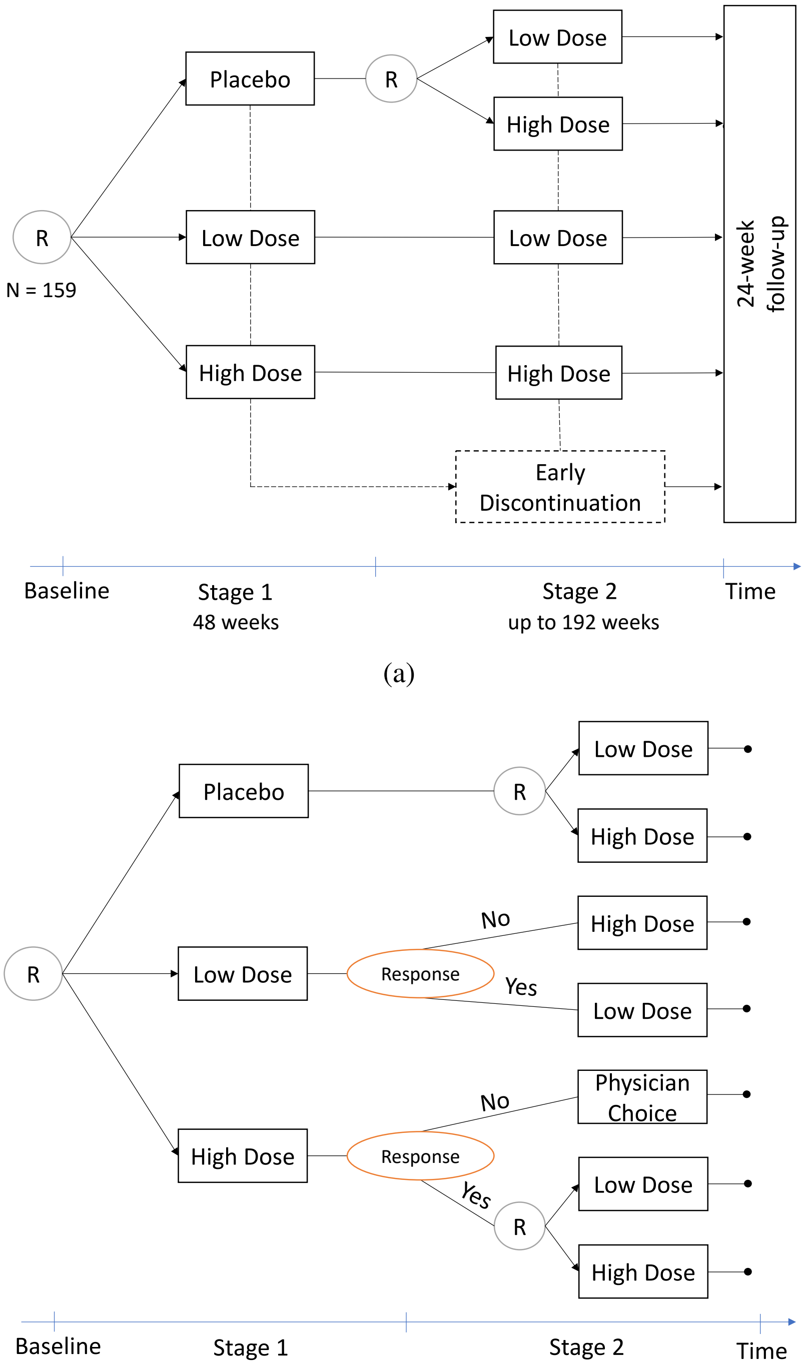
Existing snSMART designs and methods do not incorporate external control data. In DMD, the Duchenne Natural History Study (DNHS) conducted by the Cooperative International Neuromuscular Research Group (CINRG), which along with the PRO-DMD-01 Prospective Natural History Study (NCT01753804) and the University College London Natural History Study (NCT02780492) provide rich, external control information that can be combined with concurrent trial placebo data. Motivated by the DMD setting, we aim to fill this gap with an innovative snSMART design and Bayesian model as an alternative to the current practice of DMD and rare disease drug development. The new approach proposes three key improvements to current snSMART and rare disease design and methods: a) use of external control data to reduce the sample size of the placebo arm, b) allow stage 1 low dose nonresponders to receive higher dose in stage 2, similar to the design proposed by Fang et al. (2021, 2023), and c) use a Bayesian hierarchical model that dynamically incorporates external control data and borrows information across both trial stages. These features are extremely attractive in the rare disease setting where sample sizes and opportunities to perform clinical trials are limited. Our proposed design and methods assume a) the study condition remains relatively stable throughout the trial period (i.e., when there is no treatment effect, patients' primary outcomes do not fluctuate dramatically), b) an adequate washout period between the two trial stages, c) exchangeable treatment effects between stages, and d) similar external control and placebo effects. An snSMART design is not appropriate for conditions that are not stable during the trial period, and we offer modifications to the model and sensitivity analyses to address other violations of these model assumptions.

2.2.1 Proposed Modification for SPITFIRE Trial Design

Our proposed snSMART design is shown in Figure 2.2b, where eligible patients are randomized unequally (e.g., 1:2:2 or 1:3:3) between placebo, low dose, and high dose (e.g., of RO7239361) in stage 1. After 48 weeks, participants are re-assigned or re-randomized to either the same or a different dose of treatment depending on their initial treatment and post-baseline NSAA total score. Here, we define a participant as a treatment responder at week 48 if their post-baseline NSAA total score increases, stays the same, or does not decrease by more than 3.1 (Muntoni et al. 2018). Participants who received placebo in stage 1 are re-randomized with equal probability to either the low dose or high dose treatment arm in stage 2, regardless of their stage 1 response. This is beneficial to participants in the trial because everyone receives a dose of the treatment. This design differs slightly from the ones proposed by Fang et al. (2021, 2023) in stage 2 for those who received low or high dose in stage 1. These slight modifications may make the design more patient-centered. Participants who received low dose in stage 1 are assigned to stay on low dose if they responded in stage 1 or switch to high dose if they did not respond. Participants who first received high dose and responded are re-randomized equally between low and high doses. However, the nonresponders to high dose are discontinued in stage 2. In most settings, the re-randomization of high dose responders is a viable design option because low dose may continue to be effective and possibly more tolerable. On the other hand, when the high dose proves ineffective for some participants, it is unlikely that a lower dose will yield better results for them. Additionally, administering an ineffective high dose that potentially poses higher toxicity is not ethical. Thus, we chose to exclude stage 1 high-dose nonresponders in stage 2.

2.3 Methods

The following notation is used in this section. Y_{jk} and θ_{jk} denote the observed and underlying true mean change in outcome (e.g., NSAA score) from the baseline values for stage $j = 1, 2$ and treatment $k = p$ (placebo), l (low dose), h (high dose), respectively. Y_i' and θ_i' ($i = 1, 2, ..., n_c$) denote the observed and true placebo mean change in outcome from baseline for external control data. For example, in the SPITFIRE trial, θ_{1p}, θ_{1l}, and θ_{1h} represent the true mean change from baseline in the NSAA total score or 6MWD for placebo, low, and high dose groups, respectively. The key estimands of interest are the stage 1 differences in effects between each dose and placebo ($\theta_{1l} - \theta_{1p}$, $\theta_{1h} - \theta_{1p}$). Traditional analyses for Y_{1k} (change from baseline in NSAA or 6MWD) in SPITFIRE or similar studies include analysis of variance (ANOVA), analysis of covariance (ANCOVA), or a conjugate Bayesian model (normal data and normal prior). Note that a traditional analytic method excludes external control data and uses only stage 1 outcomes.



(a)

(b)

Figure 2.2: (a) Study design of the SPITFIRE trial (NCT03039686). (R) denotes randomization. (b) Study design of the proposed snSMART design that formally incorporates external control data. Participants are randomized (R) with 1:2:2 or 1:3:3 chances of receiving placebo, low dose, or high dose, respectively, in stage 1. At the end of stage 1, participants are assigned or re-randomized to their stage 2 treatment based on their stage 1 treatment and response status. Outcomes are collected at the end of stage 1 and 2.

2.3.1 Bayesian joint stage model

A modified version of the existing Bayesian joint stage model (BJSM) by Fang et al. (2023) may be used to analyze our snSMART design. Though it uses data from both stages, the existing BJSM has not previously been presented to formally incorporate external data. One indirect or crude way to incorporate external control data is to use informative normal distribution priors for the model parameter associated with the placebo effect. For the BJSM, normal distribution prior parameters may be derived using the method of moments approximation using external data, where the variance is further adjusted to ensure the desired prior effective sample size.

Here, we use the summary-level DNHS data to demonstrate the derivation of normal distribution prior parameters for BJSM:

Data source: Carefully matched DNHS data as external controls.
Mean test outcome change (from baseline to week 48):

	NSAA Score	6MWD
Mean	-1.04	-22.36
SD	0.77	27.98

Step 1:

According to method of moments approximation, the normal priors for true placebo mean change in outcome (θ_{1p}) should be $N(\mu = -1.04, sd = 0.77)$ under NSAA Score, and $N(\mu = -22.36, sd = 27.98)$ under 6MWD.

Step 2:

Decide on an effective sample size (ESS) we want to use. If ESS = E, then the final normal priors for placebo should be $N(\mu = -1.04, sd = 0.77/\sqrt{E})$ under NSAA Score, and $N(\mu = -22.36, sd = 27.98/\sqrt{E})$.

2.3.2 Meta-analytic combined (MAC) approaches

The meta-analytic combined, or MAC-snSMART, approach is a novel and unified analytic framework that incorporates "all" relevant information for efficient estimation of stage 1 treatment effects, θ_{1p}, θ_{1l}, and θ_{1h}. We use a Bayesian hierarchical model framework to dynamically borrow information from different sources (e.g., external control, stage 2). The framework consists of a series of models that link different sets of parameters (as shown in Figure 2.3). However, implementation of such a framework in the snSMART setting requires innovation to a) handle heterogeneity between different sources of placebo data, b) account for the potential conflict between internal and external placebo data despite careful selection of external data, and c) account for selection bias due to

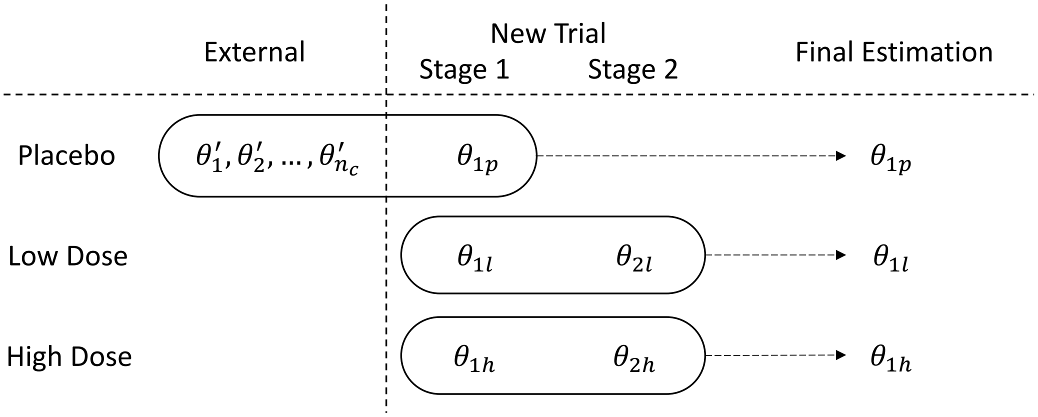


Figure 2.3: Borrowing among treatment effects (θ) in MAC-snSMART

non-randomized treatment assignments in stage 2 for participants who received low dose or did not respond to high dose in stage 1.

2.3.2.1 Use of external information for placebo

Controlling for potential bias is a primary concern when considering external control data in a clinical trial. To handle bias due to potential unmeasured confounders in the design phase and increase the validity of results, statistical methods should be prespecified. To account for measured confounders, a first and critical step is careful selection of relevant external control data. Pocock (1976) and Lim et al. (2018) proposed a set of criteria to assess the similarity between external and trial control data based on key aspects, including inclusion/exclusion criteria, endpoint definition, relevance of control treatment, distribution of demographic criteria, etc. Furthermore, matching techniques (e.g., propensity score matching) can be conducted to select external control patients used in the analysis to ensure appropriate balance of important prognostic factors. Irrespective of careful selection, the clinical outcome of interest may be variable across different sources due to associated intrinsic and extrinsic factors. A model-based approach is required to address these issues.

For illustration we use summary level, external control data for this paper. However, the methodology is also applicable to patient-level data with important baseline covariates. Let, Y'_i be the observed mean change from baseline for external placebo data source i. Furthermore, $Y'_i|\theta'_i \sim N(\theta'_i, s'^2_i)$, where $i = 1, 2, ...n_c$, and s'_i is known or derived from the n_c external control data samples. Similarly, we assume $Y_{1p}|\theta_{1p} \sim N(\theta_{1p}, s^2_{1p})$. For dynamic borrowing across different sources, we need a model that connects θ'_i and θ_{1p}. A generic approach that connects θ'_i and θ_{1p} is a generalized hierarchical model which assumes exchangeability between internal and external control. This implies $\theta'_i|\mu_p, \tau'_i \sim N(\mu_p, \tau'^2_i)$ and $\theta_{1p}|\mu_p, \tau_{1p} \sim N(\mu_p, \tau^2_{1p})$. If patient-level information about important predictors is available, the model can be extended via meta-regression. In contrast to the classical exchangeability assumption which assumes the same variance for θ'_i and θ_{1p}, this structure accounts for external control data outliers by allowing for different between-trial standard deviations. However, the exchangeability assumption results in overly optimistic borrowing and causes biased estimates when information from different sources conflicts.

The amount of information borrowed from external control data while estimating the treatment effects for control, low, and high dose groups can be quantified. This amount of borrowed information can be expressed as the ESS. Here, we use the expected local-information-ratio, which fulfills a basic predictive consistency requirement, as introduced by Neuenschwander et al. (2020). We, like Neuenschwander et al. (2020), recommend that the ESS of each external control data set should not exceed the number of participants on the placebo arm in the trial.

2.3.2.2 Robustification

To avoid overly optimistic borrowing, we include a mixture model adopted from Neuenschwander et al. (2016) for θ'_i and θ_{1p}. The proposed model allows for non-exchangeability of any of the parameters associated with the placebo effect:

$$\theta'_i \sim p_i * N(\mu_p, \tau'^2_i) + (1 - p_i) * N(m'_i, v'^2_i)$$
$$\theta_{1p} \sim p_{1p} * N(\mu_p, \tau^2_{1p}) + (1 - p_{1p}) * N(m_{1p}, v^2_{1p})$$

p_i and p_{1p} are fixed a priori to reflect confidence about the similarity between external and concurrent control data. The values of m'_i, m_{1p}, v'_i and v_{1p} are chosen based on expert knowledge. Weakly informative priors, such as priors worth approximately one observation (Kass & Wasserman 1995), are used for non-exchangeability parameters (m'_i, m_{1p}, v'_i, and v_{1p}).

The mixture model introduced in here is a mixture of normal priors. For θ_{1p}, we have $f(\theta_{1p}|p_{1p}, \mu_p, \tau_{1p}, m_{1p}, v_{1p}) = p_{1p}N(\theta_{1p}|\mu_p, \tau^2_{1p}) + (1 - p_{1p})N(\theta_{1p}|m_{1p}, v^2_{1p})$. Given that only summary-level treatment effect Y_{1p} is used to conduct Bayesian inference, the likelihood function is the same as above. Here, we introduce the latent variable Z_{1p}, such that $P(Z_{1p} = 1) = p_{1p}$ and $P(Z_{1p} = 0) = 1 - p_{1p}$, where p_{1p} is the probability that θ_{1p} is fully exchangeable with external controls, and $1 - p_{1p}$ is the probability that θ_{1p} is non-exchangeable with external controls. The robustification component for stage 1 placebo arm can be expressed as $\theta_{1p}|(Z_{1p} = 1) \sim N(\mu_p, \tau^2_{1p})$ and $\theta_{1p}|(Z_{1p} = 0) \sim N(m_{1p}, v^2_{1p})$. Therefore, Z_{1p} provides the label indicating the mixture components from which the observations have been generated. Now the posterior probability that θ_{1p}

has been generated from $N(\mu_p, \tau_{1p}^2)$ is:

$$P(Z_{1p} = 1|\theta_{1p}, \mu_p, \tau_{1p}, m_{1p}, v_{1p}) = \frac{p_{1p}N(\mu_p, \tau_{1p}^2)}{p_{1p}N(\mu_p, \tau_{1p}^2) + (1 - p_{1p})N(m_{1p}, v_{1p}^2)}.$$

Similarly, we could create latent variables Z_i for θ_i, where $i = 1, 2, ...n_c$, and use the same structure as above to obtain the posterior probability that θ_i' has been generated from $N(\mu_p, \tau_i'^2)$.

Based on conjugate priors, the posterior distribution of θ_i' and θ_{1p} are mixtures of normal distributions. Specifically, for θ_{1p}, when equation 1 is not considered, the posterior mean is:

$$p_{1p}(\eta_{1p}\hat{\mu}_p + (1 - \eta_{1p})Y_{1p}) + (1 - p_{1p})\frac{m_{1p}(s_{1p}^2 + \tau_{1p}) + Y_{1p}v_{1p}^2}{s_{1p}^2 + \tau_{1p} + v_{1p}^2}$$

where $\hat{\mu}_p = (\omega_{1p}Y_{1p} + \sum_i \omega_i' Y_i')/\omega_{sum}$ is the posterior mean of μ_p, $\omega_i' = (s_i'^2 + \tau_i'^2)^{-1}$ and $\omega_{1p} = (s_{1p}^2 + \tau_{1p})^{-1}$ are inverse variance weights, ω_{sum} is the sum of all inverse variance, and $\eta_{1p} = s_{1p}^2/(s_{1p}^2 + \tau_{1p})$.

We refer to the method as "robust MAC-snSMART" if the robustification component is included in addition to the structure outlined in Section 2.3.2.1. If this component is not included ($p_i = p_{1p} = 1$), the method is simply referred to as "MAC-snSMART".

2.3.2.3 Combining evidence for low and high dose groups

According to the proposed snSMART design, a participant may follow one of seven different treatment sequences: $(1p, 2l)$, $(1p, 2h)$, $(1l, 2l)$, $(1l, 2h)$, $(1h, 2l)$, $(1h, 2h)$, and ($1h$, *No stage 2 treatment*). For each of the first six groups, the joint distribution of outcomes from stages 1 and 2, Y_{1k} and $Y_{2k'}$, follows:

$$\begin{bmatrix} Y_{1k} \\ Y_{2k'} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \theta_{1k} + B \\ \theta_{2k'} \end{bmatrix}, \begin{bmatrix} s_{1k}^2 & s_{kk'}s_{1k}s_{2k'} \\ s_{kk'}s_{1k}s_{2k'} & s_{2k'}^2 \end{bmatrix} \right), \tag{2.1}$$

where $k = p, l, h$; $k' = l, h$. s_{jk} can be estimated based on observed data in the snSMART. $s_{kk'}$ denotes the correlation between stage 1 treatment k outcomes and stage 2 treatment k' outcomes. Note, B is a selection bias correction term. This selection bias is due to design since those who receive low dose in stage 1 are not re-randomized and those who do not respond to high dose in stage 1 are excluded in stage 2. Given that we know which treatment sequences the participants follow, to account for the difference from the stage 1 mean, the bias correction term is defined as: $B = I(k = l, k' = l)B_{ll} - I(k = l, k' = h)B_{lh} - I(k = h)B_h$. Explicitly, B_{ll} denotes the expected difference between the stage 1 mean treatment effect of group $(1l, 2l)$ and the overall stage 1 low dose mean treatment effect, B_{lh} denotes the expected difference between the stage 1 mean treatment effect of group $(1l, 2h)$ and the overall stage 1 low dose mean treatment effect, and B_h denotes the expected difference between the stage 1 mean treatment effect of high dose responders and the overall stage 1 high dose mean treatment effect. b_{ll}, b_{lh} and b_h, the observed differences corresponding to B_{ll}, B_{lh} and B_h, follow normal distributions $N(B_{ll}, \xi_{ll}^2)$, $N(B_{lh}, \xi_{lh}^2)$ and $N(B_h, \xi_h^2)$, respectively. A similarly structured but more complex bias-corrected meta-analysis model has been proposed by Verde (2021).

Under the MAC-snSMART approach, the conditional distribution of θ_{jk} follows $\theta_{1k}, \theta_{2k} \sim N(\mu_k, \tau_k^2)$, where $k = l, h$, given that we assume stage 1 and stage 2 expected outcomes for the same treatment follow the same normal distribution. That is, the treatment effects of the same dose level are fully exchangeable across trial stages. Since the snSMART design is most appropriate for relatively stable conditions and requires a washout period between stages, we believe this assumption of exchangeable treatment effects is valid in rare disease settings where an snSMART would be applied. In cases where stage-wise non-exchangeability is likely to occur, the robustification component described in Section 2.3.2.2 can be easily incorporated into $\theta_{2k'}$ in Formula 2.1 to account for non-exchangeable treatment effects across trial stages.

2.3.2.4 Prior specification

We suggest generalizable, weakly informative normal distributions as priors for μ_p and μ_k, i.e., priors that are worth approximately one observation (Kass & Wasserman 1995). To ensure the identifiability of the bias parameters B_{ll}, B_{lh} and B_h, we use weakly informative uniform distributions or normal distributions that cover all possible bias values as their priors (Verde 2021). We suggest using non-informative $Unif(-1, 1)$ prior distributions for $s_{kk'}$ since the correlation between stage 1 treatment k and stage 2 treatment k' outcomes ranges from -1 to 1 and is uncertain.

Half-normal priors with standard deviations roughly equal to $s_i'/2$, $s_{1p}/2$ and $s_{jk}/2$ for τ_i', τ_{1p} and τ_k, respectively, are used to cover very small to large between-trial heterogeneity (Spiegelhalter et al. 2004). According to the bias-corrected meta-analysis model proposed by Verde (2021), roughly four participants worth of information is provided by the priors of the bias parameter. Hence, we recommend half-normal priors with standard deviation roughly equal to $s_{1l}/4$, $s_{1l}/4$ and $s_{1h}/4$ for ξ_{ll}, ξ_{lh} and ξ_h, respectively. Then, the treatment effects of low dose in stage 1 and stage 2 follow the same normal distribution and are therefore exchangeable. Similarly, the high dose treatment effects in stages 1 and 2 are exchangeable.

2.4 Simulation Settings

We assess the sensitivity of our proposed MAC-snSMART methods under various data generating settings, treatment effects, sample sizes, and lack of exchangeability between external control and current snSMART data. We assess two different data generation processes, the first one matches with the proposed model, and the second one allows violation of exchangeability between stage 1 and stage 2. The first data generation process follows the assumption of the MAC-snSMART so that θ_{1k} is set to a predetermined treatment effect and θ_{2k} is randomly generated from $N(\theta_{1k}, 0.25^2)$. Thus, the summary level stage 1 treatment outcomes are generated based on $N(\theta_{jk}, 0.5^2)$, the stage 2 outcomes are randomly generated according to formula 2.1, and $s_{kk'}$ are randomly chosen within a certain range, i.e., $s_{pl} \in (-0.20, 0.20)$, $s_{ph} \in (-0.15, 0.15)$, $s_{ll} \in (0.70, 1)$, $s_{lh} \in (-0.50, 0)$, $s_{hl} \in (-0.30, 0.30)$, and $s_{hh} \in (0.70, 1)$. In the second data generation process θ_{1k} and θ_{2k} are not exchangeable. We set θ_{1k} to a predetermined treatment effect, and let $\theta_{2k} = \theta_{1k} + 1$. Formula 2.1 is used to randomly generate stage 1 and stage 2 treatment outcomes. This type of data may result from an snSMART where a washout period between stages 1 and 2 is inadequate.

In addition to testing different data generation processes, we investigate the performance of our proposed models considering four treatment effect scenarios. Here, considering the DMD context, we use an NSAA score of four as the threshold for categorizing responders and nonresponders at the end of stage 1. In scenario 1, we assume that the new drug is ineffective ($\theta_{1p} = \theta_{1l} = \theta_{1h} = 0$). In scenario 2, we assume that only the high dose is effective ($\theta_{1p} = \theta_{1l} = 0 < \theta_{1h} = 6$). For scenario 3, the low dose has a small, but not clinically meaningful treatment effect and high dose has a clinically meaningful treatment effect ($\theta_{1p} < \theta_{1l} = 2 < \theta_{1h} = 6$). In contrast, both low and high doses are assumed to have a clinically meaningful effect for scenario 4 ($\theta_{1p} = 0 < \theta_{1l} = 4 < \theta_{1h} = 8$). The last two scenarios assess the sensitivity of our model to the alignment of external control and current data (scenario 5: $\theta_i' \neq \theta_{1p}$ for some i; scenario 6: $\theta_i' \neq \theta_{1p}$ for all i). A summary of all simulation scenarios can be found in Table 2.1.

Under each data generating process, 10,000 realizations per scenario were simulated. Each realization includes five sets of summary-level (mean and standard deviation) external control data and one concurrent snSMART data set. Estimations from the MAC-snSMART, robust MAC-snSMART, traditional analysis, and BJSM are compared. We calculate coverage rate (CR) , rMSE, average bias, and average width of the 95% credible interval (CI) of each estimator. Note that the Monte Carlo error for all average bias and CI width is less than 0.005. Finally, type I error (under scenario 1) and power (under scenarios 2-6) of different methods are also calculated using the probability that the 95% credible intervals of $\widehat{\theta}_{1l} - \widehat{\theta}_{1p}$ and $\widehat{\theta}_{1h} - \widehat{\theta}_{1p}$ do not include 0. Results are provided for sample size of 25 and 50 randomized with a 1:2:2 ratio to the placebo, low dose, and high dose arms respectively. All computations are done via the R function `jags` in R package

Table 2.1: Simulation parameters and scenarios.

θ'_i *denote the expected mean placebo treatment effect in external control data i, where* $i = 1, 2, ...n_c$. θ_{jk} *is treatment effect of treatment* k *in stage* j*, where* $j = 1, 2, k = p, l, h$*,* p *= placebo,* l *= low dose, and* h *= high dose. External control data and patient-level treatment outcomes in both stages are generated according to the formula introduced in the Method section of the manuscript.*

	Data Generating Process 1	Data Generating Process 2
Scenario 1	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 0, \theta_{2l} \sim N(0, 0.25),$ $\theta_{1h} = 0, \theta_{2h} \sim N(0, 0.25)$	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 0, \theta_{2l} = 1,$ $\theta_{1h} = 0, \theta_{2h} = 1$
Scenario 2	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 0, \theta_{2l} \sim N(0, 0.25),$ $\theta_{1h} = 6, \theta_{2h} \sim N(6, 0.25)$	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 0, \theta_{2l} = 1,$ $\theta_{1h} = 6, \theta_{2h} = 7$
Scenario 3	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 2, \theta_{2l} \sim N(2, 0.25),$ $\theta_{1h} = 6, \theta_{2h} \sim N(6, 0.25)$	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 2, \theta_{2l} = 3,$ $\theta_{1h} = 6, \theta_{2h} = 7$
Scenario 4	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} \sim N(4, 0.25),$ $\theta_{1h} = 8, \theta_{2h} \sim N(8, 0.25)$	$\theta'_i \sim N(0, 0.25), \theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} = 5,$ $\theta_{1h} = 8, \theta_{2h} = 9$
Scenario 5	$\theta'_1, \theta'_2, \theta'_3 \sim N(0, 0.25),$ $\theta'_4, \theta'_5 \sim N(1, 3),$ $\theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} \sim N(4, 0.25),$ $\theta_{1h} = 8, \theta_{2h} \sim N(8, 0.25)$	$\theta'_1, \theta'_2, \theta'_3 \sim N(0, 0.25),$ $\theta'_4, \theta'_5 \sim N(1, 3),$ $\theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} = 5,$ $\theta_{1h} = 8, \theta_{2h} = 9$
Scenario 6	$\theta'_i \sim N(3, 1), \theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} \sim N(4, 0.25),$ $\theta_{1h} = 8, \theta_{2h} \sim N(8, 0.25)$	$\theta'_i \sim N(3, 1), \theta_{1p} = 0,$ $\theta_{1l} = 4, \theta_{2l} = 5,$ $\theta_{1h} = 8, \theta_{2h} = 9$

Table 2.2: Simulated type I error. Note: The type I error of all presented methods is defined as the probability that the credible intervals of $\widehat{\theta}_{1l} - \widehat{\theta}_{1p}$ and $\widehat{\theta}_{1h} - \widehat{\theta}_{1p}$ do not include 0 when there are no treatment effect differences between low dose and placebo and high dose and placebo. Data generating process 1 generates data sets under the assumption that the expected treatment effect of the same dose level follows the same normal distribution across stages, and data generating process 2 generates data sets without a hierarchical structure or exchangeability between stages. N denotes the total number of participants in the trial. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) are compared against the traditional method and the Bayesian Joint Stage Model (BJSM)

Sample Size	Traditional		BJSM		MS		RMS	
	Low	High	Low	High	Low	High	Low	High
Data generating process 1								
$N = 50$	0.070	0.059	0.032	0.028	0.039	0.018	0.042	0.022
$N = 25$	0.077	0.076	0.035	0.031	0.056	0.026	0.059	0.030
Data generating process 2								
$N = 50$	0.057	0.059	0.031	0.026	0.061	0.056	0.065	0.060
$N = 25$	0.076	0.080	0.036	0.030	0.070	0.053	0.075	0.058

`rjags` (Plummer 2022).

2.4.1 Results

For the data generation process where assumptions of exchangeability are upheld (data generating process 1), the rMSE, average bias, CR , and average CI width for estimators of the expected treatment effects are shown in the left columns of Figure 2.4, 2.5, 2.6, and 2.7 respectively. The estimators from the robust MAC-snSMART method have smaller rMSEs than the traditional method and BJSM. The robust MAC-snSMART method provides similar coverage compared with the traditional method while having smaller 95% CI width. Even though the robust MAC-snSMART method has slightly higher average biases, the biases are negligible.

A comparison of all metrics among the traditional method, BJSM, and the MAC-snSMART methods clearly shows that estimation of the placebo effect is improved by including external control data, even when external control data are not entirely aligned with the placebo data in the current snSMART ($\theta_i' \neq \theta_{1p}$ for some i). However, when external control data are not completely aligned with the current trial data (simulation scenarios 5 and 6), the placebo effect estimate is less biased under the robust MAC-snSMART method. Scenario 6 verifies that the robustification component introduced in Section 2.3.2.2 is effective in avoiding overly optimistic external control borrowing, even when the external controls are completely unaligned with the current trial data. The robust MAC-snSMART and the MAC-snSMART methods estimate low and high dose effects adequately under all scenarios.

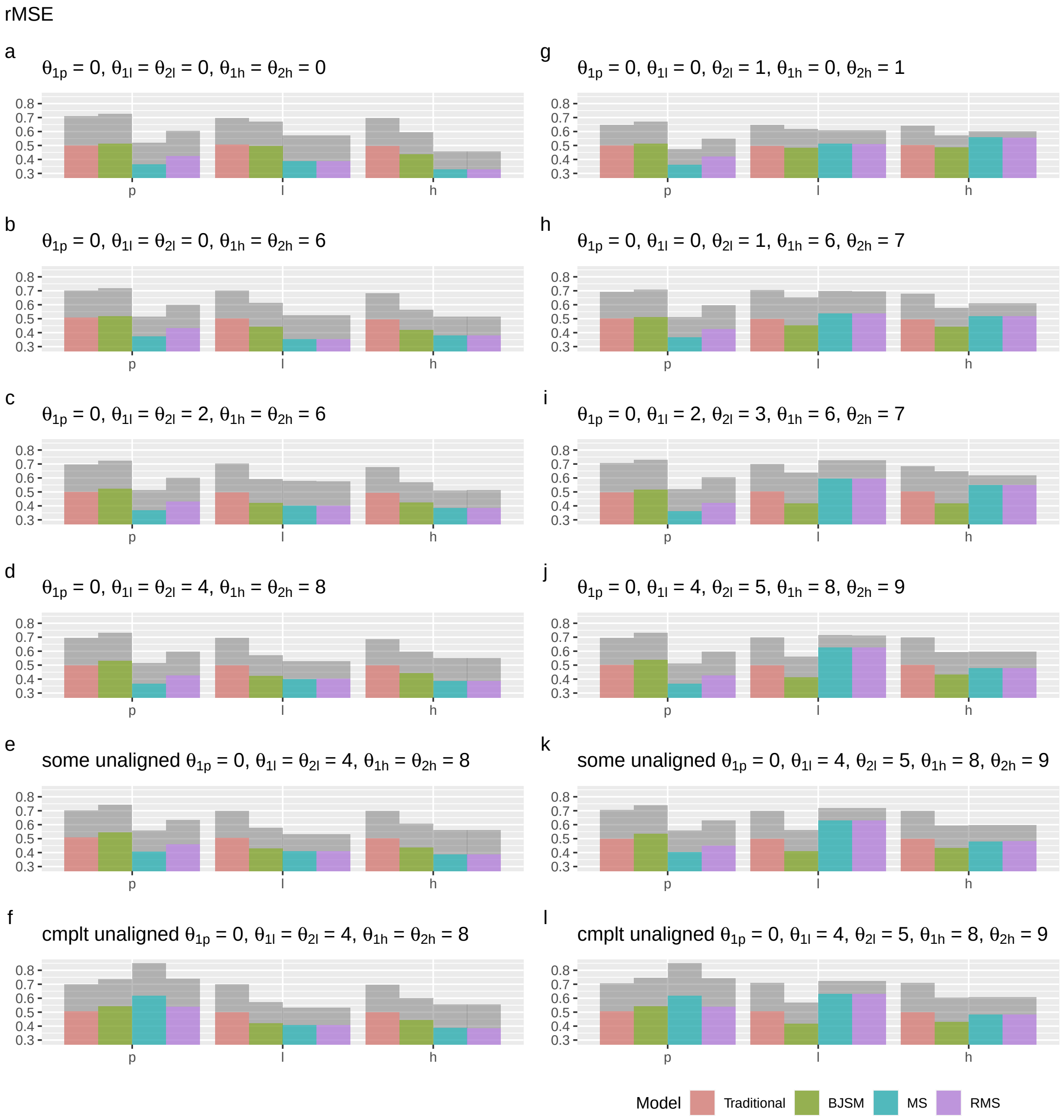


Figure 2.4: Simulated root-mean-square error (rMSE) for the estimators of θ_{1k}
θ_{jk} is the stage j treatment effect of treatment k, where $j = 1,2, k = p,l,h$, p = placebo, l = low dose, and h = high dose. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) methods are compared against the traditional method and the Bayesian Joint Stage Model (BJSM). The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying gray bars. The simulation settings are described on the top of each graph, where θ_{jk} denotes the true value of the expected treatment effects of treatment k in stage j, $j = 1,2$, $k = p,l,h$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. This figure appears in color in the electronic version of this article, and color refers to that version.

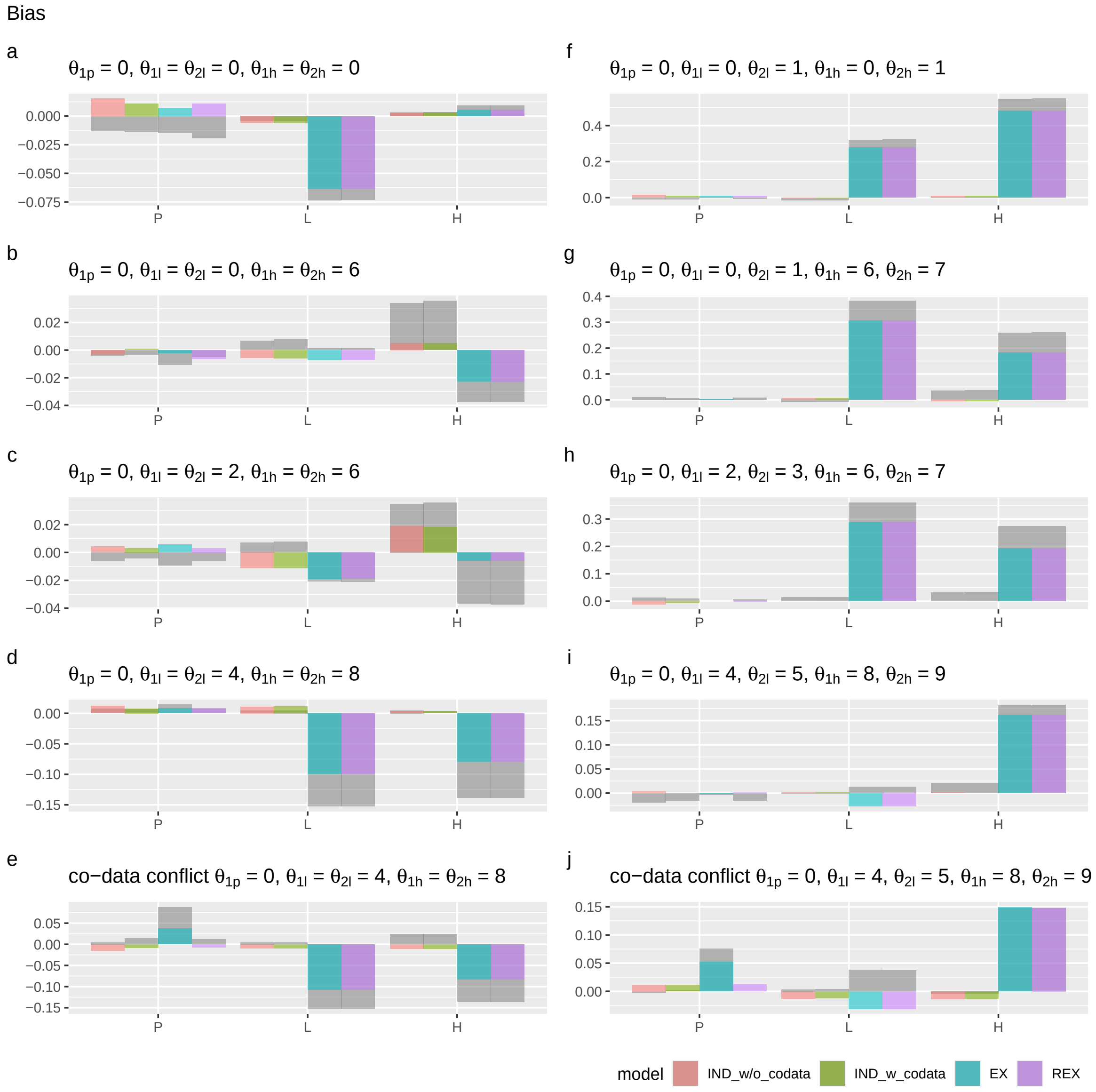


Figure 2.5: Simulated bias for the estimators of θ_{1k}.
θ_{jk} is the stage j treatment effect of treatment k, where $j = 1, 2, k = p, l, h$, p = placebo, l = low dose, and h = high dose. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) are compared against the traditional method and the Bayesian Joint Stage Model (BJSM). The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where θ_{jk} denotes the true value of the expected treatment effects of treatment k in stage j, $j = 1, 2$, $k = p, l, h$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

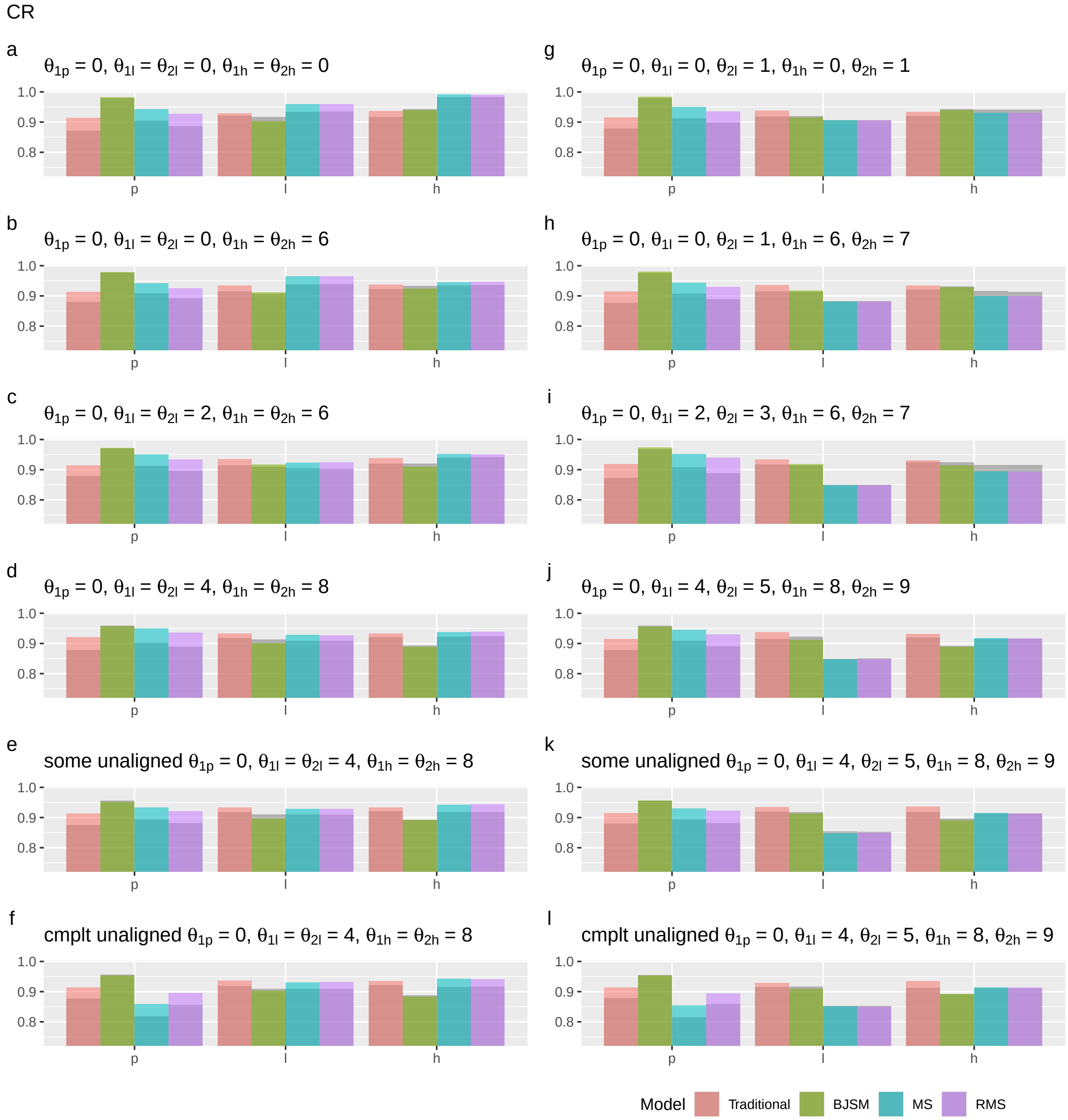


Figure 2.6: Simulated 95% coverage rate (CR) for the estimators of θ_{1k}.
θ_{jk} is the stage j treatment effect of treatment k, where $j = 1, 2, k = p, l, h$, p = placebo, l = low dose, and h = high dose. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) are compared against the traditional method and the Bayesian Joint Stage Model (BJSM). The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where θ_{jk} denotes the true value of the expected treatment effects of treatment k in stage j, $j = 1, 2$, p, l, h, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial.

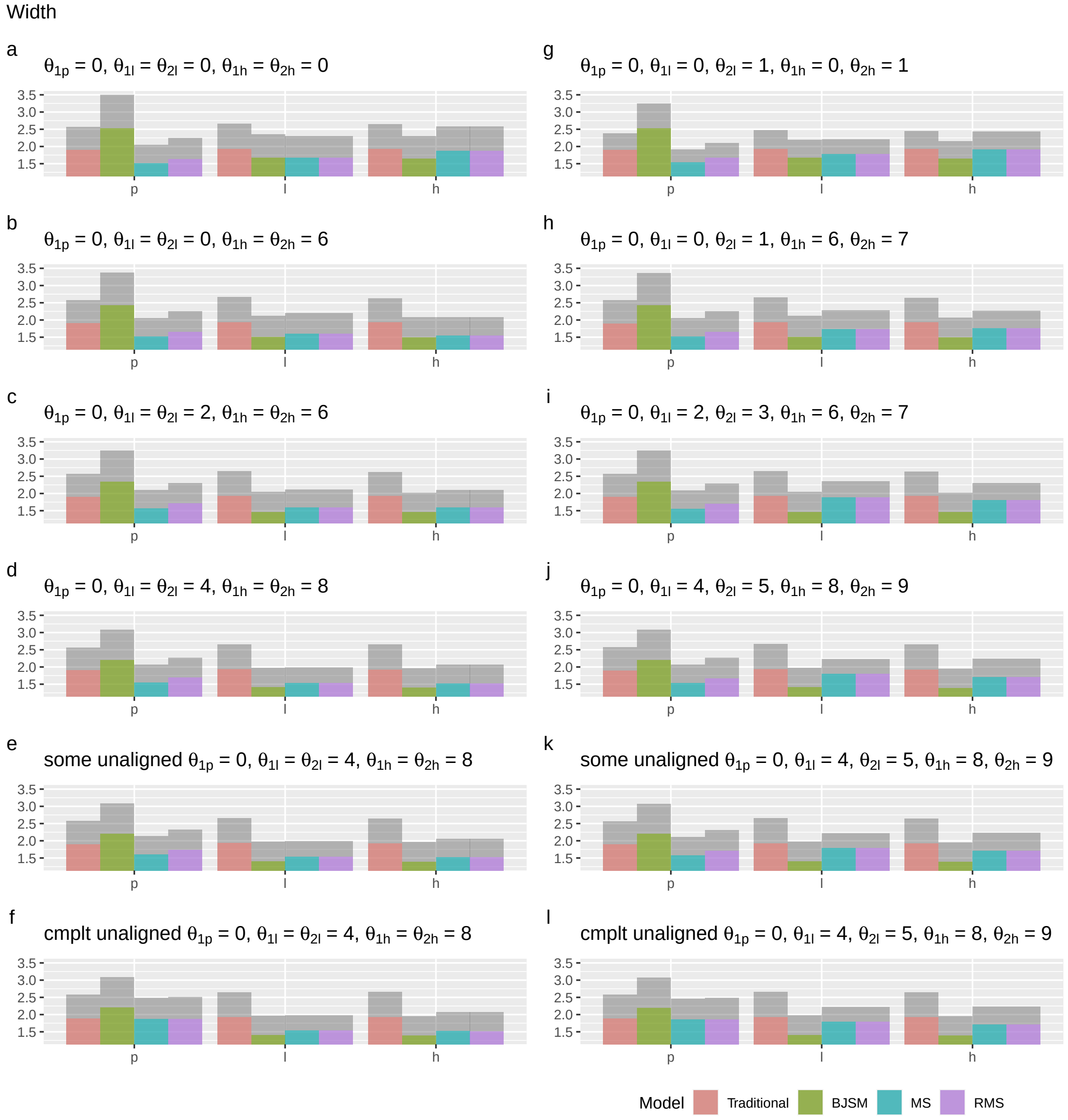


Figure 2.7: Simulated width for the estimators of θ_{1k}.
θ_{jk} is the stage j treatment effect of treatment k, where $j = 1, 2, k = p, l, h$, p = placebo, l = low dose, and h = high dose. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) are compared against the traditional method and the Bayesian Joint Stage Model (BJSM). The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where θ_{jk} denotes the true value of the expected treatment effects of treatment k in stage j, $j = 1, 2$, $k = p, l, h$, and *external unaligned* means the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

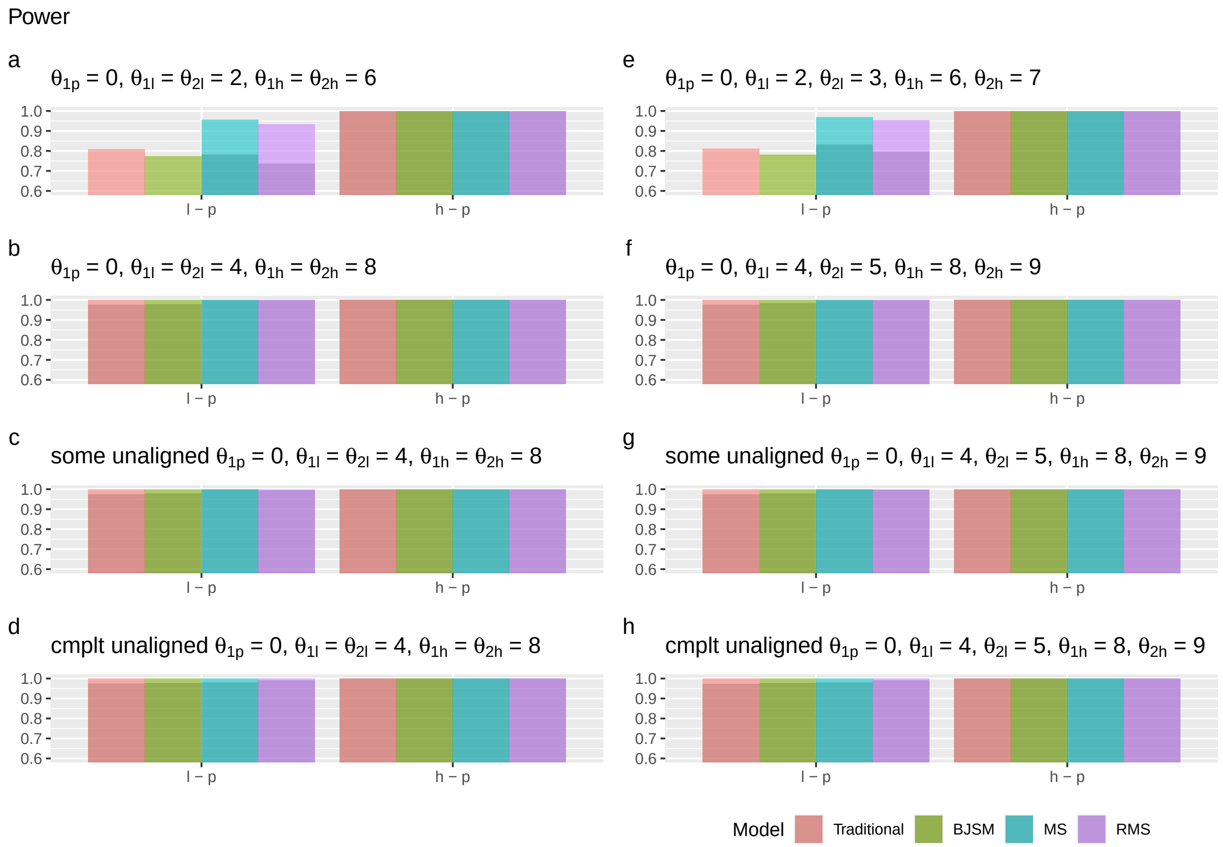


Figure 2.8: Simulated power of the MAC-snSMART methods.
In this paper, the power is defined as the probability that the credible intervals of $\hat{\theta}_{1l} - \hat{\theta}_{1p}$ and $\hat{\theta}_{1h} - \hat{\theta}_{1p}$ do not include 0 when there are treatment effect difference between low dose and placebo and high dose and placebo. Two hierarchical models: MAC-snSMART (MS) and robust MAC-snSMART (RMS) are compared against the traditional method and the Bayesian Joint Stage Model (BJSM). The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph. θ_{jk} denotes the true value of the expected treatment effects of treatment k in stage j, where $j = 1, 2$, $k = p, l, h$, p = placebo, l = low dose, and h = high dose. *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial.

The right columns of Figure 2.4, 2.5, 2.6, and 2.7 present the rMSE, average bias, CR , and average CI width for estimators of the treatment effects when data are generated violating the exchangeability assumption between stage 1 and 2 (data generating process 2). Due to information borrowing across both stages, the MAC-snSMART leads to larger positive average biases and subsequently lower coverage rates and higher rMSE values compared to the traditional method. The decrease in CR can be mitigated by setting a larger standard deviation for the half-normal prior of τ_k or incorporating the robustification component (Section 2.3.2.2) into the distribution of $\theta_{2k'}$ in

Formula 2.1. The BJSM is less susceptible to this issue since a "shift" parameter α is incorporated in the model (Fang et al. 2023). Under this data generation setting, the traditional method provides the best placebo treatment effect estimators among the presented methods, and the BJSM provides the best low and high dose treatment effect estimators. Conclusions were consistent for sample size 25 under both data generation processes and all scenarios.

Table 2.2 and Figure 2.8 present the type I error and power of each model. The MAC-snSMART methods have smaller type I errors than the traditional method when its assumption of full exchangeability between stages is upheld, and the type I error is still reasonable (below 0.1) when this assumption is violated. The BJSM has the smallest type I error when the exchangeability assumption between two stages is violated. When the treatment effect is clinically meaningful (i.e., $\geqslant 4$), all methods have power close to 1 under all scenarios and sample sizes. The MAC-snSMART methods have a significant increase in power compared to the traditional method and the BJSM when the treatment effect is equal to 2. The power gain is even greater for a smaller sample size ($n = 25$). Overall, the robust MAC-snSMART is the best-performing model when exchangeability of stage-wise treatment effects is upheld.

2.5 Re-analysis of SPITFIRE trial

To illustrate the practical utility of the proposed snSMART design and MAC-snSMART methods, we conducted a reanalysis of the SPITFIRE study. Given that the trial stopped for futility after stage 1, we only have summary-level NSAA total scores and 6MWD at baseline and week 48 (details in Table 2.3). To create a data set that matches our proposed trial design, we simulated stage 1 patient-level data by randomly drawing from normal distributions based on the SPITFIRE data. NSAA outcome data for placebo ($n = 30$), low dose ($n = 29$) and high dose ($n = 33$) were generated using $N(-2.99, 0.65^2 \times \sqrt{30})$, $N(-3.44, 0.67^2 \times \sqrt{29})$, and $N(-2.41, 0.64^2 \times \sqrt{33})$, respectively. As per study protocol, we set a change of $\geqslant -3.1$ points from baseline as the threshold for a clinically meaningful treatment effect (Muntoni et al. 2018) and to categorize stage 1 responders and nonresponders. Stage 2 patient-level treatment outcomes were again randomly generated using formula 2.1 with $s_{kk'}$ randomly chosen between -1 and 1, and low and high dose outcomes drawn from $N(-3.44, 0.67^2\sqrt{n_{2l}})$ and $N(-2.41, 0.64^2\sqrt{n_{2h}})$, respectively. Our proposed snSMART design and randomly generated stage 1 outcomes dictated n_{2l} and n_{2h}. We followed the same procedure to simulate stage 1 and stage 2 6MWD data.

We used CINRG DNHS data as the source of external control in this re-analysis. Data from DNHS participants who met the SPITFIRE trial eligibility criteria and had NSAA total score or 6MWD records were used for the external control group. Our model assumption of exchangeability between external and current trial controls seems valid due to careful selection of control data and

Table 2.3: The SPITFIRE trial study measures. Summary statistics are reported as mean (SD) for NSAA and 6MWD.

	Placebo	RO7239361 Low Dose	RO7239361 High Dose
Baseline			
Sample Size	56	55	55
NSAA	23.1 (6.4)	24.5 (5.5)	22.7 (6.7)
6MWD	388.33 (69.59)	399.73 (68.35)	370.73 (93.35)
Week 48			
Sample Size - NSAA	30	29	33
Changes in NSAA	-2.99 (0.65)	-3.44 (0.67)	-2.41 (0.64)
Sample Size - 6MWD	29	25	31
Changes in 6MWD	-41.3 (8.7)	-39.6 (9.0)	-30.0 (8.7)

similarity of demographics and disease severity in patients. Since the participant visit schedule of DNHS was different from the SPITFIRE trial, for each participant, we picked the test record with "days from baseline" closest to 336 (48×7) as their "Week 48" record. Alternatively, we could impute "Week 48" observations, a method that will be described in Chapter 3. In the end, for NSAA total score, data from 25 participants were used for the external control data with a mean NSAA total score change from baseline being -1.04 and its standard deviation being 0.77. The same data were used for 6MWD with a mean 6MWD change from baseline being -22.36 and its standard deviation being 27.98.

Considering the wide variation in outcomes, 30,000 realizations were simulated for both 6MWD and NSAA total score to assess model performance. When implementing the robust MAC-snSMART method, we carefully followed the prior specification rules outlined in Section 2.3.2.4. For example, based on the SPITFIRE stage 1 NSAA total scores, most of the observed biases b_{ll}, b_{lh} and b_h should range between 0 and 7. Therefore, we used a conservative uniform distribution $Unif(0, 15)$ as the priors for B_{ll}, B_{lh}, and B_h to cover the range. The details of all other prior specifications can be found in the R code provided.

We fitted the traditional analytic method, BJSM, and robust MAC-snSMART method, with results shown in Table 2.4. Notice that the estimators obtained from the robust MAC-snSMART method and BJSM are consistent with each other and have significantly smaller CI widths than the traditional method because of the efficient use of data across both stages. Thus, even though the BJSM and the robust MAC-snSMART reached the same conclusion as the SPITFIRE trial, i.e., failing to reject the null hypothesis, more precise treatment effects estimations were provided by these two methods.

Table 2.4: Example data analysis result comparison. Note: $\widehat{\theta}_{1k}$ is the estimated stage 1 treatment effect or change from baseline to 48 weeks in the NSAA or 6MWD for treatment k, where $k = p, l, h$, p = placebo, l = low dose, and h = high dose. Four analytic methods: original SPITFIRE trial results, traditional analytic method, robust MAC-snSMART (RMS), and Bayesian joint stage modeling (BJSM) are compared. "NSAA" stands for "North Star Ambulatory Assessment total score", and "6MWD" stands for "6-minute walk distance". The 95% confidence or credible intervals of the estimates are shown in the parenthesis.

Estimator	Original Result	Traditional	RMS	BJSM
NSAA				
$\widehat{\theta}_{1p}$	-2.99 (-4.26, -1.71)	-2.98 (-4.24, -1.73)	-2.89 (-3.68, -2.10)	-2.91 (-4.01, -1.82)
$\widehat{\theta}_{1l}$	-3.44 (-4.75, -2.13)	-3.44 (-4.74, -2.14)	-3.40 (-4.49, -2.31)	-3.38 (-4.34, -2.41)
$\widehat{\theta}_{1h}$	-2.41 (-3.66, -1.16)	-2.41 (-3.65, -1.17)	-2.49 (-3.56, -1.42)	-2.42 (-3.35, -1.48)
$\widehat{\theta}_{1l} - \widehat{\theta}_{1p}$	-0.45 (-2.17, 1.27)	-0.46 (-2.27, 1.36)	-0.52 (-1.88, 0.83)	-0.46 (-1.91, 0.98)
$\widehat{\theta}_{1h} - \widehat{\theta}_{1p}$	0.58 (-1.10, 2.26)	0.58 (-1.20, 2.35)	0.40 (-0.94, 1.74)	0.50 (-0.93, 1.92)
6MWD				
$\widehat{\theta}_{1p}$	-41.3 (-58.4, -24.2)	-41.3 (-58.2, -24.5)	-41.0 (-52.5, -29.3)	-41.1 (-54.4, -27.8)
$\widehat{\theta}_{1l}$	-39.6 (-57.2, -22.0)	-39.6 (-57.0, -22.2)	-39.4 (-53.2, -26.1)	-39.3 (-51.6, -27.6)
$\widehat{\theta}_{1h}$	-30.0 (-47.1, -12.9)	-30.0 (-46.7, -13.0)	-30.2 (-42.1, -18.0)	-29.7 (-40.8, -18.7)
$\widehat{\theta}_{1l} - \widehat{\theta}_{1p}$	1.7 (-21.1, 24.6)	1.8 (-22.6, 26.0)	1.6 (-15.6, 19.1)	1.8 (-16.6, 19.5)
$\widehat{\theta}_{1h} - \widehat{\theta}_{1p}$	11.3 (-11.0, 33.6)	11.5 (-12.5, 35.4)	10.9 (-6.4, 28.1)	11.4 (-5.8, 28.6)

2.6 Discussion

In this paper, we were motivated by the current drug development paradigm for DMD and other similar rare diseases to present an alternative design and Bayesian methods with efficient use of all available evidence including data from both stages and external control. We have proposed a new snSMART design and robust MAC-snSMART method to estimate the treatment effect of placebo, low and high doses with a continuous outcome of interest. The robust MAC-snSMART method provides accurate and robust estimators when the expected treatment effect of the same dose level across stages are similar. Our proposed snSMART design and robust MAC-snSMART methods are aligned with the mission of US Food and Drug Administration (FDA)'s Complex Innovative Design program (2020). We have provided guidelines for prior distributions and alternative models for sensitivity analyses in practical implementation. At the planning stage, it is critical to consider a wide number of scenarios (like those presented in Section 2.4) to understand the impact of model assumptions, prior choice, and sample size for the proposed design and analytic method. This exercise is crucial for sponsors and regulators to understand the practical efficiency and robustness of the model.

The proposed robust MAC-snSMART method assumes treatment effects from stages 1 and 2 are exchangeable, which relies on this assumption of stable disease (and an adequate washout period) to assume stable treatment effects across stages. Diseases like DMD used as motivation here, corticobasal degeneration (CBD), and familial Mediterranean fever (FMF), which are slower to progress, are good candidates for the proposed snSMART design and analytic methods. If first and second stage treatment effects are not similar or there is not an adequate washout period, a multi-stage design and robust exchangeable model may not be appropriate and lead to biased estimation of treatment effects. At the end of the trial, sensitivity analyses that compare results from a traditional analytic method and the robust MAC-snSMART method can be conducted to assess these assumptions.

The proposed model is tested with extensive simulation studies across various scenarios. The scenarios include incorporating external control data which are consistent and inconsistent with the current trial placebo arm. Note that, this is not equivalent to a simple pooled analysis of external and concurrent control data as the MAC-snSMART model takes into account between-trial heterogeneity. For example, in the SPITFIRE trial re-analysis, the ESS for the robust MAC-snSMART analysis is 12, which is less than the sample size of external placebo data ($n = 25$). Thus, more data contributing to the placebo will not always be useful for decreasing the number of patients needed on the placebo arm. It depends on the degree of heterogeneity between different data sources.

While, this manuscript concentrates on enriching the control arm, the MAC framework permits enrichment of treatment arms. However, we believe relevant treatment data are less likely to be available for use in dose-finding studies. An exception may be the use of adult data in pediatric drug development, but there is debate and uncertainty surrounding the validity and reliability of extrapolating safety and efficacy data from adult populations to pediatric populations.

It is possible that even a high dose of the investigative treatment cannot provide a clinically meaningful treatment effect by the end of stage 1. Under this scenario, it is not ethical to continue the trial for those who received high dose in stage 1. Hence, in practice, we recommend a stopping rule such that if less than 30% of patients respond to high dose in stage 1, stage 2 will not be conducted, and all treatment effects will be calculated based on stage 1 data only. There may be additional trade-offs between efficacy and toxicity which are beyond the scope of this manuscript and the subject of some of our future work.

The SPITFIRE trial and many other DMD trials incorporate participant demographic and baseline characteristic covariates into their analysis. In the future, we hope to extend the robust MAC-snSMART method to include patient-level covariates. The use of patient-level covariates is discussed in Kotalik et al. (2021) as a way to assess covariate-adjusted exchangeability. The study uses a linear model and the existing multi-source exchangeability models framework to enable borrowing even when marginal treatment effects are different, but covariate-adjusted exchangeability is maintained. Integrating this feature into the snSMART approach would broaden borrowing opportunities. In addition, data in DMD trials are usually collected in a longitudinal manner with 3 or more visits. In our study, like in SPITFIRE, we employed the commonly used “change from baseline” as the primary endpoint, which is in line with regulatory standards. The placebo group in our study allows us to assess the possible “regression to the mean” effect. Although commonly used, it’s worth noting that “change from baseline” may not always be the most appropriate measure. Our approach remains valid if using the absolute outcome at the end of each stage, and the performance of our method is not affected. If absolute outcomes are used, we can examine the credible intervals in the difference of the differences among the high dose, low dose, and placebo groups. Alternatively, it is important for future work to investigate ways to incorporate longitudinal data into our analytic methods.

CHAPTER 3

Integrating Inverse Probability Weighted External Control Data into a Bayesian Longitudinal Small Sample, Sequential, Multiple Assignment Randomized Trial (snSMART) Design

3.1 Introduction

DMD is a severe genetic disorder inherited in an X-linked manner, affecting approximately 19.8 per 100,000 live male births (Crisafulli et al. 2020). Characterized by a progressive loss of muscle function, DMD leads to severe physical disability, loss of mobility, and premature death due to cardiac or respiratory failure. The average life expectancy for patients with DMD hovers around 30 years, underscoring a pressing need for effective therapeutic interventions.

The complexity of conducting clinical research in rare diseases such as DMD is multifaceted. Traditional trial designs face substantial hurdles due to the limited pool of available participants, the wide variability in disease progression, and the ethical dilemmas posed by requiring long-term placebo control follow-ups. These challenges underline the necessity for innovative approaches to trial design that can accommodate the unique constraints of rare disease research, ensuring that potential therapies are evaluated efficiently and ethically.

This paper introduces a novel methodological framework aimed at enhancing the efficiency of clinical trials for DMD and similar rare diseases. Our approach leverages the integration of external controls within longitudinal study designs, addressing key obstacles traditionally associated with this realm of research. Drawing inspiration from the tadalafil trial (NCT01865084)—a randomized, double-blind, placebo-controlled study that assessed the efficacy, safety, and side effects of tadalafil in treating DMD —this paper elucidates how to effectively address heterogeneities between data sources and trial stages in order to provide a viable solution to the ethical, logistical, and analytical challenges of conducting rigorous clinical trials in rare disease populations.

The tadalafil trial, one of the largest DMD clinical studies to date, enrolled 331 patients at baseline. During the first 48 weeks, participants were randomized to receive either a placebo, a low dose, or a high dose of tadalafil. After week 48, those initially given a placebo were randomized to receive either a high or low dose, and those who began with either the low or high dose of tadalafil were not allowed to adjust their dose if it proved ineffective. The primary analysis of the trial planned to focus solely on data from the first 48 weeks, thereby overlooking valuable insights gathered after week 48. We suggest enhancements to the trial design, such as facilitating dose adjustments and reducing the number of patients assigned to the placebo group by integrating external control data. The efficiency of the primary analysis can also be enhanced by adopting a statistical model that considers data from all current trial stages and external controls.

In this paper, we propose a robust Bayesian analytic method that leverages external control data alongside current trial data to accurately estimate the effects of experimental drugs tested in snSMART as outlined by Wang et al. (2023). We build upon the MAC-snSMART approach proposed by Wang et al. by incorporating patient baseline characteristics. This enhancement allows for a more nuanced comparison of heterogeneities between the external controls and the current trial, and facilitates the use of longitudinal data for a more precise estimation of treatment effects. Our methodology not only addresses potential discrepancies between external and current trial controls but also navigates the issue of "drift" in clinical outcomes across different stages of the trial.

3.2 Motivating example: a phase 3 study of tadalafil for DMD

Our motivating example is a randomized, double-blind, placebo-controlled, parallel 3-arm, phase 3 trial of tadalafil for DMD (NCT01865084, Figure 3.1a). This trial aims to explore whether tadalafil can effectively slow the progression of walking impairment in boys with DMD . During the initial 48 weeks (stage 1), participants were given either one of two doses of tadalafil or a placebo (1:1:1). Following this stage, they had the opportunity to join open-label extension (OLE) (stage 2), which included two periods. In the first period of OLE, all participants were treated with tadalafil for 48 weeks. Patients who were given tadalafil in the double-blind treatment stage started OLE on the same dosage of tadalafil. Those who received a placebo in the double-blind treatment stage were randomly assigned to one of the tadalafil doses for the OLE period. Those who completed the first period of OLE proceeded to the second period, receiving tadalafil for a minimum of another 48 weeks. The primary outcome was change from baseline in 6-minute walk distance (6MWD), which was measured twice at baseline and once every 12 weeks thereafter. Secondary outcomes included the North Star Ambulatory Assessment (NSAA) and other timed function tests (Victor et al. 2017). The trial's primary analysis only used data gathered during stage 1. To assess the primary null hypothesis, which stated no difference between each treatment and placebo, a repeated measures analysis was conducted using a mixed-effects model of repeated measures (MMRM). This trial represents a typical development scenario for new drugs in many rare diseases. The conclusion of the trial is that 48 weeks of tadalafil treatment did not show improvement in the 6MWD or other indicators of walking ability in boys aged 7 to 14 with DMD who were receiving standard glucocorticoid therapy (Victor et al. 2017). It is important to clarify that, in this paper, our objective is not to critique or assess the clinical efficacy of the treatments used, nor to comment on decisions made by the sponsors or regulatory authorities regarding the trial or the drug involved. Instead, the goal is to use its trial setting and published results to illustrate the effectiveness of our proposed design modification and analytical methods.

Just like the SPITFIRE design (NCT03039686) discussed in Wang et al. (2023), the tadalafil trial is similar to an snSMART (Tamura et al. 2016) (Figure 2.1) but includes no re-randomization for patients started on low or high dose of tadalafil after the end of stage 1 and used only the stage 1 data in the primary analysis. Similar trials could gain advantages from incorporating external control data, enabling more patients to receive the investigational drug, thereby assisting in recruitment and retention. However, there is one aspect present in the tadalafil trial that snSMART researchers have not yet investigated: the analysis of longitudinal trial data. We aim to fill this gap with a Bayesian longitudinal piecewise model as an alternative to the traditional MMRM. In recent years, there has been considerable advancement in snSMART methods (Wei et al. 2018, 2020, Hartman et al. 2021) that are now implemented in an R package (Wang & Kidwell 2022). Further extensions of snSMART include a group-sequential design for early termination of an arm (Chao, Braun, Tamura & Kidwell 2020) and the inclusion of various dose levels (Fang et al. 2021, 2023). In Wang et al. (2023), an snSMART design is proposed that formally incorporates external control data, and a robust MAC approach, termed robust MAC-snSMART, is developed to provide accurate and robust estimators of treatment effects. A detailed literature review on incorporating external control data is provided by Wang et al. (2023). In short, clinical trials can utilize external data through Bayesian methods like power priors (Ibrahim & Chen 2000) and dynamic methods (Duan 2005, Neuenschwander et al. 2009, 2010, 2016, Hobbs et al. 2011, Schmidli et al. 2014), as well as frequentist approaches such as propensity score matching (Rosenbaum & Rubin 1983, Lin et al. 2018).

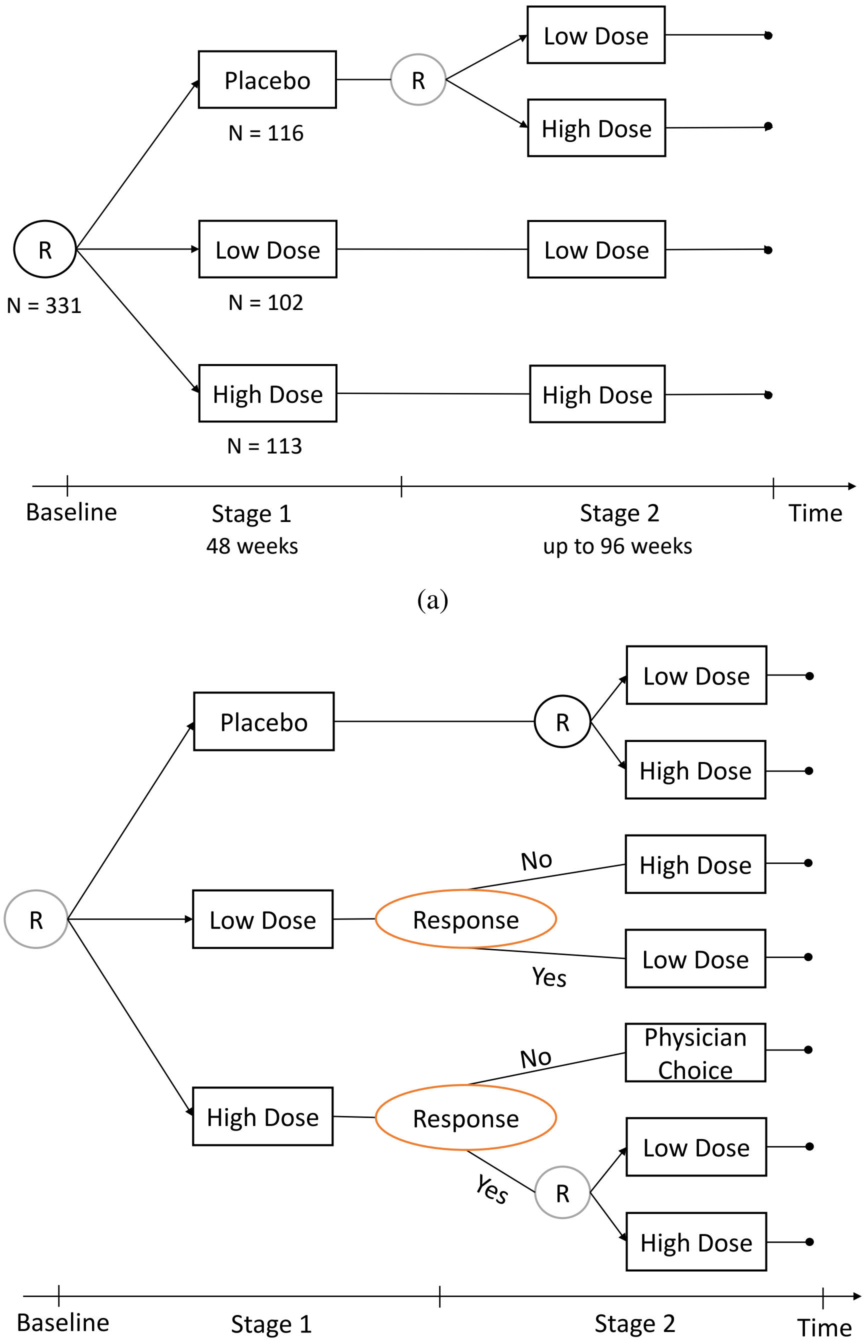
Our proposed approach in this manuscript represents a significant enhancement of the robust MAC-snSMART method, offering major improvements to the existing snSMART research: 1) enabling the analysis of longitudinal data, 2) allowing for the inclusion of patient baseline characteristics, 3) handling missing data through multiple imputation, 4) limiting the heterogeneity between data sources through PS, and 5) addressing the possible stage-wise treatment effect non-exchangeability caused by unexpected disease progression or ineffective washout period in-between stages. These enhancements significantly broaden the applicability of the innovative snSMART design and increase efficiency in rare disease drug development. However, to model the disease progression of rare diseases like DMD longitudinally and across multiple trial stages, innovative modeling methods need to be developed.

In recent years, there has been a surge in interest in modeling non-traditional treatment effects and disease progression for more effective and interpretable estimators. Raket (2022) introduced progression models for repeated measures (PMRM), non-linear mixed-effects models for analyzing time-based treatment effects like slowing disease progression. For rare diseases, Lennie et al. (2020) proposed a latent process model to study DMD progression using 6MWD results, employing a Bayesian hierarchical nonlinear mixed-effects model. Quintana et al. (2019) utilized a Bayesian approach for longitudinal data from GNE myopathy patients, developing a model that aligns subjects by latent disease age to predict long-term progression. Some studies have also focused on incorporating natural history studies data or other types of external control data in addition to innovatively modeling disease progression. For instance, Fouarge et al. (2021) adopted a hierarchical Bayesian mixed-effects model to reliably simulate the progression of centronuclear myopathy patients over time and to compare the simulated trajectories with actual observed post-treatment outcomes probabilistically. Zhou & Ji (2021) proposed a method for incorporating external data into the analysis of clinical trials using Bayesian additive regression trees (BART), specifically aimed at estimating the conditional or population average treatment effect. Kiran Chandra et al. (2021) proposed a novel Bayesian nonparametric mixture model to create synthetic control groups with electronic health records (EHR) in single-arm treatment-only clinical trials. Our method is specific to the snSMART design and, therefore, differs from the papers discussed above.

Like many of the abovementioned methods, we also adopt a Bayesian statistical method. Our approach not only flexibly models disease progression across trial stages in a piecewise manner but also greatly enhances the efficiency of estimating treatment effects through the use of trial data across various stages. Our snSMART design and methodology presuppose a) consistent study conditions throughout the trial (meaning, in the absence of treatment effects, patients' primary outcomes remain largely unchanged, or decline in a stable manner), b) a sufficient washout period between the trial's two stages, and c) exchangeable treatment effects across these stages. The snSMART design is unsuitable for conditions that lack stability during the trial. We propose model adjustments and sensitivity analyses to accommodate potential deviations from these assumptions.

3.2.1 Proposed Modification for Tadalafil Trial Design

In the snSMART design (Wang et al. 2023), illustrated in Figure 3.1b, patients are initially randomized in a 1:2:2 ratio to receive either a placebo, a low dose, or a high dose (e.g., of tadalafil)



(a)

(b)

Figure 3.1: (a) Study design of the tadalafil trial (NCT01865084). (R) denotes randomization. (b) Study design of the snSMART design (Wang et al. 2023) that formally incorporates external control data. Participants are randomized (R) with 1:2:2 chances of receiving placebo, low dose, or high dose, respectively, in stage 1. At the end of stage 1, participants are assigned or re-randomized to their stage 2 treatment based on their stage 1 treatment and response status. Outcomes are collected at the end of stage 1 and 2.

in stage 1. After 48 weeks, participants are re-assigned or re-randomized based on their initial treatment and post-baseline 6MWD. According to the protocol of the tadalafil trial, a participant is considered a responder at week 48 if their post-baseline 6MWD does not decrease by more than 10%. Those who received a placebo in stage 1 are equally re-randomized to either a low or high dose in stage 2, independent of their stage 1 response, ensuring all participants receive treatment. For participants initially on a low dose, responders remain on it, while non-responders switch to a high dose. High dose responders in stage 1 are equally re-randomized to either low or high dose, considering the potential efficacy and tolerability of a lower dose. However, non-responders to the high dose are excluded from stage 2 to avoid unnecessary toxicity and ethical concerns, as a lower dose is unlikely to be effective if a higher dose proved ineffective. This design choice aims to optimize treatment effectiveness and ethical considerations in the trial.

3.3 Methods

3.3.1 Mixed Model Repeated Measures (MMRM)

Traditionally, longitudinal data are most commonly analyzed using a MMRM analysis to test the primary null hypothesis of no difference between each treatment (i.e., low and high dose tadalafil) and placebo in mean change in treatment outcomes. The MMRM model typically includes a fixed effect for baseline primary outcomes, in our case, 6MWD, as well as fixed effects for treatment group assignment, visit, treatment group-by-visit interaction, and a random effect for subjects. An unstructured covariance matrix is typically used to model the within-subject errors. Note that a traditional analytic method does not incorporate external control data and uses only stage 1 outcomes in the analysis.

3.3.2 Bayesian joint stage model (BJSM)

The snSMART design can be analyzed with a customized BJSM. Unlike the dose level BJSM proposed by Fang et al. (2023), which uses data from two stages without incorporating external data, our approach involves integrating external control data using informative normal distribution priors for the placebo effect parameter. This involves estimating the prior's parameters through a method of moments approximation from external data and adjusting the variance to reach the desired effective sample size (ESS).

3.3.3 Bayesian longitudinal piecewise meta-analytic combined (BLPM) approaches

The primary analysis using our proposed BLPM approach involves a two-step procedure: 1) calculating the PS (Rosenbaum & Rubin 1983) for all external control patients by fitting a logistic regression model to all patients from external control data and current trial data, and 2) fitting the BLPM model to external controls and current trial data to estimate the treatment effects. The structure of the BLPM approach is illustrated in Figure 3.2. More details on each step will be provided in the following sections.

3.3.3.1 Notation

The following notation is used in this section. $Y_{d(i)jk}$ and $\mu_{d(i)jk}$ denote the observed and underlying true outcome (e.g., NSAA score or 6MWD) for patient i at visit $j = 1, ..., n_j$ in stage $k = 1, 2$ of trial $d = 1, ..., n_d, \star$. Here, we use $d = \star$ to denote current trial, and use $d = 1, ..., n_d$ to denote the n_d external controls we collected. We use l and h to denote low dose and high dose, respectively. Thus, $Y_{\star(10)31}$ denotes the observed outcome of current trial patient 10 at third visit in stage 1, and $Y_{3(1)41}$ denotes the observed outcome of patient 1 in the 3rd external control data at fourth visit in stage 1. We use $\boldsymbol{X}$ or $\boldsymbol{Z}$ to denote the baseline covariate matrix, and use βs and bs to denote the fixed effect coefficients and random effects, respectively. T_{1i} and T_{2i} are used to denote stage 1 treatment and stage 2 treatment received by subject i in current trial $\star$. ω_is stands for the IPTW of subjects.

3.3.3.2 PS and IPTW calculation

Controlling for potential bias is essential when using external control data in clinical trials. To combat bias, especially from unmeasured confounders, it's important to prespecify statistical methods. Addressing measured confounders starts with carefully selecting external control data. Pocock (1976) and Lim et al. (2018) have outlined criteria for assessing the similarity between external and trial control data.

In our context, after excluding subjects from the external controls who do not meet the inclusion and exclusion criteria of the current trial, and after collecting all baseline covariates for those retained in the external controls and those enrolled in the current trial, the PS can be calculated. The PS is defined as the probability of a subject being from the current study rather than from external data sources. The logistic regression used to obtain the PS estimates can be formulated as: $logit\{P(StudyIndicator_{di} = 1)\} = \boldsymbol{X}_{di}\boldsymbol{\beta}$, where $StudyIndicator_{\star i} = 1$ if the patient is from the current trial, and $StudyIndicator_{di} = 0$ if patient i is from one of the external control datasets

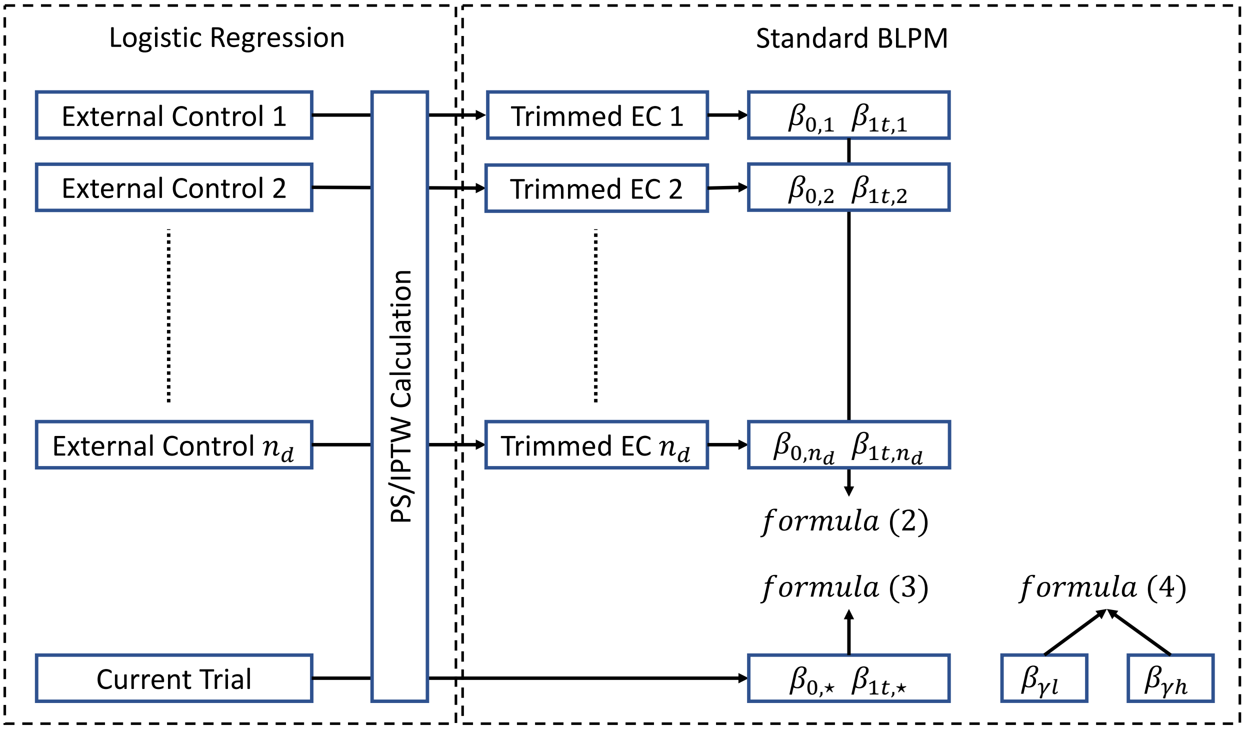


Figure 3.2: Structure of the standard BLPM approach

d. Specifically for the tadalafil trial, covariates may include variables such as *age*, *race*, *ethnicity*, *height*, *weight*, *baseline NSAA* and *baseline 6MWD*.

The IPTW for participants is determined using the stabilized weighting method, which facilitates the calculation of the average treatment effect in the treated (ATT). In this approach, subjects from the current trial are assigned a weight of 1 to maximize the utilization of the current trial data. For subjects in the external control group, weights are computed using the formula $\omega'_{di} = PS_{di}/(1 - PS_{di})$, and then normalized to $\omega_{di} = \omega'_{di}/(\sum_{i \in trial\, d} \omega'_{di}/N_d)$, where N_d represents the total number of subjects in trial d.

Next, we trim the external controls to retain only those subjects whose PS fall within the range of the PS of all subjects in the current trial. This step further reduces heterogeneity between data sources. However, despite careful selection and PS trimming, clinical outcomes may vary across sources due to various intrinsic and extrinsic factors, necessitating a model-based approach for proper analysis.

3.3.3.3 Structure the Bayesian longitudinal piecewise model across trial stages

Given the relatively stable rate of decrease in observed 6MWD and NSAA over time in the tadalafil trial (Victor et al. 2017), we adopt a piecewise linear regression model (Nahum-Shani et al.

2020). This model fits a separate line segment for different trial stages, allowing the linear trend in the outcome during stage 1 to differ from that during stage 2. If the rate of decrease varies significantly between visits, shorter time intervals for each line segment can be applied. Under the proposed snSMART design, a subject in the current trial could follow one of seven distinct treatment sequences: (p then l), (p then h), (l then l), (l then h), (h then l), (h then h), and (h with *No stage 2 treatment*). In contrast, a subject in the external controls can be considered as being in the placebo arm in stage 1 only. Consequently, our piecewise linear model for both current trial and external control subjects can be formulated as follows:

$$Y_{d(i)jk} \sim N(\mu_{d(i)jk}, \sigma_d^2/\omega_{di}),$$

$$\begin{aligned} \mu_{d(i)jk} =& \beta_{0,d} + \beta_{1t,d} \times t_d + b_{1i,d} + b_{2i,d} \times t_d + \mathbf{Z}_{i,d}\boldsymbol{\beta}_{z,d} + (\beta_{1l} I[T_{1di} = l] + \beta_{1h} I[T_{1di} = h]) \times t_1 + \\ & (\beta_{2l1} I[T_{1di} = p, T_{2di} = l] + \beta_{2h1} I[T_{1i} = p, T_{2i} = h] + \beta_{2l2} I[T_{1i} = l, T_{2i} = l] + \\ & \beta_{2h2} I[T_{1i} = l, T_{2i} = h] + \beta_{2l3} I[T_{1i} = h, T_{2i} = l] + \beta_{2h3} I[T_{1i} = h, T_{2i} = h]) \times t_2, \end{aligned} \tag{3.1}$$

where $\mathbf{Z}$s are the baseline covariates, and ω_{di}, introduced previously in Section 3.3.3.2, is included to further reduce the impact of external control subjects who are dissimilar to those in the current trial. $\beta_{0,d}$ and $\beta_{1t,d}$ are trial-specific and represent the model intercept and the natural DMD disease progression time (placebo) effect, respectively. β_{1l}, and β_{1h} represent the treatment effect difference between stage 1 low dose and the time (placebo) effect, and the treatment effect difference between stage 1 high dose and the time (placebo) effect, respectively. Similarly, β_{2l1} denotes the treatment effect difference between stage 2 low dose and the time (placebo) effect when the subject received placebo in stage 1, β_{2h1} denotes the treatment effect difference between stage 2 high dose and the time (placebo) effect when the subject received placebo in stage 1, β_{2l2} denotes the treatment effect difference between stage 2 low dose and the time (placebo) effect when the subject received low dose in stage 1, and so on. The key estimands of interest are the stage 1 differences in treatment effects between each dose and placebo (β_{1l}, β_{1h}). In this model, $t_\star = t_1 + t_2$, where $t_1 = 1, \ldots, n_j$ and $t_2 = 0, \ldots, n_j$. For example, when an observation gathered at stage 1 visit 1 is fitted, $t_1 = 1$, $t_2 = 0$, and $t_\star = 1$; similarly, when observations gathered at stage 2 visit 2 are fitted, $t_1 = 4$, $t_2 = 2$, and $t_\star = 6$. To account for between-individual heterogeneity in the model, both a trial-specific subject-level random slope, denoted as $b_{2i,d}$, and a random intercept, $b_{1i,d}$, are included. $(b_{1i,d}, b_{2i,d})$ are zero-mean bivariate Gaussian variables with variances $\sigma_{b1,d}^2$ and $\sigma_{b2,d}^2$ respectively, and a correlation coefficient $\rho_{1,d}$.

3.3.3.4 Use of external information for $\beta_{0,\star}$, $\beta_{1t,\star}$, $\sigma_{b1,\star}$ and $\sigma_{b2,\star}$

To more effectively utilize the external control data, we employ the robust MAC approach, detailed in Neuenschwander et al. (2016) and Wang et al. (2023), to assist in estimating the coefficients $\beta_{0,\star}$, $\beta_{1t,\star}$, $\sigma_{b1,\star}$, and $\sigma_{b2,\star}$. The robust MAC structure for $\beta_{0,\star}$ and $\beta_{1t,\star}$ of current trial and $\beta_{0,d}$ and $\beta_{1t,d}$ of external control d is specified as below:

$$\begin{bmatrix}\beta_{0,d}\\ \beta_{1t,d}\end{bmatrix} \sim p_d \times BVN\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \boldsymbol{V}(\tau_{0,d},\tau_{1,d})\right) + (1-p_d)\times BVN\left(\begin{bmatrix}m_{0,d}\\ m_{1,d}\end{bmatrix}, \boldsymbol{V}(v_{0,d},v_{1,d})\right), \quad (3.2)$$

$$\begin{bmatrix}\beta_{0,\star}\\ \beta_{1t,\star}\end{bmatrix} \sim p_\star \times BVN\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \boldsymbol{V}(\tau_{0,\star},\tau_{1,\star})\right) + (1-p_\star)\times BVN\left(\begin{bmatrix}m_{0,\star}\\ m_{1,\star}\end{bmatrix}, \boldsymbol{V}(v_{0,\star},v_{1,\star})\right), \quad (3.3)$$

where

$$\boldsymbol{V}(\tau_{0,d},\tau_{1,d}) = \begin{bmatrix}\tau_{0,d}^2 & \rho_{01\tau,d}\tau_{1,d}\tau_{0,d}\\ \rho_{01\tau,d}\tau_{1,d}\tau_{0,d} & \tau_{1,d}^2\end{bmatrix}, \boldsymbol{V}(v_{0,d},v_{1,d}) = \begin{bmatrix}v_{0,d}^2 & \rho_{01v,d}v_{1,d}v_{0,d}\\ \rho_{01v,d}v_{1,d}v_{0,d} & v_{1,d}^2\end{bmatrix},$$

$$\boldsymbol{V}(v_{0,\star},v_{1,\star}) = \begin{bmatrix}v_{0,\star}^2 & \rho_{01,\star}v_{1,\star}v_{0,\star}\\ \rho_{01,\star}v_{1,\star}v_{0,\star} & v_{1,\star}^2\end{bmatrix}.$$

p_d and $p_\star$ determine the resemblance between $\beta_{0,d}$ and $\beta_{1t,d}$ of the external control dataset d and the current trial $\beta_{0,\star}$ and $\beta_{1t,\star}$. The selection of values for $m_{0,d}, m_{1,d}, m_{0,\star}, m_{1,\star}, v_{0,d}, v_{1,d}, v_{0,\star}$, and $v_{1,\star}$ relies on expert insights. If no expert insights are available, weakly informative values can be used (Kass & Wasserman 1995). In doing so, we further down-weight the non-exchangeable external data based on variability between data sources. In addition to $\beta_{0,\star}$ and $\beta_{1t,\star}$, a similar mixture of bivariate normal distributions can also be used to incorporate external information into the estimation of $\sigma_{b1,\star}$ and $\sigma_{b2,\star}$.

3.3.3.5 Combining evidence for low and high dose groups

Given that our model assumes exchangeable treatment effects across trial stages, and since the original purpose of designing stage 2 of the current trial is to gather more data to assess the treatment effect in stage 1, it is logical to add a component to the method that combines the data from all low dose and high dose groups to achieve better estimation efficiency. Therefore, we assume that

$$\beta_{\gamma_l} \sim N(\mu_l, \tau_{\gamma_l}^2); \quad \beta_{\gamma_h} \sim N(\mu_h, \tau_{\gamma_h}^2), \quad (3.4)$$

where $\gamma_l = 1l, 2l1, 2l2, 2l3$ and $\gamma_h = 1h, 2h1, 2h2, 2h3$. The snSMART design, suitable for stable conditions with a required washout period between stages, supports the assumption of exchangeable treatment effects. If stage-wise non-exchangeability is a concern, the robustification (mixture) component outlined in Section 3.3.3.4 can be integrated into the distribution of treatment effect βs to address non-exchangeable treatment effects across trial stages. We name this version of the model as the robust BLPM model. In this paper, we implement both the standard BLPM method and the robust BLPM in simulation studies (Section 3.4) and the reanalysis of tadalafil trial (Section 3.5).

3.3.3.6 Missing data imputation

Sometimes, data from the external controls have missing values due to different visiting schedules from the current trial or have been lost to follow-up because the external control trial was stopped. Although the BLPM model code can run smoothly with missing data, we recommend conducting multiple imputations to obtain more convincing estimators when the data is missing completely at random (MCAR) or missing at random (MAR). We perform multiple imputation using a Bayesian linear mixed model (LMM) incorporating fixed effects for the intercept, received treatment, visit, baseline outcomes, and various other baseline covariates. Additionally, this model includes subject-level random effects for visits and random intercepts.

We generate imputed datasets using the posterior predictive distribution (PPD) technique (Gelman et al. 2014). The PPD illustrates the distribution of potential unobserved values derived from the observed values, and it adheres to the following structure:

$$p(y_{pred}|y) = \int p(y_{pred}, \theta|y)d\theta = \int p(y_{pred}|\theta, y)p(\theta|y)d\theta = \int p(y_{pred}|\theta)p(\theta|y)d\theta,$$

This formulation takes advantage of the conditional independence between the observed data y, unobserved data y_{pred}, and model parameters θ. The PPD can be conceptualized as an averaging of conditional predictions across the posterior distribution of θ. For each sampled θ from the posterior distribution, a corresponding sample of y_{pred} is acquired.

In total, more than 100 imputed datasets should be generated, and to obtain the final estimators, we should combine all the sampled draws from the Markov chain Monte Carlo (MCMC) run on each imputed dataset (Zhou & Reiter 2010).

3.3.3.7 Prior specification

We suggest generalizable, weakly informative normal distributions as priors for all the μ parameters, i.e., priors that are worth approximately one observation (Kass & Wasserman 1995). We suggest using non-informative $Unif(-1, 1)$ prior distributions for all the correlation parameters (ρ).

Half-normal priors with standard deviations roughly equal to half of the standard deviation of the estimated parameter, i.e., $\beta_{0,\star}$, $\beta_{1t,\star}$, $\sigma_{b1,\star}$ and $\sigma_{b2,\star}$, are recommended for all the τs, aiming to cover very small to large between-trial heterogeneity. More details about how to specify prior for τ can be found in Spiegelhalter et al. (2004) and Gelman (2006). We suggest weakly informative normal and half-normal distributions for covariates in the $\boldsymbol{\beta}_{z,d}$ vector and σ in Formula 3.1, respectively.

3.4 Simulation settings

We assess the sensitivity of our proposed BLPM and robust BLPM methods similarly to the simulation studies conducted by Wang et al. (2023). The simulation scenarios cover a range of data generating settings, treatment effects, sample sizes, and instances of non-exchangeability between external control and current snSMART data. We evaluate two distinct data generation processes: the first adheres to the proposed model, while the second allows for deviations from exchangeability between stages 1 and 2. The first data generation process assumes stage-wise exchangeability: at the subject level, the placebo, low dose, and high dose treatment effects between visits in stage 1 are generated through normal distributions with predetermined mean and standard deviation values. For stage 2, the low dose and high dose treatment effects for subjects between visits are generated using the same distributions as in stage 1. In the second data generation process, treatment effects for stage 1 and stage 2 are not exchangeable. Specifically, the stage 2 low dose and high dose treatment effects are generated using the same normal distributions as stage 1, but with their means set 1 unit higher than those in stage 1. This type of data could occur in an snSMART where the washout period between stages 1 and 2 is inadequate or unexpected disease progression is observed. In addition to treatment effects, we also generated subject-level baseline age, baseline outcome, and the time effect on disease progression using normal distributions. To generate longitudinal data at each visit, it is necessary to sum up all the baseline effects, accumulated time effects, and accumulated treatment effects by each visit. In the simulated datasets, 4 visits are included in both stage 1 and stage 2.

Alongside evaluating various data generation processes, we also explore the efficacy of our proposed models across six distinct treatment effect scenarios, particularly in the context of DMD. In this setting, a NSAA score of 4 serves as the criterion to distinguish responders from nonresponders at the conclusion of stage 1. For this purpose, we define δ_{1p}, δ_{1l}, and δ_{1h} as the absolute accumulated treatment effects (e.g., NSAA score) for placebo, low dose, and high dose, respectively, by the end of stage 1, and δ_d represents the absolute placebo treatment effect in external control d. Scenario 1 posits the new drug's ineffectiveness ($\delta_d = \delta_{1p} = \delta_{1l} = \delta_{1h} = 0$). In scenario

Table 3.1: Simulation parameters and scenarios. Note: δ_d denotes the mean absolute placebo treatment effect (e.g., NSAA score) in external control d at the end of stage 1, where $d = 1, 2, ...n_d$. δ_{kT} is the mean absolute treatment effect of treatment T at the end of stage k, where $T = p, l, h$, $k = 1, 2$, p = placebo, l = low dose, and h = high dose.

	Data Generating Process 1	Data Generating Process 2
Scenario 1	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 0$, $\delta_{1h} = \delta_{2h} = 0$	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = 0$, $\delta_{2l} = 1$, $\delta_{1h} = 0$, $\delta_{2h} = 1$
Scenario 2	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 0$, $\delta_{1h} = \delta_{2h} = 6$	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = 0$, $\delta_{2l} = 1$, $\delta_{1h} = 6$, $\delta_{2h} = 7$
Scenario 3	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 2$, $\delta_{1h} = \delta_{2h} = 6$	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = 2$, $\delta_{2l} = 3$, $\delta_{1h} = 6$, $\delta_{2h} = 7$
Scenario 4	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 4$, $\delta_{1h} = \delta_{2h} = 8$	$\delta_d = \delta_{1p} = 0$, $\delta_{1l} = 4$, $\delta_{2l} = 5$, $\delta_{1h} = 8$, $\delta_{2h} = 9$
Scenario 5	$\delta_d = 1\, or\, 0$, $\delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 4$, $\delta_{1h} = \delta_{2h} = 8$	$\delta_d = 1\, or\, 0$, $\delta_{1p} = 0$, $\delta_{1l} = 4$, $\delta_{2l} = 5$, $\delta_{1h} = 8$, $\delta_{2h} = 9$
Scenario 6	$\delta_d = 1$, $\delta_{1p} = 0$, $\delta_{1l} = \delta_{2l} = 4$, $\delta_{1h} = \delta_{2h} = 8$	$\delta_d = 1$, $\delta_{1p} = 0$, $\delta_{1l} = 4$, $\delta_{2l} = 5$, $\delta_{1h} = 8$, $\delta_{2h} = 9$

2, effectiveness is attributed solely to the high dose ($\delta_d = \delta_{1p} = \delta_{1l} = 0 < \delta_{1h} = 6$). Scenario 3 suggests a modest effect for the low dose and a clinically significant effect for the high dose ($\delta_d = \delta_{1p} = 0 < \delta_{1l} = 2 < \delta_{1h} = 6$). Lastly, scenario 4 assumes both low and high doses yield clinically meaningful effects ($\delta_d = \delta_{1p} = 0 < \delta_{1l} = 4 < \delta_{1h} = 8$). We also created two more scenarios to evaluate our model's sensitivity regarding the alignment between external control and current data (scenario 5: $\delta_d \neq \delta_{1p}$ for some d; scenario 6: $\delta_d \neq \delta_{1p}$ for all d). A comprehensive overview of all simulation scenarios is available in Table 3.1.

Under each data generating process, 1,000 realizations per scenario were simulated. Each realization includes one external control data set and one current snSMART data set. Estimations from the traditional MMRM analysis, BJSM, BLPM, and robust BLPM are compared. We calculate CR, rMSE, average bias, and average width of the 95% CI of each estimator. Note that the Monte Carlo error for all average bias and CI width is less than 0.005. Finally, type I error (under scenario 1) and power (under scenarios 3-6) of different methods are also calculated using the probability that the 95% credible intervals of $\hat{\beta}_{1l}$ and $\hat{\beta}_{1h}$ do not include 0. Results are provided for sample

Table 3.2: Simulated type I error. Note: The type I error of all presented methods is defined as the probability that the credible intervals of $\widehat{\beta}_{1l}$ and $\widehat{\beta}_{1h}$ do not include 0 when there are no treatment effect differences between low dose and placebo and high dose and placebo. Data generating process 1 generates data sets under the assumption that the expected treatment effect of the same dose level is exchangeable across trial stages, and data generating process 2 generates data stage-wise non-exchangeability. N denotes the total number of participants in the trial. We compared results obtained from the traditional MMRM, the Bayesian Joint Stage Model (BJSM), the standard BLPM, and the robust BLPM.

Sample Size	MMRM		BJSM		BLPM		RBLPM	
	Low	High	Low	High	Low	High	Low	High
Data generating process 1								
$N = 250$	0.05	0.03	0.06	0.05	0.06	0.04	0.06	0.04
$N = 50$	0.04	0.05	0.04	0.03	0.06	0.04	0.06	0.04
$N = 25$	0.07	0.08	0.04	0.04	0.06	0.05	0.06	0.05
Data generating process 2								
$N = 250$	0.04	0.03	0.04	0.03	0.04	0.04	0.04	0.04
$N = 50$	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.04
$N = 25$	0.07	0.07	0.06	0.04	0.06	0.05	0.05	0.04

size of 25, 50 and 250 randomized with a 1:2:2 ratio to the placebo, low dose, and high dose arms respectively. All computations are done via the R function `jags` in R package `rjags` (Plummer 2022).

3.4.1 Results

In data generation process 1, which satisfies stage-wise exchangeability, the left columns of Figure 3.3, as well as Figures 3.4, 3.5, and 3.6, display the rMSE, average bias, CR, and average CI width for $\hat{\beta}_{1l}$ and $\hat{\beta}_{1h}$. Compared to MMRM and BJSM, estimators from the BLPM methods exhibit notably lower rMSEs. The BLPM methods achieve similar coverage rates to traditional methods, with a slightly lower CR than MMRM at $n = 50$ and a slightly higher CR when $n = 25$, while offering significantly narrower 95% CI widths in all scenarios. Despite the BLPM methods displaying marginally increased average biases in scenario 6, due to completely non-exchangeable external controls, these biases are negligible.

A comparison across all metrics between the traditional method, BJSM , and the BLPM methods demonstrates the superior efficiency and accuracy of the BLPM methods. Incorporating external control data and stage 2 data from the current trial significantly enhances the estimation of treatment effects. The BLPM models adeptly manage the heterogeneity between external controls and current trial data, as evidenced by the minimal increase in rMSEs and average biases in Scenarios 5 and 6 compared to Scenario 4. This confirms the effectiveness of PS trimming, IPTW weighting,

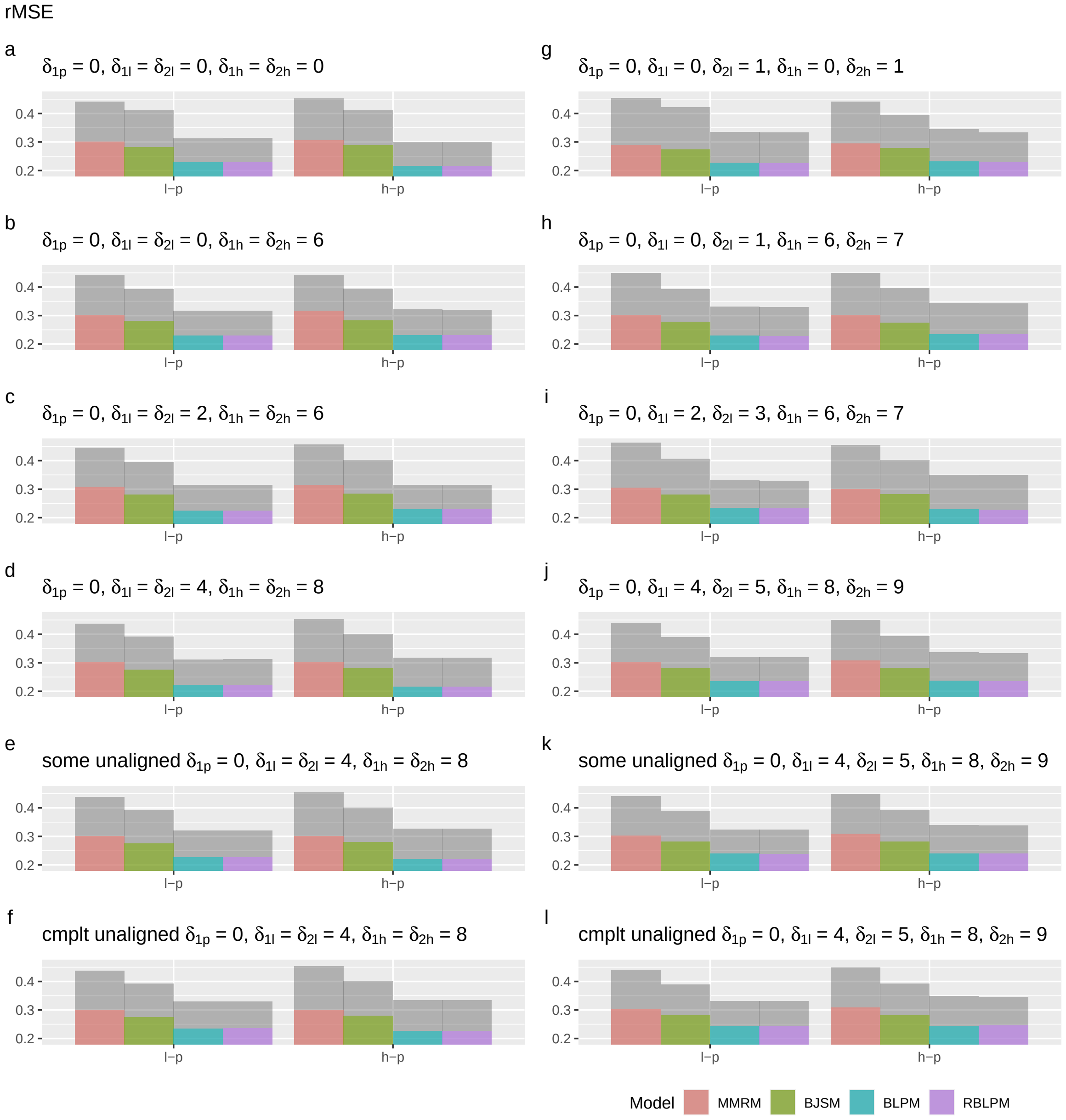


Figure 3.3: Simulated root-mean-square error (rMSE) for the estimators of β_{1l} and β_{1h}. Note: β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM. The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

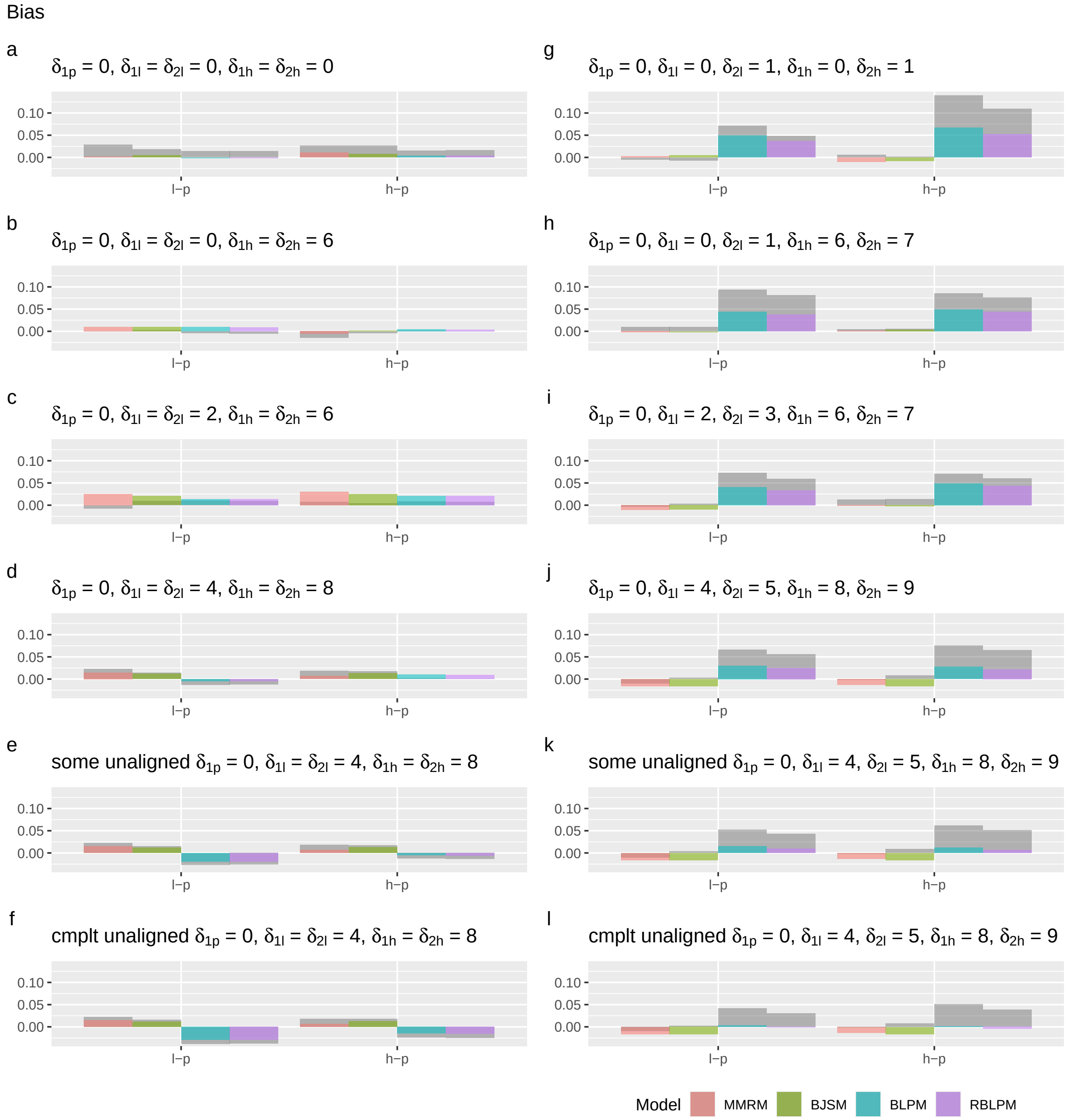


Figure 3.4: Simulated average bias for the estimators of β_{1l} and β_{1h}
β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM. The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

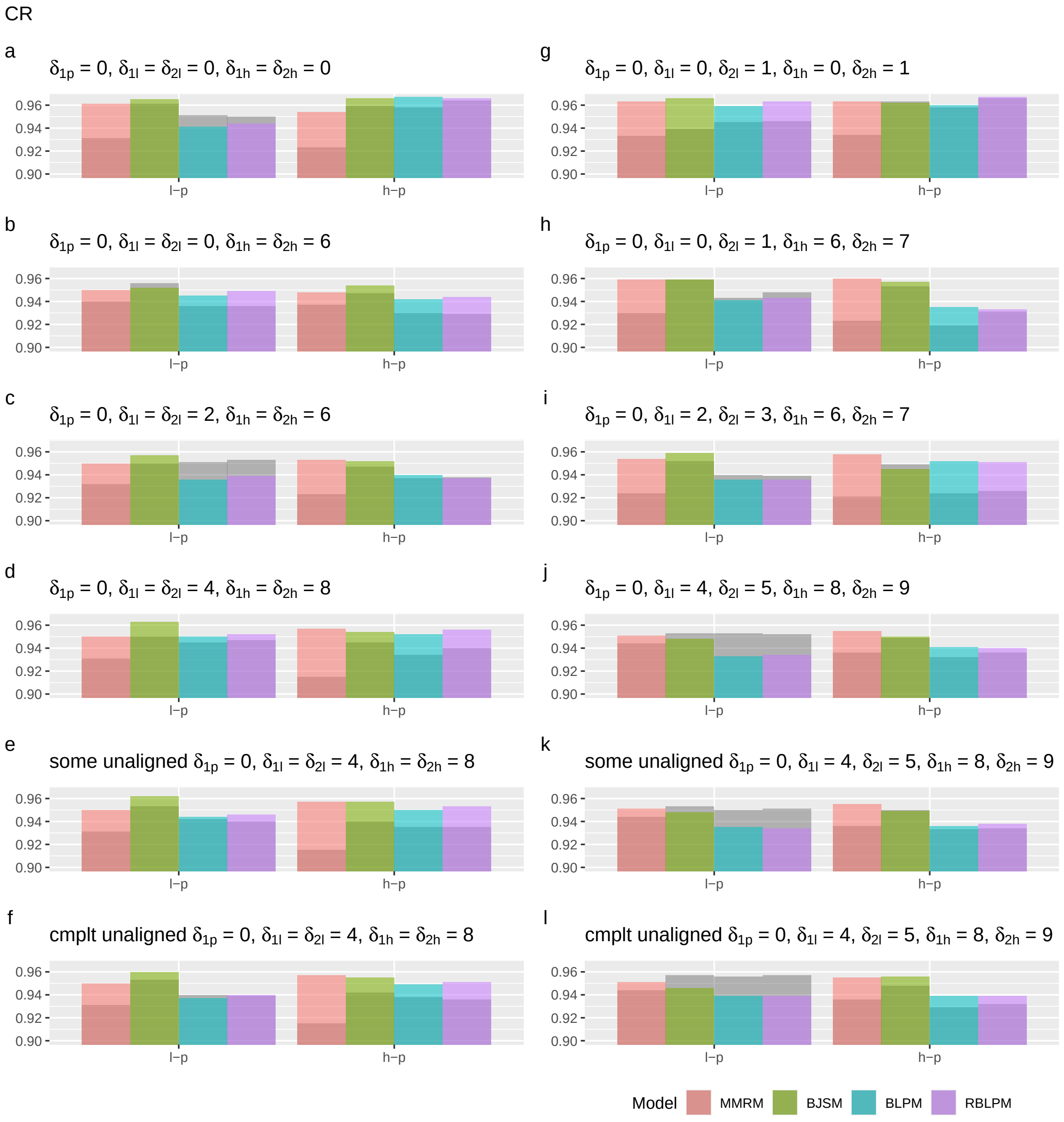


Figure 3.5: Simulated 95% coverage rate (CR) for the estimators of β_{1l} and β_{1h}
β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM. The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

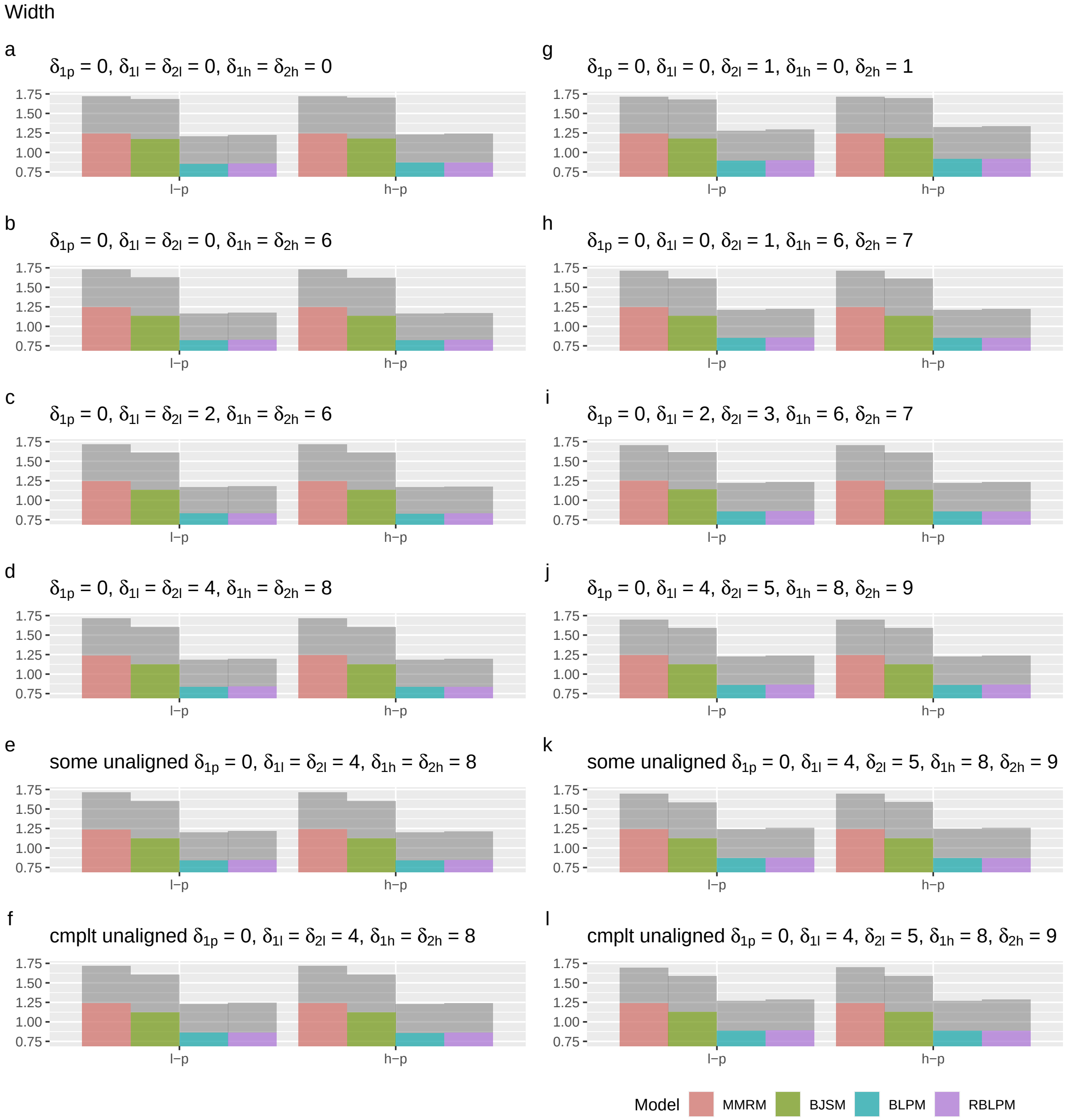


Figure 3.6: Simulated average credible interval width for the estimators of β_{1l} and β_{1h} β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM. The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

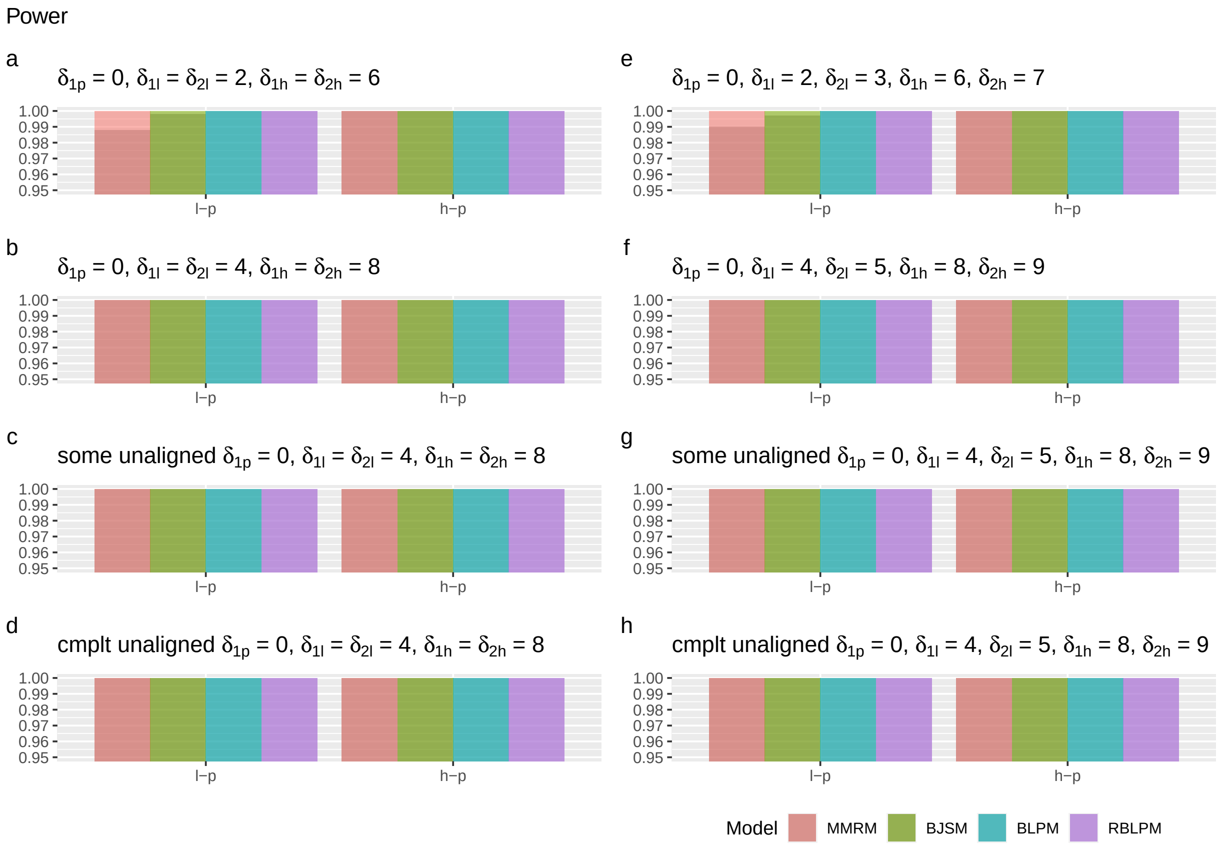


Figure 3.7: Simulated power of the traditional MMRM, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM methods. In this paper, the power is defined as the probability that the credible intervals of $\hat{\beta}_{1l}$ and $\hat{\beta}_{1h}$ do not include 0 when there are treatment effect difference between low dose and placebo and high dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The results of total sample size 50 are shown as the colored bars, while the results of total sample size 25 are shown as the overlaying grey bars. The simulation settings are described on the top of each graph. δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, where $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial.

and the MAC borrowing structure in addressing discrepancies between data sources, ensuring that the borrowing of external controls remains prudent, even when external controls are completely misaligned with the data from the current trial.

The right columns of Figure 3.3, 3.4, 3.5, and 3.6 reveal the rMSE, average bias, CR, and average CI width for β_{1l} and β_{1h} under data generation process 2, where the assumption of exchangeability between stages 1 and 2 is breached. Owing to the cross-stage information borrowing, BLPM methods exhibit larger positive average biases and, as a consequence, reduced CR. Nonetheless, their rMSE values remain notably lower than those of MMRM and BJSM. The impact on the CR can be lessened by choosing a larger standard deviation for the half-normal prior of τs, which allows the model to accommodate greater stage-wise discrepancies. Compared to the robust BLPM approach, the standard BLPM method displays greater biases under data generation process 2, aligning with expectations and underscoring the MAC structure's effectiveness in managing stage-wise non-exchangeability. While the enhancements from the robust BLPM are subtle under our simulation, due to the already high performance of the standard BLPM model, it is anticipated to show significant improvement under more pronounced non-exchangeability. The BJSM model that incorporates a "shift" parameter α (Fang et al. 2023), shows resilience to stage-wise non-exchangeability. Consequently, under data generating process 2, we still recommend the BLPM methods for their consistently lower rMSE and negligible biases. This conclusion holds true across all tested sample sizes, with results for $N = 250$ detailed in Figure 3.8, 3.9, 3.10, 3.11, and 3.12.

Table 3.2 displays the Type I error rates for each model. Across all tested scenarios, the Type I error rates for all models approximate 0.05. We have also calculated the power for each model (Figure 3.7 and 3.12). When the treatment effect is clinically significant (i.e., $\geqslant 4$), all methods demonstrate power nearing 1, regardless of the scenario or sample size. The BLPM methods and the BJSM show a notable increase in power compared to the traditional method when the treatment effect is equal to 2 and $N = 25$.

3.5 Re-analysis of tadalafil trial

To demonstrate the practical application of the snSMART design and BLPM methods, we conducted a reanalysis of the tadalafil study, obtaining patient-level trial data through Vivli. The summary-level NSAA total scores and 6MWD at baseline and week 48 are shown in Table 3.3. Considering that stage 2 of the tadalafil trial assigned or re-randomized treatments differently from an snSMART design, and the trial was stopped for futility after stage 1, it was necessary to reassign or rerandomize subjects in stage 2 according to the snSMART trial design and impute stage 2 data based on the new treatment assignment. Following Zhou & Reiter (2010), 100 sets of stage 2 data were imputed using the methods outlined in Section 3.3.3.6. We assumed that the treatment effects in stage 2 and

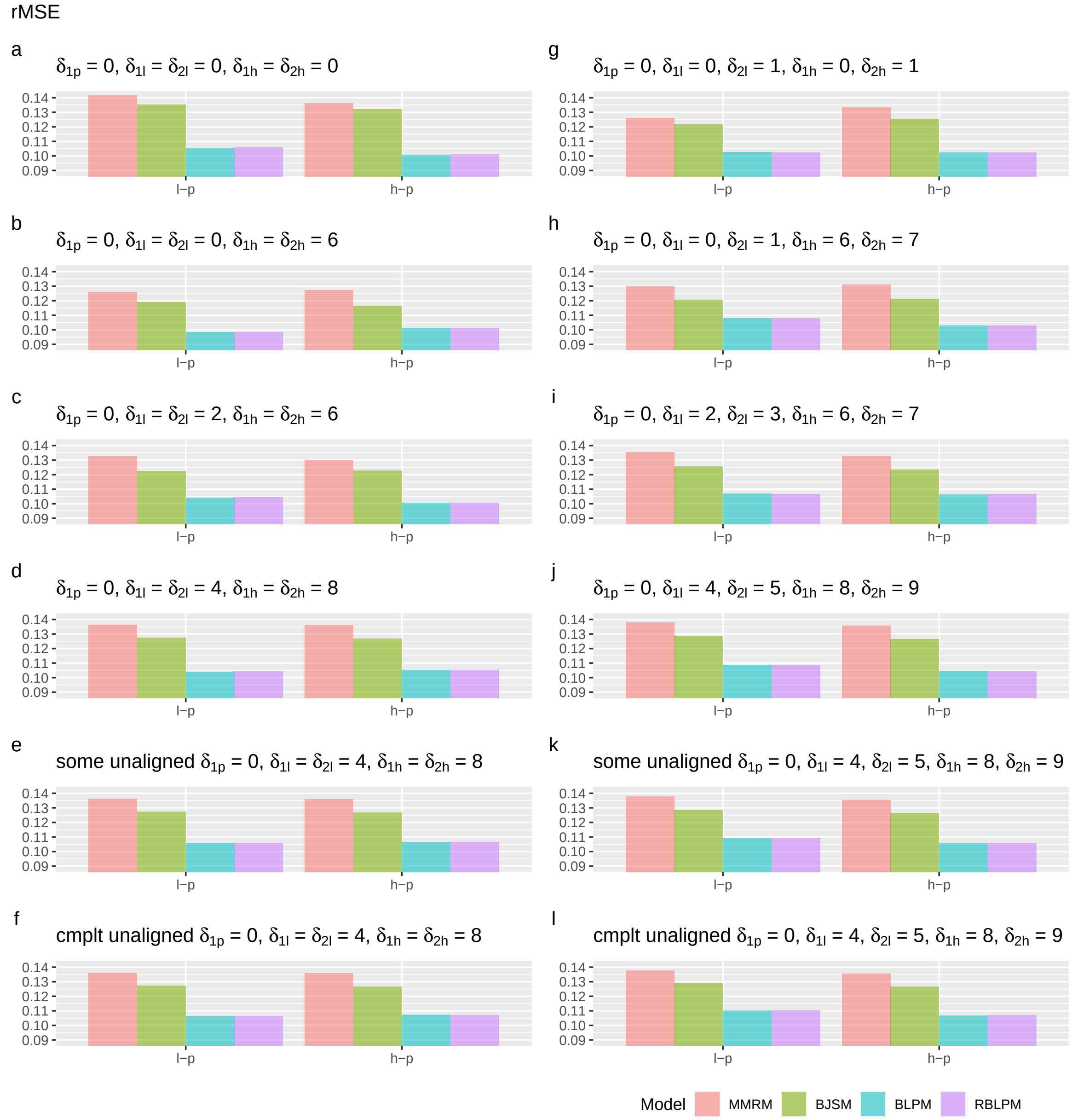


Figure 3.8: Simulated root-mean-square error (rMSE) for the estimators of β_{1l} and β_{1h} under $N = 250$.

Note: β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

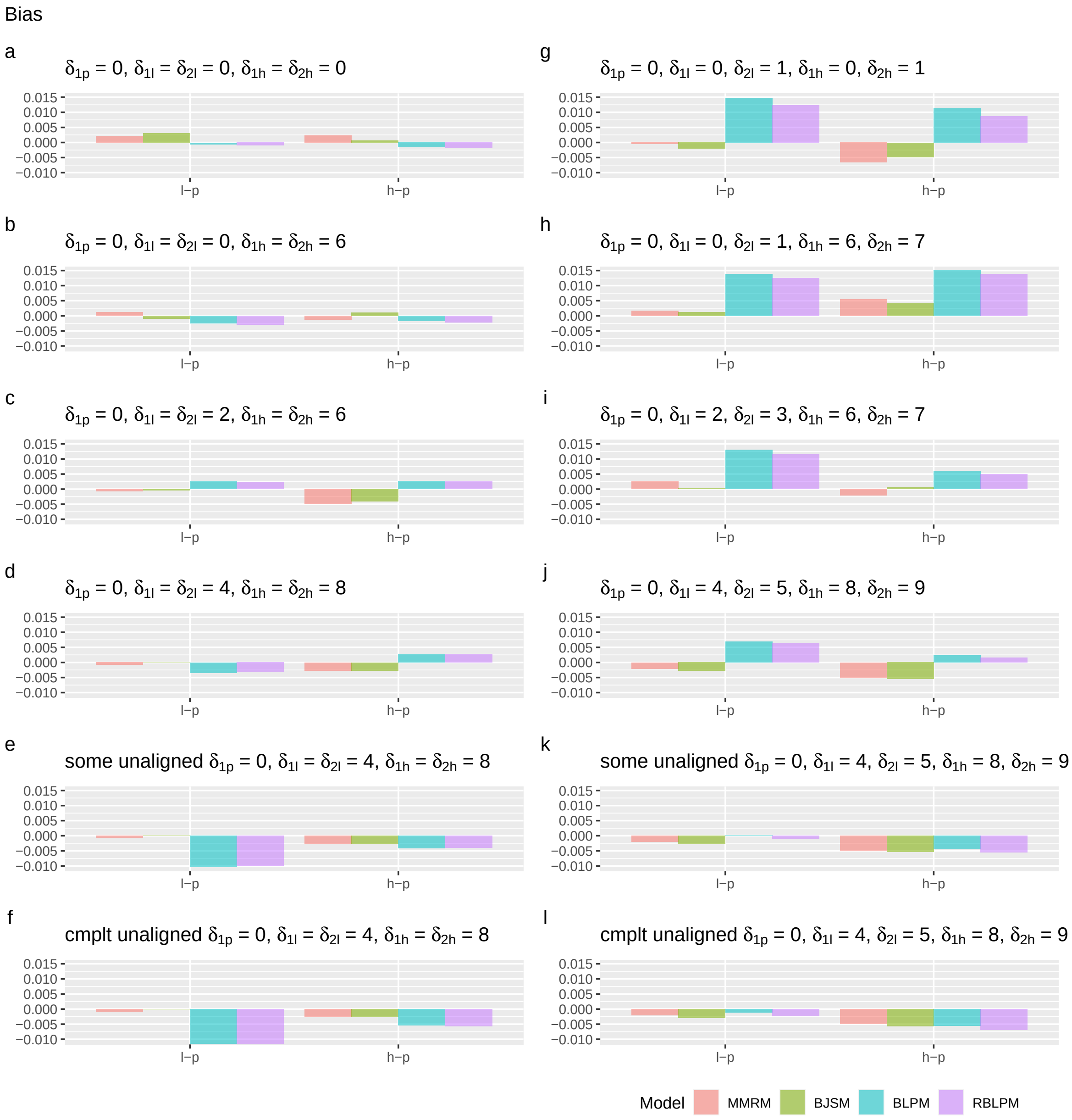


Figure 3.9: Simulated average bias for the estimators of β_{1l} and β_{1h} under $N = 250$ β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

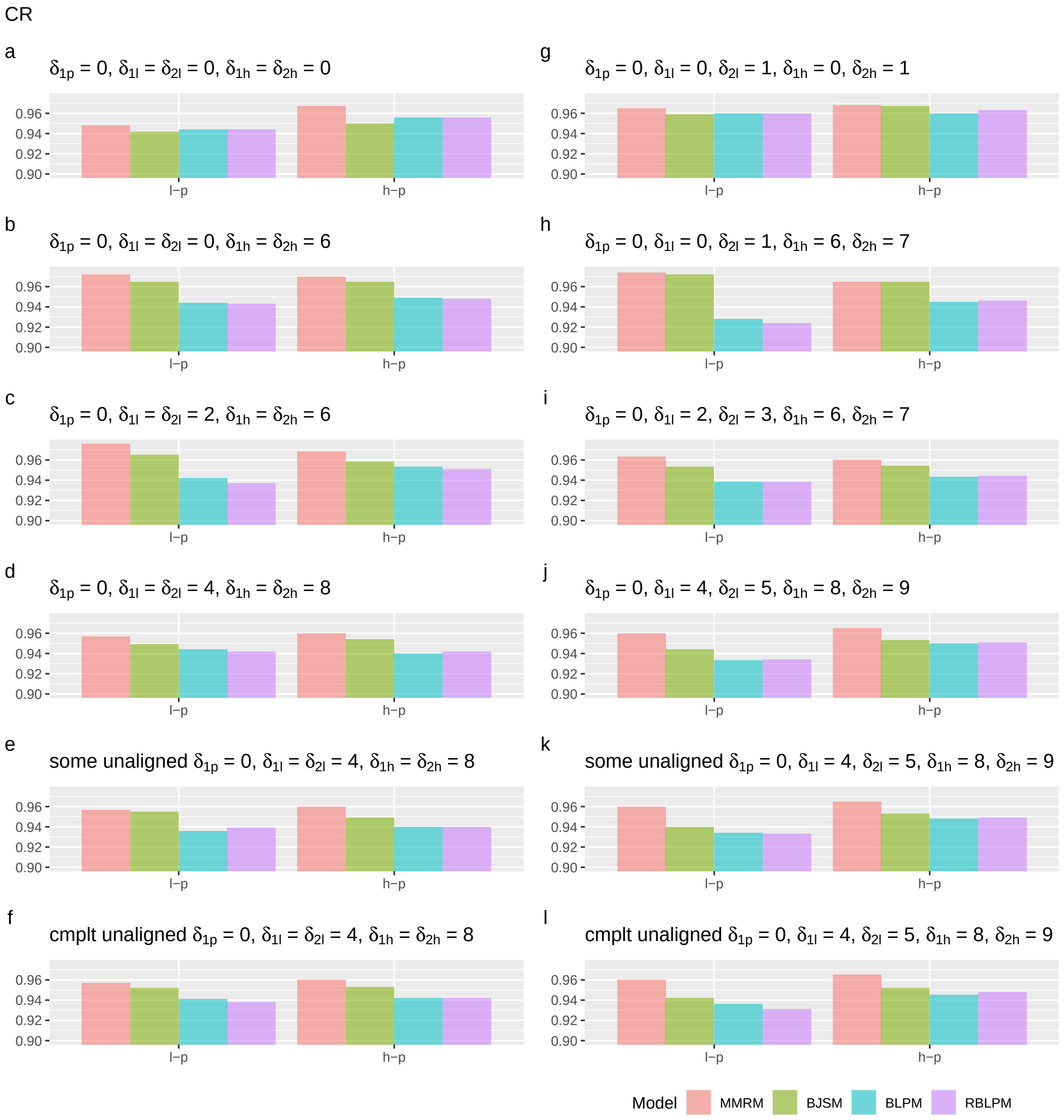


Figure 3.10: Simulated 95% coverage rate (CR) for the estimators of β_{1l} and β_{1h} under $N = 250$ β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

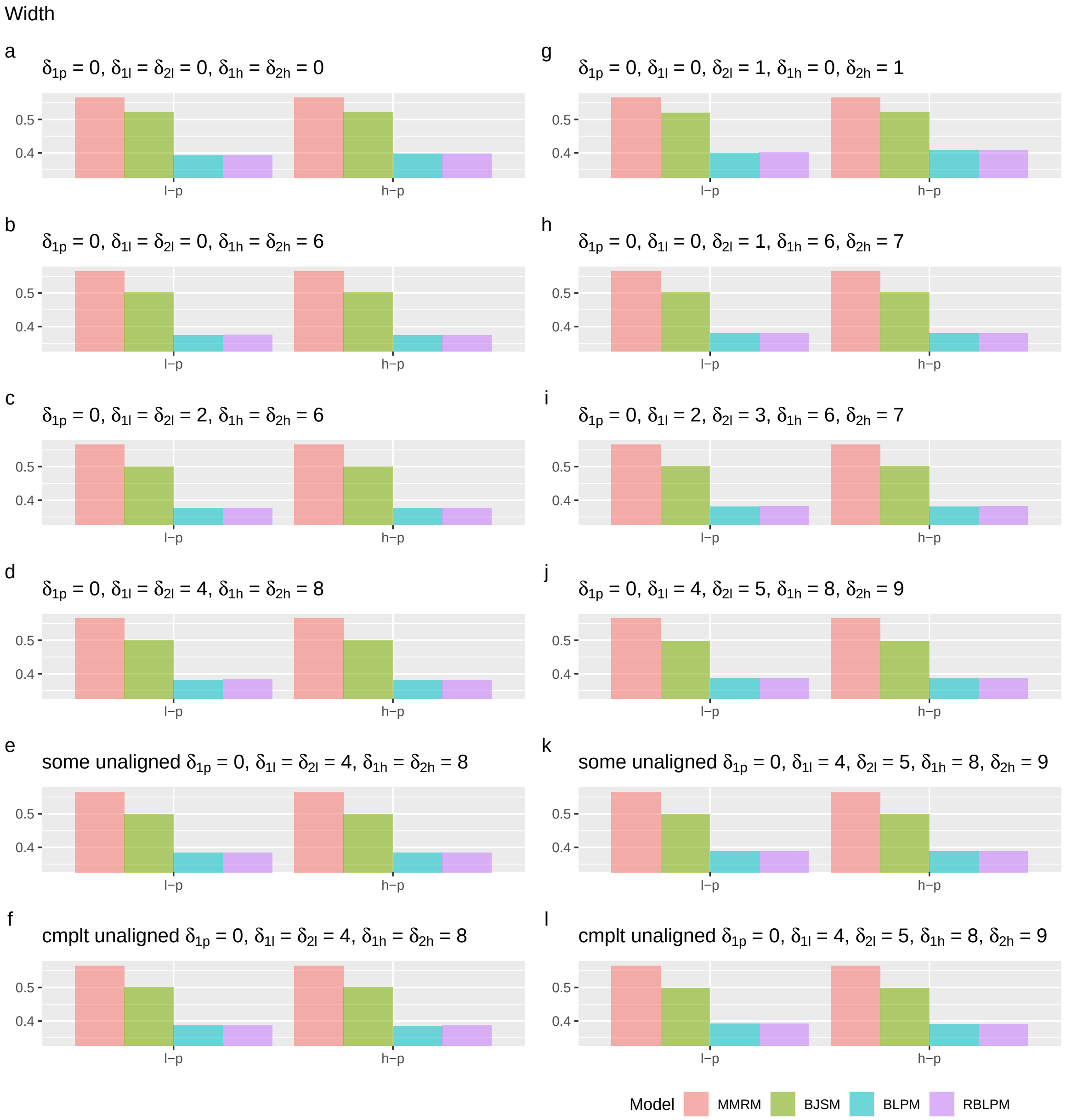


Figure 3.11: Simulated average credible interval width for the estimators of β_{1l} and β_{1h} under $N = 250$

β_{1l} (β_{1h}) is the treatment effect difference between stage 1 low (high) dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The simulation settings are described on the top of each graph, where δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial. Monte Carlo error (MCE) is smaller than 0.005 for all estimators.

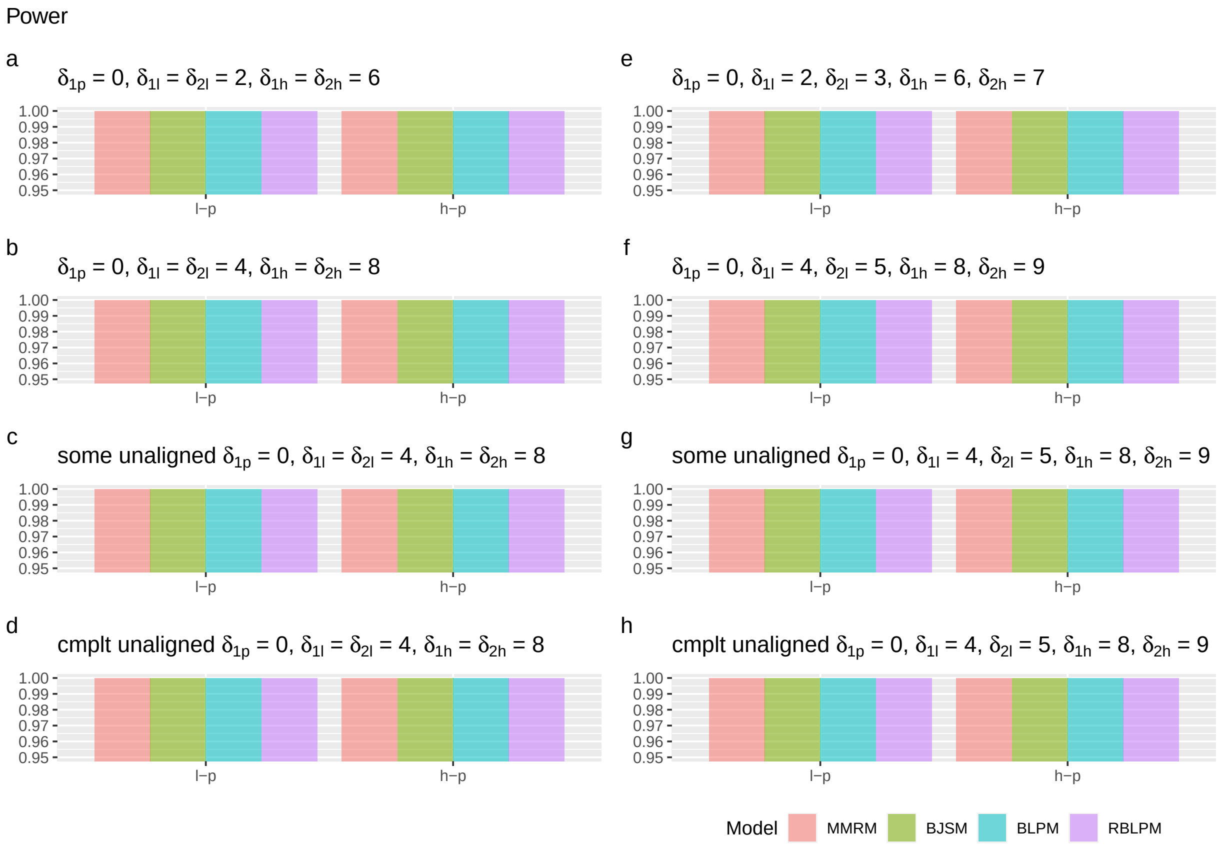


Figure 3.12: Simulated power of the traditional MMRM, the Bayesian joint stage model (BJSM), the standard BLPM, and the robust BLPM methods under $N = 250$. In this paper, the power is defined as the probability that the credible intervals of $\hat{\beta}_{1l}$ and $\hat{\beta}_{1h}$ do not include 0 when there are treatment effect difference between low dose and placebo and high dose and placebo. We compared the traditional MMRM model, the Bayesian joint stage model (BJSM), the standard BLPM and the robust BLPM. The simulation settings are described on the top of each graph. δ_{kT} denotes the true value of the expected treatment effects of treatment T in stage k, where $T = p, l, h$, $k = 1, 2$, and *some/cmplt unaligned* means some/all of the placebo treatment effects in external control data are inconsistent with the placebo treatment effect in the current trial.

Table 3.3: The tadalafil trial study measures. Summary statistics are reported as mean (SD) for 6MWD and NSAA.

	Placebo	Tadalafil Low Dose	Tadalafil High Dose
Baseline			
Sample Size	116	102	113
6MWD	337.5 (51.2)	323.4 (56.1)	327.0 (58.6)
NSAA	20.2 (7.0)	20.1 (7.6)	19.8 (7.2)
Week 48			
Sample Size - 6MWD	113	101	111
Changes in 6MWD	-51.0 (9.3)	-64.7 (9.8)	-59.1 (9.4)
Sample Size - NSAA	116	102	112
Changes in NSAA	-8.8 (1.1)	-9.3 (1.2)	-9.0 (1.1)

Table 3.4: Tadalafil trial 6MWD and NSAA percentage of missingness at each visit

	Baseline	Visit 1	Visit 2	Visit 3	Visit 4
6MWD	0.00%	3.93%	8.46%	12.39%	18.13%
NSAA	0.60%	2.42%	3.63%	4.83%	6.65%

stage 1 are exchangeable.

In line with the study protocol, we identified a change in 6MWD of $\geqslant -10\%$ from baseline as a clinically significant treatment effect for distinguishing between stage 1 responders and non-responders. We imputed missing 6MWD data in both external controls and the current trial, applying the same method to impute stage 1 and stage 2 NSAA data. In the tadalafil trial, the percentages of missingness for the 6MWD and NSAA at each visit are shown in Table 3.4.

For external control in this re-analysis, we used data from the Cure Duchenne project (NCT01753804) accessed via Vivli. We selected participants from NCT01753804 who matched the eligibility criteria of the tadalafil trial and had records of NSAA total score or 6MWD for the external control group. Ultimately, 49 subjects were included in the external control data. The baseline and week 48 summary statistics of these 49 subjects are shown in Table 3.5. Our assumption of exchangeability between external and current trial controls appears valid, given the careful selection of control data and the similarity in demographics and disease severity among patients (Table 3.3 and 3.5). Since the visit schedule for participants in NCT01753804 was less frequent than in the tadalafil trial, we imputed two missing visits for each participant. Given that the snSMART design requires a randomization ratio of 1:2:2 in stage 1, we randomly removed half

Table 3.5: Summary statistics (mean & SD) for 6-minute walk distance (6MWD) and North Star Ambulatory Assessment (NSAA) score, derived from external control data of 49 subjects in NCT01753804.

	6MWD	NSAA
Baseline	331.1 (47.0)	20.8 (5.0)
Week 48	280.4 (89.9)	15.7 (6.7)

Table 3.6: Example Data Analysis Results. Note: $\widehat{\beta}_{1k}$ represents the estimated difference in the stage 1 treatment effect between dose level k and placebo at week 48 for either NSAA or 6MWD, where k can be either l for low dose or h for high dose. The analysis compares four methods: traditional MMRM, Bayesian joint stage modeling (BJSM), BLPM, and robust BLPM (RBLPM). "6MWD" refers to the 6-minute walk distance, and "NSAA" refers to the total score of the North Star Ambulatory Assessment. The 95% confidence or credible intervals of the estimates are displayed in parentheses.

Estimator	MMRM	BJSM	BLPM	RBLPM
6MWD				
$\widehat{\beta}_{1l}$	-0.4 (-22.1, 21.3)	-4.9 (-20.0, 10.5)	-1.2 (-16.1, 14.3)	-0.6 (-15.9, 15.0)
$\widehat{\beta}_{1h}$	-7.1 (-28.5, 14.2)	-11.8 (-27.4, 3.8)	-7.9 (-23.2, 7.7)	-7.3 (-22.6, 8.0)
NSAA				
$\widehat{\beta}_{1l}$	0.3 (-1.0, 1.5)	0.3 (-0.8, 1.3)	0.3 (-0.8, 1.3)	0.3 (-0.8, 1.3)
$\widehat{\beta}_{1h}$	0.0 (-1.2, 1.2)	0.1 (-1.0, 1.2)	0.1 (-1.0, 1.2)	0.1 (-1.0, 1.1)

of the subjects from the placebo arm of the tadalafil trial before applying the BLPM methods.

We applied the traditional MMRM, BJSM, and BLPM methods, with results displayed in Table 3.6. It is noticeable that the estimators obtained from the BLPM methods and MMRM are consistent with each other, and BLPM methods exhibit narrower CI widths than the traditional method due to the efficient use of data across both stages. The BJSM method's 6MWD estimations slightly deviate from those of the MMRM and BLPM methods. We believe this deviation is due to the significant variations in patient-level 6MWD decrease across stages, and the BJSM method's sensitivity to irregular stage-wise non-exchangeability. Thus, although the BJSM and BLPM methods arrived at the same conclusion as the tadalafil trial, i.e., failing to reject the null hypothesis, these two methods provided more precise estimations of treatment effects.

3.6 Discussion

This paper introduces a significant advancement in statistical methods for analyzing data from an snSMART in the context of drug development for DMD and similar rare diseases, with a focus on integrating external control data. The proposed BLPM methods adeptly manage the heterogeneity between external controls and the current trial. The BLPM methods also leverage longitudinal data and baseline characteristics of subjects, a capability previously unavailable in the snSMART setting. Moreover, BLPM addresses concerns about stage-wise non-exchangeability, a common issue for practitioners. We offer guidance on selecting prior distributions and alternative models for sensitivity analyses, emphasizing the importance of assessing a broad spectrum of scenarios during the planning stage. This evaluation is essential for comprehending the model's assumptions, the impact of prior choices, and the effect of sample size on the design and analytical methods, thereby aiding sponsors and regulators in appreciating the model's practical efficiency and robustness.

Although our method can manage non-exchangeability of treatment effects across different stages, we advise applying the snSMART design to conditions with stable disease progression (including a sufficient washout period). This ensures relatively consistent treatment effects throughout the stages, especially since the aim of stage 2 in an snSMART is to collect additional data to refine the estimation of stage 1 treatment effects. At the end of the trial, conducting sensitivity analyses to compare outcomes from traditional MMRM with those from the BLPM methods will help evaluate the validity of our assumptions regarding stage-wise treatment effect exchangeability.

We suggest allowing stage 1 high dose responders the opportunity to be re-randomized to a lower dose in stage 2, based on the possibility that a patient responding to a high dose might also respond to a lower, less toxic dose. If this assumption proves unrealistic, the snSMART design could be adapted to maintain high dose treatment in stage 2 for all stage 1 high dose responders. The BLPM methods would seamlessly adapt to this modified trial design.

The BLPM methods utilize a MAC framework that models all data sources collectively. Alternatively, the MAP framework could be employed to generate informed priors based on external controls for current trial estimates. The MAP and MAC equivalence is further elaborated by Schmidli et al. (2014).

Our proposed BLPM methods employ PS trimming and IPTW weighting prior to fitting the main BLPM model. As an alternative, we could implement propensity score matching. This approach would further assist in selecting external control patients for analysis, ensuring a balanced representation of key prognostic factors.

In this version of the BLPM methods, we treated time as a continuous variable. This decision was informed by observations from the tadalafil trial, where the decline in 6MWD and NSAA remained relatively stable throughout the trial period. However, there are instances where trials are conducted on diseases with outcomes that exhibit significant fluctuations at each visit. In such cases, we propose two potential modifications: first, employing categorical time in the model, and second, adjusting the piecewise treatment effect to include shorter segments, potentially as frequent as each visit. The primary challenge with this approach is the increased complexity due to the additional parameters required in the model. Evaluating the effectiveness of this modified model is an area of future research.

While the BLPM methods effectively address data source heterogeneity, the extent of this heterogeneity affects the ESS of external controls, which is crucial for ensuring trial power. Future work will develop methods to assess the ESS of external controls in an snSMART, guiding the quantity of external control data needed to further enhance trial outcomes.

CHAPTER 4

Dynamic Prediction of Landmark Survival Time in Cancer Clinical Trials Using a Joint Modeling Framework

4.1 Introduction

In recent years, many cancer drugs have received regulatory approvals based on PFS as the primary endpoint. Although PFS is a well-accepted surrogate endpoint for OS in many cancer types, improvement in survival remains the clinical gold standard for assessing a patient's benefit (Tang et al. 2007, Driscoll & Rixe 2009, Methy et al. 2010, Grigore et al. 2020). However, planning a statistically well-powered OS analysis in practice often presents challenges due to longer survival times in early-stage cancers, patients switching to alternative treatments after progression, starting other anti-cancer therapies, or being lost to follow-up. The timing of the primary analysis for trials with PFS as the primary endpoint is determined by the number of patients progressed, and survival data is often not mature enough for meaningful statistical inference due to a low number of deaths. Hence, if PFS is statistically significant in the final analysis, regulatory agencies often request one or more updated OS analyses once the survival data are more mature. Updated OS analyses are crucial for market access and reimbursement decisions. Therefore, proper planning is essential to ascertain when mature OS data will be available.

Model-based prediction of survival time for trial participants aids research teams in efficiently allocating resources, accurately planning future OS analyses, and understanding the potential survival benefit of an experimental drug. It also assists in designing and committing to Phase 3 trials based on disease progression data observed in Phase 2 studies. Moreover, it helps clinicians and scientists understand the mechanism of a new compound that improves patients' survival while reducing disease burden. Besides drug development, mature OS data are crucial for decision-making by both patients and clinicians, and understanding a drug's cost-effectiveness. Sborov et al. (2019) demonstrated that when oncologists inaccurately predict OS, patients with advanced cancer are more likely to receive aggressive end-of-life care, often contradicting the patients' wishes. Mackillop & Quirt (1997) stated that accurate OS prediction facilitates the effective use of limited healthcare resources and helps patients make suitable plans for their remaining lives. Henderson et al. (2001) further provides examples of how accurate survival predictions have financial implications for insurance programs and health authorities.

PFS is defined as the time from randomization until PD or death from any cause. In oncology studies, progression is evaluated using the Response Evaluation Criteria in Solid Tumors (RECIST) (Eisenhauer et al. 2009). RECIST provides standardized definitions and procedures for documenting how subjects progress, respond, or remain stable in terms of their disease burden during a treatment course. RECIST offers methods for assessing solid tumor responses using X-ray, CT, and MRI scans and is recommended by the National Cancer Institute. Note, PFS is viewed as a composite outcome with four components: 1) measurement of the target lesion, which is captured as longitudinal continuous data, 2) time to non-target lesion progression, 3) time until the emergence of a new lesion, and 4) time to death from any cause. Depending on the cancer type, each of these components has a different degree of predictivity for OS. For example, Stein et al. (2013) showed that the progression of non-target lesions or the appearance of new lesions is more predictive of OS benefit than the sum of the longest tumor diameters of the target lesions for patients with metastatic renal cell carcinoma. The current practice of OS prediction often overlooks the disease progression process and models only the number of deaths.

In this manuscript, we evaluated several model-based approaches to forecast the death times of trial participants "still alive", leveraging information about disease progression along with important baseline factors. The relationship between PFS and OS has garnered interest in both statistical and clinical literature over time. Notably, Fleischer et al. (2009) employed exponential time-to-event distributions to describe the dependency structure between OS and PFS. This is further expanded by Weber & Titman (2019), Fu et al. (2013), Meller et al. (2019) by utilizing copula and multi-state models to jointly analyze PFS and OS without strict parametric assumptions for the marginal survival distributions of PFS and OS. Another important work by Shukuya et al. (2016) investigated the correlation between median PFS and median OS, concluding that both tumor response and PFS are significant predictors of OS. Other methodologies for OS prediction include joint modeling of longitudinal tumor size data of the target tumor and survival data (Claret et al. (2009, 2013), Wang et al. (2009), Bruno et al. (2014), Zecchin et al. (2016), Lim et al. (2019)), which incorporates baseline survival data into the OS prediction. Meanwhile, Yu et al. (2020) jointly modeled the dynamics of target lesions and the progression of non-target lesions to predict PFS. However, most of the literature in this area focuses primarily on improving the estimation of OS rather than predicting the future survival time of patients ongoing in the trial. Moreover, to our current knowledge, there is no existing methodology that formally combines all three components of disease progression, i.e., target lesion, non-target lesion and new lesion. In this paper, we propose a joint modeling framework to address these gaps and further enhance the copula model and the multi-state model for predicting the survival times of ongoing patients.

The rest of the paper is structured as follows: A motivating example from a Phase 3 renal cell carcinoma study is used to establish the problem in Section 4.2. This is followed by a set of proposed models that explore the correlation between disease progression and survival as presented in Section 4.3. The simulation studies conducted to test the properties of our method are presented in Section 4.4. The renal cell carcinoma example is revisited in Section 4.5 to illustrate the practical utility of the proposed methodology. Finally, we conclude with a discussion that summarizes the key messages and lessons learned.

4.2 Motivating Example: Phase 3 Study in Renal Cell Carcinoma

The methods outlined in this paper are inspired by a phase 3, randomized, open-label, parallel-arm study. A total of 800 treatment-naive adult participants with advanced renal cell carcinoma were randomized (1:1) to receive either the experimental drug or the standard of care. A key inclusion criterion was the presence of at least one measurable lesion, as defined by RECIST version 1.1, which had not been previously irradiated. Tumor assessments were performed using computed tomography or magnetic resonance imaging at baseline, every six weeks post-randomization for the first 18 months, and subsequently every 12 weeks until confirmed disease progression. Important baseline demographic and disease characteristics were evenly distributed between the two treatment groups. The study had two primary endpoints: PFS and OS among patients with programmed death ligand 1 (PD-L1)–positive tumors. The primary analysis was scheduled when at least 397 PFS events occurred (Figure 4.1). At the time of the primary PFS analysis only 146 deaths were observed which was immature to draw any reasonable conclusion for OS. Therefore, an updated analysis of OS was planned after 341 deaths. The trial findings indicated that patients administered the experimental drug experienced a notably extended PFS compared to those given the standard of care. It is crucial to note that while we have utilized the data structure from a real-life trial, the actual values presented here are simulated based on published summary statistics, both for proprietary considerations and to ensure proper protection of patient data.

We aim to use the observed data to construct a predictive tool that provides a reliable estimate of the time of the *n*th death in the trial. Specifically, we utilize data on participants' baseline characteristics, treatment group, OS, and the processes underlying PFS, including measurements of target lesions and the progression of disease in non-target and new lesions. This approach is

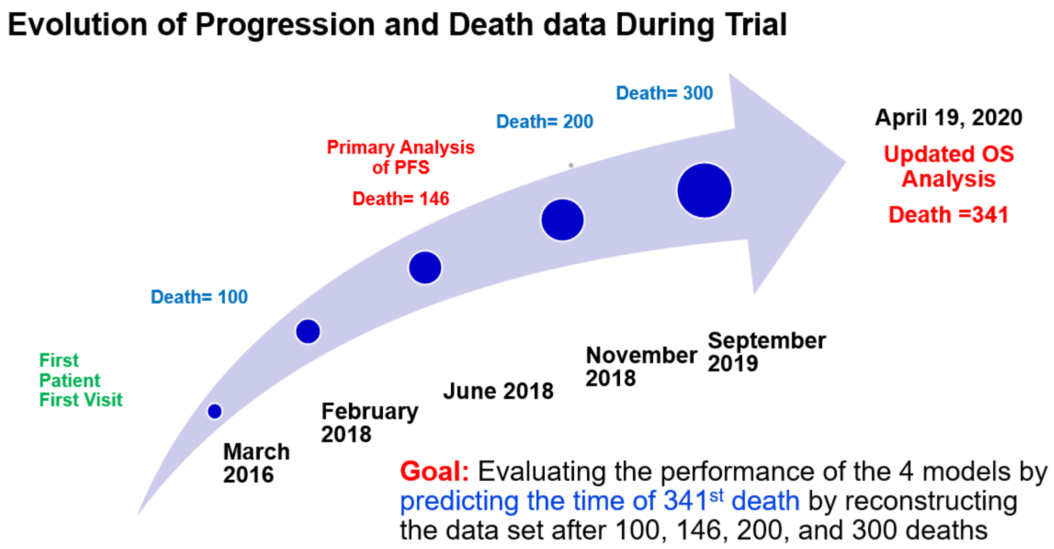

Evolution of Progression and Death data During Trial
Death= 300
Death= 200
April 19, 2020
Updated OS Analysis
Death =341
Primary Analysis of PFS
Death= 146
Death= 100
First Patient First Visit
March 2016
February 2018
June 2018
November 2018
September 2019
Goal: Evaluating the performance of the 4 models by predicting the time of 341st death by reconstructing the data set after 100, 146, 200, and 300 deaths

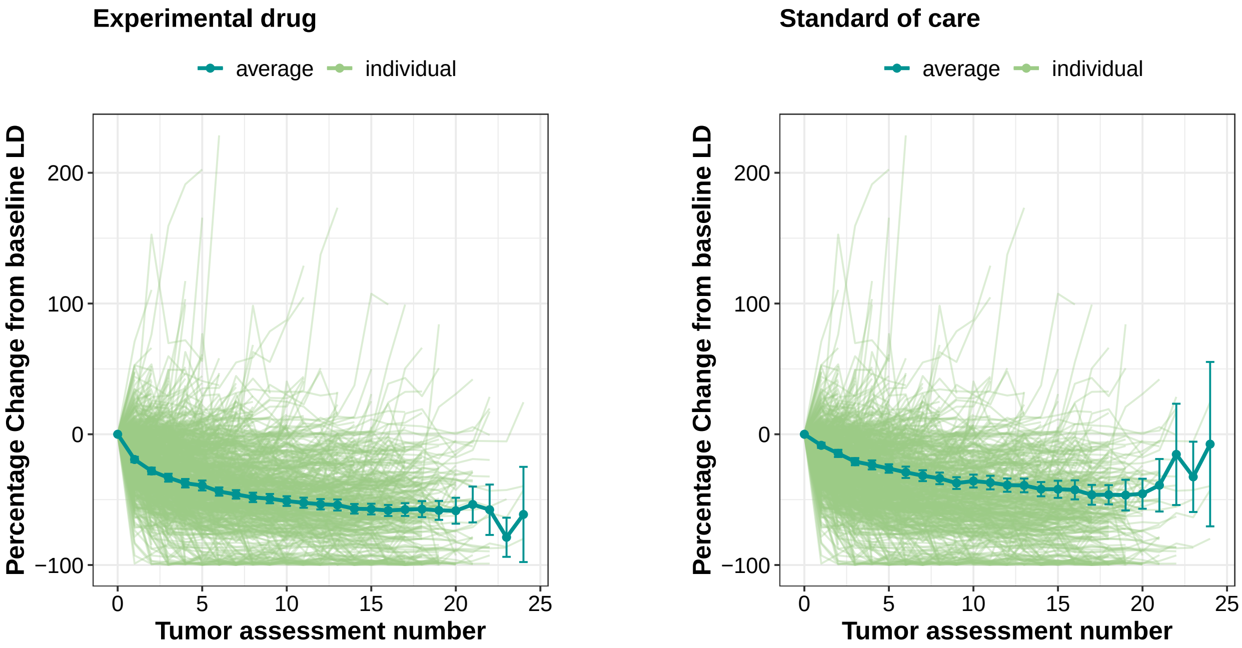
Figure 4.1: Evolution of progression and death data during trial

justified as PFS is known to be a strong predictor of OS in advanced renal cell carcinoma (Heng et al. 2011)

Figure 4.2 demonstrates significant variability among individuals in the sum of the longest diameters of target lesions, making it difficult to draw any conclusions about the treatment without systematic modeling of this variability. Moreover, Kaplan-Meier plots for non-target lesions and new lesions are shown in Figures 4.3b and 4.3c, respectively. From both plots, we observe differences between the two treatment arms: patients who received the experimental drug exhibited a longer time to PD when considering assessments of non-target lesions and new lesions. A similar trend is also observed in the Kaplan-Meier plots for OS (Figure 4.3d) and PFS (Figure 4.3e)

4.3 Method

In this manuscript, the general process of using Bayesian statistical models to generate OS predictions is as follows: At any given point during the clinical trial, we collect the current data and input it into our Bayesian models. Subsequently, a Markov chain Monte Carlo (MCMC) process is utilized to estimate the parameters. These estimations are then used to generate predicted OS outcomes by leveraging the posterior predictive distribution (PPD) technique. This process allows us to create a predictive framework that is both dynamic and reflective of the evolving nature of trial data, providing timely and robust predictions for OS.



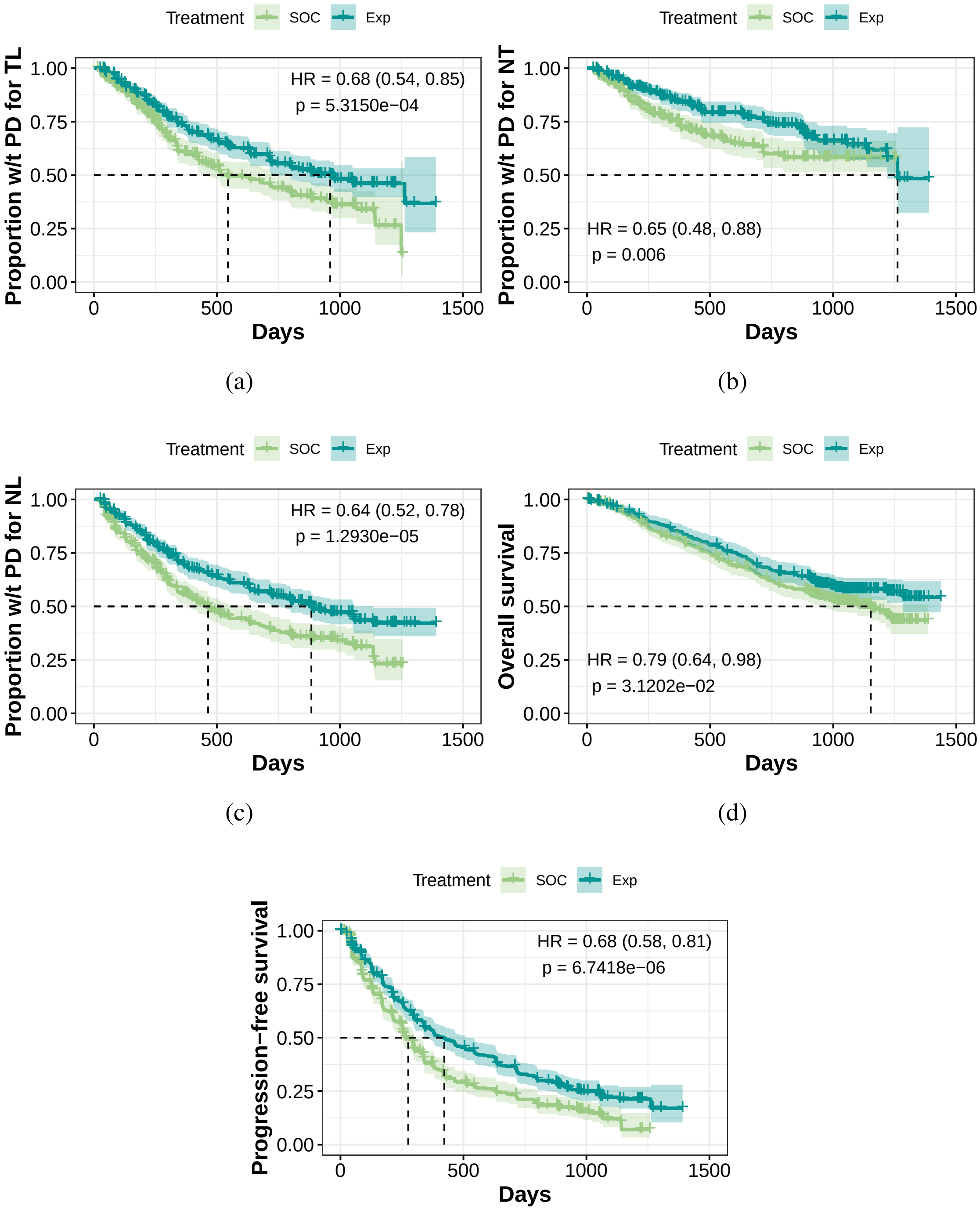
(a)

(b)

Figure 4.2: Profile plot of the sum of the longest diameter of target lesions

4.3.1 Bayesian Model-Averaged Joint Modeling Approach

In this section, we outline our proposed model by establishing three distinct joint models between each of the processes of disease progression and OS. These three joint models include: 1) the joint model between the sum of the longest diameter of target lesions and OS, 2) the joint model between the time to non-target lesion PD and OS, and 3) the joint model between the time to new lesion PD and OS. Furthermore, we incorporate a marginal model directly modeling OS. Three main steps are included in our model. First, we mitigate "information loss" by considering all four granular components of PFS. By constructing joint, marginal models for each component, we derive real-time OS predictions from each trial participants "still alive". This results in four sets of intermediate predictions. Second, our proposed methods fully encapsulate the association between progression and death by incorporating random processes. Third, we enhance the accuracy of OS predictions from PFS by deriving the final OS prediction from all four models' intermediate predictions simultaneously using BMA (Hoeting et al. 1999). Using the BMA technique, we blend these different predictions to create a final predicted OS, which combines all four individual forecasts with appropriate weights. Thus, joint modeling represents a promising scientific approach to overcome the challenges associated with OS prediction, in comparison to existing methods that will be discussed in the subsequent sections. The structure of the proposed multivariate joint



(a)

(b)

(c)

(d)

(e)

Figure 4.3: Kaplan-Meier plots at the time of updated OS analysis considering target lesions progressive disease (PD) (a), non-target lesions PD (b), new lesions PD (c), overall survival (d), and progression-free survival (e).

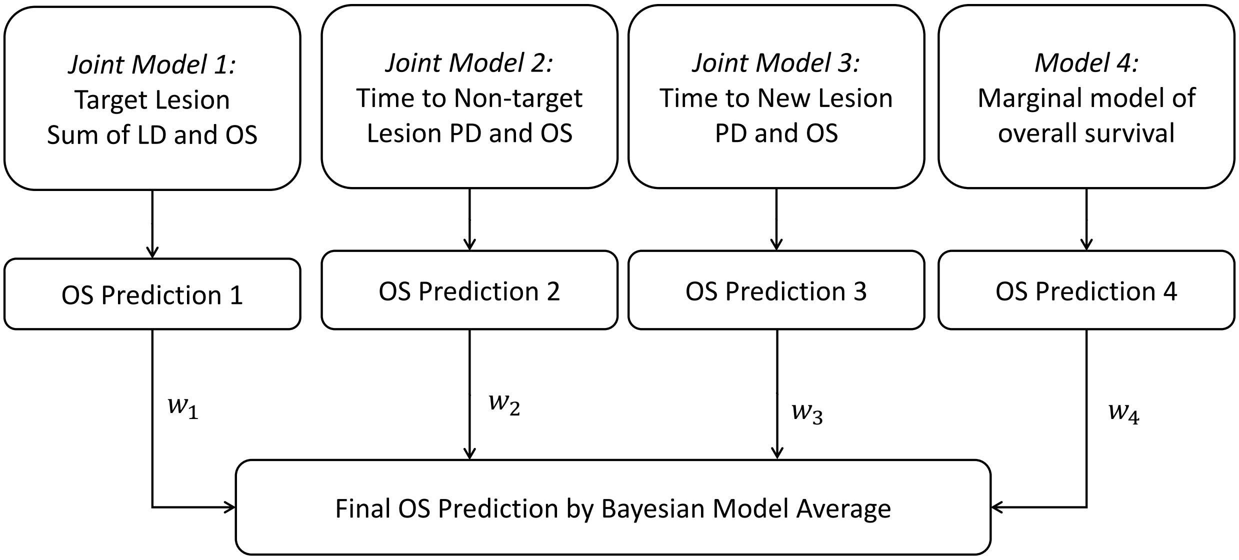


Figure 4.4: Multivariate joint modeling structure.

modeling approach is illustrated in Figure 4.4, with details provided in the following subsections.

4.3.1.1 Joint Model for the Association between Target Lesion and OS

We consider a set of n subjects followed over an interval of time [0,τ]. Let z_{ij} be the target lesion measurement (or % change from baseline) at time t_{ij} (for the i-th participant at the j-th visit), where $i = 1, ..., n$ and $j = 1, ..., k_i$. We assume the target lesion measurement time t_{ij} is non-informative so that it is independent of the longitudinal measurement and event-time processes for OS. We define a latent zero-mean Gaussian process $W_i(t)$ for the i-th participant, independent of different participants. The following sub-models link the joint model of lesion measurements and OS:

The observed lesion measurement $Y_i(t) = \mu_i(t) + \epsilon_i$, where ϵ_i is the measurement error, and the lesion measurement process μ_{i1}, μ_{i2},... is modeled by the linear model $\mu_i(t) = \mathbf{X}_i(t)'\boldsymbol{\beta}_\mu + W_i(t)$. Here, $\mathbf{X}_i(t)$ is the time-dependent covariate matrix for subject i's lesion measurement. $W_i(t)$ is a Gaussian process that has the form $W_i(t) = b_{1i} + b_{2i}t$, where (b_{1i}, b_{2i}) are zero-mean bivariate Gaussian variables with variances σ_1^2 and σ_2^2 respectively, and correlation coefficient ρ.

The OS process at time t is modeled by $\lambda(t) = \lambda_0(t)\exp(\mathbf{Z}_i'\boldsymbol{\beta}_{\text{OS}} + \lambda\mu_i(t) + b_{3i})$. Under a Weibull baseline hazard, $\lambda(t)$ has the form $\lambda(t) = \alpha_{\text{OS,TL}}\gamma_{\text{OS}}t^{\alpha_{\text{OS,TL}}-1}\exp(\mathbf{Z}_i'\boldsymbol{\beta}_{\text{OS}} + \lambda\mu_i(t) + b_{3i})$, where $\mathbf{Z}_i$ is the covariate matrix for OS, γ_{OS} and $\alpha_{\text{OS,TL}}$ are the scale and shape parameters of the Weibull distribution, and $b_{3i} \sim N(0, \sigma_3)^2$ are random effects orthogonal to the measurement process. The construction of the likelihood for this joint model is detailed below:

The marginal distribution of the observed measurements μ is easily obtained. The likelihood for the observed data can be factorized as the product of this marginal distribution and the conditional distribution of OS, given the observed values of μ. Let $\boldsymbol{\theta}_1$ denote the combined vector of unknown parameters. Conditional on lesion measurements μ, OS is independent of these measurements μ, so we can write the likelihood $L = L(\boldsymbol{\theta}_1, \mu, OS)$ as $L = L_\mu(\boldsymbol{\theta}_1, \mu) \times L_{OS|\mu}(\boldsymbol{\theta}_1, OS|\mu)$, where $L_\mu(\boldsymbol{\theta}_1, \mu)$ is of standard form corresponding to the marginal multivariate normal distribution of μ and

$$L_{OS|\mu}(\boldsymbol{\theta}_1, OS|\mu) = \prod_i \left\{ \left[\lambda_0(t) \exp(\mathbf{Z}_i' \boldsymbol{\beta}_{\text{OS}} + \lambda \mu_i(t)) \right]^{\delta_{OS,i}} \right.$$
$$\left. \times \exp\left[- \int_0^{y_{OS,i}} \lambda_0(t) \exp(\mathbf{Z}_i' \boldsymbol{\beta}_{\text{OS}} + \lambda \mu_i(t)) dt \right] \right\},$$

here, $y_{OS,i} = \min(t_{OS,i}, c_{OS,i})$ and $\delta_{OS,i} = I(t_{OS,i} \leq c_{OS,i})$. $c_{OS,i}$ is the censoring time for the ith participant and $t_{OS,i}$ is the true event time.

4.3.1.2 Joint Model for the Association between Time to Non-Target Lesion and OS

Copulas are commonly used for modeling the dependence between random variables. They are continuous multivariate cumulative distributions with each random variable following a uniform marginal distribution on the interval $[0, 1]$. Sklar's theorem states that any multivariate joint distribution can be written using univariate marginal distribution functions and a copula describing the variables' dependence structure (Sklar 1959). Therefore, copulas can be used to model the dependence between these survival times, allowing for more accurate prognostic prediction. Here, we can use the copula to jointly model time to non-target lesion and OS, where time to non-target lesion corresponds to the time between randomization and PD of a non-target lesion is observed. A bivariate copula function $C : [0, 1]^2 \rightarrow [0, 1]$ of the endpoints (T, O), where $T :=$ *time to non-target lesion* and $O := OS$, can be expressed by joint survival function $\boldsymbol{S}(t, o) = T(T \geq t, O \geq o) = C(\boldsymbol{S}_T(t), \boldsymbol{S}_O(o))$, where $t, o \geq 0$, $\boldsymbol{S}_T$ and $\boldsymbol{S}_O$ are the marginal survival functions of T and O, respectively (Weber 2020). To capture various dependency patterns, one can employ different copula families. Here, we opt for the Clayton copula function (Clayton 1978), which belongs to the Archimedean copula class and can capture strong dependence in the left tail. By doing so, we can express the relationship between time to non-target lesion and OS using a single parameter η_{NT}, a parameter that can be directly linked to Kendall's tau, effectively characterizing their dependence. In the subsequent paragraphs, we formally define the copula model between time to non-target lesion and OS.

Let t_{NT} and t_{OS} denote times to non-target lesion and death (OS), respectively, and let $\mathbf{Z}$ be covariates of interest. For a participant with covariates $\mathbf{Z}$, let $\lambda_{\text{NT}}(t|\mathbf{Z})$ denote the hazard function for the time to non-target lesion. Under the Cox proportional hazards model, we have $\lambda_{\text{NT}}(t_{NT}|\mathbf{Z}) = \lambda_{0,\text{NT}}(t_{NT})\exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})$. When the baseline hazard $\lambda_{0,\text{NT}}(t_{NT})$ is modeled by a Weibull distribution, the corresponding survival function is given by $S_{\mathrm{NT}}(t_{NT}|\mathbf{Z}) = \exp\{-\gamma_{\mathrm{NT}} t_{NT}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}$, such that γ_{NT} and α_{NT} are the scale and shape parameters of the Weibull distribution, respectively. Similarly, for OS we have $S_{\mathrm{OS}}(t_{OS}|\mathbf{Z}) = \exp\{-\gamma_{\mathrm{OS}} t_{OS}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}$. Hence, the Clayton copula model can be specified as $S_1(t_{\mathrm{NT}}, t_{\mathrm{OS}}|\mathbf{Z}) = \{S_{\mathrm{NT}}(t_{\mathrm{NT}}|\mathbf{Z})^{-\eta_{\mathrm{NT}}} + S_{\mathrm{OS}}(t_{\mathrm{OS}}|\mathbf{Z})^{-\eta_{\mathrm{NT}}} - 1\}^{-1/\eta_{\mathrm{NT}}}$, where $\eta_{\mathrm{NT}} > 0$ measures the correlation between time to non-target lesion and OS. To take into account the heterogeneity of the patient population, we introduce random effects for the shape parameter of the Weibull model denoted as b_{NT_i} and b_{OS_i}. Therefore, we have $\alpha_{NT_i} = \alpha_{NT} + b_{NT_i}$, $\alpha_{OS_i} = \alpha_{OS} + b_{OS_i}$. The relationship between Kendall's τ and the correlation parameter η_{NT} is $\tau_{\mathrm{NT}} = \eta_{\mathrm{NT}}/(2 + \eta_{\mathrm{NT}})$. A large value of η_{NT} represents a high correlation. When η_{NT} goes to 0, the correlation approaches 0, and when η_{NT} goes to ∞, the correlation converges to 1. For non-target lesion, we define $y_{\mathrm{NT}} = \min(t_{\mathrm{NT}}, c_{\mathrm{NT}})$ and $\delta_{\mathrm{NT}} = I(t_{\mathrm{NT}} \leq c_{\mathrm{NT}})$, where c_{NT} and $I(\cdot)$ denote the censoring time and the indicator function, respectively; and define y_{OS} and δ_{OS} similarly for OS.

Depending on the censoring pattern, the observed data for the ith participant falls into one of the following mutually exclusive cases: 1) both t_{NT} and t_{OS} are observed (δ_{NT}=1, δ_{OS}=1), 2) t_{NT} is observed and t_{OS} is censored (δ_{NT}=1, δ_{OS}=0), 3) t_{NT} is censored and t_{OS} is observed (δ_{NT}=0, δ_{OS}=1), and 4) both t_{NT} and t_{OS} are censored (δ_{NT}=0, δ_{OS}=0). Based on these four scenarios, we can derive the likelihood of the copula model with the following steps:

Let $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_{\mathbf{NT}}, \alpha_{\mathrm{NT}}, \lambda_{\mathrm{NT}}, \boldsymbol{\beta}_{\mathbf{OS,NT}}, \alpha_{\mathrm{OS,NT}}, \lambda_{\mathrm{OS}}, \eta_{\mathrm{NT}})$, then the likelihood for the ith patient with $\mathrm{Data}_i = (y_{\mathrm{NT}}, y_{\mathrm{OS}}, \delta_{\mathrm{NT}}, \delta_{\mathrm{OS}}, \mathbf{Z})_i$ is given by

$$L(\boldsymbol{\theta}_2|\mathrm{Data}_i) = L_1^{\delta_{\mathrm{NT}}\delta_{\mathrm{OS}}} L_2^{\delta_{\mathrm{NT}}(1-\delta_{\mathrm{OS}})} L_3^{(1-\delta_{\mathrm{NT}})\delta_{\mathrm{OS}}} L_4^{(1-\delta_{\mathrm{NT}})(1-\delta_{\mathrm{OS}})},$$

where

$$\begin{aligned}
L_1 &= \frac{\partial^2 S_1(t_{\mathrm{NT}}, t_{\mathrm{OS}}|\mathbf{Z})}{\partial t_{\mathrm{NT}} \partial t_{\mathrm{OS}}}\Big|_{t_{\mathrm{NT}}=y_{\mathrm{NT}}, t_{\mathrm{OS}}=y_{\mathrm{OS}}} \\
&= \left(\eta_{\mathrm{NT}} + 1\right)\left(\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}\exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}\right)^{-(\eta_{\mathrm{NT}}+1)} \\
&\times \left(\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}^{-\eta_{\mathrm{NT}}} + \exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}^{-\eta_{\mathrm{NT}}} - 1\right)^{-\frac{2\eta_{\mathrm{NT}}+1}{\eta_{\mathrm{NT}}}} \\
&\times \gamma_{\mathrm{NT}} \alpha_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}-1} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}}) \exp\{-\gamma_{\mathrm{NT}} t^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\} \\
&\times \gamma_{\mathrm{OS}} \alpha_{\mathrm{OS,NT}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}-1} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}}) \exp\{-\gamma_{\mathrm{OS}} t^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\},
\end{aligned}$$

and

$$\begin{aligned} L_2 &= -\frac{\partial S_1(t_{\mathrm{NT}}, t_{\mathrm{OS}}|\mathbf{Z})}{\partial t_{\mathrm{NT}}}\bigg|_{t_{\mathrm{NT}}=y_{\mathrm{NT}}, t_{\mathrm{OS}}=y_{\mathrm{OS}}} \\ &= -\left(\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}\right)^{-(\eta_{\mathrm{NT}}+1)} \\ &\times \left(\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}^{-\eta_{\mathrm{NT}}} + \exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}^{-\eta_{\mathrm{NT}}} - 1\right)^{-\frac{\eta_{\mathrm{NT}}+1}{\eta_{\mathrm{NT}}}} \\ &\times \gamma_{\mathrm{NT}}\alpha_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}-1} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\exp\{-\gamma_{\mathrm{NT}} t^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}, \end{aligned}$$

and

$$\begin{aligned} L_3 &= -\frac{\partial S_1(t_{\mathrm{NT}}, t_{\mathrm{OS}}|\mathbf{Z})}{\partial t_{\mathrm{OS}}}\bigg|_{t_{\mathrm{NT}}=y_{\mathrm{NT}}, t_{\mathrm{OS}}=y_{\mathrm{OS}}} \\ &= -\left(\exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}\right)^{-(\eta_{\mathrm{NT}}+1)} \\ &\times \left(\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}^{-\eta_{\mathrm{NT}}} + \exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}^{-\eta_{\mathrm{NT}}} - 1\right)^{-\frac{\eta_{\mathrm{NT}}+1}{\eta_{\mathrm{NT}}}} \\ &\times \gamma_{\mathrm{OS}}\alpha_{\mathrm{OS,NT}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}-1} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\exp\{-\gamma_{\mathrm{OS}} t^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}, \end{aligned}$$

and

$$\begin{aligned} L_4 &= S_1(t_{\mathrm{NT}}, t_{\mathrm{OS}}|\mathbf{Z})\bigg|_{t_{\mathrm{NT}}=y_{\mathrm{NT}}, t_{\mathrm{OS}}=y_{\mathrm{OS}}} \\ &= \{\exp\{-\gamma_{\mathrm{NT}} y_{\mathrm{NT}}^{\alpha_{\mathrm{NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{NT}})\}^{-\eta_{\mathrm{NT}}} + \exp\{-\gamma_{\mathrm{OS}} y_{\mathrm{OS}}^{\alpha_{\mathrm{OS,NT}}} \exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{OS,NT}})\}^{-\eta_{\mathrm{NT}}} - 1\}^{-1/\eta_{\mathrm{NT}}}. \end{aligned}$$

For a given subject, L1, L2, L3 and L4 correspond to the likelihood components that both NT and OS are observed, NT is observed but OS is censored, NT is censored but OS is observed, and both NT and OS are censored, respectively.

4.3.1.3 Joint Model for Time to New Lesion and OS

Similar to Section 4.3.1.2, for the relationship between time to new lesion and OS, we model the bivariate time-to-event data using the Clayton (1978) model, $S_2(t_{\mathrm{NL}}, t_{\mathrm{OS}}|\mathbf{Z}) = \{S_{\mathrm{NL}}(t_{\mathrm{NL}}|\mathbf{Z})^{-\eta_{\mathrm{NL}}} + S_{\mathrm{OS}}(t_{\mathrm{OS}}|\mathbf{Z})^{-\eta_{\mathrm{NL}}} - 1\}^{-1/\eta_{\mathrm{NL}}}$. The derivation of this copula model's likelihood is similar to that specified in section 4.3.1.2.

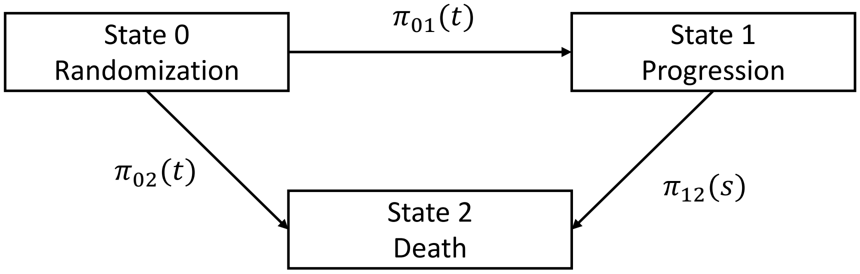


Figure 4.5: The three-state illness-death model

4.3.1.4 Marginal Model for OS

As a standard way of modeling time to event data, we model OS with a simple Weibull baseline hazard model that has the form $\lambda(t) = \alpha_{\text{OS,M}}\gamma_{\text{OS}}t^{\alpha_{\text{OS,M}}-1}\exp(\mathbf{Z}_i'\boldsymbol{\beta_M})$, where $\mathbf{Z}_i$ is the covariate matrix for modeling OS.

4.3.2 Semi-Markov Three-State Illness-Death Model

At the time this paper is written, no literature has covered survival time prediction based on a multi-state model; here, we fill this gap in this section. A multi-state model is another common approach for modeling the event history of participants in clinical trials (Andersen & Keiding 2002, Meira-Machado et al. 2009, Putter et al. 2007). Among the various multi-state models, the one that finds the widest application is the three-state illness-death model, which includes a singular immediate state denoting "illness" (Figure 4.5). In this paper, we employ the homogeneous semi-Markov assumption (Cox & Miller 1977) for the illness-death model, which implies that the hazard of death after progression depends on time since progression rather than time since randomization. Therefore, for every patient, we examine two distinct periods: the time between randomization and progression, and the duration from progression to death. Both of these intervals are treated as separate components and modeled individually. A third scenario occurs when a patient passes away without experiencing progression. In the following paragraphs, we explain the specific mathematical details of the illness-death model.

For the illness-death model, transition probabilities represent the probabilities of transition from one state to another over a given time period. Here, we denote the transition probability as $P_{kl}(t_1, t_2)$ from time t_1 to t_2, where k, l describes the state with $k \in 0, 1, 2$ and $l \in 0, 1, 2$. The expression of the transition probabilities are given by (Meira-Machado et al. 2009):

$$P_{00}(t_1, t_2) = S_0(t_2 - t_1) = exp(-\prod_{01}(t_2 - t_1) - \prod_{02}(t_2 - t_1)),$$

$$P_{11}(t_1,t_2) = S_1(t_2-t_1) = exp(-\prod_{12}(t_2-t_1)),$$

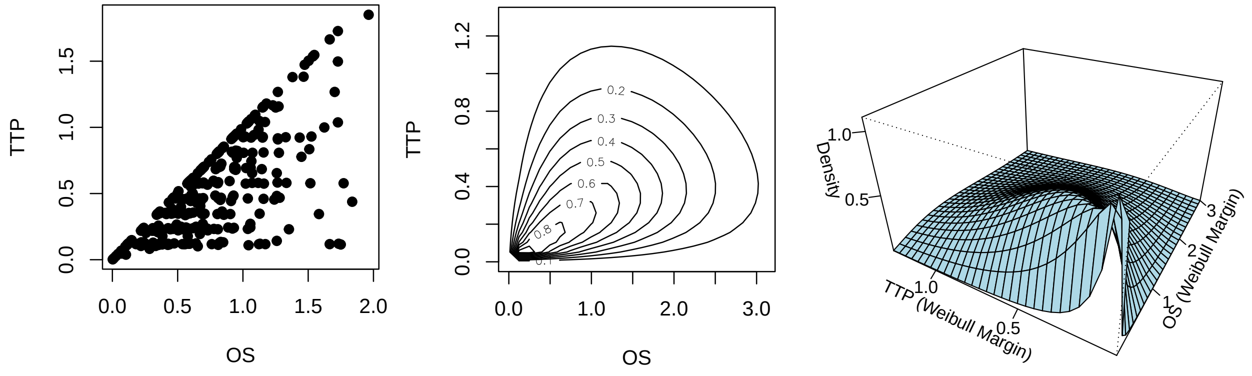
$$P_{12}(t_1,t_2) = S_1(t_2-t_1) = \int_{t_1}^{t_2} P_{11}(t_1,u)\pi_{12}(u;\mathcal{F}_u)P_{22}(u,t_2)du,$$

where $\prod_{kl}(t_1,t_2) = \int_{t_1}^{t_2} \pi_{kl}(t,\mathcal{F}_t)dt$ is the cumulative transition intensity between states k and l, where $k \leq l$. If we consider the transition intensities $\pi_{kl}(t;\mathcal{F}_t)$ to follow Weibull distributions, they can be expressed as $\pi_{01}(t) = \alpha\left(\frac{1}{\gamma_{01}}\right)^{\alpha} t^{\alpha-1}$, $\pi_{02}(t) = \alpha\left(\frac{1}{\gamma_{02}}\right)^{\alpha} t^{\alpha-1}$, and $\pi_{12}(s) = \alpha\left(\frac{1}{\gamma_{12}}\right)^{\alpha} s^{\alpha-1}$, where γ_{01}, γ_{02} and γ_{03} are the scale parameters, α is the shape parameter, t refers to time since randomization and s refers to time since progression. γ_{01}, γ_{02} and γ_{03} for each patient are further defined as $\gamma_{01,i} = \tilde{\gamma}_{01}exp(\boldsymbol{\beta}_{01}\boldsymbol{X}_i)$, $\gamma_{02,i} = \tilde{\gamma}_{02}exp(\boldsymbol{\beta}_{02}\boldsymbol{X}_i)$, and $\gamma_{12,i} = \tilde{\gamma}_{12}exp(\boldsymbol{\beta}_{12}\boldsymbol{X}_i)$, respectively, where $i \in N$, $\boldsymbol{X}_i$ is the vector of covariates, and $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{02}$ and $\boldsymbol{\beta}_{12}$ are the regression coefficients. $\tilde{\gamma}_{01}$, $\tilde{\gamma}_{02}$, and $\tilde{\gamma}_{12}$ are the baseline hazards. To guarantee the convergence of MCMC , a uniform shape parameter α is employed for all three Weibull functions. The multi-state model's likelihood construction is as follows:

Here, we aim to estimate the parameter vector $\theta = (\alpha, \gamma_{01}, \gamma_{02}, \gamma_{12})$ using maximum likelihood estimation. Firstly, we need to model the survival experiences of individuals, which can be characterized by four distinct cases based on a patient's progression through the illness-death model: 1) Patients progress and are then censored, 2) Patients progress and subsequently die, 3) Patients die without prior progression, 4) Patients are censored without experiencing either progression or death. If for every individual i in the set $(1, ..., N)$, we let t_{i1} indicate the duration from the initial state to either progression or death and let t_{i2}, which is conditioned on progression, represents the time from progression to death, then the individual likelihood for these 4 experiences, denoted as $L_i^{(k)}(\theta)$ where k varies from 1 to 4, can be described as: $L_i^{(1)}(\theta) = f_1(t_{i1})S_2(t_{i1})S_3(t_{i2})$, $L_i^{(2)}(\theta) = f_1(t_{i1})S_2(t_{i1})f_3(t_{i2})$, $L_i^{(3)}(\theta) = S_1(t_{i1})f_2(t_{i1})$, and $L_i^{(4)}(\theta) = S_1(t_{i1})S_2(t_{i1})$. Here, $f_1(\cdot)$ and $S_1(\cdot)$ are density and survival function for time to progression, $f_2(\cdot)$ and $S_2(\cdot)$ are density and survival function for time to death without progression, and $f_3(\cdot)$ and $S_3(\cdot)$ are density and survival function for time to death post progression. Therefore, comprehensive log-likelihood for all subjects can be formulated as:

$$\begin{aligned}\log(L(\theta)) = \sum_{i=1}^{N}[&(d_1)(1-d_2)(1-d_3)L_i^{(1)}(\theta) + (d_1)(1-d_2)(d_3)L_i^{(2)}(\theta)\\ &+ (1-d_1)(d_2)L_i^{(3)}(\theta) + (1-d_1)(1-d_2)L_i^{(4)}(\theta)],\end{aligned}$$

where d_1 is the indicator for progression, d_2 is the indicator for death without preceding progression, and d_3 is the indicator for death post progression. Note: here, an indicator value of 0 represents censoring, whereas a value of 1 indicates the event's occurrence.



(a) (b) (c)

Figure 4.6: (a) The scatterplot between TTP and OS, (b) the contour plot of TTP vs. OS fitted with a Clayton copula, (c) the surface plot of TTP vs. OS fitted with a Clayton copula.

4.3.3 Copula Model for Time-to-Progression (TTP) and OS

TTP models disease progression similarly to PFS but specifically excludes deaths from any cause. If a patient dies without documented disease progression, their TTP is censored at the last follow-up. Similar to the model defined in 4.3.1.2, we use $\lambda_{\text{TTP}}(t|\mathbf{Z})$ to denote the hazard function for TTP. Under the Cox proportional hazards model, we have $\lambda_{\text{TTP}}(t|\mathbf{Z}) = \lambda_{0,\text{TTP}}(t)\exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{TTP}})$, and the corresponding survival function is given by $S_{\text{TTP}}(t|\mathbf{Z}) = \exp\{-\gamma_{\text{TTP}}t^{\alpha_{\text{TTP}}}\exp(\mathbf{Z}'\boldsymbol{\beta}_{\mathbf{TTP}})\}$, such that γ_{TTP} and α_{TTP} are the scale and shape parameters of the Weibull distribution, respectively. The Clayton model between time-to-progression (TTP) and OS is specified as

$$S_1(t_{\text{TTP}}, t_{\text{OS}}|\mathbf{Z}) = \{S_{\text{TTP}}(t_{\text{TTP}}|\mathbf{Z})^{-\eta_{\text{TTP}}} + S_{\text{OS}}(t_{\text{OS}}|\mathbf{Z})^{-\eta_{\text{TTP}}} - 1\}^{-1/\eta_{\text{TTP}}},$$

where $\eta_{\text{TTP}} > 0$ measures the correlation. For TTP, define $y_{\text{TTP}} = \min(t_{\text{TTP}}, c_{\text{TTP}})$ and $\delta_{\text{TTP}} = I(t_{\text{TTP}} \leq c_{\text{TTP}})$, where c_{TTP} and $I(\cdot)$ denote the censoring time and the indicator function, respectively; and define y_{OS} and δ_{OS} similarly for OS. Depending on the censoring pattern, the observed data for the ith participant falls into one of the 4 mutually exclusive cases: (δ_{TTP}=1, δ_{OS}=1), (δ_{TTP}=1, δ_{OS}=0), (δ_{TTP}=0, δ_{OS}=1), and (δ_{TTP}=0, δ_{OS}=0). Figure 4.6 illustrates the Clayton copula fitted between TTP and OS.

4.3.4 Standard Statistical Analysis

Most commonly, PFS is modeled using a Kaplan-Meier curve or Cox regression. For example, just like what we discussed previously in Section 4.3.1.4, a simple Weibull baseline hazard model has the form $\lambda(t) = \alpha\gamma t^{\alpha-1}\exp(\mathbf{Z}'\boldsymbol{\beta})$, where $\mathbf{Z}_i$ is the covariate matrix for modeling PFS, α and γ are the shape and scale parameters of the Weibull distribution, and $\boldsymbol{\beta}$ is the vector of coefficients. Note that a standard analytic method does not consider the dynamics of the component processes of PFS and does not take random effects into account.

4.3.5 Prediction

4.3.5.1 Posterior Predictive Distribution (PPD)

Within each of the models outlined in Sections 4.3.4 through 4.3.1, we utilize the PPD technique (Gelman et al. 2014) to generate OS predictions. The PPD represents the distribution of potential unobserved values based on the observed values, and it follows this structure: $p(y_{pred}|y) = \int p(y_{pred}, \theta|y)d\theta = \int p(y_{pred}|\theta, y)p(\theta|y)d\theta = \int p(y_{pred}|\theta)p(\theta|y)d\theta$, This formulation leverages the conditional independence between y (observed data), y_{pred} (unobserved data), and θ (parameters). The PPD can be understood as an average of conditional predictions over the posterior distribution of θ. During the MCMC process, for each sampled θ from the posterior distribution, a corresponding y_{pred} sample is obtained. To produce PPD samples for OS using the survival time modeling methods discussed, we used the `JAGS` software (Plummer 2017) along with the `rjags` package for implementation in the R environment.

4.3.5.2 Bayesian Model Averaging (BMA)

In Section 4.3.1, we have established the three models governing the relationships between the target lesion, non-target lesion, new lesion, and OS, along with the marginal model for OS in our joint model, a BMA approach can be adopted to link all four models under the multivariate joint modeling approach together, enhancing the overall prediction capability.

In prognostic modeling, it is standard to generate predictions by selecting a single model based on certain metrics. In the context of oncology, different PFS components may be more correlated with OS under various tumor types or patient groups, thus offering more accurate OS predictions. Here, we use BMA to calculate the weighted average of the predicted OS under the three joint models and one marginal model for OS defined above. The weights indicate the plausibility of each model being the most predictive model. We denote the models defined in Section 4.3.1.1, 4.3.1.2, 4.3.1.3, and 4.3.1.4 by M_T, M_{NT}, M_{NL}, and M_{OS}, respectively. Then by BMA, the final predicted probability of OS follows

$$\begin{aligned}P(Y_{pred}|Data) &= P(Y_{pred.t}|M_T, Data)P(M_T|Data)\\ &\quad + P(Y_{pred.nt}|M_{NT}, Data)P(M_{NT}|Data)\\ &\quad + P(Y_{pred.nl}|M_{NL}, Data)P(M_{NL}|Data)\\ &\quad + P(Y_{pred.os}|M_{OS}, Data)P(M_{OS}|Data),\end{aligned}$$

where $P(M_T|Data)$, $P(M_{NT}|Data)$, $P(M_{NL}|Data)$, and $P(M_{OS}|Data)$ are the model weights. In general, if we have a total of Q models, the model weight for model q at the tth MCMC iteration is

$$w_q^{(t)} = P(M_q|Data, \theta^{(t)}) = \frac{P(M_q, Data, \theta^{(t)})}{P(Data, \theta^{(t)})} = \frac{P(Data|M_q, \theta^{(t)})P(\theta^{(t)}|M_q)P(M_q)}{P(Data, \theta^{(t)})},$$

where $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_Q^{(t)})$. If we set $P(\theta_h|M_h, Data) = g_h$, and $q \neq h$, then according to Congdon (2007),

$$P(\theta^{(t)}|M_q) = P(\theta_q^{(t)}|M_q) \prod_{h=1, h\neq q}^{Q} P(\theta_h^{(t)}|M_q) = P(\theta_q^{(t)}|M_q) \prod_{h=1, h\neq q}^{Q} g_h.$$

Therefore,

$$w_q^{(t)} = P(M_q|Data, \theta^{(t)}) = \frac{P(Data|M_q, \theta^{(t)})P(\theta_q^{(t)}|M_q) \prod_{h=1, h\neq q}^{Q} g_h P(M_q)}{\sum_{k=1}^{Q} P(Data|M_k, \theta^{(t)})P(\theta_k^{(t)}|M_k) \prod_{h=1, h\neq k}^{K} g_h P(M_k)}.$$

We implement the BMA approach using the above formula and calculate the final predicted OS.

4.3.6 Prior Specification

In this section, we provide recommendations for specifying priors for the parameters in the models discussed in Section 4.3. Our aim is to promote the use of generalizable, weakly informative, or non-informative priors for all model parameters to ensure robust and interpretable results. We also emphasize the importance of selecting appropriate priors to maintain parameter identifiability.

4.3.6.1 General Guidelines

For the $\boldsymbol{\beta}$ vector of coefficients, we recommend using weakly informative normal distributions with a mean of 0. We suggest weakly informative exponential distributions for the shape parameters (α) in the Weibull distribution. All random effect parameters (b) should follow weakly informative normal distributions, denoted as $N(0, \sigma^2_{b_k})$, where k stands for different random variables under different models. Weakly informative half-normal priors are recommended for the standard deviations (σ_{b_k}) associated with the random effects. For all Clayton copula parameters (η), we recommend using weakly informative exponential distributions. These priors can accommodate a wide range of plausible values. In the following subsections, we provide recommendations for parameters specific to each model:

4.3.6.2 Three-State Illness-Death Model

In the Three-State Illness-Death Model, we advise using weakly informative half-normal priors for all $\tilde{\gamma}$ parameters. These priors allow flexibility while constraining extreme values.

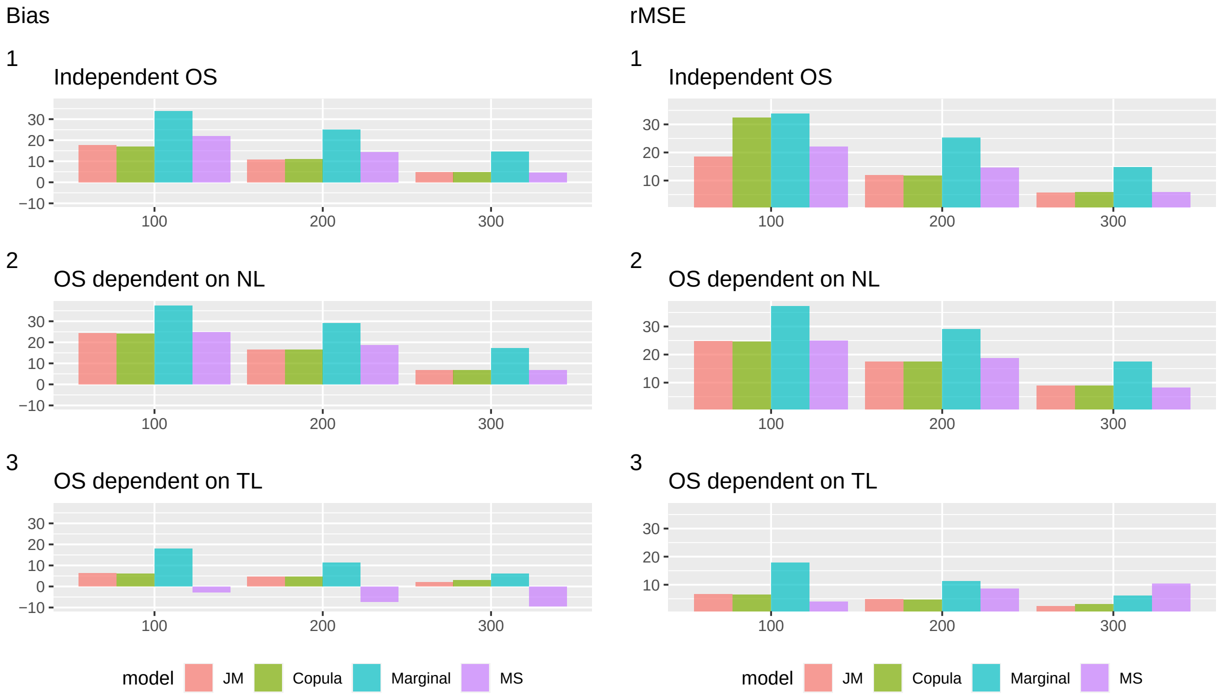
4.3.6.3 Multivariate Joint Modeling Approach

In the model assessing the association between target lesion and OS, which is a component of our proposed multivariate joint modeling approach, we suggest the following priors: For the parameter λ, a weakly informative normal distribution with a mean of 0 is recommended. The correlation coefficient (ρ) should have a non-informative prior, such as $Unif(-1, 1)$, to avoid biasing the estimation. Additionally, we assume that all 4 submodels are equally weighted before fitting BMA, that is, $P(M_q) = 1/Q$, for all $q = 1, ..., Q$.

4.4 Simulation

4.4.1 Design

We performed extensive simulations to evaluate the performance of the proposed multivariate joint modeling approach. All simulated datasets mimic the structure of the renal cell cancer trial data. Specifically, we assume 400 subjects and 25 visits. Longitudinal measurements of target lesion and new lesion status are recorded at each visit. For simplicity purposes, non-target lesion status is not included. Survival status is updated whenever deaths occur. The follow-up visits are scheduled for every two months within the first two years, then change to every three months until the end of the 5th year. We censor the subject's subsequent target lesion measurements and new lesion status whenever that subject's death occurs. In addition, we have assumed three scenarios: 1) OS is independent of target lesion measurements and time to new lesion, 2) OS is correlated with time to new lesion only, and 3) OS is correlated with target lesion measurements only. In all scenarios, target lesion measurements and new lesion status are randomly and independently simulated, but OS is generated under different conditions. We assume a negative correlation between OS and



(a)

(b)

Figure 4.7: The bias (a) and rMSE (b) of the last death date predictors of four models: the multivariate joint modeling approach (JM), a copula model between TTP and OS (Copula), the marginal Weibull baseline hazard model of OS (Marginal), and a three-state illness-death model (MS), under three OS scenarios. The x-axis denotes the number of death at cutoff. The y-axis denotes the number of months.

changes in tumor measurements. For each scenario, 100 datasets are generated. To simplify the simulation process, only $treatment$ is included in the covariate matrices.

We compare predictions under four models: our proposed multivariate joint model (Section 4.3.1), a copula model for TTP and OS (Section 4.3.3), a semi-markov three-state illness-death model (Weber 2020) (Section 4.3.2), and a marginal Weibull baseline hazard model of OS (Section 4.3.4).

To emulate real trial scenarios and assess the prediction performance of our proposed method, we generate snapshots of the data such that for each simulated dataset, we have snapshots of what the data would have looked like if only 100, 200, and 300 death events had occurred. Then, we predict the time of the last (400th) death event under each snapshot dataset. Note that predictions may be performed at any time during the trial, and 100, 200, and 300 were selected for illustration purposes. Predictions from the four methods are compared. Finally, we calculate each predictor's CR, rMSE, bias, and width of the 90% CI. The 90% CI is the narrowest interval that includes 90%

Table 4.1: The coverage rate (CR) and credible interval (CI) width for predictors of the time of the last death. Comparing the results from the multivariate joint modeling approach (JM), a copula model between TTP and OS (Cop), a three-state illness-death model (MS), and the marginal Weibull baseline hazard model of OS (Mrgl). The unit for CI width is month.

	Independent OS				OS dependent on NL				OS dependent on TL			
	JM	Cop	MS	Mrgl	JM	Cop	MS	Mrgl	JM	Cop	MS	Mrgl
100 death event												
CR	0.49	0.49	0.01	0.00	0.11	0.11	0.03	0.00	0.91	0.91	1.00	0.00
Width	35.9	35.8	24.9	3.4	29.2	29.5	26.8	2.9	24.4	25.4	41.8	3.2
200 death event												
CR	0.89	0.86	0.33	0.00	0.45	0.45	0.24	0.00	0.85	0.98	1.00	0.00
Width	35.7	35.3	26.0	5.9	32.2	32.2	26.8	5.6	15.2	17.1	39.6	4.1
300 death event												
CR	0.98	0.98	0.98	0.00	0.90	0.89	0.95	0.01	0.97	0.98	1.00	0.00
Width	30.5	30.5	29.6	9.3	32.5	32.6	31.5	9.9	12.0	11.2	36.7	4.6

of the posterior distribution of the predictor.

4.4.2 Results

The bias and rMSE of the predicted time of the last death are shown in Figures 4.7a and 4.7b, and the corresponding CR and CI width are shown in Table 4.1. The weights of all submodels of the multivariate joint modeling approach under each simulation scenario are shown in Figure 4.8.

Under scenario 1, where OS is generated independently, the bias and rMSE of all predictors become smaller as we gather more information from the trial, i.e., more death events are observed. Overall, the marginal model performs worse than the other three models. This is likely due to the conservativeness of the Weibull model and the fact that it only utilizes OS data, causing predictions to concentrate near the observed maximum OS. Given that the proposed model assigned all the weight to the copula model between new lesion and OS, it is reasonable that the copula model and the multivariate joint modeling approach have similar bias and rMSE. Having evaluated all the metrics, the multivariate joint modeling approach and the copula model are the best-performing method under scenario 1.

In scenario 2, where OS is generated based on new lesion status, the multivariate joint modeling approach, the copula model, and the three-state illness-death model exhibit similar bias and rMSE values. Once again, the marginal model lags behind the other models in performance. As with scenario 1, as the number of observed death events increases, the bias and rMSE of all predictors decrease. Compared to scenario 1, the improved performance in bias and rMSE of the three-state

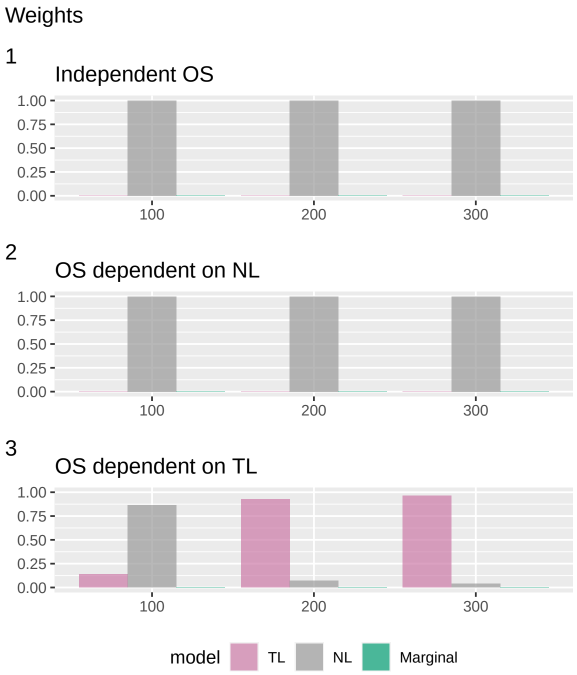


Figure 4.8: The weights of all submodels of the multivariate joint modeling approach for each snapshot dataset in the simulation studies.

illness-death model can largely be attributed to the fact that the OS observations generated under scenario 2 more closely resemble a Weibull distribution. However, when only 100 or 200 death events are observed, the three-state illness-death model has a lower CR than the other two. This might be due to the conservatives of the standard Weibull model. Consequently, for scenario 2, we would still recommend either the joint modeling approach or the copula model.

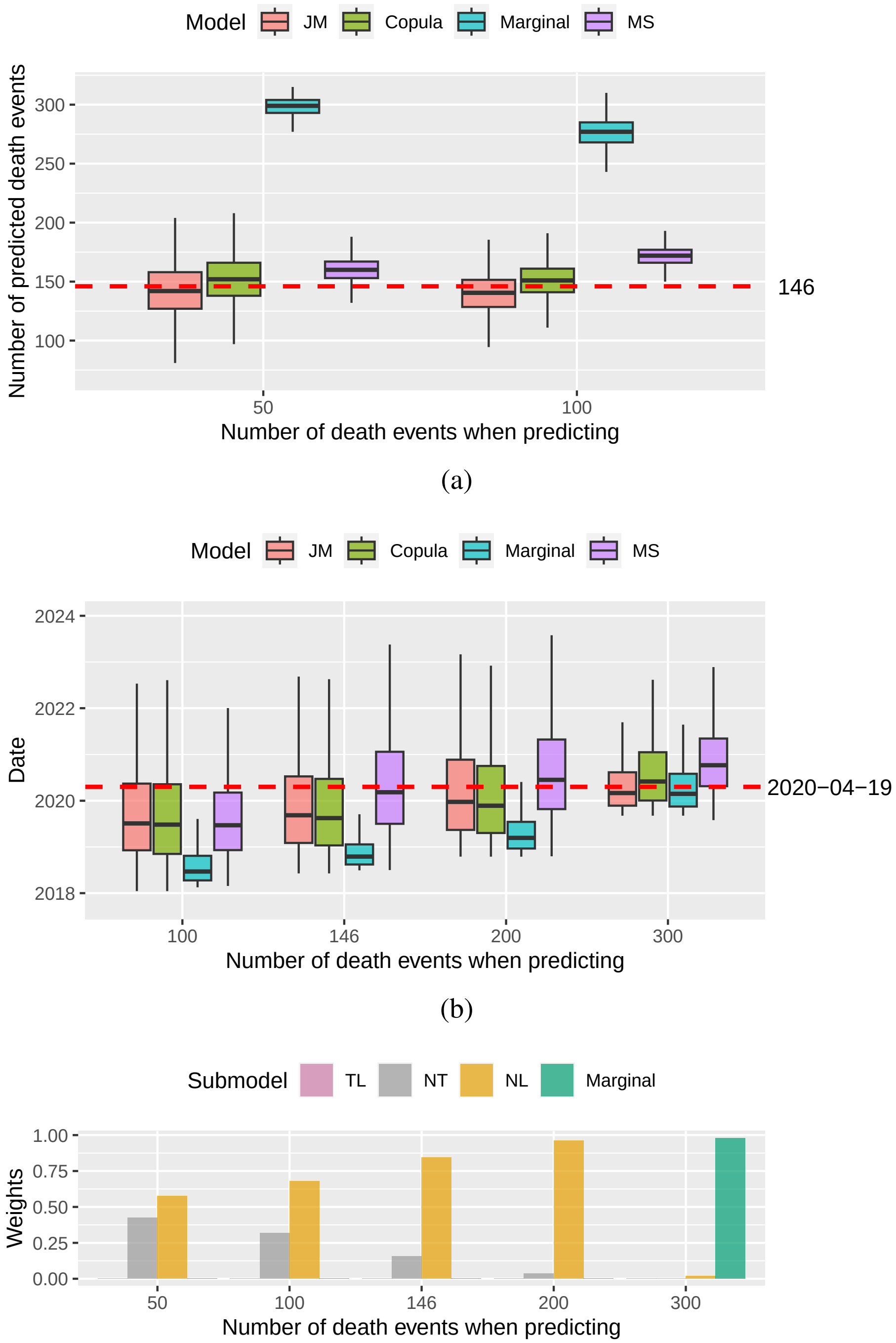
In scenario 3, where OS is derived based on target lesion measurements, the proposed multivariate joint modeling approach begins to surpass the copula model as more death events are observed. Simultaneously, as the count of observed death events increases, the multivariate joint modeling approach assigns greater weight to the joint model between the target lesion and OS, which is anticipated. This is expected as the multivariate joint modeling approach is the only method that captures the relationship between OS and the target lesion. Differing from the other two scenarios, the bias and rMSE of predictors from the three-state illness-death model do not consistently diminish with the increase in observed death events. Moreover, the three-state model displays the widest CI among all the models. This can be attributed to the alteration in OS distributions, as it is generated based on the target lesion distribution, which does not conform to a Weibull distribution. The inflexibility of the standard Weibull baseline hazard model, devoid of random effects, further contributes to this.

According to Figure 4.7, across all scenarios, the proposed multivariate joint modeling approach performs either the best or very similar to the copula model, if the copula model is the best-performing one. This is reasonable as the BMA could assign $w_q = 1$ to submodel q when submodel q is significantly better than the other $q - 1$ submodels. If this is the case, the multivariate joint modeling approach predictions will be the same as the predictions from its submodel q. The multivariate joint modeling approach and the copula model are less sensitive to changes in the OS distribution and still provide relatively reasonable predictions even when OS prediction does not closely resemble a Weibull distribution. In summary, the multivariate joint modeling approach provides the most reliable predictions across all tested scenarios.

4.5 Analysis of the Renal Cell Carcinoma Data

We return to the renal cell carcinoma dataset introduced in Section 4.2. Our objective here is to predict the timing of the last death in this clinical trial by fitting the multivariate joint model we developed in Section 4.3.1 to all available tumor measurements. We generate snapshots of the trial data at the 100th, 146th (the time of primary analysis), 200th, and 300th death, and compare predictions under four models: the multivariate joint model, a copula model for TTP and OS, a three-state illness-death model, and a marginal model of OS.

We included *gender, age, nephrectomy at baseline, Heng prognostic criteria at baseline, and ECOG Performance Status* into the covariates matrix $\mathbf{X}_i$ and $\mathbf{Z}_i$. Then, we fit each model using two MCMC chains, each with at least 10,000 MCMC iterations, 1,000 burn-in, and 8,000 adaptation iterations. Convergence diagnostic tests and autocorrelation plots did not show indications of convergence failure. Figure 4.9a displays the predicted number of deaths at the time of the primary analysis. Figure 4.9b presents boxplots of the posterior distributions for the predictors of the last death date from all tested models compared to the true last death date. The median values of the boxplots serve as point predictions. Figure 4.9c displays the weights of all submodels for each recreated dataset. In this specific case study, the new lesion is observed to have the strongest correlation with OS. Therefore, it makes sense that greater weights are assigned to the new lesion copula submodel, and the copula model between TTP and OS also performs well. While the marginal model exhibits significantly smaller CIs, its predictors prove to be less accurate than those of the multivariate joint modeling approach and the copula model, especially when we observe 100, 146, and 200 deaths. These latter models consistently demonstrate the most reliable performance improvements as the number of observed deaths increases. When we have 300 deaths observed and the marginal model performs the best among all submodels, the BMA procedure naturally assigns more weight to it, as indicated in Figures 4.9b and 4.9c. Considering that the posterior samples of the three-state illness-death model are derived by combining predictions from two distinct time periods - randomization to progression and progression to death — it is reasonable for the three-state



(a)

(b)

(c)

Figure 4.9: (a) Boxplots of the predicted number of deaths at primary analysis when 50 or 100 deaths are observed in the trial, with the true number of deaths being “146”.(b) Boxplots of the posterior distributions of the last death date by the proposed multivariate joint modeling approach, the copula model between TTP and OS, the marginal Weibull baseline hazard model of OS, and the three-state illness-death model. Date “2020-04-19” is the true last death date. (c) The final weights of each submodel under the proposed multivariate joint modeling approach.

Table 4.2: Bias and rMSE of the predictors of multivariate joint modeling approach (JM), copula model, marginal model, and three-state illness-death model (MS).

No. of events		Metric	JM	Copula	Marginal	MS
Observed	Predicted					
100	241	Bias	0.67	1.71	9.46	1.35
		rMSE	3.71	3.94	10.56	3.73
146	195	Bias	1.41	2.11	9.13	2.54
		rMSE	2.56	3.13	9.65	3.14
200	141	Bias	0.91	2.15	7.89	3.11
		rMSE	2.06	2.91	8.19	3.65
300	41	Bias	6.23	2.97	6.32	4.13
		rMSE	8.24	6.14	8.28	7.31

model to exhibit the widest CIs.

Table 4.2 presents the bias and rMSE of the predicted OS when 100, 146, 200, or 300 death events are observed. In the first three scenarios, the multivariate joint modeling approach performs similarly to the copula model and the three-state model. In the last scenario, the joint model performs similarly to the marginal model, consistent with the submodel weights shown in Figure 4.9c. Additionally, the joint model exhibits smaller bias and rMSEs compared to the marginal model under all scenarios. While the multivariate joint modeling approach may not always have the smallest bias in this particular case study dataset, its prediction reliability is consistently verified.

We used all four models to forecast the OS of patients with censored OS in the case study data set. Subsequently, we generated predicted Kaplan-Meier plots for each model (Figure 4.10). Finally, we estimated the potential gain or loss in life expectancy by calculating the difference in the area under the curve between the experimental drug and the standard of care (Pak et al. 2017, Uno et al. 2022). The results from Figure 4.10 indicate that all models consistently project a substantial improvement in life years. In addition, the multivariate joint modeling approach and the copula model, whose reliability was previously established in this case study, both forecast a roughly 7.5-month gain in life years for patients who receive the experimental drug, as opposed to the standard of care.

4.6 Discussion

In this paper, we were motivated by an advanced renal cell carcinoma clinical trial to present how real-time OS predictions from joint models with different components of PFS can be dynamically combined via BMA. This multivariate joint modeling approach provides reliable estimates of the time of the *n*th death in a trial using all available tumor assessment data based on RECIST 1.1. This is valuable for clinical trial planning and patients' end-of-life medical care. Furthermore, the proposed multivariate joint model method provides the most reliable and robust predictions in the case study and across all the scenarios we tested in the simulation studies. This approach can be easily applied to other solid tumor types beyond renal cell carcinoma.

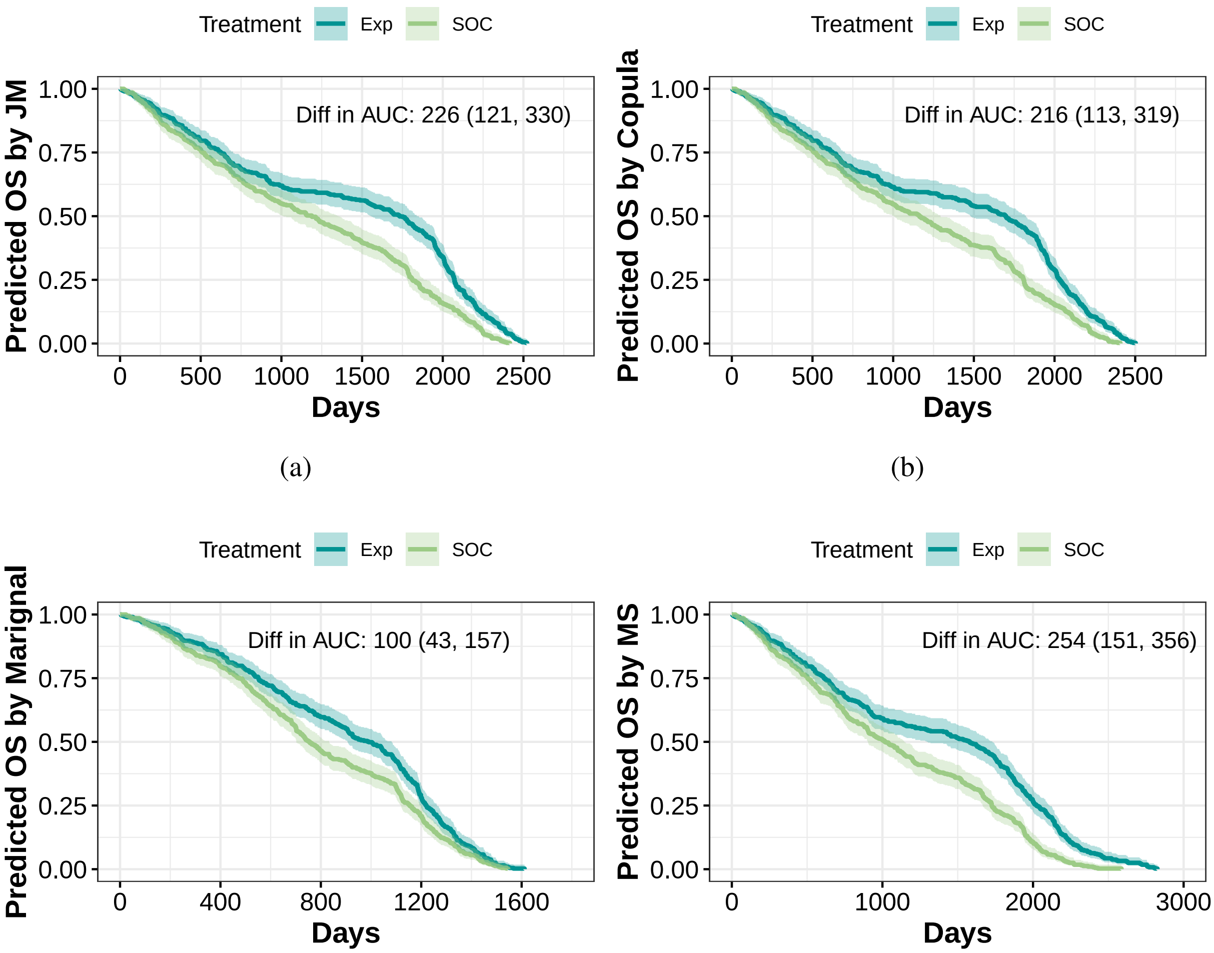
The multivariate joint modeling approach is flexible and can easily incorporate a linear model with time-dependent covariates or a nonlinear mixed-effects model. Although BMA methods optimally combine multiple models, they typically require extended training time due to their complexity. In the example data analysis, only five covariates were included. Incorporating additional baseline characteristics is possible, but doing so would expand the covariate matrix and further increase the computational demands of the multivariate joint modeling approach. Considering that the covariates utilized in each submodel may vary, it is beneficial to determine the most suitable covariates for each submodel before implementation. This covariates selection process can be guided by expert opinion or variable selection models, potentially reducing the covariate matrix's size.

Based on the simulation studies and the example data analysis, it is evident that the copula model for TTP and OS ranks as the second most reliable model among all those tested. If the TTP and OS data are well-fitted by Weibull distributions with random effects, the copula model typically performs better than the marginal OS model. Therefore, fitting only a copula model with random effects is an excellent option to reduce model training time. Based on our simulations, without including random effects, high variability between individuals will cause the copula model's predictions to be too heavy-tailed.

For a marginal OS model to perform better, it would be beneficial to incorporate more predictive covariates or use a model offering greater flexibility than the Weibull baseline hazard model, such as piecewise exponential models. Similarly, to improve the performance of the multi-state model, we could replace the Weibull baseline hazard model with more flexible models.

During the simulation studies, the data for non-target lesions was not simulated, yet it was included in the example data analysis. This does not undermine the outcome of our simulation studies, given that both the non-target lesion and new lesion were modeled using the exact same joint model, and each component of PFS was modeled separately. If the non-target lesion data had been included in the simulation dataset, we would anticipate longer training times for the multivariate joint modeling approach without observing any relative improvement in performance given the scenarios we defined.

Apart from applying the same model weights for all the patients in the dataset, BMA can potentially provide personalized model weights, where each patient may have different weights for



(a) (b)

(c) (d)

Figure 4.10: Kaplan-Meier plots of the observed and predicted overall survival by the multivariate joint modeling approach (a), the copula model (b), the marginal model (c), and the three-state illness-death model (d). The difference in the area under the curve (AUC) and its credible intervals for the experimental drug (Exp) and the standard of care (SOC) are also included.

each model. This could further improve the prediction accuracy. Future work can explore how to implement personalized model weights in a time-efficient manner under the setting of this study.

CHAPTER 5

Summary and Future Work

Motivated by two phase 2/3 DMD trials, SPITFIRE (NCT03039686) and tadalafil (NCT01865084), this dissertation proposes innovative trial designs and Bayesian methods for rare diseases, such as DMD, to formally incorporate external data. Additionally, inspired by a renal cell carcinoma trial, it introduces Bayesian methods for landmark survival time prediction in oncology trials. Thus, this work aims to provide promising alternatives to the current drug development paradigm in rare diseases and a robust method for predicting patient survival times in ongoing trials.

Chapter 2 presents an snSMART design comparing two dose levels with a continuous outcome, incorporating external controls formally to reduce the number of subjects needed in the placebo arm. The proposed robust MAC-snSMART method yields the most accurate and robust estimators among all tested methods when the assumption of stage-wise treatment effect exchangeability is met. This model leverages summary statistics from both external controls and current trials, employing "change from baseline" as the outcome measure. This design and method efficiently utilize "all" available data sources, namely external controls and all stages of the current trial. It supports subject recruitment and retention and aligns with the goals of the FDA's Complex Innovative Design program.

Building on the trial design proposed in Chapter 2, we significantly expand the application area of snSMART by introducing a Bayesian longitudinal piecewise meta-analytic combined model. This model not only dynamically incorporates external control data but also utilizes subject-level longitudinal outcomes and baseline characteristics to further enhance the power and efficiency of trial analysis. Moreover, the proposed BLPM methods address both between-trial non-exchangeability and stage-wise treatment effect non-exchangeability. This significantly improves the model's ability to tackle real-life challenges of trial analysis. We compared the BLPM methods with the BJSM proposed by Fang et al. (2023) and traditional MMRM methods through simulation studies and example data analysis. Among all tested scenarios, the BLPM provides the most efficient estimators.

Chapter 4 proposes a multivariate joint modeling approach that provides real-time OS predictions based on joint models between different components of PFS and OS. This approach has proven to provide reliable and robust predictions in our case study and has been validated across various scenarios in simulation studies. Such predictive power is invaluable for planning oncology trials and managing end-of-life care for patients. It can also be readily adapted to other types of solid tumors beyond renal cell carcinoma. Furthermore, this chapter enhances the copula model and the multi-state model to predict survival times for patients currently undergoing treatment, an application not previously explored before this work was published. Our proposed innovative modeling approach holds significant promise for benefiting the drug development process in the realm of solid tumors.

Regarding Chapters 2 and 3, future work may consider adapting the proposed models to handle various types of outcomes, such as binary or categorical outcomes. Additionally, the proposed snSMART design could be expanded to include interim analyses that allow for early termination of the study. Given the enrichment of the placebo arm with external controls, developing a method to accurately calculate the ESS of external controls is crucial for assisting in the determination of the sample size needed for the trial.

Regarding Chapter 4, future work may consider incorporating individual-level weights into the model to provide personalized predictions for each subject. The current model, implemented in *JAGS* and run under the *rjags* package, converges quite slowly. Future efforts could explore strategies to accelerate model convergence, thereby making its use more time efficient.

In conclusion, this dissertation introduces innovative Bayesian methods for rare disease clinical trial design and survival time prediction in solid tumor clinical trials. We hope that our proposed trial design will lead to more approved drugs for patients with rare diseases, inspire future innovative trial designs, assist in planning oncology clinical trials, and ultimately improve patient well-being.

BIBLIOGRAPHY

Andersen, P. K. & Keiding, N. (2002), 'Multi-state models for event history analysis', *Statistical methods in medical research* **11**(2), 91–115.

Bruno, R., Mercier, F. & Claret, L. (2014), 'Evaluation of tumor size response metrics to predict survival in oncology clinical trials', *Clinical Pharmacology & Therapeutics* **95**(4), 386–393.

Chao, Y.-C., Braun, T. M., Tamura, R. N. & Kidwell, K. M. (2020), 'A Bayesian group sequential small n sequential multiple-assignment randomized trial', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69**(3), 663–680.

Chao, Y.-C., Trachtman, H., Gipson, D. S., Spino, C., Braun, T. M. & Kidwell, K. M. (2020), 'Dynamic treatment regimens in small n, sequential, multiple assignment, randomized trials: an application in focal segmental glomerulosclerosis', *Contemporary Clinical Trials* **92**, 105989.

Claret, L., Girard, P., Hoff, P. M., Van Cutsem, E., Zuideveld, K. P., Jorga, K., Fagerberg, J. & Bruno, R. (2009), 'Model-based prediction of phase iii overall survival in colorectal cancer on the basis of phase ii tumor dynamics', *Journal of Clinical Oncology* **27**(25), 4103–4108.

Claret, L., Gupta, M., Han, K., Joshi, A., Sarapa, N., He, J., Powell, B. & Bruno, R. (2013), 'Evaluation of tumor-size response metrics to predict overall survival in western and chinese patients with first-line metastatic colorectal cancer', *Journal of Clinical Oncology* **31**(17), 2110–2114.

Clayton, D. G. (1978), 'A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence', *Biometrika* **65**(1), 141–151.

Congdon, P. (2007), 'Model weights for model choice and averaging', *Statistical Methodology* **4**(2), 143–157.

Cox, D. R. & Miller, H. D. (1977), *The theory of stochastic processes*, Vol. 134, CRC press.

Crisafulli, S., Sultana, J., Fontana, A., Salvo, F., Messina, S. & Trifirò, G. (2020), 'Global epidemiology of Duchenne muscular dystrophy: An updated systematic review and meta-analysis', *Orphanet Journal of Rare Diseases* **15**(1), 1–20.

Driscoll, J. J. & Rixe, O. (2009), 'Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials', *The Cancer Journal* **15**(5), 401–405.

Duan, Y. (2005), A modified Bayesian power prior approach with applications in water quality evaluation, PhD thesis, Virginia Polytechnic Institute and State University.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M. et al. (2009), 'New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)', *European journal of cancer* **45**(2), 228–247.

Fang, F., Hochstedler, K. A., Tamura, R. N., Braun, T. M. & Kidwell, K. M. (2021), 'Bayesian methods to compare dose levels with placebo in a small n, sequential, multiple assignment, randomized trial', *Statistics in Medicine* **40**(4), 963–977.

Fang, F., Tamura, R., Braun, T. M. & Kidwell, K. M. (2023), 'Comparing dose levels to placebo using a continuous outcome in a small n, sequential, multiple assignment, randomized trial (snsmart)', *Statistics in Biopharmaceutical Research* **15**(3), 502–509.

Fleischer, F., Gaschler-Markefski, B. & Bluhmki, E. (2009), 'A statistical model for the dependence between progression-free survival and overall survival', *Statistics in Medicine* **28**(21), 2669–2686.

Fouarge, E., Monseur, A., Boulanger, B., Annoussamy, M., Seferian, A. M., De Lucia, S., Lilien, C., Thielemans, L., Paradis, K., Cowling, B. S. et al. (2021), 'Hierarchical bayesian modelling of disease progression to inform clinical trial design in centronuclear myopathy', *Orphanet Journal of Rare Diseases* **16**(1), 1–11.

Fu, H., Wang, Y., Liu, J., Kulkarni, P. M. & Melemed, A. S. (2013), 'Joint modeling of progression-free survival and overall survival by a bayesian normal induced copula estimation model', *Statistics in medicine* **32**(2), 240–254.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)'.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2014), 'Bayesian data analysis (vol. 2)'.

Grigore, B., Ciani, O., Dams, F., Federici, C., de Groot, S., Möllenkamp, M., Rabbe, S., Shatrov, K., Zemplenyi, A. & Taylor, R. S. (2020), 'Surrogate endpoints in health technology assessment: an international review of methodological guidelines', *Pharmacoeconomics* **38**(10), 1055–1070.

Hartman, H., Tamura, R. N., Schipper, M. J. & Kidwell, K. M. (2021), 'Design and analysis considerations for utilizing a mapping function in a small sample, sequential, multiple assignment, randomized trials with continuous outcomes', *Statistics in Medicine* **40**(2), 312–326.

Henderson, R., Jones, M. & Stare, J. (2001), 'Accuracy of point predictions in survival analysis', *Statistics in medicine* **20**(20), 3083–3096.

Heng, D. Y., Xie, W., Bjarnason, G. A., Vaishampayan, U., Tan, M.-H., Knox, J., Donskov, F., Wood, L., Kollmannsberger, C., Rini, B. I. et al. (2011), 'Progression-free survival as a predictor of overall survival in metastatic renal cell carcinoma treated with contemporary targeted therapy', *Cancer* **117**(12), 2637–2642.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. & Sargent, D. J. (2011), 'Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials', *Biometrics* **67**(3), 1047–1056.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors', *Statistical science* **14**(4), 382–417.

Ibrahim, J. G. & Chen, M.-H. (2000), 'Power prior distributions for regression models', *Statistical Science* pp. 46–60.

Kass, R. E. & Wasserman, L. (1995), 'A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion', *Journal of the American Statistical Association* **90**(431), 928–934.

Kiran Chandra, N., Sarkar, A., de Groot, J. F., Yuan, Y. & Müller, P. (2021), 'Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials', *arXiv e-prints* pp. arXiv–2201.

Kotalik, A., Vock, D. M., Donny, E. C., Hatsukami, D. K. & Koopmeiners, J. S. (2021), 'Dynamic borrowing in the presence of treatment effect heterogeneity', *Biostatistics* **22**(4), 789–804.

Lennie, J. L., Mondick, J. T. & Gastonguay, M. R. (2020), 'Latent process model of the 6-minute walk test in duchenne muscular dystrophy', *Journal of Pharmacokinetics and Pharmacodynamics* **47**(1), 91–104.

Lim, H.-S., Sun, W., Parivar, K. & Wang, D. (2019), 'Predicting overall survival and progression-free survival using tumor dynamics in advanced breast cancer patients', *The AAPS journal* **21**(2), 1–12.

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N. et al. (2018), 'Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities', *Therapeutic Innovation & Regulatory Science* **52**, 546–559.

Lin, J., Gamalo-Siebers, M. & Tiwari, R. (2018), 'Propensity score matched augmented controls in randomized clinical trials: A case study', *Pharmaceutical Statistics* **17**(5), 629–647.

Mackillop, W. J. & Quirt, C. F. (1997), 'Measuring the accuracy of prognostic judgments in oncology', *Journal of clinical epidemiology* **50**(1), 21–29.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. & Andersen, P. K. (2009), 'Multi-state models for the analysis of time-to-event data', *Statistical methods in medical research* **18**(2), 195–222.

Meller, M., Beyersmann, J. & Rufibach, K. (2019), 'Joint modeling of progression-free and overall survival and computation of correlation measures', *Statistics in medicine* **38**(22), 4270–4289.

Methy, N., Bedenne, L. & Bonnetain, F. (2010), 'Surrogate endpoints for overall survival in digestive oncology trials: which candidates? a questionnaires survey among clinicians and methodologists', *BMC cancer* **10**(1), 1–9.

Muntoni, F., Manzur, A., Mayhew, A., Signorovitch, J., Sajeev, G., Yao, Z. et al. (2018), ‘Minimal detectable change in the North Star Ambulatory Assessment (NSAA) in Duchenne muscular dystrophy (DMD)’, *Neuromuscular Disorders* **28**, S121–S121.

Muntoni, F., Signorovitch, J., Sajeev, G., Goemans, N., Wong, B., Tian, C. et al. (2022), ‘Real-world and natural history data for drug evaluation in Duchenne muscular dystrophy: Suitability of the North Star Ambulatory Assessment for comparisons with external controls’, *Neuromuscular Disorders* **32**(4), 271–283.
URL: *https://www.sciencedirect.com/science/article/pii/S096089662200061X*

Nahum-Shani, I., Almirall, D., Yap, J. R., McKay, J. R., Lynch, K. G., Freiheit, E. A. & Dziak, J. J. (2020), ‘Smart longitudinal analysis: A tutorial for using repeated outcome measures from smart studies to compare adaptive interventions.’, *Psychological Methods* **25**(1), 1.

Neuenschwander, B., Branson, M. & Spiegelhalter, D. J. (2009), ‘A note on the power prior’, *Statistics in Medicine* **28**(28), 3562–3566.

Neuenschwander, B., Capkun-Niggli, G., Branson, M. & Spiegelhalter, D. J. (2010), ‘Summarizing historical information on controls in clinical trials’, *Clinical Trials* **7**(1), 5–18.

Neuenschwander, B., Roychoudhury, S. & Schmidli, H. (2016), ‘On the use of co-data in clinical trials’, *Statistics in Biopharmaceutical Research* **8**(3), 345–354.

Neuenschwander, B., Weber, S., Schmidli, H. & O’Hagan, A. (2020), ‘Predictively consistent prior effective sample sizes’, *Biometrics* **76**(2), 578–587.

Ouma, L. O., Grayling, M. J., Wason, J. M. & Zheng, H. (2022), ‘Bayesian modelling strategies for borrowing of information in randomised basket trials’, *Journal of the Royal Statistical Society. Series C: Applied Statistics* .

Pak, K., Uno, H., Kim, D. H., Tian, L., Kane, R. C., Takeuchi, M., Fu, H., Claggett, B. & Wei, L.-J. (2017), ‘Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio’, *JAMA oncology* **3**(12), 1692–1696.

Plummer, M. (2017), ‘Jags version 4.3. 0 user manual [computer software manual]’, *Retrieved from sourceforge. net/projects/mcmc-jags/files/Manuals/4. x* **2**.

Plummer, M. (2022), *rjags: Bayesian graphical models using MCMC*. R package version 4-13, `https://CRAN.R-project.org/package=rjags` (accessed November 10, 2022).

Pocock, S. J. (1976), ‘The combination of randomized and historical controls in clinical trials’, *Journal of Chronic Diseases* **29**(3), 175–188.

Putter, H., Fiocco, M. & Geskus, R. B. (2007), ‘Tutorial in biostatistics: competing risks and multi-state models’, *Statistics in medicine* **26**(11), 2389–2430.

Quintana, M., Shrader, J., Slota, C., Joe, G., McKew, J., Fitzgerald, M., Gahl, W., Berry, S. & Carrillo, N. (2019), ‘Bayesian model of disease progression in gne myopathy’, *Statistics in Medicine* **38**(8), 1459–1474.

Raket, L. L. (2022), ‘Progression models for repeated measures: Estimating novel treatment effects in progressive diseases’, *arXiv preprint arXiv:2203.15555* .

Rosenbaum, P. R. & Rubin, D. B. (1983), ‘The central role of the propensity score in observational studies for causal effects’, *Biometrika* **70**(1), 41–55.

Sborov, K., Giaretta, S., Koong, A., Aggarwal, S., Aslakson, R., Gensheimer, M. F., Chang, D. T. & Pollom, E. L. (2019), ‘Impact of accuracy of survival predictions on quality of end-of-life care among patients with metastatic cancer who receive radiation therapy’, *Journal of Oncology Practice* **15**(3), e262–e270.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O’Hagan, A., Spiegelhalter, D. & Neuenschwander, B. (2014), ‘Robust meta-analytic-predictive priors in clinical trials with historical control information’, *Biometrics* **70**(4), 1023–1032.

Shukuya, T., Mori, K., Amann, J. M., Bertino, E. M., Otterson, G. A., Shields, P. G., Morita, S. & Carbone, D. P. (2016), ‘Relationship between overall survival and response or progression-free survival in advanced non–small cell lung cancer patients treated with anti–pd-1/pd-l1 antibodies’, *Journal of thoracic oncology* **11**(11), 1927–1939.

Sklar, M. (1959), Fonctions de répartition à n dimensions et leurs marges, *in* ‘Annales de l’ISUP’, Vol. 8, pp. 229–231.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004), *Bayesian approaches to clinical trials and health-care evaluation*, Vol. 13, John Wiley & Sons.

Stein, A., Bellmunt, J., Escudier, B., Kim, D., Stergiopoulos, S. G., Mietlowski, W., Motzer, R. J., Group, R.-. T. S. et al. (2013), ‘Survival prediction in everolimus-treated patients with metastatic renal cell carcinoma incorporating tumor burden response in the record-1 trial’, *European urology* **64**(6), 994–1002.

Tamura, R. N., Krischer, J. P., Pagnoux, C., Micheletti, R., Grayson, P. C., Chen, Y.-F. et al. (2016), ‘A small n sequential multiple assignment randomized trial design for use in rare disease research’, *Contemporary Clinical Trials* **46**, 48–51.

Tang, P. A., Bentzen, S. M., Chen, E. X. & Siu, L. L. (2007), ‘Surrogate end points for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy’, *Journal of Clinical Oncology* **25**(29), 4562–4568.

Uno, H., Tian, L., Horiguchi, M., Cronin, A., Battioui, C. & Bell, J. (2022), *survRM2: Comparing Restricted Mean Survival Time*. R package version 1.0-4.
URL: *https://CRAN.R-project.org/package=survRM2*

U.S. Food and Drug Administration (2020), ‘Interacting with the FDA on complex innovative trial designs for drugs and biological products’.

Verde, P. E. (2021), ‘A bias-corrected meta-analysis model for combining, studies of different types and quality’, *Biometrical Journal* **63**(2), 406–422.

Victor, R. G., Sweeney, H. L., Finkel, R., McDonald, C. M., Byrne, B., Eagle, M., Goemans, N., Vandenborne, K., Dubrovsky, A. L., Topaloglu, H. et al. (2017), 'A phase 3 randomized placebo-controlled trial of tadalafil for duchenne muscular dystrophy', *Neurology* **89**(17), 1811–1820.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S. et al. (2014), 'Use of historical control data for assessing treatment effects in clinical trials', *Pharmaceutical Statistics* **13**(1), 41–54.

Wadsworth, I., Hampson, L. V. & Jaki, T. (2018), 'Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods', *Statistical Methods in Medical Research* **27**(2), 398–413.

Wang, S. & Kidwell, K. (2022), *snSMART: small n sequential multiple assignment randomized trial methods*. R package version 0.2.2, https://CRAN.R-project.org/package=snSMART (accessed April 4, 2023).

Wang, S., Kidwell, K. M. & Roychoudhury, S. (2023), 'Dynamic enrichment of bayesian small-sample, sequential, multiple assignment randomized trial design using natural history data: a case study from duchenne muscular dystrophy', *Biometrics* **79**(4), 3612–3623.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13887*

Wang, Y., Sung, C., Dartois, C., Ramchandani, R., Booth, B., Rock, E. & Gobburu, J. (2009), 'Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development', *Clinical Pharmacology & Therapeutics* **86**(2), 167–174.

Weber, E. M. (2020), *Statistical models to capture the association between progression-free and overall survival in oncology trials*, Lancaster University (United Kingdom).

Weber, E. M. & Titman, A. C. (2019), 'Quantifying the association between progression-free survival and overall survival in oncology trials using kendall's τ', *Statistics in medicine* **38**(5), 703–719.

Wei, B., Braun, T. M., Tamura, R. N. & Kidwell, K. (2020), 'Sample size determination for Bayesian analysis of small n sequential, multiple assignment, randomized trials (snSMARTs) with three agents', *Journal of Biopharmaceutical Statistics* **30**(6), 1109–1120.

Wei, B., Braun, T. M., Tamura, R. N. & Kidwell, K. M. (2018), 'A Bayesian analysis of small n sequential multiple assignment randomized trials (snSMARTs)', *Statistics in Medicine* **37**(26), 3723–3732.

Yu, J., Wang, N. & Kågedal, M. (2020), 'A new method to model and predict progression free survival based on tumor growth dynamics', *CPT: pharmacometrics & systems pharmacology* **9**(3), 177–184.

Zecchin, C., Gueorguieva, I., Enas, N. H. & Friberg, L. E. (2016), 'Models for change in tumour size, appearance of new lesions and survival probability in patients with advanced epithelial ovarian cancer', *British journal of clinical pharmacology* **82**(3), 717–727.

Zhou, T. & Ji, Y. (2021), 'Incorporating external data into the analysis of clinical trials via bayesian additive regression trees', *Statistics in Medicine* **40**(28), 6421–6442.

Zhou, X. & Reiter, J. P. (2010), 'A note on bayesian inference after multiple imputation', *The American Statistician* **64**(2), 159–163.