Estimation Methods and Clinical Trial Design in Small n, Sequential, Multiple-Assignment, Randomized Trials

by

Yan-Cheng Chao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2021

Doctoral Committee:

Professor Thomas M. Braun, Co-Chair
Associate Professor Kelley M. Kidwell, Co-Chair
Professor Veerabhadran Baladandayuthapani
Assistant Professor Daniel L. Hertz

Yan-Cheng Chao

ycchao@umich.edu

ORCID iD: 0000-0002-5021-0415

ACKNOWLEDGEMENTS

I couldn't have imagined that I would eventually get a PhD degree in biostatistics six years ago. As a student majoring in Biology, everything I knew about statistics back then was as simple as the mean, standard deviation or linear regression. Without this opportunity from the University of Michigan Biostatistics, I could not have learned so much new knowledge about statistics, let alone devoting my effort to contributing to the field of clinical trials. During the last four years of pursuing my doctoral degree, I was fortunate to receive a lot of help from many people, including my advisors, committee members, friends and family.

First, my deepest gratitude goes to my two thesis advisors, Drs. Kelley Kidwell and Thomas Braun, for their support and guidance. Dr. Kidwell brought me into the world of research in clinical trial studies, and she kept teaching me identifying interesting new topics and developing methods to address the issues that we encounter during our research. Dr. Braun, as an expert in clinical trial researches, also provided me with a plenty of great ideas coming from his profound knowledge and extensive experience. What impressed me the most was that they held a great balance of flexibility and guidance as my advisors. They always encouraged me to search for any potential solution to a question, but they would also give me some suggestions such that I would not deviate from our original goal. In addition to their help in my research, I also appreciate their continuous weekly meetings with me throughout my entire four years as a doctoral student, even during the COVID-19 pandemic. It was a great experience that we could share our interesting stories with each other, and I

also learned a lot from their personal experiences and passion for their careers, which would definitely be helpful for my future professional development.

Next, I would also like to thank my dissertation committee members, Dr. Hertz and Dr. Baladandayuthapani. It was great to have an opportunity to work with Dr. Hertz on collaborative projects, and I also appreciate suggestions from both Drs. Hertz and Baladandayuthapani on my dissertation. Furthermore, I would also like to thank Drs. Xiang Zhou and Jian Kang for offering me chances to work with them as a research assistant when I was still a master student. Without their guidance, I probably would not discover my interests in biostatistics research.

I would also like to thank my friends in the department. Over the last six years in Michigan Biostatistics, I received a lot of help and insightful suggestions from them. Without their generosity, I might not be able to complete my work so smoothly, and my life here in Ann Arbor could be more challenging.

Last but not least, I could not have any chance to complete my doctoral degree without the support of my family. My wife, Jie, is a person who can always come up with brilliant ideas to tackle any difficulty we have. My parents, Chien-An and Li-Ling, always support my decision and give me confidence whenever I feel depressed. Thank you for your love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

The application of a small n, sequential, multiple-assignment randomized trial (snSMART) to rare disease studies remains an active research area. In this dissertation, we present methods that estimate dynamic treatment regimens (DTRs), or tailored sequences of treatments, for rare diseases, such as focal segmental glomerulosclerosis. We also develop an snSMART design that allows for removing an inferior treatment arm. Moreover, we summarize methods and develop new approaches to incorporate data from both stages in this estimation of the first stage treatment effect in an snSMART through the use of power priors.

Following an introduction of an snSMART and its potential application in Chapter I, in Chapter II, we propose a Bayesian joint stage model and a joint stage regression model, first developed by *Wei et al.* (2018). These models can be applied to estimate DTRs by combining information across stages. We show that the estimates from these two methods are more efficient than that of a standard SMART analysis of weighted and replicated regression (*Nahum-Shani et al.*, 2012). In addition, we introduce a sample size calculation method for our snSMART design when implementing the joint stage regression model with Dunnett's correction.

In Chapter III, we are motivated by an ongoing snSMART, ARAMIS (NCT02939573), focusing on the evaluation of three drugs for isolated skin vasculitis. We propose an alternative design by formulating an interim decision rule for removing one of the treatments, using Bayesian modelling and the resulting posterior distributions to provide sufficient evidence that one treatment is inferior to the other treatments. By doing so, we can remove the worst performing treatment at an interim analysis and

prevent subsequent participants from receiving the removed treatment. In addition, by adjusting the decision rule criteria for the posterior probabilities, we can control the probability of incorrectly removing a treatment, a Bayesian counterpart of Type I or Type II error rate used in frequentist methods.

In Chapter IV, we develop a novel method to incorporate outcomes from both stages in an snSMART to estimate the first stage treatment effects using power prior models. Here, we consider the first stage outcomes from an snSMART as the primary, or current, data and second stage outcomes as supplemental, or historical. We apply existing power prior models to snSMART data, and develop new extensions of power prior models. All methods are compared to each other and to the Bayesian joint stage model (BJSM) via simulation studies. By comparing the biases and the efficiency of the response rate estimates among all proposed power prior methods, we suggest application of Fisher's exact test or the Bhattacharyya's overlap measure to an snSMART to estimate the treatment effect in an snSMART, which both have performance mostly as good or better than the BJSM.

CHAPTER I

Introduction

1.1 Motivation

Individualized treatment strategies are encouraged for the treatment of long-term chronic diseases, such as cancers or psychological disorders. Patients' pre-existing conditions are taken into account when treatments are assigned, and their response to earlier treatments may be evaluated and serve as guidance for later treatments. The concept of a dynamic treatment regimen (DTR), also known as an adaptive treatment strategy, was introduced by *Lavori et al.* (2000), and further described by *Lavori and Dawson* (2000), *Murphy* (2003), and *Lavori and Dawson* (2004), to describe this clinical practice of sequences of tailored interventions, and DTRs provide guidelines for clinical decision making.

In order to develop effective DTRs and identify one that can lead to the best overall outcome, *Murphy* (2005a) proposed a clinical trial design called a sequential, multiple assignment, randomized trial (SMART) that embeds DTRs by its design. A SMART is a multi-stage design where the participants are randomized to one of the treatment arms, and the later treatment assignments depend on whether they respond to their earlier treatment. The numbers of participants enrolled in a SMART usually exceed 200 (*Estey et al.*, 1999; *Rush et al.*, 2004; *Kelleher et al.*, 2017; *Ruppert et al.*, 2019).

A RAndomized Multi-center study for Isolated Skin vasculitis (ARAMIS) trial was a SMART launched in 2016 that started recruiting patients in 2017. Because skin vasculitis is a rare disease, the expected sample size for this clinical trial is 90 patients, which is smaller than most of the other SMARTs being conducted. Since most of the existing methods are usually applied for the settings of a relatively large sample sizes, new methods need to be developed for the estimation of DTRs in a SMART with small sample sizes.

Due to the nature of multiple treatment assignments for each patient, a SMART design is particularly useful to estimate treatment effects of rare disease therapies because more information can be collected from one patient. Thus, we refer to a small sample (n) SMART, or snSMART, when the goal of a SMART is to estimate individual treatment effects in a small sample setting. *Wei et al.* (2018) has proposed models to estimate the first stage treatment effect that combine outcomes from two stages in an snSMART together by making an assumption of proportionality of first and second stage response rates. However, such an assumption may be hard to justify in practice. Thus, we present an alternative model that can be used in a more general setting, which includes the scenario where this assumption of proportionality is violated.

ARAMIS is a fixed snSMART design, such that no adaptation or interim analyses were performed. We consider a modification to the original snSMART design to allow for the dropping of an inferior treatment arm, which potentially appeal to patients. Many rare disease trials implement adaptation for better recruitment or early termination.

1.2 Summary of Objectives

With the focus on tackling the issues described above, we present the main objectives of the next three chapters and introduce briefly how we achieve those goals.

In Chapter II, we apply and modify the Bayesian joint stage model and joint stage regression model developed by *Wei et al.* (2018) to estimate the response rates of DTRs embedded in an snSMART. We perform simulation studies to compare the performance, in terms of biases and root mean squared errors of estimation, of these two models to the existing weighted and replicated regression model used for larger sample SMART studies. In addition, we develop a simulation-based method to calculate the required sample size of an snSMART when there is a control arm.

In Chapter III, we introduce a modified snSMART design incorporating a group sequential feature, or a group sequential snSMART. In this design, investigators can decide if a treatment arm should be removed from the trial due to inferiority during the interim analyses. This new design allows more participants to be assigned to the better performing treatments since the worst treatment arm tends to be removed. The probabilities of correctly or incorrectly removing an arm during the interim analyses are presented.

In Chapter IV, we propose a power prior model approach for the estimation of treatment effects in an snSMART. Different ways of estimating power parameters are presented, including existing likelihood based methods and a novel application of measures of closeness. Through simulation studies, we compare these power prior models to the existing Bayesian joint stage model.

CHAPTER II

Frequentist and Bayesian Methods to Estimate Dynamic Treatment Regimens in a Small n, Sequential, Multiple-Assignment, Randomized Trial

2.1 Introduction

Focal segmental glomerulosclerosis (FSGS) is a rare kidney disease with an annual incidence of 0.2-1.8 cases per 100,000 individuals (*Rosenberg and Kopp*, 2017). FSGS has traditionally been diagnosed in patients with persistent proteinuria based on characteristic lesions in a kidney biopsy specimen (*D'Agati et al.*, 2011). The identification of an effective treatment for FSGS is generally by trial and error, i.e., try a therapy, assess response, move to alternate treatment option in treatment failures, and then repeat these steps. There is little evidence to guide the choice of initial therapy or the selection of subsequent therapies dependent on initial treatment response patterns.

Nephrologists are confronted with several questions when caring for patients with FSGS. How should treatments targeting a specific mechanism of disease be selected, implemented, and assessed? Which of these treatments can provide the best short- and long-term response rate? What is the best sequence to introduce therapy when

there are several options? These questions are not unique to FSGS, but similar types of questions are shared across many rare diseases and have proved difficult to answer in clinical trials with small samples.

One clinical trial design that could address these questions in FSGS and other rare diseases is the small n sequential, multiple assignment, randomized trial design (snSMART) (*Tamura et al.*, 2016; *Wei et al.*, 2018). An snSMART design is a multi-stage trial where participants are first randomized to one of the treatment arms and those who do not respond to the initial treatment are re-randomized to one of the other treatment options.

Wei et al. (2018) proposed both a frequentist Joint stage regression model (JSRM) and a Bayesian joint stage model (BJSM) to estimate the treatment effects in an snSMART where the outcome of interest is a binary indicator of response to treatment. We show here that 1) snSMART designs may be used in settings with two active treatments and a standard of care, 2) snSMART designs may also be used to estimate and compare DTRs, and 3) sample size can be calculated via simulation study when the frequentist model is used.

Several methods exist for comparing DTRs embedded in a SMART. For continuous and binary data, a weighted and replicated regression approach allows for simultaneous estimation of all embedded DTRs while controlling for covariates (*Nahum-Shani et al.*, 2012; *Kidwell et al.*, 2017), which will be described in more detail in Section 2.2.3. Additional semiparametric methods for estimating and comparing DTRs include G-estimation (*Robins*, 2004), regret-regression (*Henderson et al.*, 2010) adapted from Murphy's iterative minimization of regrets method (*Murphy*, 2003) and Q-learning (*Murphy*, 2005b). Most of the existing methods, however, are based on large sample theory, which means that the methods may not provide reliable estimation when the sample size is small.

Several Bayesian approaches have been developed in an attempt to obtain efficient and/or unbiased results for DTRs in small and large samples. One approach uses data from an observational study to infer the joint posterior predictive distribution of all baseline covariates, potential first-stage outcomes and second-stage outcomes of each possible subgroup that an individual may follow (*Zajonc*, 2012). From this joint posterior, Zajonc simulated new samples to determine the best DTR using a pre-specified utility function. In a similar vein, *Saarela et al.* (2015) proposed a Bayesian model for estimating DTRs that incorporates inverse probability of treatment weighting to deal with the potential confounding. Two methods have also been proposed for DTRs with time-to-event outcomes (*Thall et al.*, 2007; *Xu et al.*, 2016). A common characteristic for both methods is the estimation of DTR-specific mean failure times based on the sum of the estimated transition times of different stages, which avoids the need to consider weighting.

In Section 2.2, we present both frequentist and Bayesian models to estimate and compare first-stage treatment effects sharing information across stages and to estimate and compare DTRs where interest is in the longer-term course of care. In Section 2.3, bias and efficiency of the presented models are compared via simulations, which are motivated by the FSGS setting. In addition, we demonstrate the method to find the sample size required in an snSMART to compare novel treatments to a standard of care implementing the frequentist model with Dunnett's correction. Lastly, we close with a discussion in Section 2.4.

2.2 Bayesian and Frequentist Analyses of an snSMART

We present Bayesian and frequentist models that can estimate both first stage treatment effects and DTR effects using data from both the first and second stages of an snSMART. Specifically, we present a Bayesian joint stage model (BJSM) and a

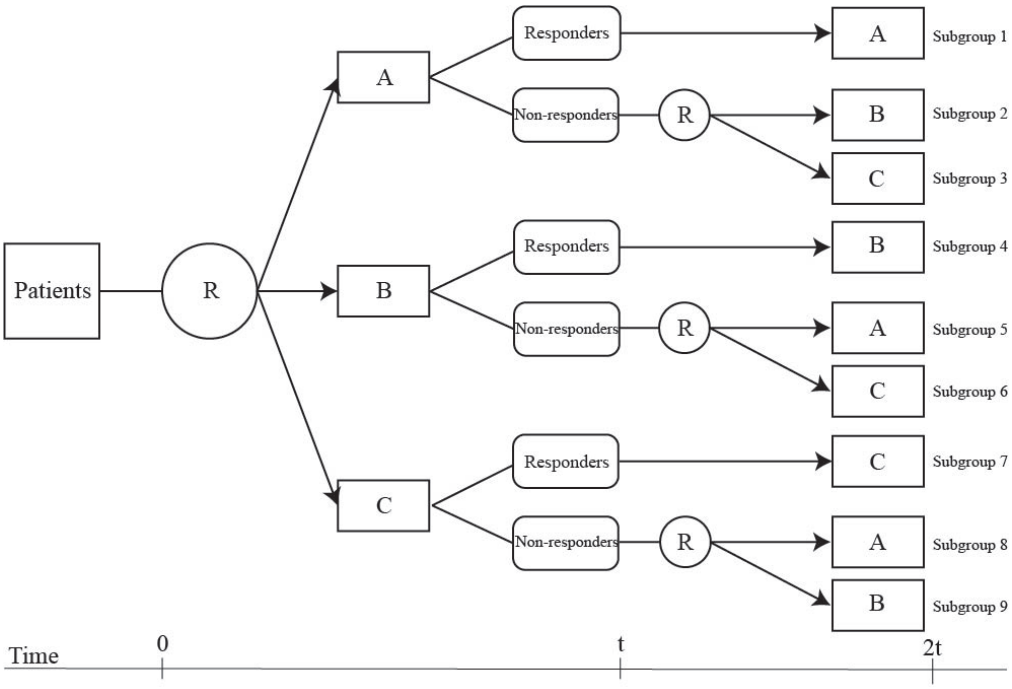

Figure 2.1: A small n, sequential, multiple assignment, randomized trial (snSMART) design. Subjects are allocated to one of the three first stage treatment groups A, B, C at time 0. R represents equal randomization to the following treatments. Based on the response status at time t, patients either continue the initial treatment or are re-randomized to one of the other two treatments. Subgroups 1 through 9 denote the treatment paths that any one patient may follow. Second stage responses can be obtained at time $2t$. The combination of two treatment paths, one for responders and another for non-responder sharing the same first stage treatment defines a DTR.

frequentist joint stage regression model (JSRM) using generalized estimating equations (GEE). These models are extensions of those in *Wei et al.* (2018) allowing for estimates of DTRs in the presence of potential first and second stage treatment interactions.

2.2.1 Bayesian Joint Stage Model

For subject $i = 1, 2, \ldots, N$, treatment $j = A, B, C$, and stage $k = 1, 2$, we let Y_{ik}^{j} be an indicator of response for subject i receiving treatment j in stage k. The first-stage response rate to treatment j is denoted by π_j. The second-stage response rate of the first-stage responders to treatment j is denoted by $\beta_{1j}\pi_j$, and the second-stage

response rate of the non-responders to first-stage treatment j who receive treatment j' in the second stage is denoted by $\beta_{0j}\pi_{j'}$. Here, β_{1j} and β_{0j} are called linkage parameters since they link the first stage response rate to the second stage response rate (*Wei et al.*, 2018). We assume that the first-stage non-responders are less likely to respond to either of the two other treatments in second stage ($\beta_{0j} < 1$). We also assume that linkage parameters only depend on the first stage treatment.

We are interested in estimating the second stage response rates of all DTRs, denoted by $\pi_{jjj'}$, where first j indicates the first stage treatment, second j indicates the second stage treatment as a first-stage responder, and j' indicates the second stage treatment as a first-stage non-responder. For example, the DTR "AAB" encompasses all patients who received treatment A in the first stage and then would receive either treatment A or B in the second stage depending on whether the patients respond to the first stage treatment. The second stage response rate of the DTR "AAB" is denoted by π_{AAB}. We first obtain the posterior draws of π_j, β_{1j}, and β_{0j} through the BJSM (*Wei et al.*, 2018) as follows:

$$Y_{i1}^{j}|\pi_j \sim \text{Bernoulli}(\pi_j) \tag{2.1}$$

$$Y_{i2}^{j'}|Y_{i1}^{j}, \pi_j, \pi_{j'}, \beta_{1j}, \beta_{0j} \sim \text{Bernoulli}\Big((\beta_{1j}\pi_j)^{Y_{i1}^{j}}(\beta_{0j}\pi_{j'})^{1-Y_{i1}^{j}}\Big) \tag{2.2}$$

$$\pi_j \sim \text{Beta}(\theta_1, \delta_1) \tag{2.3}$$

$$\beta_{0j} \sim \text{Beta}(\theta_2, \delta_2) \tag{2.4}$$

$$\beta_{1j} \sim \text{Gamma}(\theta_3, \delta_3) \tag{2.5}$$

Equations 2.1 and 2.2 show the distributions of the first and second stage responses. The prior distributions for the parameters π_j, β_{0j} and β_{1j} are given in Equations 2.3, 2.4 and 2.5. The hyperparameters of the prior distributions should be based on prior

knowledge from investigators. Specifically, for π_j, we assigned the values of $\theta_1 = 0.4$ and $\delta_1 = 1.6$, which gives a prior mean of 0.2 for the response rates since we believe that an ineffective treatment or standard of care would have a response rate of 20%. Similarly, for β_{0k}, we have assigned the values of $\theta_2 = 1.6$ and $\delta_2 = 0.4$, so that the average prior response rate for the second stage treatment for non-responders was assumed *a priori* to be reduced by 20% compared to the first-stage response rate of the same treatment. For β_{1j}, we assigned the values of $\theta_3 = 2$ and $\delta_3 = 2$, so that the prior mean of 1 indicates that the first stage responders in the first stage are assumed to have the same response rate to the same treatment in the second stage. We note the change in the prior distribution of β_{1j} from *Wei et al.* (2018). We made the prior distribution more flexible here such that β_{1j} can range from zero to infinity as opposed to one to infinity.

Next, we compute the posterior draws for each DTR $\pi_{jjj'}$ from the following equation using the the posterior draws of β_{0j}, β_{1j}, and π_j:

$$\pi_{jjj'} = \pi_j(\pi_j\beta_{1j}) + (1 - \pi_j)(\pi_{j'}\beta_{0j}) \tag{2.6}$$

As a result, it is easy to calculate the means and standard deviations of $\pi_{jjj'}$ from their posterior draws.

2.2.2 Joint Stage Regression Model

A joint-stage regression model (JSRM) is a frequentist modeling approach that incorporates the responses of both stages as repeated measurements for each subject. Hence, generalized estimating equations (GEE) are used to estimate the response rates of each treatment and from these estimates, we can compute the marginal response rates for each DTR. For binary outcomes, the logit link is most commonly applied to estimate the response rate (*Lei et al.*, 2012; *Kidwell et al.*, 2017). However,

in small samples, the standard errors of parameters tend to be underestimated if we fit the model with the logit link function (*Mancl and DeRouen*, 2001). Instead of applying the bias-corrected variance estimator of *Mancl and DeRouen* (2001), we use a log link in our GEE model (*Williamson et al.*, 2013).

For subject $i = 1, \ldots, N$, and stage $k = 1, 2$, we let Y_{ik} be the response of subject i in stage k. The JSRM with six linkage parameters is as follows:

$$\begin{aligned}\log(P(Y_{ik} = 1)) = {} & \alpha_1 \mathbb{1}(j_{ik} = A) + \alpha_2 \mathbb{1}(j_{ik} = B) + \alpha_3 \mathbb{1}(j_{ik} = C) \\ & + [\alpha_4 \mathbb{1}(Y_{i1} = 1) + \alpha_5 \mathbb{1}(Y_{i1} = 0)] \mathbb{1}(j_{i1} = A, k = 2) \\ & + [\alpha_6 \mathbb{1}(Y_{i1} = 1) + \alpha_7 \mathbb{1}(Y_{i1} = 0)] \mathbb{1}(j_{i1} = B, k = 2) \\ & + [\alpha_8 \mathbb{1}(Y_{i1} = 1) + \alpha_9 \mathbb{1}(Y_{i1} = 0)] \mathbb{1}(j_{i1} = C, k = 2)\end{aligned} \tag{2.7}$$

where j_{ik} is the treatment indicator of subject i in stage k and $\mathbb{1}(\cdot)$ is an indicator function. In Equation 2.7, α_1 to α_3 correspond to the first stage response rates of the treatments A, B, and C, or π_A, π_B, and π_C, respectively. The second stage response rates start with the first stage response rates, but are then augmented by an amount depending on both (i) the treatment received in stage 1, and (ii) whether or not a response occurred in stage 1. For example, consider the individuals who receive treatment B in stage 2 after not responding to treatment A in stage 1. Their second stage response rate would be a function of α_2, in the first line of Equation 2.7, and then augmented by α_5 to reflect the non-response to treatment A in stage 1.

We assume that the dependency of Y_{i2} on Y_{i1} is already taken into account with the covariates in the JSRM, so we fit this GEE model with the independence working covariance structure and use a robust "sandwich" estimator to estimate $\text{Cov}(\hat{\boldsymbol{\alpha}})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_9)^\top$. Even if there is additional correlation of Y_{i1} and Y_{i2} that is not captured by the working covariance matrix, the use of sandwich estimator guarantees the unbiasedness of parameter estimates, and it is flexible to change the

working covariance matrix to other type, such as compound symmetry, if a strong additional within-subject correlation can be identified.

To estimate the first stage response rates denoted by π_j, $j = A, B, C$, we exponentiate the first three coefficients in Equation 2.7. Specifically, $\hat{\pi}_A = \exp(\hat{\alpha}_1)$, $\hat{\pi}_B = \exp(\hat{\alpha}_2)$, and $\hat{\pi}_C = \exp(\hat{\alpha}_3)$. To estimate the response rates of the six embedded DTRs, we use linear combinations of estimates from Equation 2.7. For example, from Figure 2.1, DTR "AAB" contains subgroups 1 and 2 and the estimated response rate is $\hat{\pi}_{AAB} = \hat{\pi}_A \times \hat{\beta}_{1A}\hat{\pi}_A + (1 - \hat{\pi}_A) \times \hat{\beta}_{0A}\hat{\pi}_B$, where $\hat{\pi}_A = \exp(\hat{\alpha}_1)$, $\hat{\pi}_B = \exp(\hat{\alpha}_2)$, $\hat{\beta}_{1A} = \exp(\hat{\alpha}_4)$, $\hat{\beta}_{0A} = \exp(\hat{\alpha}_5)$. The response rates of the other DTRs can be estimated through a similar approach. The standard error of each estimated DTR response rate can be obtained from $\text{Cov}(\hat{\boldsymbol{\alpha}})$ using the Delta method.

2.2.3 Weighted and Replicated Regression Model

We briefly review the existing weighted and replicated regression model (WRRM) to estimate the response rate of DTRs embedded in a SMART design; more details can be found in *Nahum-Shani et al.* (2012). We note that WRRM uses only second stage responses to estimate the response rates of DTRs, in contrast to the previous two methods which use both first and second stage outcomes. As previously described in Section 2.1, the subjects are weighted before the model is fit using weights based on the inverse-probability-of-treatment. To estimate the response rates of different DTRs simultaneously using standard software, we then need to implement replication. In general, the second stage outcomes for subjects who are consistent with more than one DTR are replicated. For example, the subjects who respond to A in the first stage and continue the same treatment are consistent with two DTRs "AAB" and "AAC", meaning that their second stage responses are used in estimation of both DTRs "AAB" and "AAC". Thus, we replicate the data of these subjects who are consistent

with two DTRs and assign these two sets of data to two DTRs, respectively. For non-responders to the first stage treatments, they are only consistent with one DTR, so no replication is required for them, and their second stage outcomes are only used for the estimation of the corresponding DTRs. As a result of replication of data for the first stage responders, the data are now considered as repeated measurements, which is the reason that the estimation of DTR effects is conducted under the framework of GEE (*Nahum-Shani et al.*, 2012).

Parallel to the JSRM approach, we use a log link function in our model. We follow a model parametrization so that dummy variable coding indicates the first and second stage treatments, where the DTR "AAB" is chosen as a reference DTR. Thus, if we let Y_{i2r}, $r = 1, 2$, be the second-stage response of subject i, the WRRM is:

$$\begin{aligned}\log(P(Y_{i2r} = 1)) &= \alpha'_0 + \alpha'_1 \mathbb{1}(j_i = B) + \alpha'_2 \mathbb{1}(j_i = C) \\ &+ \alpha'_3 \mathbb{1}(j_i = A, j'_{ir} = C) + \alpha'_4 \mathbb{1}(j_i = B, j'_{ir} = C) + \alpha'_5 \mathbb{1}(j_i = C, j'_{ir} = B)\end{aligned} \tag{2.8}$$

where j'_{ir} is the r-th second-stage treatment of the subject i. For the non-responders to first stage treatment who are consistent with only one DTR, $r = 1$, and for the responders to first stage treatment who are consistent with two DTRs, $r = 1, 2$. We note that WRRM uses only second stage responses to estimate the response rates of DTRs, in contrast to the previous two methods which use both first and second stage outcomes.

After the model is fit, we estimate the response rates for each DTR by considering linear combinations of the regression parameters. DTR "AAB" is estimated by $\exp(\hat{\alpha}'_0)$, DTR "AAC" is estimated by $\exp(\hat{\alpha}'_0 + \hat{\alpha}'_3)$, DTR "$BBA$" is estimated by $\exp(\hat{\alpha}'_0 + \hat{\alpha}'_1)$, DTR "$BBC$" is estimated by $\exp(\hat{\alpha}'_0 + \hat{\alpha}'_1 + \hat{\alpha}'_4)$, DTR "$CCA$" is estimated by $\exp(\hat{\alpha}'_0 + \hat{\alpha}'_2)$, and DTR "$CCB$" is estimated by $\exp(\hat{\alpha}'_0 + \hat{\alpha}'_2 + \hat{\alpha}'_5)$. The variances of the estimated response rates for each DTR are calculated using the

delta method, where the variances and covariances of the estimated parameters in the models can be computed by the robust "sandwich" variance estimators.

Although both JSRM and WRRM are GEE-based models, they are conceptually different. The repeated measurements in JSRM represent the first- and second-stage outcomes of one subject after receiving the treatment in the first and second stages. However, the repeated measurements in WRRM are the second-stage outcomes for responders for their associated consistent DTRs in order to use standard software to simultaneously estimate the DTRs.

2.3 Simulation Results

2.3.1 Scenarios

In order to compare the performance of the methods described in Section 2.2, we conducted simulation studies where we estimated the response rates of DTRs and their variances from the two-stage design shown in Figure 2.1. We outline the data generation process here. Each arm in stage 1 contains exactly one-third of the subjects. Subject responses to first-stage treatments A, B and C are generated from Bernoulli distributions with parameters π_A, π_B and π_C, respectively. Second-stage responses for the responders to first-stage treatments are generated from Bernoulli distributions with parameters specified as the products of first-stage response rates and the corresponding linkage parameters: $\beta_{1A}\pi_A, \beta_{1B}\pi_B$, and $\beta_{1C}\pi_C$. Similarly, second-stage responses for the non-responders to first-stage treatments are generated from Bernoulli distributions with parameters specified as the products of β_{0j} and $\pi_{j'}$, $j = A, B, C$. For example, second-stage responses for the subjects who do not respond to treatment A in the first stage and receive the treatment B in the second stage (subgroup 2 in Figure 2.1) are generated from Bernoulli($\beta_{0A}\pi_B$).

Four sets of scenarios are considered in our study with different true treatment response rates. In scenarios 1a-c and 2a-c, two potentially active treatments A and B have the same response rates, but the treatments A and B in scenarios 2a-c have even higher response rate. The true response rates in these two sets of scenarios resemble potential settings in FSGS. In scenarios 3a-c, only one of the potentially active treatments is truly better than the standard of care in terms of response rate. In scenarios 4a-c, both potentially active treatments have higher response rates than that of the standard of care, but A has a even higher response rate. In each set of scenarios, there are three different combinations of linkage parameters β_{0j} and β_{1j}, $j = A, B, C$, with different assumptions. The true parameter values of each scenario are shown in Table 2.1. In the scenarios ending with a, β_{1j}, $j = A, B, C$, are assumed equal, and β_{0j}, $j = A, B, C$, only depend on the first stage treatment. For example, for the non-responders to treatment A, their linkage parameters are 0.8 regardless of which alternative treatments they receive in the second stage. In the scenarios ending with b, both β_{0j} and β_{1j} depend on the first stage treatements. In the scenarios ending with c, the linkage parameters for non-responders depend on both first and second stage treatments, which violates the assumption on β_{0k} of the BJSM and JSRM.

2.3.2 Estimation with Bayesian and Frequentist Methods

We evaluate the response rate estimates of treatments obtained using different methods within each set of scenarios. Since the models here differ from those in *Wei et al.* (2018), we compare the estimates from the BJSM, JSRM and first stage maximum likelihood estimate (FSMLE, the MLE of response rates using only first stage outcomes). For each scenario, we simulate 1,000 replications and obtain first stage treatment effect estimates using the BJSM, JSRM and FSMLE. Table 2.2 shows the biases and rMSEs of the estimated treatment response rates in all twelve scenarios with the given sample sizes. The sample sizes used in these scenarios were calculated

Scenario		π_A	π_b	π_C	β_{1A}	β_{1B}	β_{1C}	β_{0AB}	β_{0AC}	β_{0BA}	β_{0BC}	β_{0CA}	β_{0CB}
1	a	0.40	0.40	0.20	1.0	1.0	1.0	0.8	0.8	0.6	0.6	0.4	0.4
	b				1.5	1.0	0.5	0.8	0.8	0.6	0.6	0.4	0.4
	c				1.5	1.0	0.5	0.65	0.75	0.7	0.6	0.75	0.45
2	a	0.45	0.45	0.20	1.0	1.0	1.0	0.8	0.8	0.6	0.6	0.4	0.4
	b				1.5	1.0	0.5	0.8	0.8	0.6	0.6	0.4	0.4
	c				1.5	1.0	0.5	0.65	0.75	0.7	0.6	0.75	0.45
3	a	0.45	0.20	0.20	1.0	1.0	1.0	0.8	0.8	0.6	0.6	0.4	0.4
	b				1.5	1.0	0.5	0.8	0.8	0.6	0.6	0.4	0.4
	c				1.5	1.0	0.5	0.65	0.75	0.7	0.6	0.75	0.45
4	a	0.45	0.30	0.20	1.0	1.0	1.0	0.8	0.8	0.6	0.6	0.4	0.4
	b				1.5	1.0	0.5	0.8	0.8	0.6	0.6	0.4	0.4
	c				1.5	1.0	0.5	0.65	0.75	0.7	0.6	0.75	0.45

Table 2.1: Response rates and linkage parameter values used to generate data for all scenarios. π_k $(k = A, B, C)$ is the response rate of treatment K in the first stage. β_{1k} is the linkage parameter for first stage responders to k, and $\beta_{0kk'}$ is the linkage parameter for first stage non-responders to k who receive k' in the second stage.

using the method in Section 2.3.3.

In all scenarios, the BJSM has the largest biases among the three models, while the biases for JSRM and FSMLE are negligible. The only exceptions are scenario c's where the biases for JSRM are larger than that of the FSMLE due to the fact that model assumptions on the linkage parameters for non-responders are violated, i.e., the linkage parameters for non-responders depend on both first and second stage treatments. When looking at the rMSEs, we find that BJSM performs slightly better than JSRM, and the rMSE of the FSMLE is the highest, which can be expected because only first-stage outcomes are used. Thus, the BJSM is more efficient than the other methods, but efficiency comes at a price of computational intensiveness and some biases. The BJSM and JSRM are preferred over the FSMLE based on efficiency. The choice between the BJSM and JSRM may depend on the bias-variance tradeoff and computational resources.

We can evaluate the response rate estimates of DTRs obtained using different methods within each set of scenarios as well. We compare the estimates from the BJSM, JSRM

and WRRM. The expected DTR response rates are shown in Table 2.3. Figure 2.2 shows the absolute values of mean biases and mean rMSEs of the estimated DTR response rates in all twelve scenarios with the calculated sample sizes. The detailed results for Scenarios 1a-c through 4a-c are tabulated in Table A.1 in Appendix A. Estimates from the BJSM have the largest biases among the three methods, while the biases from the other two methods are negligible. However, estimates from BJSM have the smallest rMSEs, and estimates from WRRM have the largest rMSEs because only second stage outcomes are used. Similar to the results from first stage treatment effect estimation, BJSM and JSRM are preferred over the WRRM based on the efficiency, and the choice between these two methods need to account for the bias-variance tradeoff and computational resources.

2.3.3 Sample Size Calculation via Dunnett's Method

A sample size calculation for an snSMART is available for comparing the first stage response rates (*Tamura et al.*, 2016). This sample size calculation, however, is based on a frequentist method that does not use all of the second stage data and thus is not efficient (*Wei et al.*, 2018). Here, we present a simulation-based sample size calculation using the JSRM with six linkage parameters when interest is in comparing two active treatments to a control or standard of care with a specified family wise error rate and power. We use the JSRM for sample size calculation based on its computation speed and frequentist operating characteristics.

We apply Dunnett's approach under GEE (*Orelien et al.*, 2002; *Hsu*, 1992) to identify a significant difference between the two drugs of interest, in our setting the novel antifibrotic drug (A) and novel anti-inflammatory drug (B), with the standard of care (C). The detailed steps of the approach can be found in *Orelien et al.* (2002). Simulations are conducted to obtain the total sample size to achieve a family-wise

Scenario		BJSM		JSRM		FSMLE	
		Bias	rMSE	Bias	rMSE	Bias	rMSE
1a	π_A	-0.031	0.068	-0.001	0.069	0.000	0.072
	π_B	-0.021	0.065	0.001	0.069	0.000	0.072
	π_C	-0.009	0.047	-0.002	0.052	-0.003	0.058
1b	π_A	-0.020	0.062	-0.001	0.069	0.000	0.072
	π_B	-0.020	0.065	0.001	0.069	0.000	0.072
	π_C	-0.014	0.048	-0.002	0.052	-0.003	0.058
1c	π_A	-0.001	0.056	0.020	0.071	0.000	0.072
	π_B	-0.040	0.072	-0.021	0.071	0.000	0.072
	π_C	-0.016	0.049	-0.002	0.052	-0.003	0.058
2a	π_A	-0.040	0.083	-0.001	0.086	-0.001	0.089
	π_B	-0.029	0.080	0.001	0.086	0.000	0.089
	π_C	-0.010	0.057	-0.003	0.064	-0.004	0.071
2b	π_A	-0.026	0.075	-0.001	0.086	-0.001	0.089
	π_B	-0.029	0.080	0.001	0.086	0.000	0.089
	π_C	-0.016	0.057	-0.003	0.064	-0.004	0.071
2c	π_A	-0.004	0.068	0.022	0.088	-0.001	0.089
	π_B	-0.047	0.086	-0.023	0.087	0.000	0.089
	π_C	-0.017	0.058	-0.003	0.064	-0.004	0.071
3a	π_A	-0.042	0.077	-0.002	0.076	0.000	0.078
	π_B	-0.009	0.051	0.000	0.057	0.000	0.062
	π_C	-0.009	0.049	-0.002	0.056	-0.003	0.062
3b	π_A	-0.030	0.069	-0.002	0.076	0.000	0.078
	π_B	-0.009	0.051	0.000	0.057	0.000	0.062
	π_C	-0.015	0.050	-0.002	0.056	-0.003	0.062
3c	π_A	-0.009	0.060	0.014	0.076	0.000	0.078
	π_B	-0.019	0.052	-0.013	0.056	0.000	0.062
	π_C	-0.017	0.050	-0.004	0.055	-0.003	0.062
4a	π_A	-0.039	0.076	-0.002	0.075	0.000	0.078
	π_B	-0.016	0.061	0.000	0.067	-0.001	0.071
	π_C	-0.009	0.049	-0.002	0.055	-0.003	0.062
4b	π_A	-0.028	0.069	-0.002	0.075	0.000	0.078
	π_B	-0.016	0.061	0.000	0.067	-0.001	0.071
	π_C	-0.015	0.050	-0.002	0.055	-0.003	0.062
4c	π_A	-0.007	0.060	0.018	0.077	0.000	0.078
	π_B	-0.031	0.065	-0.018	0.067	-0.001	0.071
	π_C	-0.017	0.051	-0.004	0.055	-0.003	0.062

Table 2.2: The bias and root mean squared error (rMSE) of the treatment response rate estimates using Bayesian Joint Stage Model (BJSM), Joint Stage Regression Model (JSRM), and first stage MLE (FSMLE). The sample sizes for scenarios 1a-c, 2a-c, 3a-c, and 4a-c, are 135, 90, 120 and 120, respectively.

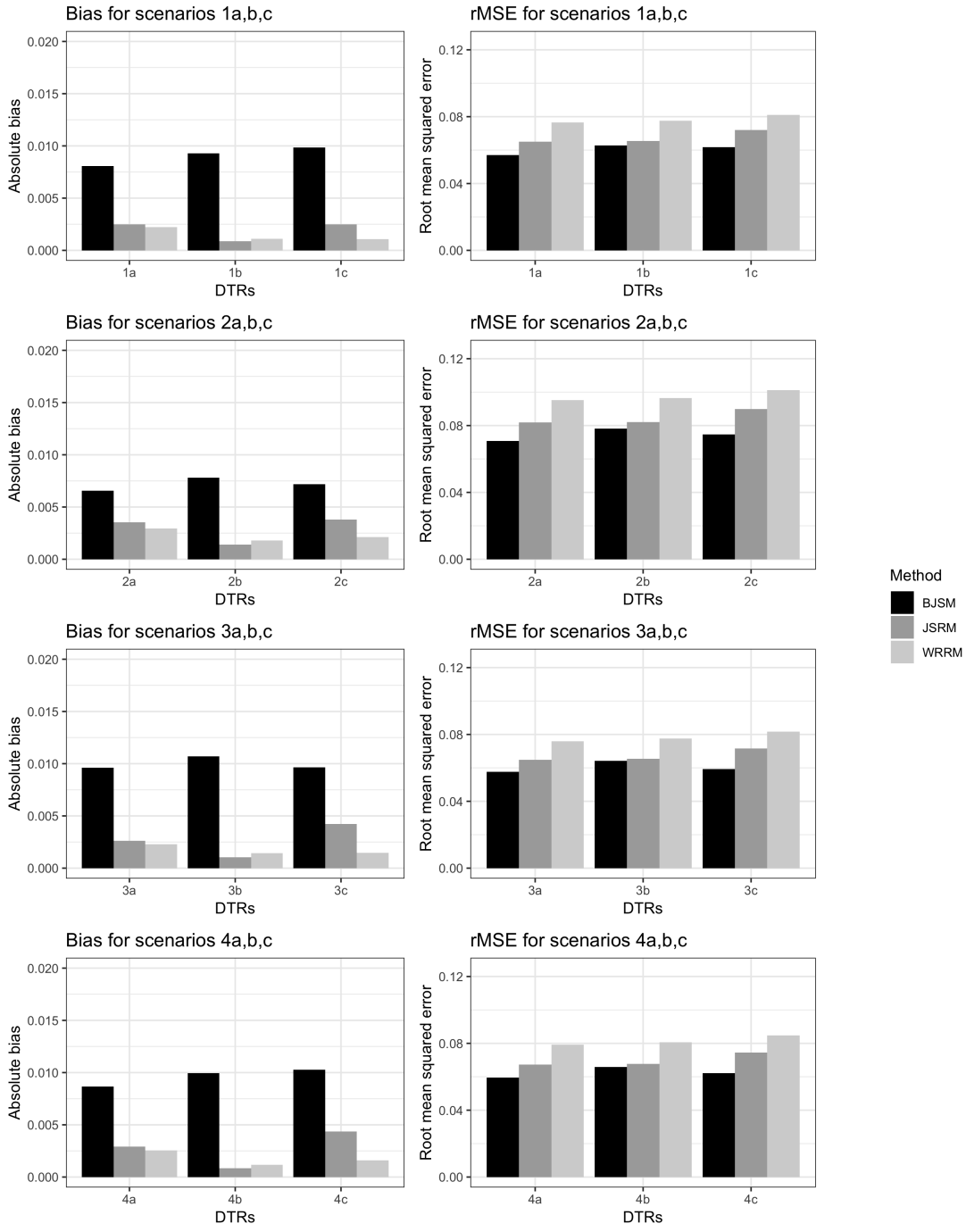

Figure 2.2: Left column: The absolute values of means of bias of DTR response rate estimates across Scenarios 1 to 4. Right column: The means of root mean squared error (rMSE) of DTR response rate estimates across Scenarios 1 to 4.

Scenario		π_{AAB}	π_{AAC}	π_{BBA}	π_{BBC}	π_{CCA}	π_{CCB}
1	a	0.352	0.256	0.304	0.232	0.168	0.168
	b	0.432	0.336	0.304	0.232	0.148	0.148
	c	0.396	0.330	0.328	0.232	0.260	0.164
2	a	0.401	0.291	0.351	0.268	0.184	0.184
	b	0.502	0.392	0.351	0.268	0.164	0.164
	c	0.465	0.386	0.376	0.268	0.290	0.182
3	a	0.291	0.291	0.256	0.136	0.184	0.104
	b	0.392	0.392	0.256	0.136	0.164	0.084
	c	0.375	0.386	0.292	0.136	0.290	0.092
4	a	0.334	0.291	0.279	0.174	0.184	0.136
	b	0.436	0.392	0.279	0.174	0.164	0.116
	c	0.411	0.386	0.310	0.174	0.290	0.128

Table 2.3: The expected response rate of dynamic treatment regimens (DTRs) for each scenario in Table 2.1. π_{AAB} corresponds to DTR "AAB", and the rests are similar.

type I error rate (α) of 10% and 80% power. Since we are performing two pair-wise comparisons (A vs. C and B vs. C), type I error rate is defined as the probability that either or both of the two p-values are smaller than the nominal α when all three drugs have same response rates, and power is defined as the probability that either or both of the two p-values are smaller than the nominal α if both drugs of interest truly have higher response rates than the that of the standard of care. One thousand replicates have been performed for each sample size. We show power curves in Figure 2.3 under scenarios $1a$, $2a$, $3a$ and $4a$ given in Table 2.1. We find that the appropriate total sample sizes for these four scenarios are about 135, 90, 120 and 120, respectively. The total sample sizes for scenarios ending in "b" and "c" resemble that of the corresponding scenarios ending in "a" (results not shown). The result indicates that an snSMART comparing two active treatments to a control is feasible for rare disease studies because the sample size can be controlled at the level of about 100 individuals, and the comparison of DTRs can be performed simultaneously. Specifically, if the difference in the response rates between active treatments and the control is 0.25, the sample size of this snSMART can be as small as 90.

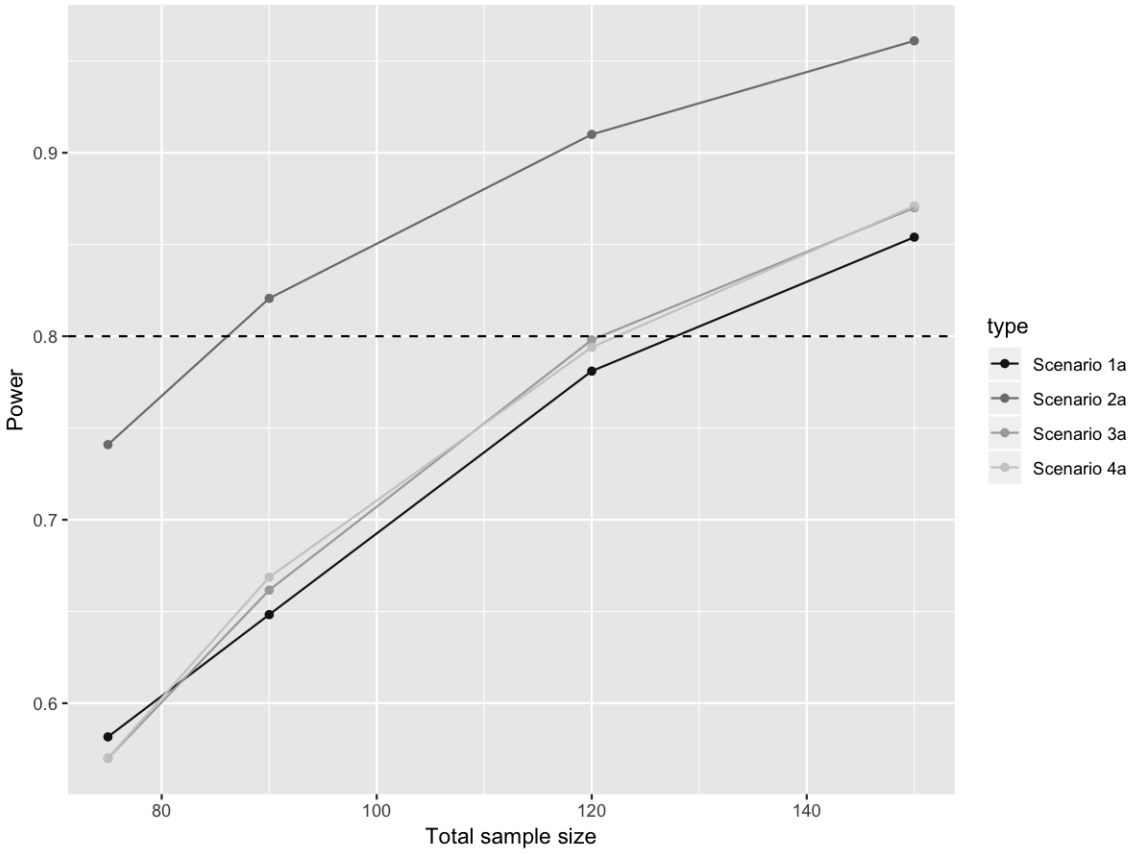

Figure 2.3: Power curve using JSRM with Dunnett's approach. Two pair-wise comparisons (active treatment A vs. standard of care C and active treatment B vs. standard of care C) are performed for each run. Power is estimated by the proportion of runs in which one or both of the p values from the two pair-wise comparisons after Dunnett's correction are smaller than the nominal α.

2.4 Discussion

Using FSGS as an illustrative example, we have outlined how an snSMART design can be implemented to test the efficacy of novel therapies for rare diseases. An snSMART design can address important clinical issues regarding the optimal agent for the disease and the individual patient as well as how treatments can be sequenced and tailored to produce long-term responses. Moreover, an snSMART design maximizes the amount of information that can be learned from each patient and is likely to enhance acceptance of clinical research by patients and their families, and therefore promote participation in clinical trials.

We extended the BJSM and JSRM beyond the work of *Wei et al.* (2018) and focused on estimating both the first stage response rates and DTR effects from models that included six linkage parameters. In general, the estimators of response rates from the BJSM and JSRM are slightly biased relative to estimators from WRRM because the WRRM does not involve any assumptions on the linkage parameters. However, when we consider rMSE, which involves both bias and variance, the BJSM is the best among the all three models under every scenario and sample size. The BJSM and JSRM may be preferred over WRRM for studying the rare diseases under a SMART design because of the general low bias and high efficiency. The estimators of response rates of WRRM are least efficient because only second-stage responses are used.

An assumption of this snSMART design is that the disease of interest should be relatively stable, and that the disease status of individuals does not wax and wane dramatically if there is no change in intervention. In some diseases where this assumption might be violated, this snSMART design, as well as other multi-stage designs, may not be appropriate because the observed outcomes from an individual might reflect the random fluctuation of the disease status rather than the actual treatment effects.

Future work includes improving the BJSM and JSRM to include baseline and time-varying covariates. The WRRM can include baseline and/or intermediate variables to potentially improve the efficiency of the estimated effects. However, the BJSM does not easily lend itself to controlling for covariates, and the JSRM model can only control for baseline measures. Future research will focus on applying precision medicine in the Bayesian and JSRM methods so that variables, such as age, sex or adherence to the initial treatment, can be successfully incorporated into our models.

Bayesian analysis has been a recommended approach for trials in the rare disease setting since the analysis incorporates prior information, more can be gained from

smaller sample sizes. Bayesian analysis, however, requires a shift in the expectations of results such that p-values are not generated at the end of a Bayesian analysis. The analysis instead can provide estimates of the response rates, credible intervals (similar to confidence intervals) and probabilities that the treatments differ in their efficacy (e.g., the probability that the standard of care results in 20% less efficacy than the anti-fibrotic therapy is 90%). The results from Bayesian analyses are often more intuitive and interpretable than frequentist results.

CHAPTER III

A Bayesian Group Sequential Small n, Sequential, Multiple-Assignment, Randomized Trial

3.1 Introduction

As an alternative to a traditional trial design, a small sample (n), sequential, multiple assignment, randomized trial (snSMART) can be used for efficient estimation of treatment effects in rare diseases (*Tamura et al.*, 2016). An snSMART is a multi-stage design where participants can be re-randomized at an interim timepoint based on their responses to initial treatment. A Randomized Multicenter Study for Isolated Skin Vasculitis (ARAMIS) is an ongoing snSMART of 90 participants designed to compare the effects of three active treatments for skin vasculitis (NCT02939573), and the motivating design for our proposed methods.

In contrast, a traditional sequential, multiple assignment, randomized trial (SMART), first proposed by *Lavori and Dawson* (2000) and *Murphy* (2005a), is a multi-stage design used to evaluate the effects of tailored intervention sequences for treating disease, or dynamic treatment regimens (*Murphy*, 2003, 2005a), with a relatively large number of participants. Thus, although an snSMART may seem similar to a traditional SMART, the two designs differ significantly in both their objective and assumed sample size.

Like traditional clinical trials, investigators may prefer a design that allows for the potential to remove an inferior treatment arm at an interim point during the trial. Adapting the snSMART design to allow for removing a treatment arm may also be favorable to participants because they are expected to receive a more effective treatment if the worst treatment is removed during the trial. Currently, no formal group sequential methods exist for an snSMART design, although many such methods exist for more traditional designs.

Frequentist interim analysis methods for clinical trials have been proposed by *Stallard and Todd* (2003), *Stallard and Friede* (2008), and *Magirr et al.* (2012). However, those methods assume that the study has a control arm, and any treatment that is not superior to the control is removed. However, in our motivating snSMART design, there is no control arm, but rather three active treatment arms. *Shih and Lavori* (2013) did propose an alternative method in which they determine the current observed best treatment at each interim analysis, and all treatments shown to be inferior to the current best treatment are removed.

Bayesian approaches also exist for group sequential designs. *Rosner and Berry* (1995) focused on the posterior distribution of the difference in the treatment response rates to determine superiority at each interim analysis. However, they artificially divided their four treatments into two groups and performed two within-pair comparisons and one between-pair comparison, which is a limitation for application to a more general scenario of comparing multiple treatments. *Yin et al.* (2012) used the posterior predictive probability of treatment difference to decide early stopping boundaries in their Bayesian group sequential design. However, similar to many of the frequentist methods, *Yin et al.* (2012) also selected one treatment as the standard to which all other treatments were compared. *Zhu et al.* (2017) and *Shi and Yin* (2019) developed methods to control the overall Type I error rate in their Bayesian group sequential test, but only in the scenario of two treatment arms.

In our current work, we propose a Bayesian group sequential design that allows for removal of a worst performing treatment in an snSMART. Similar to a conventional group sequential design, before the start of an snSMART, we specify the number of interim analyses (looks) and the criteria for removing an arm at each interim analysis so that we control the overall probability of removing an arm under the scenario when three treatments have the same response rate. We describe our method in Section 3.2 and demonstrate the results of our approach via simulation in Section 3.3. We close with a discussion in Section 3.4.

3.2 Design

3.2.1 Standard snSMART design

3.2.1.1 General setup

The two-stage design of our motivating trial ARAMIS is shown in Figure 3.1(a); the original design had no interim analyses. In stage 1, participants are randomized equally to one of the three active treatments and then followed for six months, during which response to treatment may occur. In stage 2, stage 1 responders continue with the same treatment, while non-responders are re-randomized to one of the other two treatments that they did not initially receive. Participants are then followed for an additional six months for the occurrence of response to treatment. The length of stage 1 is the same as the length of stage 2, and stage 2 begins immediately after stage 1 ends. We emphasize that the term "stage" refers to the fixed period of time from a participant's receipt of a treatment to the end of their follow-up for response to that treatment.

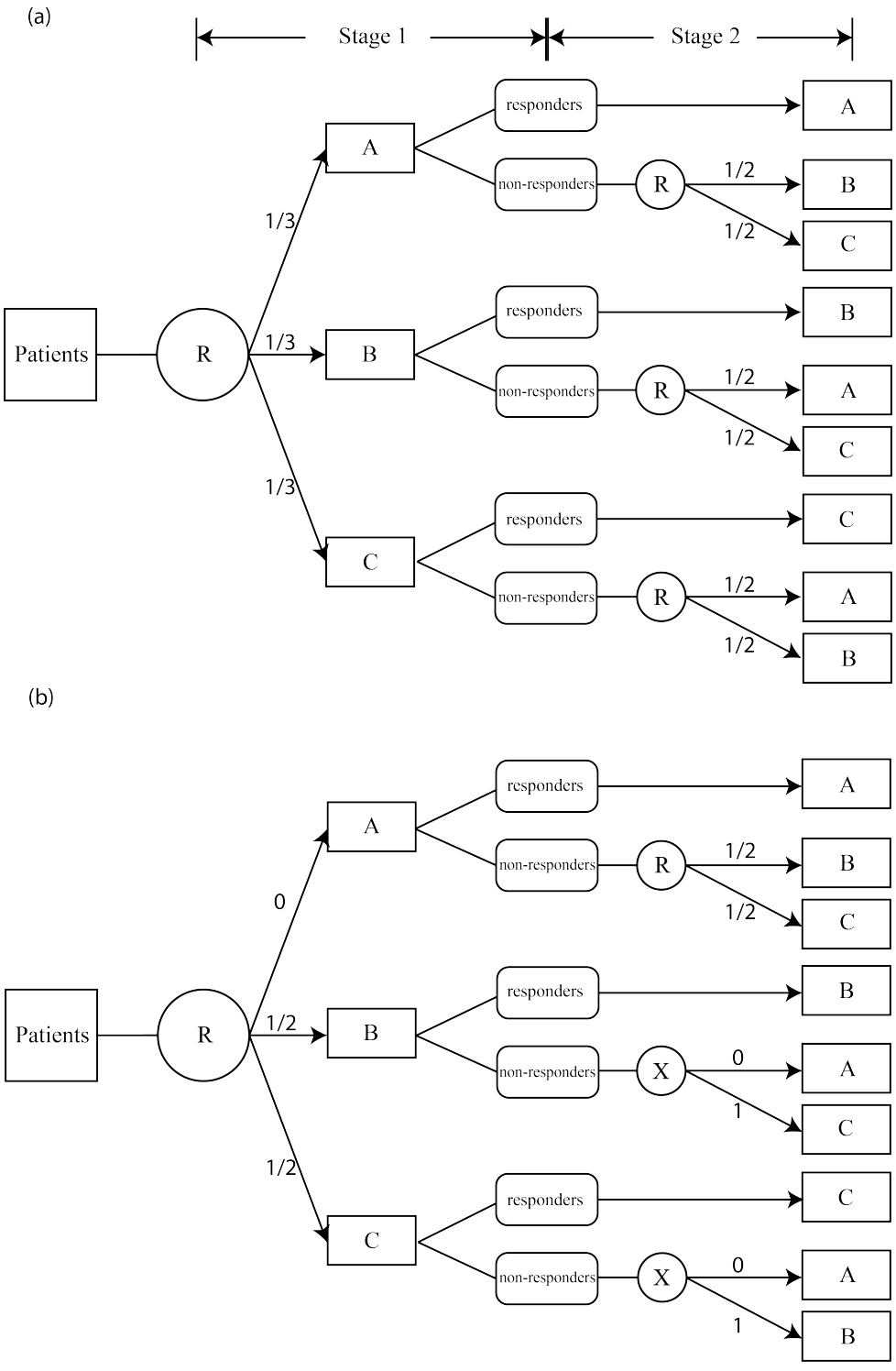

Figure 3.1: (a) A group sequential small n sequential multiple assignment randomized trial (snSMART) design before an arm being removed, which is also an snSMART design without interim analysis. (b) A group sequential snSMART design after treatment A is removed. The numbers around the arrows indicate the probabilities that a participant is assigned to the treatment. R represents randomization to the following treatments. X represents deterministic assignment to the following treatment.

3.2.1.2 Bayesian Joint Stage Model (BJSM) for an snSMART

Wei et al. (2018) developed a Bayesian Joint Stage Model (BJSM) to estimate the response rates of three treatments in an snSMART with binary outcomes. We briefly present the BJSM here because it is used in both the decision rule mentioned in Section 3.2.2.2 and the estimation of response rates at the end of a trial. For participant $i = 1, 2, \ldots, N$, where N is the number of participants, treatment $j = A, B, C$, and stage $k = 1, 2$, we let Y_{ik}^{j} be an indicator of response for participant i receiving treatment j in stage k. The stage 1 response rate to treatment j is denoted by π_j.

We then let $\beta_{1j}\pi_j$ denote the stage 2 response rate of the stage 1 responders to treatment j, with the assumption that $\beta_{1j} > 1$, so that if a participant responds in stage 1, they are at least as likely to respond again to the same treatment in stage 2. For stage 1 non-responders to treatment j, we let $\beta_{0j}\pi_{j'}$ denote the reponse rate to treatment j' in stage 2, with the assumption that $\beta_{0j} < 1$, i.e. stage 1 non-responders are less likely to respond to either of the two other treatments in stage 2. *Wei et al.* (2018) referred to β_{1j} and β_{0j} as linkage parameters because they link the stage 1 response rates to the stage 2 response rates.

The Bayesian Joint Stage Model (BJSM) estimates the response rates of three treatments as follows:

$$Y_{i1}^{j} | \pi_j \sim \text{Bernoulli}(\pi_j) \tag{3.1}$$

$$Y_{i2}^{j'} | Y_{i1}^{j}, \pi_j, \pi_{j'}, \beta_{1j}, \beta_{0j} \sim \text{Bernoulli}\Big((\beta_{1j}\pi_j)^{Y_{i1}^{j}}(\beta_{0j}\pi_{j'})^{1-Y_{i1}^{j}}\Big) \tag{3.2}$$

$$\pi_j \sim \text{Beta}(\theta_1, \delta_1) \tag{3.3}$$

$$\beta_{0j} \sim \text{Beta}(\theta_2, \delta_2) \tag{3.4}$$

$$\beta_{1j} \sim \text{Pareto}(1, c) \tag{3.5}$$

Beta priors are used for π_j and β_{0j} because we assume that they range from 0 to 1, while Pareto$(1, c)$ is used for β_{1j} because it requires $\beta_{1j} > 1$. For more details about the specification of hyperparameters, see *Wei et al.* (2018). The response rate for each treatment is estimated from the posterior distribution of π_j using Markov Chain Monte Carlo (MCMC).

3.2.2 Group Sequential snSMART

3.2.2.1 General Setup

In stage 1, randomization will assign equal numbers of participants to each treatment; in contrast, the number of participants assigned to each treatment in stage 2 will depend upon the proportion of responders in stage 1. Thus, even without interim analyses, more participants are expected to receive the better treatments in an snSMART. We now wish to determine if we can further increase the number of participants assigned to the better treatments if we allow for the removal of an inferior arm.

In a group sequential snSMART, treatment effects are estimated and compared at each interim analysis (or look) $l = 1, 2, \ldots, L$, where L is the maximum number of interim analyses performed during a trial. Here we will assume that $L = 2$ so that there are at most two looks in the snSMART. If an interim analysis suggests that one treatment is inferior to the others, then the treatment is removed and subsequent participants entering the trial no longer receive the removed treatment. If none of the treatments is considered inferior after look L, all three treatments are kept to the end of the trial. We note that "stage" and "look" are two different concepts in our group sequential snSMART design. Stage refers to a period of time specific to when each *participant* is followed for a response, while "look" refers to a period of time specific to the *entire study* when the accrued data are analyzed in an interim analysis.

If an interim analysis suggests removal of a treatment, the trial continues such that stage 1 non-responders to that inferior treatment are randomized equally to the two non-inferior treatments, while stage 1 non-responders to each of the non-inferior treatments are deterministically switched to the non-inferior treatment they had not received. In addition, stage 1 responders continue to receive the same treatment in stage 2 regardless of whether or not the treatment has been removed. An example of a two-stage snSMART design after treatment A is removed at look l is demonstrated in Figure 3.1(b).

In order to better describe the process of the trial, we demonstrate an example of a group sequential snSMART with two interim analyses, in Figure 3.2. Here we assume that three participants are enrolled in the trial every month, and recruitment continues for thirty months. The interim analyses are planned after the 30th and 60th patients have completed stage 1. When the stage 1 outcome from the 30th participant is collected (marked by the first dashed box at month 16 in Figure 3.2), the first look occurs and response rates are estimated using the BJSM, and consideration of removing a treatment is based on the decision rule presented in Figure 3.3, the details for which are found in Section 3.2.2.2. We note that the stage 2 outcomes from some early participants are available for model fitting when the interim analysis is conducted, but not all participants will have stage 2 outcomes.

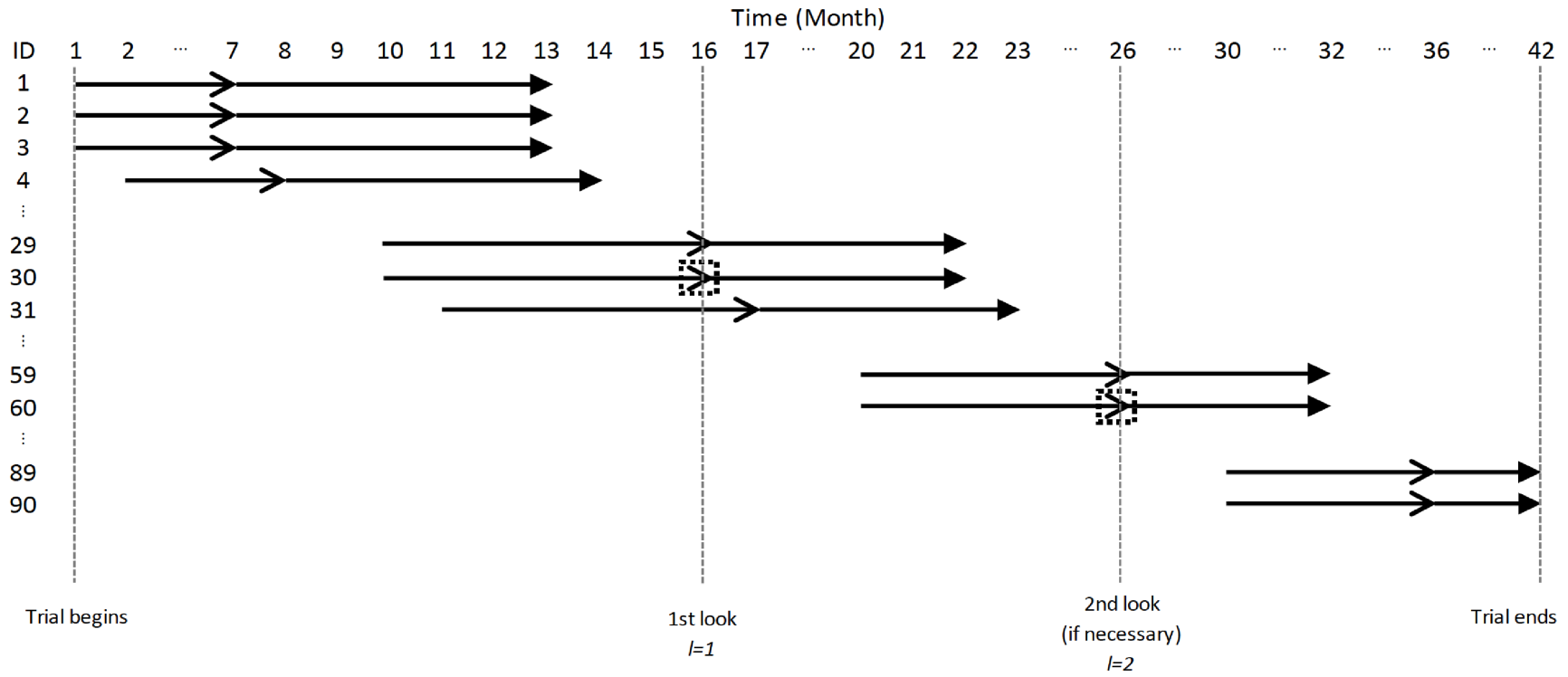

Figure 3.2: An illustration of how a group sequential snSMART with at most two interim analyses (looks) proceeds. Each row is an ID of a participant, and each column is the number of months after the trial begins. The participants are enrolled in the study at the rate of 3 people per month. Enrollment takes 30 months in total. The outcomes in each stage can be obtained from participants six months after the treatment assignment and the second stage treatments are assigned to participants immediately after their first stage outcomes are obtained. $\longrightarrow$ shows the time duration when the participants are in the first stage, and $\longrightarrow$ shows the time duration when the participants are in the second stage. Two dashed boxes indicate the events when interim analyses are conducted. Although the arrows of some participants may be aligned at the same start and end points, it represents that they start and end in the same months, not necessarily the same days.

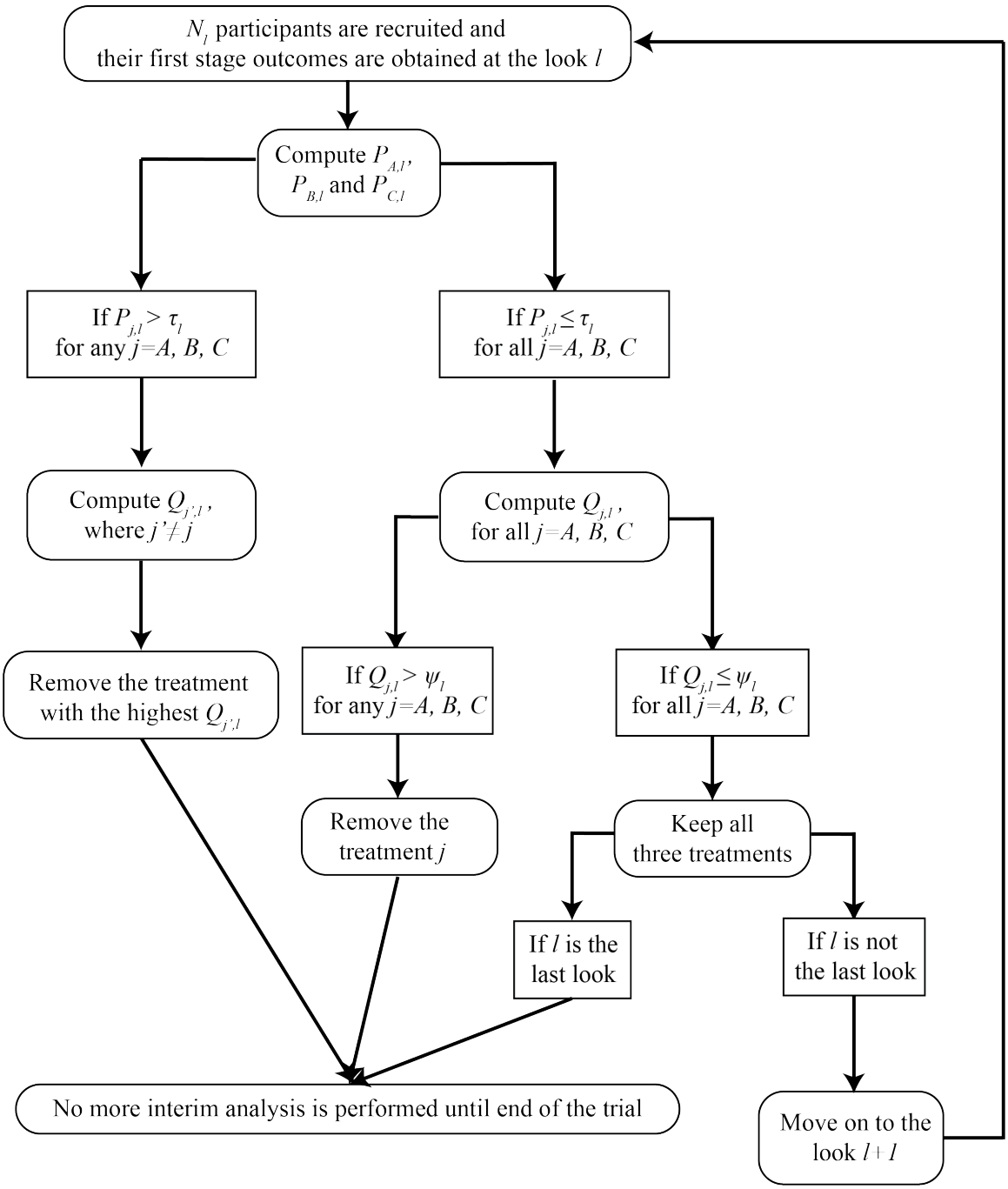

Figure 3.3: The detailed procedure of the proposed two-step Bayesian decision rule performed at an interim analysis l. If an one-step rule is applied, then the procedure starts from computing $Q_{j,l}$, $j = A, B, C$.

If a treatment is removed at the first look, the second look would not occur. If no arm is removed at the first look, the second look would occur when the stage 1 outcome

from the 60th participant is collected (marked by the second dashed box at month 26 in Figure 3.2). At this point, whether an arm is removed depends on the result from the BJSM and the decision rule, but no more looks would be conducted until the final data analysis at the end of the trial. After the trial ends, we apply the BJSM to estimate the response rates of the three treatments using the stage 1 and stage 2 response indicators from all participants. Note that if the trial had been designed with only one look, that look could be conducted when the stage 1 outcome from the 45th participant was collected.

3.2.2.2 Bayesian Decision Rules

To consider the removal of a treatment arm, we introduce a two-step decision rule based on the posterior distributions of the response rates at each interim look l. The sample size for each look l is N_l, which is a cumulative number of all the accrued participants until look l, and the total sample size for the snSMART is denoted by N_T. In our design, an equal number of participants is accrued between looks, i.e., $N_l - N_{l-1} = N_T/(L+1)$. At each look, the BJSM is able to produce posterior draws of the response rates of all treatments even though stage 2 outcomes may be missing from some participants. In this case, the participants that provide Y_{i2}^j are a subset of the participants that provide Y_{i1}^j.

We let $P_{j,l} = P_l(\pi_j > \pi_{j'} \text{ for all } j' \neq j|\text{Data}_l)$ denote the interim posterior probability that treatment j has the *greatest* response rate given the data up to the look l, and the posterior probability $Q_{j,l} = P_l(\pi_j < \pi_{j'} \text{ for all } j' \neq j|\text{Data}_l)$ denote the interim posterior probability that treatment j has the *smallest* response rate given the data up to the look l, where Data_l are all available Y_{i1}^j and Y_{i2}^j for all $j = A, B, C$ at look l. The first step of the decision rule is based on $P_{j,l}$ and the second step is based upon $Q_{j,l}$, conditional upon the value of $P_{j,l}$. A visual presentation of the detailed two-step

decision rule is shown in the Figure 3.3.

Specific steps are:

1. For each treatment $j = A, B, C$, compute $P_{j,l}$ and compare to the pre-specified cutoff τ_l.

2. (a) If $P_{j,l} > \tau_l$ for *any* of the $j = A, B, C$, then compute $Q_{j',l}$ for treatments $j' \neq j$ and remove the treatment with higher $Q_{j',l}$.

   (b) If $P_{j,l} \leq \tau_l$ for *all* $j = A, B, C$, then compute $Q_{j,l}$ for all j and compare the posterior probability $Q_{j,l}$ with the pre-specified cutoff ψ_l. If $Q_{j,l} > \psi_l$ for any of the $j = A, B, C$, then remove treatment j. Otherwise, keep all three treatments.

Our two-step approach is quite intuitive. If enough evidence shows that one treatment is best (Step 2(a)), then one of the two inferior treatments should be removed. Similarly, if no single best treatment is identified, but there is enough evidence that one treatment is worst (Step 2(b)), then the worst treatment should be removed. Since we want to guarantee that at least two treatments remain until the end of the trial, at most one treatment can be removed at an interim analysis, after which, no more interim analyses would be conducted. Thus, when we refer to a design with L looks in the following sections, we mean that at most L looks may take place. If a treatment arm is removed at an early look, the total number of looks may be smaller than L.

The thresholds τ_l and ψ_l used in Steps 1 and 2 can be selected by a user through a grid search as follows. First, consider a "null" setting in which all three treatments have the same response rate ($\pi_A = \pi_B = \pi_C$). If we let α_l denote the probability of incorrectly removing an arm from the trial at look l, the overall probability of making such an incorrect decision during the trial is equal to $\alpha = \sum_{l=1}^{L} \alpha_l$. Thus, for a pre-defined value of α, we recommend assigning the same values to each τ_l and to

each ψ_l in a range from 0.98 to 0.80 with a step size of 0.02. Simulations are then run with these pre-assigned τ_l and ψ_l under the "null" scenario and the resulting value of α is recorded to obtain an approximate range of values assigned to τ_l and ψ_l that all result in our pre-specified α. We can then apply these values to new "non-null" settings in which all three treatments do not have the same response rates to assess the probability that an inferior arm is now correctly dropped.

Without loss of generality, we assume that $\pi_A \leq \pi_B \leq \pi_C$. There are four possible scenarios for the values of these response rates. We describe how our two-step decision rule works in each of these scenarios.

(1) $\pi_A = \pi_B = \pi_C$: $P_{j,l} > \tau_l$ is unlikely to be true for $j = A, B, C$, meaning that none of the arms is superior, then $Q_{j,l} > \psi_l$ is also unlikely to be true. The rule results in keeping all three arms.

(2) $\pi_A < \pi_B = \pi_C$: $P_{j,l} > \tau_l$ is unlikely to be true because $P_{B,l}$ and $P_{C,l}$ should be close, but $Q_{A,l} > \psi_l$ is likely to be true. The rule results in removing arm A.

(3) $\pi_A = \pi_B < \pi_C$: $P_{C,l} > \tau_l$ is likely to be true. The rule results in removing either arm A or arm B with nearly identical probabilities.

(4) $\pi_A < \pi_B < \pi_C$: $P_{C,l} > \tau_l$ is likely to be true. The rule results in removing arm A more often than arm B because $Q_{A,l} > Q_{B,l}$ is more likely to be true.

Although our decision rule is comprised of two steps, we could modify the rule to only have one step based solely on each $Q_{j,l}$. Specifically, if any of the $Q_{j,l}$ exceeds the pre-specified ψ_l, treatment j should be removed. Thus, in the one-step rule, we only consider inferiority of a treatment, whereas in the two-step rule we also consider superiority of a treatment. We investigate the operating characteristics of group sequential snSMARTs with both one-step and two-step decision rules in Section 3.3.2.

3.2.2.3 Estimation of treatment effects under the decision rule

In an snSMART without interim analyses, response rates are estimated by pooling the first and second stage outcomes using the BJSM. We will show that due to the sequential randomization, each response rate obtained from the BJSM is an unbiased estimate of the true treatment response rate. In our group sequential snSMART, it is possible that stage 2 randomization is not conducted for some first stage responders because one treatment arm is removed. We now justify that an unbiased estimate of the response rate can be obtained even when the second stage treatment allocation is deterministic for some non-responders.

To distinguish from the observed first and second stage outcomes Y_1^j and $Y_2^{j'}$ (subscript i is omitted here for simplicity), respectively, we denote the counterfactual outcomes for first stage treatment j and second stage treatment j' by $Y_1(j)$ and $Y_2(j, j')$. We also denote the first and second stage treatment assignments by J_1 and J_2. Under the consistency assumption, the individual with observed treatment $J_1 = j$ or $(J_1, J_2) = (j, j')$ has the observed outcomes Y_1^j and $Y_2^{j'}$ equal to his counterfactual outcomes $Y_1(j)$ and $Y_2(j, j')$. In addition, randomization guarantees that the assignment of treatment is independent of the counterfactual outcomes, or $J_1 \perp Y_1(j)$, $J_1 \perp Y_2(j, j')$ and $J_2 \perp Y_2(j, j')$. For the first stage outcomes, under the consistency assumption and randomization:

$$\begin{aligned} P(Y_1^j = 1 | J_1 = j) &= P(Y_1(j) = 1 | J_1 = j) \quad \text{(consistency)} \\ &= P(Y_1(j) = 1) \quad \text{(first stage randomization)} \\ &= \pi_j \end{aligned}$$

The observed response rate of participants who did not respond to j in the first stage and receive j' in the second stage can be expressed by $P(Y_2^{j'} = 1 | J_1 = j, Y_1^j = 0, J_2 =$

j'). Thus, under the consistency assumption and randomization:

$$
\begin{aligned}
P(Y_2^{j'} = 1 | J_1 = j, Y_1^j = 0, J_2 = j') &= P(Y_2(j, j') = 1 | J_1 = j, Y_1^j = 0, J_2 = j') \quad \text{(consistency)} \\
&= P(Y_2(j, j') = 1 | J_1 = j, Y_1^j = 0) \quad \text{(second stage randomization)} \\
&= P(Y_2(j, j') = 1 | J_1 = j, Y_1(j) = 0) \quad \text{(consistency)} \\
&= P(Y_2(j, j') = 1 | Y_1(j) = 0) \quad \text{(first stage randomization)} \\
&= \beta_{0j}\pi_{j'}
\end{aligned}
$$

The relationship of observed and true second stage response rates for first stage responders to treatment j can be derived using a similar approach. Thus, valid inference can be made for π_j with the observed response rates from both stages using the BJSM in an snSMART without interim analysis, meaning that the estimated response rates from BJSM are unbiased.

In a group sequential snSMART, if arm A is removed after an interim analysis, the subsequent participants are not randomized to A, and the non-responders to B (or C) in the first stage are assigned C (or B) in the second stage deterministically (Figure 3.1(b)). The failure to conduct second stage randomization may undermine the above derivation such that $P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0, J_2 = C) \neq P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0)$. However, in this specific case, we see that the condition "$J_2 = C$" is equivalent to the condition "$J_1 = B$ and $Y_1^B = 0$", and this idea can be generalized to situations where other second stage response rates are of interest. Thus, $P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0, J_2 = C) = P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0)$ is valid for group sequential snSMART even if the second stage randomization does not occur for some first stage non-responders, leading to the conclusion that the second stage response rate of C obtained from the observed outcomes, $P(Y_2^C = 1 | J_1 = B, Y_1^B = 0, J_2 = C)$ is still an unbiased estimate of the true second stage response rate, $\beta_{0B}\pi_C$.

3.3 Simulation

3.3.1 Data generation

We conducted simulation studies to examine the impact of interim analyses in an snSMART in four specific scenarios: (1) $\pi_A = \pi_B = \pi_C = 0.25$; (2) $\pi_A = 0.25, \pi_B = \pi_C = 0.5$; (3) $\pi_A = \pi_B = 0.25, \pi_C = 0.5$; (4) $\pi_A = 0.25, \pi_B = 0.45, \pi_C = 0.65$. For analysis with the BJSM, we let $\beta_{1A} = \beta_{1B} = \beta_{1C} = 1.5$ and $\beta_{0A} = \beta_{0B} = \beta_{0C} = 0.8$. The prior distributions for π_j, β_{1j} and β_{0j} are Beta$(0.4, 1.6)$, Pareto$(1, 3)$ and Beta$(1.6, 0.4)$, respectively, which have respective prior means of 0.2, 1.5, and 0.8. The hyperparameters of the prior distributions were chosen based on the prior knowledge of the stage 1 and stage 2 treatment effects motivated by ARAMIS.

We examined a group sequential snSMART that uses a maximum of 1 look, one that uses a maximum of 2 looks, as well as a traditional snSMART with no interim analyses. The interim analyses will be based on both the one-step and two-step decision rules described in Section 3.2.2.2. We also examine accrual rates of 2, 3, and 5 participants per month. In all trials, the number of participants was $N_T = 90$ and values for τ_l and ψ_l in the decision rule were chosen such that the probability of dropping a treatment in scenario 1 is close to a pre-specified value of $\alpha = 0.1$.

3.3.2 Simulation results

Table 3.1 presents a summary of the simulations for all four scenarios when three participants accrue each month. In this table we wish to see how operating characteristics first change as a function of the decision rule, and then how they change as a function of the number of interim analyses.

By comparing the top two rows of Table 3.1 to the middle two rows, we find that the probability of correctly removing an arm in scenario 2 is relatively unaffected whether

Looks	Steps	τ_l	ψ_l	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
				P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w
$N_T = 90$															
1	1	NA	0.90	0.10	NA	NA	0.52	1.00	1.00	0.21	1.00	1.00	0.56	1.00	0.99
		NA	0.89	0.10	NA	NA	0.55	1.00	1.00	0.22	1.00	1.00	0.58	1.00	0.99
1	2	0.91	0.95	0.10	NA	NA	0.46	0.96	0.96	0.54	1.00	1.00	0.78	1.00	0.93
		0.95	0.91	0.10	NA	NA	0.55	0.97	0.97	0.47	1.00	1.00	0.78	1.00	0.94
2	2	0.96, 0.96	0.96, 0.96	0.10	NA	NA	0.57	0.97	0.97	0.55	0.99	0.99	0.80	1.00	0.91
		0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.61	0.97	0.97	0.60	1.00	1.00	0.84	1.00	0.91

Table 3.1: The proportion of runs that drop an arm (P_{drop}), the proportion of not dropping the best treatment if an arm is dropped ($1-P_b$), and the proportion of dropping the worst treatment if an arm is dropped (P_w) for all four scenarios listed in Section 3.3.2 with different type of dropping rule (one-step or two-step), different number of interim analyses (one look or two looks) and dropping threshold. Accrual rate is 3 people/ month and accrual time is 30 months for a total of 90 participants. For each case, 1000 runs are conducted.

one step or two steps are used in the decision rule. However, in scenarios 3 and 4, we see that the two-step rule performs better than the one-step rule, with an increase of 20-30 percentage points in the probability of removing a treatment arm. We note that this observed difference in probability of correctly removing a treatment arm increases as N_T increases (data not shown). Thus, a two-step rule is preferred to a one-step rule.

Next, we compare the middle two rows of Table 3.1 to the bottom two rows to assess the impact of moving from one interim analysis to two interim analyses. In all of scenarios 2, 3, and 4, we see that the probability of correctly removing a treatment arm increases when two interim analyses are performed relative to one interim analysis. When $N_T = 300$ (data not shown), the benefit of two interim analyses is no longer apparent, mostly because with such a large sample size, the probability of correctly removing a treatment arm with one look already reaches 0.95.

In Figure 3.4, we assess how interim analyses impact the number of stage 2 participants assigned to the best treatment in a group sequential snSMART. The height of each bar represents the ratio of the number of participants assigned to each treatment relative to the number of participants that would occur in an snSMART without interim analyses. In scenario 1, we see bar heights close to 1.0, indicating that interim analyses have little impact on patient allocation, relative to no interim analyses, because all three response rates are equal.

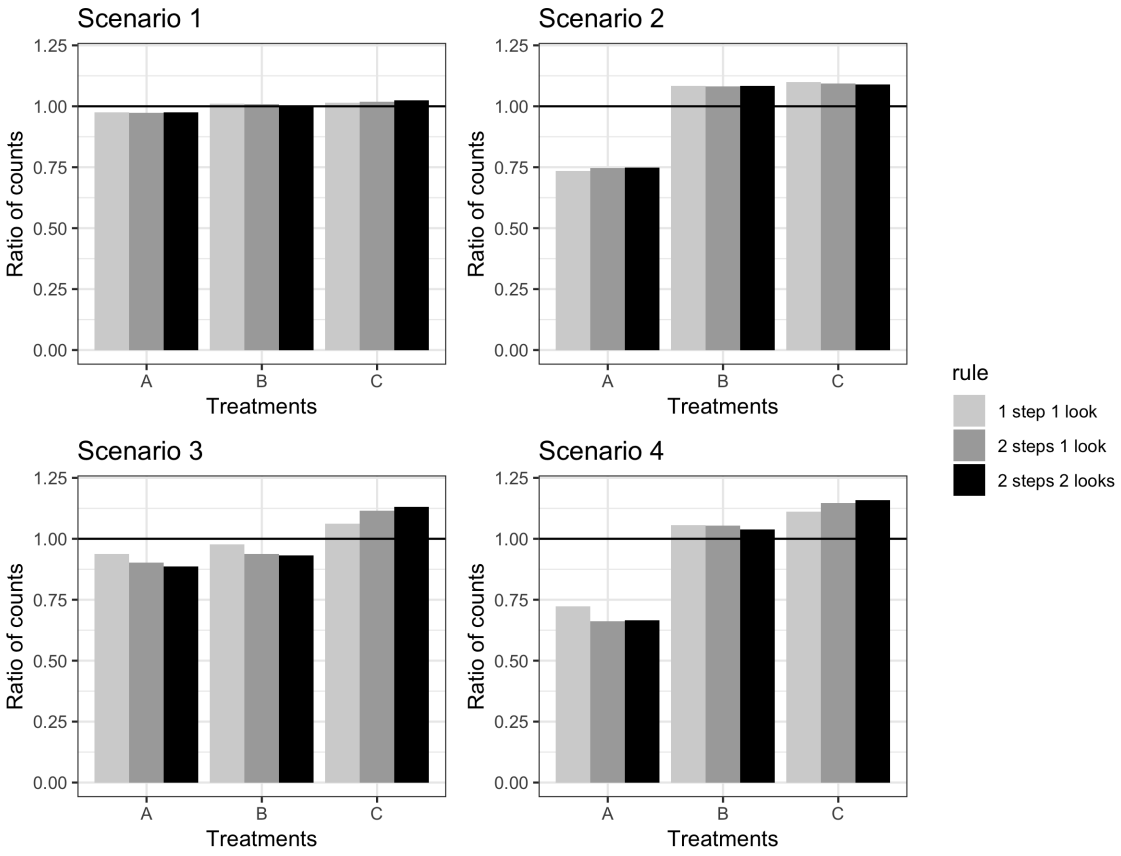

Figure 3.4: The ratio of the second stage participant count under a group sequential snSMART with the given rule (one-step or two-step) and number of maximum interim analyses (one look or two looks) to the second stage participant count under an snSMART without interim analyses. The four scenarios are listed in the Section 3.3.2. The total number of participants on trial $N_T = 90$.

In scenarios 2, 3 and 4, we see bars with heights greater than 1.0 corresponding to treatments with the highest response rate and bars with heights less than 1.0 for treatments with the lowest response rate. This indicates that including interim analyses leads to assigning more participants to the better performing treatments compared to the snSMART without interim analyses. Furthermore, the ratio for the best treatment is highest when the two-step decision rule is used with two interim analyses, which agrees with the pattern of probabilities of correctly removing a treatment arm shown in Table 3.1. We obtained a similar pattern if we focused on the stage 1 participant counts (data not shown). Thus, with regard to participant assignment,

a two-step decision rule with two interim analyses is preferred for all scenarios for $N_T = 90$.

In Table 3.2, we assess how interim analyses impact the numbers of responders to each treatment in each scenario. In scenario 1, since all response rates are equal, there are almost equal numbers of participants responding to each treatment. However, in scenarios 2, 3 and 4, we see that incorporating interim analyses leads to more responders to the treatments with higher response rates. Most importantly, when the response rates of three treatments are not equal, a group sequential design has more responders than that of a design without interim analyses. Together with the result in Figure 3.4, we conclude that group sequential snSMARTs allocate more participants to the better treatment, and more participants can benefit from their assigned treatment.

In Figure 3.5, we assess the impact of interim analyses on the bias and root mean-squared error (rMSE) of the response rates using the BJSM. We focus solely on a design with two interim analyses that use the two-step decision rule, as that design was seen to be best in terms of patient assignment. In general, the interim analysis does appear to lead to a slightly higher bias, but the overall biases still remain small compared to the true response rates. We note that the bias corresponding to the worst treatment can be higher than the bias of the other treatments, which is expected because fewer participants are assigned to the worst treatment. As with bias, rMSE is impacted to a small degree when interim analyses are incoporated in the design. Although there is a little impact on the rMSE of the best treatment, the efficiency corresponding to the worst treatment is compromised in the group sequential snSMART, again because fewer participants are assigned to this treatment when interim analyses are used. Furthermore, the conditional bias using only the simulations where a treatment arm was removed increased slightly in the scenarios where (1) P_{drop} was small or (2) the response rates of a treatment was small (results

Looks	Steps	τ_l	ψ_l	Treatment	Mean number of treatment responders in stage 2			
					Scenario 1	Scenario 2	Scenario 3	Scenario 4
				$N_T = 90$				
NA	NA	NA	NA	A	7.52	7.53	7.52	7.53
				B	7.41	14.89	7.42	13.39
				C	7.50	15.02	15.01	19.51
				Total	22.43	37.43	29.95	40.43
1	1	NA	0.89	A	7.29	4.48	6.26	4.17
				B	7.25	20.35	6.46	15.96
				C	7.24	20.28	21.10	31.86
				Total	21.78	45.11	33.81	51.99
2	1	0.95	0.91	A	7.27	4.55	6.07	3.83
				B	7.24	20.31	6.26	16.00
				C	7.28	20.14	22.08	32.84
				Total	21.79	45.01	34.40	52.67
2	2	0.96, 0.95	0.96, 0.95	A	7.27	4.50	5.86	3.78
				B	7.27	20.40	6.30	15.86
				C	7.34	19.96	22.20	33.21
				Total	21.88	44.85	34.36	52.85

Table 3.2: The average numbers of responders to the treatments in the second stage of a standard snSMART (snSMART without interim analyses) or a group sequential snSMART with the given type of rule (one-step or two-step), for a given number of interim analyses (one look or two looks) under all four scenarios listed in Section 3.3.2. The mean numbers of responders to each treatment and all treatments are listed for each design under each scenario.

not shown). This increase is expected because these biases were calculated using the results from fewer simulations and/or fewer participants assigned to a treatment. When neither of the above conditions was true, the conditional bias was almost as small as the marginal bias shown in Figure 3.5.

In Table 3.3, we examine how the probability of correctly removing a treatment is impacted by the accrual rate, as faster (slower) accrual implies a higher (lower) proportion of participants who have not completed stage 2 by the time of the interim analysis. The top two rows of Table 3.3 summarize when accrual is faster (5 participants/month), the middle two rows are the original accrual (3 participants/month), and the bottom two rows correspond to slower accrual (2 participants/month).

In scenarios 2, 3, and 4, we see generally as the accrual rate increases, there is a decrease in the probability of correctly removing a treatment arm, which is likely due to the increasing proportion of missing stage 2 outcomes. Correspondingly, when the accrual rate is slower, more stage 2 outcomes from participants can be collected for model fitting and there is an increase in the probability of correctly removing a treatment arm. Nonetheless, although the slower accrual rate leads to a slightly higher probability of correctly removing a treatment arm, the slower accrual rate also leads to a longer trial. Certainly the accrual rate will vary with the rarity of the disease and the number of sites that recruit participants, but overall, we expect that realistic rates of accrual will only slightly affect the probabilities of correctly removing a treatment arm.

3.4 Discussion

We provide a framework for incorporating interim analyses into an snSMART to potentially remove one of three treatment arms. With the proposed two-step Bayesian decision rule, a group sequential snSMART with two interim analyses may be more

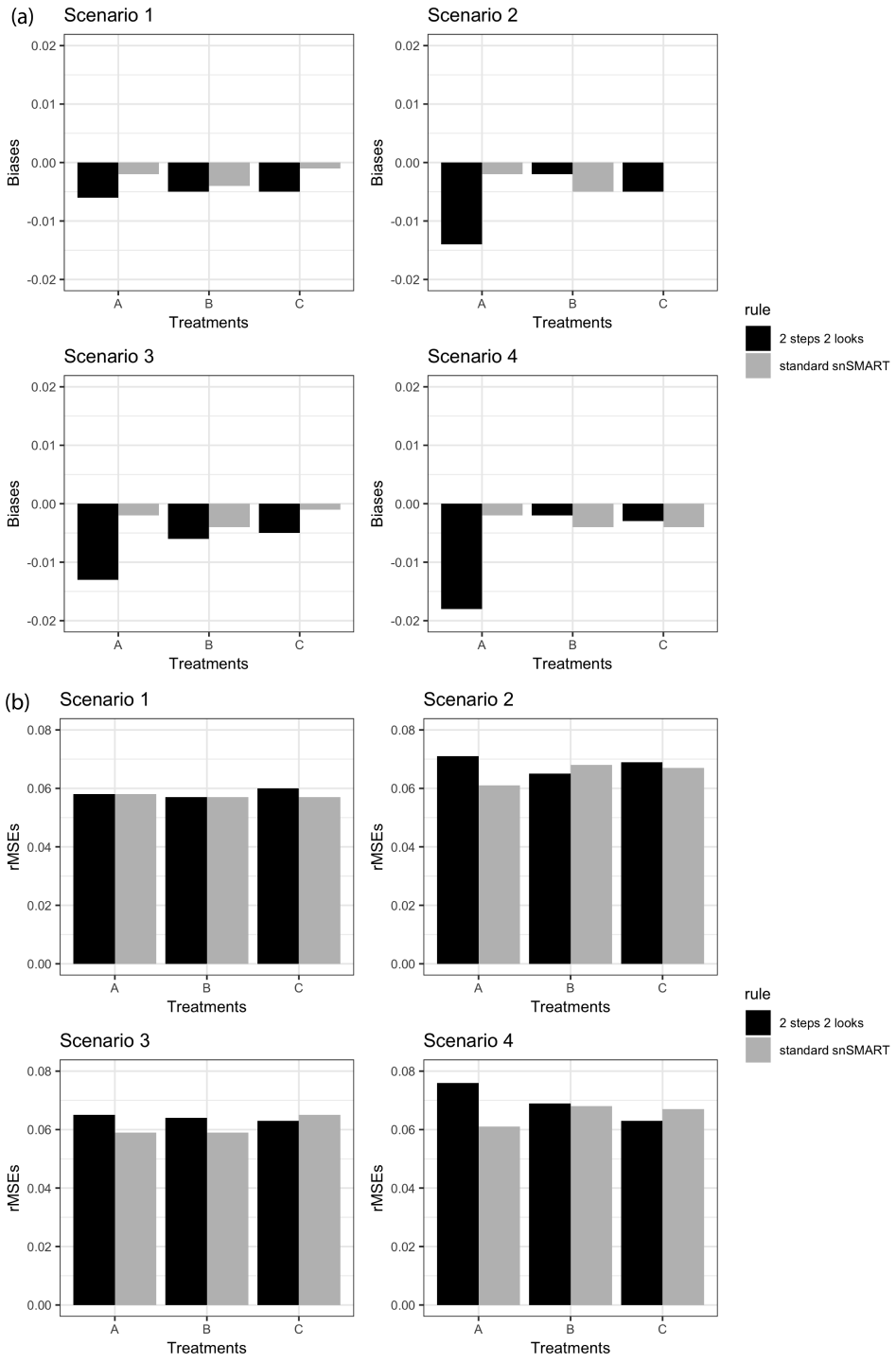

Figure 3.5: (a) The bias of the estimated response rates under the four scenarios listed in the Section 3.3.2. (b) The root mean squared error (rMSE) of the estimated response rates under the same four scenarios. "2 steps 2 looks" means the group sequential snSMART design using the two-step decision rule with at most two looks, and the "standard snSMART" means the snSMART without interim analyses. The total number of participants on trial $N_T = 90$.

Rate × Month	τ_l	ψ_l	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
			P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w
5×18	0.95, 0.96	0.95, 0.96	0.10	NA	NA	0.51	0.96	0.96	0.48	1.00	1.00	0.74	1.00	0.90
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.54	0.97	0.97	0.53	1.00	1.00	0.77	1.00	0.90
3×30	0.96, 0.96	0.96, 0.96	0.10	NA	NA	0.57	0.97	0.97	0.55	0.99	0.99	0.80	1.00	0.91
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.61	0.97	0.97	0.60	1.00	1.00	0.84	1.00	0.91
2×45	0.95, 0.96	0.95, 0.96	0.10	NA	NA	0.59	0.98	0.98	0.59	1.00	1.00	0.81	1.00	0.92
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.62	0.98	0.98	0.63	1.00	1.00	0.85	1.00	0.92

Table 3.3: The proportion of runs that drop an arm (P_{drop}), the proportion of not the best treatment if an arm is dropped ($1-P_b$), and the proportion of dropping the worst treatment if an arm is dropped (P_w) for all four scenarios listed in Section 3.3.2 with different accrual rates and times, but same total sample sizes ($N_T = 90$) and same two-step rule and number of interim analyses (two looks). For each case, 1000 runs are conducted.

appealing to both those designing the trial and those participating in the trial. In a group sequential snSMART, fewer participants are expected to receive the worst treatment and the estimation of the response rate of the best treatment is not compromised relative to an snSMART without interim analyses. Similar to traditional group sequential designs, we can control the overall probability of removing an arm under a "null" scenario when three response rates are equal by using simulations to determine the values used for the cutoff values in the decision rule.

Our group sequential snSMART design can be used more flexibly in real practice. First, the proposed decision rule can be extended if there are interactions between stage 1 and 2 treatments that vary depending upon which treatments are used. Second, we assumed that interim analyses were performed when stage 1 outcomes were collected from a fixed number of participants at equal intervals. Instead, we can easily adjust the design to accommodate interim analyses at any interval of time. Third, the prior distributions of the response rates and linkage parameters can also be changed to reflect prior beliefs in the treatment response rates and linkage parameters. We assumed a Pareto distribution for the linkage parameters β_1 because we believed that responders were more likely to respond again in stage 2 had they already responded in the stage 1. However, we can change this prior distribution to a gamma or log-normal distribution, which ranges from 0 to infinity, under different assumptions for the responders. Similarly, the other prior distributions and their hyperparameters could differ given the specific trial setting. Based upon other simulations (results not shown), even if the prior distributions are centered away from the true parameter values, estimation of the response rates shows little bias.

We note that in a traditional group sequential design, the number of interim analyses is often decided by many factors, including the total sample size, the power under the expected treatment effect difference, the effort to carry out interim analyses (*Jennison and Turnbull*, 1999). Practioners can decide an appropriate number of interim

analyses through simulation studies after the total sample size, power under expected treatment effect difference, accrual rate and maximum number of interim analyses are pre-specified in group sequential snSMART designs. In small sample scenarios, such as 90 participants in our simulations, we do not recommend more than two interim analyses. A greater number of interim analyses will not substantially enhance the probability of correctly removing an arm because insufficient information will be available for decision making at the earlier interim analyses. Furthermore, if one wants to remove an arm more quickly when some early evidence of strong inferiority can be identified, then earlier interim analysis would be desired. On the contrary, if one wants to be more conservative about making a decision to remove an arm, a late interim analysis would be preferred.

Choosing the specific values of response rates under scenario 1 is arbitrary as long as the three response rates are equal. In our simulations we chose 0.25 as the "null" response rates for all three treatments because this response rate was considered ineffective across treatments for our setting. Although different response rates for scenario 1 might change the chosen threshold values τ_l and ψ_l, we have found that the small difference in threshold values does not greatly change the operating characteristics of the group sequential snSMART in scenarios 2, 3 and 4 (data not shown). In addition, we investigated simulation studies with different true "null" response rates, where the threshold values were chosen assuming null response rates of 0.25, but true null response rates were 0.35 or 0.45. For both "null" values of 0.35 and 0.45, we found $\alpha = 0.09$, which was very close to the nominal value of 0.10.

The posterior probabilities $Q_{j',l}$ of the two-step decision rule can be equal in extremely rare cases because these two probabilities were computed using the posterior draws from MCMC. For example, in scenario 3, where treatments A and B have the same response rate that is smaller than that of C, it is possible, though very unlikely, that $Q_{A,l}$ and $Q_{B,l}$ are equal at the second step of the decision rule. As a solution, one

could randomly remove one of the two treatments or instead decide not to remove either arm and wait for a later look to make a decision.

Our group sequential snSMART is preferred for rare disease trials or trials where the accrual rate is relatively slow. If patient accrual is much faster than the timing of outcome measurements, most treatment allocations will be completed before interim analyses can be performed. In this case, the removal of a treatment arm will have a very limited effect in allocating patients to potentially better treatments.

Our two-step decision rule is currently only applicable to a three arm trial, where there is a single best or worst treatment if three treatments do not have the same response rate. Thus, future work includes the development of a more general decision rule that can be applied to an snSMART with more than three arms. Moreover, if many arms are compared at the same time, we would like to develop a decision rule that can remove more than one arm.

CHAPTER IV

Power Prior Models for Treatment Effect Estimation in a Small n, Sequential, Multiple-Assignment, Randomized Trial

4.1 Introduction

In rare disease studies, estimating treatment effects efficiently is often a challenging task because information is collected from a relatively small number of participants. Developed to meet this challenge, a small n, sequential, multiple assignment, randomized trial (snSMART) is a two-stage design where participants are given up to two treatments sequentially; whether they receive the same or different treatment in the second stage depends on how they respond to the first stage treatment (*Tamura et al.*, 2016). Primary interest in an snSMART is the first stage treatment effect, but when multiple outcomes are obtained from each participant, a method to combine the information across stages can be used to efficiently estimate the treatment effects of interest.

Frequentist and Bayesian approaches have been proposed to pool the results together for estimation. *Tamura et al.* (2016) presented a weighted Z-statistic to perform the estimation, but the Z-statistic is not based on all the collected data. To address

these limitations, *Wei et al.* (2018) and *Chao et al.* (2020) presented both a Bayesian joint stage model (BJSM) and a joint stage regression model, each of which includes parameters that link first and second stage treatment responses to provide more efficient treatment effect estimates. Here, we present an alternative approach that links data from the two stages through a power prior, which was first proposed by *Ibrahim et al.* (2000).

A power prior contains the likelihood of the historical data, power parameters that quantify the compatibility of the historical and the current data, and prior distributions for the parameters in the likelihood of the current data. The power parameters can be either fixed or random and there are numerous ways the parameters are specified or determined. Extensions of this power prior approach include modified power priors, or normalized power priors (*Duan et al.*, 2006; *Neuenschwander et al.*, 2009; *Hobbs et al.*, 2011; *Banbeta et al.*, 2019; *van Rosmalen et al.*, 2018), power prior in Bayesian hierarchical models (*Chen et al.*, 2006), commensurate power priors (*Hobbs et al.*, 2011; *van Rosmalen et al.*, 2018), power priors with an empirical Bayesian approach (*Gravestock and Held*, 2017) and power priors with a likelihood-based weight selection criterion (*Ibrahim et al.*, 2003, 2015).

Pan et al. (2017) proposed a calibrated power prior that utilizes a nonparametric Kolmogorov-Smirnov statistic to measure the compatibility of historical and current data in biosimilar designs. *Nikolakopoulos et al.* (2018) developed another calibrated power prior that quantifies the conflict of historical to current data through prior predictive p-values. *Li and Yuan* (2020) applied the notion of a power prior model to control information borrowing through Bayesian model averaging between pediatric and adult phase I oncology trials.

In previous studies, the idea of power prior models was applied to control how much information should be borrowed from historical data or earlier trials to a current

trial. However, information sharing is also crucial in a multistage clinical trial, which motivates our work. In this study, we propose a novel application of power prior models to the estimation of treatment effects in an snSMART, which is a two-stage design. In addition, we first introduce novel measures of closeness to describe the compatibility of stage 1 and 2 data in our snSMART. In our setting, we consider stage 1 responses as "current" data and stage 2 responses as "historical" data, which may seem counterintuitive. However, because a second stage outcome is obtained after a first stage outcome, second stage outcomes are conditional on the treatments received in the first stage and response to that first stage treatment. Because of this biased sampling scheme, the second stage outcomes are viewed as supplemental data, and the first stage outcomes are viewed as the primary data, since they are collected in an unbiased, randomized design.

Small sample size is another challenge when applying power prior models to the snSMART setting. In existing designs, the historical data are often assumed to come from a multitude of participants who received the same treatment. In contrast, in an snSMART, it is possible that outcomes will only be obtained from a very small number of participants in the second stage. The operating characteristics of power prior models with small samples has not been investigated before, and thus, we seek to examine their performance in the snSMART setting relative to the existing BJSM.

In our current work, we propose three different power prior models to estimate the response rates of three active treatments in an snSMART. In Section 4.2, we motivate the use of power prior models in snSMART designs and briefly describe the existing BJSM. In Section 4.3, we present the power prior models with different power parameter specification approaches. In Section 4.4, we use simulations to examine how these power prior models perform and compare them to the BJSM under different scenarios, and we close with a discussion in Section 4.5.

4.2 Motivating example and existing methods

4.2.1 ARAMIS trial

Our methods are motivated by the snSMART, A RAndomized Multicenter study for Isolated Skin vasculitis (ARAMIS) (*Micheletti et al.*, 2020), *Wei et al.* (2018) and *Chao et al.* (2020) and shown in Figure 2.1. In brief, all enrolled individuals are randomized to one of the three treatments in the first stage. During a specific period of follow-up of six months, each individual is assessed for a response. The individuals who respond in the first stage receive the same treatment in the second stage, while non-responders in the first stage are randomized to one of the alternative treatments in the second stage and followed for six more months for response.

The first stage is a traditional randomized trial; thus, we can estimate treatment effects using only the first stage data. In the proposed power prior methods, these first stage outcomes are called "current data". By contrast, the second stage outcomes alone could not be used to correctly estimate the response rates because they are conditional on first stage treatment and responses to that treatment. Thus, second stage outcomes serve as "historical" data. Inclusion of "historical" data can provide additional information and increase the efficiency of estimation of treatment effects in small samples. Thus, the application of power prior models to our setting provides a way to incorporate both stages of data such that first stage data are weighted fully, and second stage data receive partial weight through the power prior to provide more efficient treatment estimates in small samples.

4.2.2 Joint stage models

Frequentist and Bayesian joint stage models are existing approaches that estimate the treatment effects in an snSMART, where the details can be found in *Wei et al.*

(2018) and *Chao et al.* (2020). Because the results from both models are similar, we briefly present the BJSM here due to our focus on Bayesian methods.

The (first stage) response rate of a treatment k is denoted by π_k, where $k = A, B, C$. Since the response rate of a treatment in the second stage can differ from that in the first stage, and because stage 2 response rates are conditional on stage 1 treatments and responses, we denote the second stage response rates of the first stage responders to treatment k by $\beta_1\pi_k$, and the second stage response rates of the first stage non-responders to k who receive k' in the second stage by $\beta_0\pi_{k'}$. β_1 and β_0 are called linkage parameters for stage 1 responders and non-responders, respectively, because they link the first stage and second stage response rates. An assumption of the BJSM is that the linkage parameters, β_0 and β_1, do not depend on the first and second stage treatments received. The parameters, π_k, β_1 and β_0, can be estimated via Markov Chain Monte Carlo with appropriate prior distributions on these parameters.

However, we may not have *a priori* information about the possible relationship between first stage and second response rates, particularly in the rare disease settings, which may make it difficult to pre-specify prior distributions of the linkage parameters. Thus, the power prior approaches presented next provide a framework to circumvent the requirement of assuming the proportionality of response rates from the stage 1 to 2.

4.3 Methods

We first briefly review the power prior models and their associated notation. We let $\boldsymbol{\pi} = \{\pi_A, \pi_B, \pi_C\}$, where the elements are the response rates of treatments A, B, and C, respectively, and $\boldsymbol{\delta} = \{\delta_j\}$ denote power parameters for different subgroups of individuals, where $j = 1, \ldots, J$ and J is the number of subgroups. In our design, we separate the second stage data into two distinct sets: those from first stage responders

and those from first stage non-responders. The individuals in these two subgroups are assumed to share some common within-group characteristics that may affect how they respond to the second stage treatments. Thus, each subgroup can be regarded as a distinct set of "historical" data, and we assume that $J = 2$ in this study. We also made this assumption of $J = 2$ because a parsimonious model is preferred when the sample size is small, and two power parameters mimics the two linkage parameters from the BJSM. Let $n_k^{(1)}$ and $Z_k^{(1)}$ denote the number of individuals assigned to treatment k and the corresponding number of responders to k in stage 1, respectively, where $k = A, B, C$. Similarly, we let $n_{k,j}^{(2)}$ and $Z_{k,j}^{(2)}$ be the numbers of individuals in stage 2 assigned to treatment k within subgroup j and the corresponding number of responders to k in subgroup j, respectively. Let $\mathbf{D^{(1)}} = \{n_k^{(1)}, Z_k^{(1)}; k = A, B, C\}$ and $\mathbf{D^{(2)}} = \{n_{k,j}^{(2)}, Z_{k,j}^{(2)}; k = A, B, C; j = 1, \ldots, J\}$.

In its simplest form, the joint power prior distribution of the first stage response rates in our setting can be formulated as

$$p(\boldsymbol{\pi}|\mathbf{D^{(2)}}, \boldsymbol{\delta}) \propto \prod_{k=A,B,C} \left[\prod_{j=1,\ldots,J} L(Z_{k,j}^{(2)}; \pi_k)^{\delta_j} \right] p_0(\pi_k) \tag{4.1}$$

where $L(Z_{k,j}^{(2)}; \pi_k)$ is a likelihood function for second stage outcomes, $p_0(\pi_k)$ is the initial prior for π_k, and $0 \leq \delta_j \leq 1$ for all j. We interpret δ_j as a measure of compatibility of the "current" data and the "historical" data from subgroup j. When $\delta_j = 0$, the corresponding "historical" data, i.e., second stage data, from subgroup j contribute nothing to the estimation of response rates, while $\delta_j = 1$ indicates that the corresponding "historical" data from subgroup j can be pooled together with "current" data. When combining with the likelihood function of first stage outcomes, the posterior distribution of $\boldsymbol{\pi}$ is

$$q(\boldsymbol{\pi}|\mathbf{D^{(1)}}, \mathbf{D^{(2)}}, \boldsymbol{\delta}) \propto \left[\prod_{k=A,B,C} L(Z_k^{(1)}; \pi_k)\right] p(\boldsymbol{\pi}|\mathbf{D^{(2)}}, \boldsymbol{\delta})$$
$$= \prod_{k=A,B,C} \left[L(Z_k^{(1)}; \pi_k) \prod_{j=1,\ldots,J} L(Z_{k,j}^{(2)}; \pi_k)^{\delta_j}\right] p_0(\pi_k) \tag{4.2}$$

The key issue in the application of power prior models lies in the choice of δ_j. Thus, we next introduce three types of approaches for choosing δ_j and investigate to what extent stage 2 data can be incorporated with stage 1 data to estimate π_k.

4.3.1 Power prior models with likelihood-type criteria

The power parameters δ_1 and δ_2 can be taken as fixed values and determined by likelihood-type criteria, which was first proposed by *Ibrahim et al.* (2003) and extended from Bayesian Information Criterion (BIC). The rationale of utilizing likelihood-type criteria is to use both "current" and "historical" data to choose the optimal values for δ_1 and δ_2 that minimize the criteria function. Two criteria applied to power prior models are the penalized likelihood-type criterion (PLC) (*Ibrahim et al.*, 2003, 2015) and the marginal likelihood criterion (MLC) (*Ibrahim et al.*, 2015; *Gravestock and Held*, 2017), the latter of which is also referred to as the empirical Bayesian method.

For the PLC, the "current" and "historical" data are combined in the function

$$m^*(\boldsymbol{\delta}) = \int_{\boldsymbol{\pi}} \prod_k \left[L(Z_k^{(1)}; \pi_k) \prod_j L(Z_{k,j}^{(2)}; \pi_k)^{\delta_j} p_0(\pi_k)\right] d\boldsymbol{\pi}$$
$$= M \prod_k \left\{B\left(Z_k^{(1)} + \sum_j Z_{k,j}^{(2)}\delta_j + a_\pi, n_k^{(1)} - Z_k^{(1)} + \sum_j \left(n_{k,j}^{(2)} - Z_{k,j}^{(2)}\right)\delta_j + b_\pi\right)\right\} \tag{4.3}$$

where M is a constant unrelated to any of the parameters, and $B(\cdot,\cdot)$ is a beta function. The power parameters δ_1 and δ_2 can then be determined by minimizing the PLC function

$$G(\boldsymbol{\delta}) = -2\log\left[m^*(\boldsymbol{\delta})\right] + \sum_j \frac{\log(\sum_k n_{k,j}^{(2)})}{\delta_j}. \tag{4.4}$$

The penalty term $\sum_j[\log(\sum_k n_{k,j}^{(2)})/\delta_j]$ allows for the chosen δ_j being higher when the sample size of subgroup j is larger, which corresponds to more weight applied to a subgroup with a larger sample size. After the optimal $\boldsymbol{\delta}$ is determined by $\boldsymbol{\delta}^{PLC} = \arg\min_{\boldsymbol{\delta}} G(\boldsymbol{\delta})$, we then treat $\boldsymbol{\delta}^{PLC}$ as fixed and use Equation (4.2) to obtain the posterior distribution of all π_k.

For the MLC, we use the marginal likelihood of $\boldsymbol{\delta}$

$$\begin{aligned} m(\boldsymbol{\delta}) &= \frac{\int_{\boldsymbol{\pi}} \prod_k \left[L(Z_k^{(1)};\pi_k) \prod_j L(Z_{k,j}^{(2)};\pi_k)^{\delta_j} p_0(\pi_k) \right] d\boldsymbol{\pi}}{\int_{\boldsymbol{\pi}} \prod_k \left[\prod_j L(Z_{k,j}^{(2)};\pi_k)^{\delta_j} p_0(\pi_k) \right] d\boldsymbol{\pi}} \\ &= M' \frac{\prod_k \left\{ B\left(Z_k^{(1)} + \sum_j Z_{k,j}^{(2)}\delta_j + a_\pi, n_k^{(1)} - Z_k^{(1)} + \sum_j \left(n_{k,j}^{(2)} - Z_{k,j}^{(2)} \right) \delta_j + b_\pi \right) \right\}}{\prod_k \left\{ B\left(\sum_j Z_{k,j}^{(2)}\delta_j + a_\pi, \sum_j \left(n_{k,j}^{(2)} - Z_{k,j}^{(2)} \right) \delta_j + b_\pi \right) \right\}} \end{aligned} \tag{4.5}$$

where M' is a constant unrelated to any of the parameters. Values for the power parameters are determined as $\boldsymbol{\delta}^{MLC} = \arg\min_{\boldsymbol{\delta}}\{-2\log[m(\boldsymbol{\delta})]\}$.

4.3.2 Modified power prior model

The modified power prior (MPP) model proposed by *Duan et al.* (2006) treats δ_1 and δ_2 as random variables; *Banbeta et al.* (2019) applied the MPP to estimate treatment effects that incorporate control arms into a current trial. In our study, the MPP is

given by

$$p_{MPP}(\boldsymbol{\pi}, \boldsymbol{\delta}|\mathbf{D^{(2)}}) = \frac{\left[\prod_k \prod_j L(Z_{k,j}^{(2)}; \pi_k)^{\delta_j}\right] \left[\prod_j p_0(\delta_j)\right] \left[\prod_k p_0(\pi_k)\right]}{C(\boldsymbol{\delta})} \tag{4.6}$$

where

$$C(\boldsymbol{\delta}) = \int_{\boldsymbol{\pi}} \left[\prod_k \prod_j L(Z_{k,j}^{(2)}; \pi_k)^{\delta_j}\right] \left[\prod_k p_0(\pi_k)\right] d\boldsymbol{\pi} \tag{4.7}$$

and $p_0(\delta_j)$ is an initial prior distribution of δ_j. The normalizing constant $C(\boldsymbol{\delta})$ is necessary in the formulation of MPP when δ_j is random to enforce the likelihood principle (*Duan et al.*, 2006; *Banbeta et al.*, 2019).

We assume that $Z_k^{(1)}$ and $Z_{k,j}^{(2)}$ are distributed as Binomial$(n_k^{(1)}, \pi_k)$ and Binomial$(n_{k,j}^{(2)}, \pi_k)$, respectively. The initial prior distributions $p_0(\pi_k)$ and $p_0(\delta_j)$ are Beta(a_π, b_π) and Beta(a_δ, b_δ), respectively. After plugging in these distributions and likelihood functions to Equation (4.6), we can analytically derive the MPP as follows, which is a multi-parameter version of the formula derived in *Banbeta et al.*, 2019:

$$p_{MPP}(\boldsymbol{\pi}, \boldsymbol{\delta}|\mathbf{D^{(2)}}) \propto \frac{\prod_k \left\{\pi_k^{\sum_j Z_{k,j}^{(2)}\delta_j + a_\pi - 1}(1-\pi_k)^{\sum_j \left(n_{k,j}^{(2)} - Z_{k,j}^{(2)}\right)\delta_j + b_\pi - 1}\right\} \left\{\prod_j \frac{\delta_j^{a_\delta - 1}(1-\delta_j)^{b_\delta - 1}}{B(a_\delta, b_\delta)}\right\}}{\prod_k \left\{B\left(\sum_j Z_{k,j}^{(2)}\delta_j + a_\pi, \sum_j \left(n_{k,j}^{(2)} - Z_{k,j}^{(2)}\right)\delta_j + b_\pi\right)\right\}} \tag{4.8}$$

The choice of hyperparameters a_π, b_π, a_δ and b_δ reflects our belief in the response rates of treatments and the compatibility of "current" and "historical" in our snSMART. If we do not have previous knowledge about π_k and δ_j, their prior distribution can be set as Beta(1,1).

4.3.3 Power prior model with closeness measure

In addition to likelihood-based approaches, we can define a metric that describes the closeness of the posterior distributions of first stage and second stage response rates. A natural choice of such a metric is Bhattacharyya's overlap measure (BOM) (*Bhattacharyya*, 1946). If distributions from two populations are continuous with probability density functions $f_1(\theta)$ and $f_2(\theta)$, the BOM is defined as $\mathrm{O}(f_1, f_2) = \int_{-\infty}^{\infty} \sqrt{f_1(\theta) f_2(\theta)} d\theta$. The BOM is useful in our setting because it takes values in the interval $[0, 1]$, in which $\mathrm{O}(f_1, f_2) = 0$ indicates that two distributions are fully separated, while $\mathrm{O}(f_1, f_2) = 1$ means that two distributions are identical. This agrees with the interpretation of power parameters in power prior models.

We define the posterior distributions of response rates of treatment k in stage 1 and stage 2 (within a specific subgroup j) as $p_1(\pi_k|\mathbf{D^{(1)}})$ and $p_{2j}(\pi_k|\mathbf{D^{(2)}})$, respectively, where $p_1(\pi_k|\mathbf{D^{(1)}}) \propto L(Z_k^{(1)}; \pi_k) p_0(\pi_k)$ and $p_{2j}(\pi_k|\mathbf{D^{(2)}}) \propto L(Z_{k,j}^{(2)}; \pi_k) p_0(\pi_k)$. Because we assume that the prior distributions of π_k are Beta distributions, the posterior distributions p_1 and p_{2j} will also follow $\mathrm{Beta}(a_1, b_1)$ and $\mathrm{Beta}(a_{2j}, b_{2j})$, respectively. Thus, we have

$$\begin{aligned} \mathrm{O}_k(p_1, p_{2j}) &= \int_0^1 \sqrt{p_1(\pi_k|\mathbf{D^{(1)}}) p_{2j}(\pi_k|\mathbf{D^{(2)}})} d\pi_k \\ &= \int_0^1 \sqrt{\frac{\pi_k^{a_1+a_{2j}-2}(1-\pi_k)^{b_1+b_{2j}-2}}{B(a_1, b_1)B(a_{2j}, b_{2j})}} d\pi_k \\ &= \frac{B(\frac{a_1+a_{2j}}{2}, \frac{b_1+b_{2j}}{2})}{\sqrt{B(a_1, b_1)B(a_{2j}, b_{2j})}} \end{aligned} \tag{4.9}$$

where $a_1 = Z_k^{(1)} + a_\pi$, $b_1 = n_k^{(1)} - Z_k^{(1)} + b_\pi$, $a_{2j} = Z_{k,j}^{(2)} + a_\pi$, $b_{2j} = n_{k,j}^{(2)} - Z_{k,j}^{(2)} + b_\pi$. We then derive values for δ_1 and δ_2 as the average of BOM for all three treatments, or $\delta_j^{BOM} = \sum_k \mathrm{O}_k(p_1, p_{2j})/3$.

Alternatively, the two-sided p-value of a Fisher's exact test (FET) from stage 1 and

stage 2 data from subgroup j can be used to quantify the closeness of treatment response rates in both stages. Specifically, we construct a 2×2 table where the rows contain the numbers of participants from stage 1 or stage 2 subgroup j and the columns contain the numbers of responders or non-responders. The two-sided p-value is computed using all the tables that are equally or more extreme than the observed table where extremity is defined by a table's hypergeometric probability. If the response rates change across the stages, the p-value from the FET should be small, suggesting that the data from stage 1 and stage 2 subgroup j are incompatible. On the contrary, if the response rates do not change across the stages, we can expect a p-value close to 1, indicating that a higher weight should be put on the "historical" data in subgroup j. Similar to the δ_j^{BOM}, we can calculate $\delta_j^{FET} = \sum_k P_{k,j}/3$, in which $P_{k,j}$ is the p-value for subgroup j and treatment k.

4.4 Simulation studies

4.4.1 Data generation

We conducted Monte Carlo simulations to compare the performance of the power prior models described in Section 4.3. The seven scenarios that we examined are listed in Table 4.1. In all scenarios in stage 1, exactly 1/3 of participants are assigned to each of the three possible treatments. Their stage 1 responses are generated by a Bernoulli distribution with the response rates corresponding to the assigned treatments, shown in Table 4.1(a). Their stage 2 responses are also generated by a Bernoulli distribution with the response rates corresponding to the assigned stage 1 and 2 treatments, shown in Table 4.1(b). In scenarios 1-5, the first stage response rates of the treatments differ from each other, whereas these response rates are identical in scenarios 6 and 7. The last two scenarios can be used to examine the performance of estimation under the "null" cases.

(a) First stage response rates

	A	B	C
Scenario 1-5	0.2	0.3	0.4
Scenario 6-7	0.3	0.3	0.3

(b) Second stage response rates

		Stage 1 treatment			Stage 1 treatment		
		A	B	C	A	B	C
		Scenario 1			**Scenario 2**		
Stage 2 treatment	A	0.2	0.2	0.2	0.4	0.2	0.2
	B	0.3	0.3	0.3	0.3	0.6	0.3
	C	0.4	0.4	0.4	0.4	0.4	0.8
		Scenario 3			**Scenario 4**		
Stage 2 treatment	A	0.2	0.1	0.1	0.4	0.3	0.3
	B	0.15	0.3	0.15	0.45	0.6	0.45
	C	0.2	0.2	0.4	0.6	0.6	0.8
		Scenario 5			**Scenario 6**		
Stage 2 treatment	A	0.6	0.4	0.4	0.3	0.3	0.3
	B	0.6	0.6	0.15	0.3	0.3	0.3
	C	0.2	0.2	0.6	0.3	0.3	0.3
		Scenario 7					
Stage 2 treatment	A	0.2	0.2	0.2			
	B	0.3	0.3	0.3			
	C	0.4	0.4	0.4			

Table 4.1: The true first and second stage response rates for simulation scenarios 1-7. (a) The response rates of the treatments in the first stage, which is the response rates of the interest. (b) The response rates of the treatments in the second stage, which depend on the first stage treatment and whether an individual responds to it. According to the snSMART design in Figure 2.1, responders to their first stage treatment continue with the same treatments in the second stage, and the response rates of which are highlighted in gray. The non-highlighted response rates correspond to those from first stage non-responders.

The rationale of designing the scenarios is as follows:

Scenario 1 The response rates remain unchanged in stage 2; there is full compatability between stage 1 and 2 data.

Scenario 2 The stage 2 response rates double if participants respond in stage 1; there is full compatibility between stage 1 data and stage 2 data only for stage 1 non-responders.

Scenario 3 The stage 2 response rates are halved for participants who do not respond in stage 1; there is full compatability between stage 1 data and stage 2 data only for stage 1 responders.

Scenario 4 The stage 2 response rates increase, but the scale of increase differs between stage 1 responders and non-responders; there is not full compatibility between stage 1 and stage 2 data.

Scenario 5 Stage 2 response rates change with respect to both first and second stage treatments, which violates a main assumption of the BJSM; there is not full compatibility between stage 1 and stage 2 data.

Scenario 6 All stage 1 and stage 2 response rates are equal; there is full compatibility between stage 1 and 2 data.

Scenario 7 Response rates are the same in stage 1 but not stage 2, and these depend on both first and second stage treatment (this violates a main assumption of the BJSM); there is not full compatibility between stage 1 and stage 2 data.

In Section 4.4.2, we use scenarios 1-4 to investigate the impact on δ_j when a part of or the whole stage 2 data are not compatible with the stage 1 data. We expect that: (1) both δ_1 and δ_2 are close to 1 in scenario 1; (2) δ_1 should should move closer to 0 in scenario 2; (3) δ_2 should move closer to 0 in scenario 3; (4) both δ_1 and δ_2 should move closer to 0 in scenarios 4. In Section 4.4.3, we evaluate the estimation of π_k

using scenarios 4-7, with which we compare the performance either within different power prior models or between power prior models and the BJSM. We also examine whether partial borrowing of information from second stage data ($0 < \delta_j < 1$) can outperform situations when instead complete borrowing ($\delta_j = 1$) or no borrowing ($\delta_j = 0$) is applied.

The prior distribution of π_k was Beta(1,1) for all methods, and the prior distribution of δ_j was Beta(1,1) in MPP. To maximize the flexibility of the BJSM, we set the prior distributions of both linkage parameters to gamma distributions with the support of $(0, \infty)$ and the prior mean of 1. All simulation studies were performed with 10,000 runs, and the total sample size for each run was either 90 or 300.

4.4.2 Estimation of δ_1 and δ_2 for power prior models

In Table 4.2, we present the mean estimated δ_1 and δ_2 and their Monte Carlo standard errors obtained from five different power prior models in scenarios 1-4. Presently, we restrict our focus on scenarios 1-4 because these scenarios are designed to examine how $\boldsymbol{\delta}$ changes when data from the two stages become incompatible.

When $N = 90$, we first observe the differences in $\boldsymbol{\delta}$ when comparing scenarios 2-4 to scenario 1, in which the data from stages 1 and 2 are fully compatible. In MLC, the mean estimated δ_1 is 0.65 in scenario 1 compared to 0.32 in scenario 2 and 0.08 in scenario 4 where the stage 2 data from stage 1 responders is not compatible with the stage 1 data. The mean estimated δ_2 is from 0.75 in scenario 1 compared to 0.40 in scenario 3 and 0.45 in scenario 4 where the stage 2 data from stage 1 non-responders is not compatible with the stage 1 data.

Similarly, we can see the same pattern in MPP, PLC, BOM and FET, but the scale of difference varies. The differences are about 0.2 to 0.4 when comparing $\boldsymbol{\delta}$ from scenario 1 to scenarios 2-4 in FET and BOM, 0.1 to 0.2 in MPP, and less than 0.1 in PLC. The

Scenario	$N = 90$									
	MPP		PLC		MLC		BOM		FET	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
1	0.51(0.06)	0.54(0.07)	0.32(0.04)	0.23(0.02)	0.65(0.42)	0.75(0.35)	0.76(0.10)	0.81(0.11)	0.64(0.19)	0.59(0.18)
2	0.41(0.09)	0.61(0.06)	0.28(0.03)	0.23(0.02)	0.32(0.39)	0.87(0.26)	0.48(0.14)	0.81(0.11)	0.28(0.17)	0.59(0.18)
3	0.53(0.07)	0.44(0.11)	0.31(0.04)	0.30(0.17)	0.76(0.36)	0.40(0.36)	0.76(0.10)	0.64(0.15)	0.64(0.19)	0.38(0.18)
4	0.32(0.10)	0.46(0.14)	0.29(0.03)	0.22(0.02)	0.08(0.21)	0.45(0.39)	0.48(0.14)	0.66(0.16)	0.28(0.17)	0.40(0.19)
Scenario	$N = 300$									
	MPP		PLC		MLC		BOM		FET	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
1	0.52(0.06)	0.55(0.07)	0.23(0.10)	0.13(0.04)	0.67(0.42)	0.77(0.34)	0.75(0.11)	0.81(0.12)	0.57(0.18)	0.55(0.18)
2	0.25(0.10)	0.66(0.06)	0.18(0.01)	0.14(0.01)	0.13(0.22)	0.87(0.25)	0.20(0.11)	0.81(0.12)	0.08(0.09)	0.55(0.18)
3	0.60(0.07)	0.26(0.11)	0.20(0.02)	0.16(0.01)	0.87(0.26)	0.14(0.17)	0.75(0.11)	0.35(0.15)	0.57(0.18)	0.13(0.12)
4	0.10(0.04)	0.28(0.15)	0.18(0.01)	0.13(0.02)	0.00(0.01)	0.15(0.18)	0.20(0.11)	0.41(0.16)	0.08(0.09)	0.17(0.14)

Table 4.2: The means and standard errors (in parentheses) of δ_1 and δ_2 obtained from each of the three power prior approaches, which are modified power prior model (MPP), power prior model with penalized likelihood-type criterion (PLC), power prior model with marginal likelihood criterion (MLC), power prior model with Bhattacharyya's overlap measure and power prior models with Fisher's exact test. Scenarios 1-4 in Table 4.1 are used to evaluate how these δs change with different levels of compatibility between first and second stage data. All simulation studies are done at $N = 90$ or 300.

differences become larger when $N = 300$ for all methods. However, there is a trade-off between the difference in $\boldsymbol{\delta}$ across various scenarios and the standard errors of estimated $\boldsymbol{\delta}$. The estimated $\boldsymbol{\delta}$ from MLC have much larger standard errors than that of all other methods. In contrast, the estimates $\boldsymbol{\delta}$ from PLC slightly change across different scenarios, resulting in relatively small standard errors of the estimates.

In addition, we also investigated the ranges of the mean estimated $\boldsymbol{\delta}$ from different methods. When $N = 90$, the values of $\boldsymbol{\delta}$ from the BOM are close to 0.5 even when the data from two stages are incompatible, which indicates that the BOM tends to put higher weights on "historical" data, regardless of the compatibility of first and second stages data. In contrast, the values of $\boldsymbol{\delta}$ from PLC are between 0.2 and 0.35 in all scenarios, which agrees with the finding in *Ibrahim et al.* (2003) that the estimated $\boldsymbol{\delta}$ from this method is relatively small in general. For MPP, MLC and FET, the values of $\boldsymbol{\delta}$ are greater than 0.5 when data from two stages are compatible, whereas the values of $\boldsymbol{\delta}$ are smaller than 0.5 if data are incompatible.

We note that data compatibility is not the only driving force of the value of $\boldsymbol{\delta}$ for MPP. The prior distribution of $\boldsymbol{\delta}$ also plays an important role in the range of mean estimated $\boldsymbol{\delta}$. In Table 4.3, we let the prior distributions of $\boldsymbol{\delta}$ be Beta(0.4,1.6), Beta(1,1), and Beta(1.6,0.4), which correspond to the prior means of 0.2, 0.5 and 0.8, respectively. We can see that the range of $\boldsymbol{\delta}$ is centered at the prior mean of $\boldsymbol{\delta}$, especially when $N = 90$. When $N = 300$, the data have more capacity to adjust the estimated $\boldsymbol{\delta}$ in addition to the influence from the prior distributions. Thus, we conclude, similar to *Neuenschwander et al.* (2009), that the specification of the prior distribution of $\boldsymbol{\delta}$ can greatly impact the results from the MPP method. The mean estimated $\boldsymbol{\delta}$ under scenarios 5-7 for all methods can be found in the Table B.1 in Appendix B.

We further examine the distributions of estimated δ from different methods under scenarios 1-4 in Figure 4.1 when $N = 90$. The histograms from the PLC under four

Scenario	$N = 90$					
	$E(\delta) = 0.20$		$E(\delta) = 0.50$		$E(\delta) = 0.80$	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
1	0.23(0.05)	0.31(0.06)	0.51(0.06)	0.54(0.07)	0.80(0.04)	0.81(0.05)
2	0.15(0.05)	0.36(0.07)	0.41(0.09)	0.61(0.06)	0.73(0.08)	0.86(0.03)
3	0.25(0.05)	0.23(0.08)	0.54(0.07)	0.44(0.11)	0.82(0.05)	0.73(0.11)
4	0.11(0.05)	0.23(0.10)	0.33(0.10)	0.46(0.14)	0.65(0.11)	0.75(0.13)
Scenario	$N = 300$					
	$E(\delta) = 0.20$		$E(\delta) = 0.50$		$E(\delta) = 0.80$	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
1	0.25(0.05)	0.33(0.07)	0.52(0.06)	0.55(0.08)	0.81(0.04)	0.81(0.05)
2	0.08(0.04)	0.42(0.07)	0.25(0.10)	0.66(0.06)	0.50(0.15)	0.88(0.04)
3	0.34(0.07)	0.12(0.06)	0.60(0.07)	0.26(0.11)	0.85(0.04)	0.49(0.17)
4	0.03(0.02)	0.14(0.09)	0.10(0.04)	0.28(0.15)	0.22(0.11)	0.51(0.23)

Table 4.3: The means and standard errors (in parentheses) of δ_1 and δ_2 obtained from modified power prior model (MPP) with different $E(\delta)$, or prior mean of δ. Scenarios 1-4 in Table 4.1 are used to evaluate how these δs change with different levels of compatibility between first and second stage data. All simulation studies are done at $N = 90$ or 300.

scenarios do not differ much, indicating that the power parameters obtained from PLC do not vary with the changing scenarios. For MLC, the chance of choosing 0 or 1 for power parameters is extremely high, which is not a desirable property because second stage data are likely to be completely ignored even when the data across stages are fully compatible. This result suggests that the estimated δ from the power prior model with MLC is highly sensitive to slight changes in the number of responders. In particular, when the expected number of responders to a treatment in a subgroup in stage 2 is smaller than 10, which may be common in an snSMART, a change in the observed number of responders by 1 or 2 can result in a sharp decrease of the estimated δ from 1 to 0 or vice versa. The histograms from the MPP, BOM and FET are more appealing. In scenario 1, a large portion of distributions of δ_1 and δ_2 can overlap, while in other scenarios, we can easily see the move of either one or both distributions when part of or all the second stage data are not compatible with first

stage data.

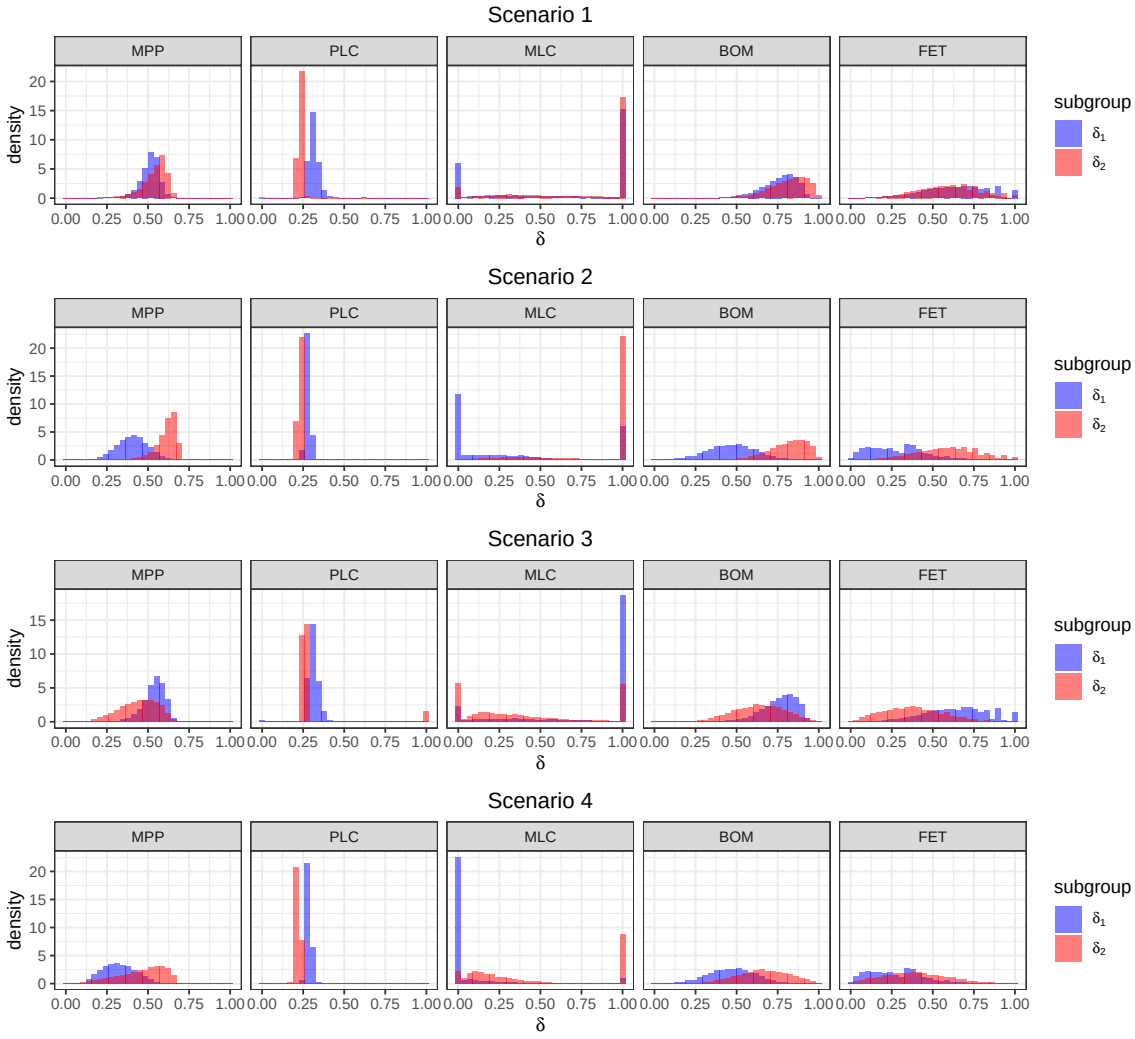

Figure 4.1: The distributions of δ_1 and δ_2 from modified power prior (MPP), power prior with penalized likelihood-type criterion (PLC), marginal likelihood criterion (MLC), Bhattacharyya's overlap measure (BOM) and measure from Fisher's exact test (FET) under scenarios 1-4. $N = 90$

When $N = 300$, the distributions for δ_1 and δ_2 in Figure 4.2. For FET, BOM and MPP, due to the increased sample size, the distributions move more when the data are incompatible, compared to the histograms in Figure 4.1. For MLC, it seems that the chance of assigning the wrong power parameters becomes lower compared to $N = 90$, but completely ignoring the second stage data is still undesirable even when the data across stages are not compatible. Borrowing some information from

incompatible second stage data may still increase efficiency given that the bias may increase as well, which we will discuss in next Section 4.4.3. The distributions of δ_1 and δ_2 under scenarios 5-7 can be found in the Figures B.1 and B.2 in Appendix B.

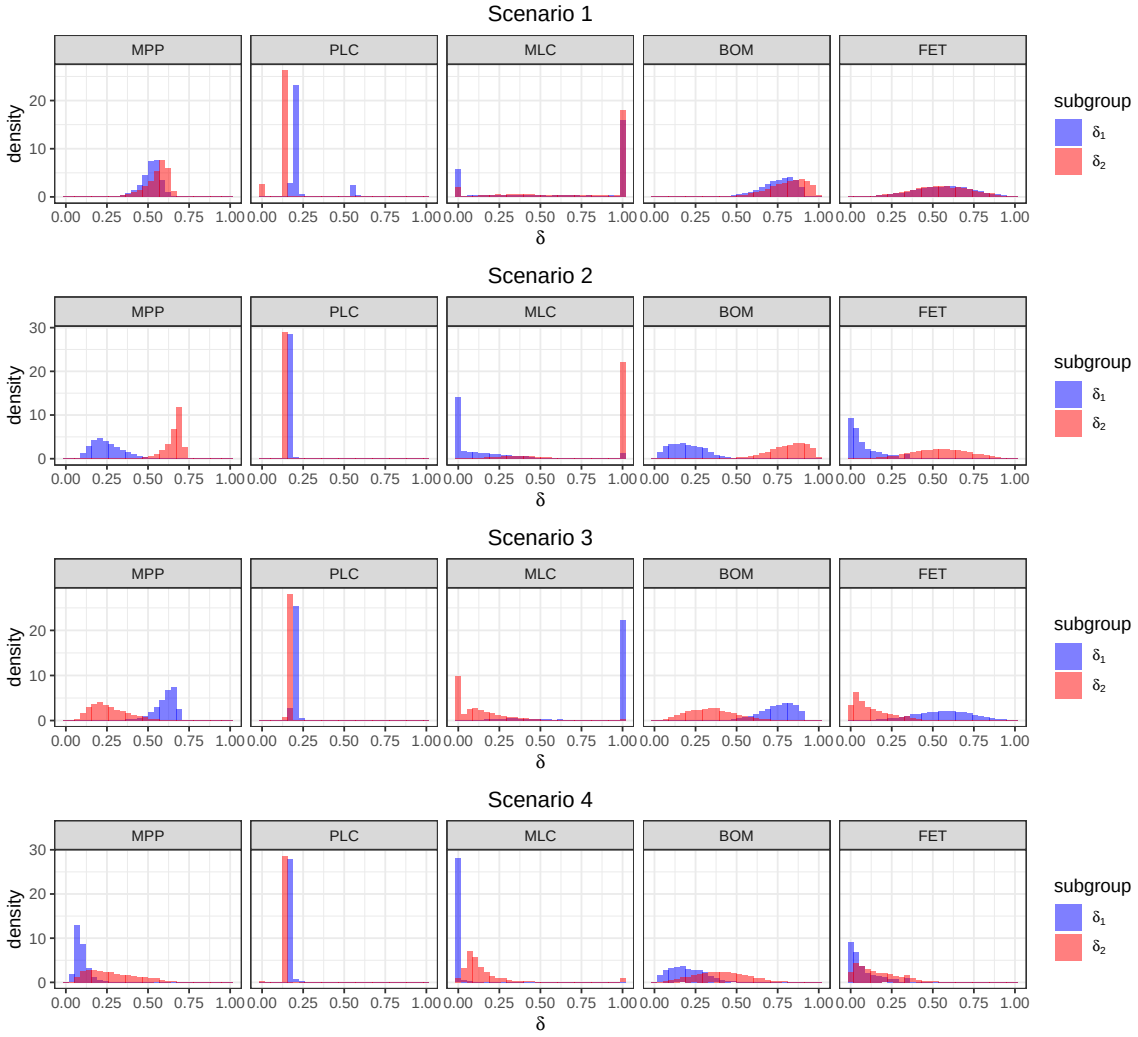

Figure 4.2: The distributions of δ_1 and δ_2 from modified power prior (MPP), power prior with penalized likelihood-type criterion (PLC), marginal likelihood criterion (MLC), Bhattacharyya's overlap measure (BOM) and measure from Fisher's exact test (FET) under scenarios 1-4. $N = 300$

4.4.3 Estimation of π

In Figure 4.3, each bar is the simulation-wide average absolute value of bias or root mean squared error (rMSE) of the three treatment response rate estimates from each

of the methods. We include results for power prior models when $\boldsymbol{\delta}$ is fixed at 0 or 1 for reference, as these two approaches only perform well in either fully compatible or highly incompatible scenarios, and are not preferred in most realistic settings.

In scenario 4, we first note that BJSM has smallest bias and rMSE among all methods because the assumption of the linkage parameters is met in this scenario. Among all the power prior methods, we expect some bias because stage 2 data are highly incompatible with stage 1 data. Although the estimation from MLC is least biased because the estimated $\boldsymbol{\delta}$ are close to 0 in a large portion of simulated runs, we see that the rMSE of MLC is close to that from MPP, PLC and FET due to the high Monte Carlo variability of the MLC estimates. In scenario 5, the power prior models are more able to appropriately weight the second stage data, leading to lower rMSE compared to the BJSM because of violation to assumptions needed for the BJSM.

In scenario 6, the data from two stages are compatible, and although the bias for all methods is small, we see that the rMSEs of BOM are smaller than other methods. This is because the distributions of δ_1 and δ_2 for BOM in Figure B.1 in Appendix B are clustered at the right half of the distribution, indicating power parameters closer to 1 compared to other histograms.

Scenario 7 is similar to scenario 5 in terms of data incompatibility and violation of an assumption of the BJSM, but the level of data incompatibility is less strong according to Figure B.1. Thus, we see that the rMSE of the power prior models is lower than the rMSE from the BJSM. The details of the bias and rMSE for all methods under scenarios 1-7 can be found in Tables B.2 and B.3 in Appendix B. We also have examined the patterns of bias and rMSE when $N = 60, 75$ or 300, and the patterns are similar to $N = 90$ (results not shown). Thus, the power prior models can still be applied to snSMARTs with even smaller sample sizes.

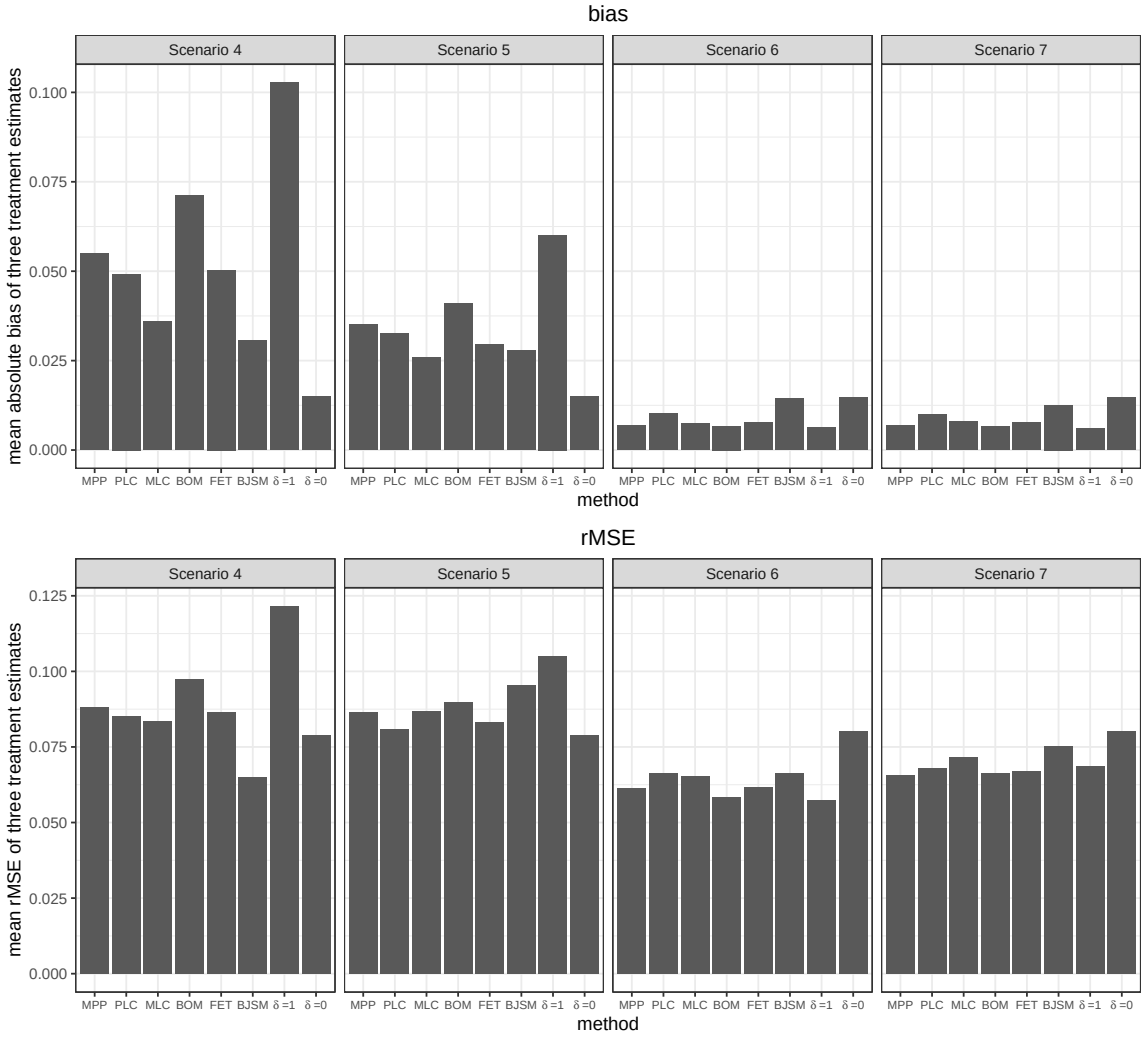

Figure 4.3: The barplots of the mean absolute biases and root mean squared errors (rMSEs) of the treatment response rate estimates under different methods. The results from scenarios 4-7 are shown. MPP=modified power prior model; PLC=power prior model with penalized likelihood-type criterion; MLC=power prior model with marginal likelihood criterion; BOM=power prior model with Bhattacharyya's overlap measure; FET=power prior models with Fisher's exact test; BJSM=Bayesian joint stage model. Power prior model is also applied with all δ fixed at 1 (or 0), meaning that the second stage data are completely used (or ignored). $N = 90$

4.5 Discussion

Overall, we do not recommend use of the power prior models with the MPP, PLC or MLC. For the MPP, the choice of $\boldsymbol{\delta}$ highly depends on its prior distribution, especially in an snSMART where the sample size is small. For the PLC, the estimated δ stay relatively constant across different scenarios in snSMARTs, regardless of whether the data from the two stages are compatible. For the MLC, the mean estimated $\boldsymbol{\delta}$ can change along with the compatibility across two stages, but the value for $\boldsymbol{\delta}$ is highly sensitive to small changes in number of responders in an snSMART. This sensitivity of the MLC leads to a high chance of choosing 0 or 1.

Therefore, we feel that PLC and MLC should not be used to estimate response rates in an snSMART because it is undesirable to choose a fixed value or extreme values of 0 or 1 with high probability. MPP is not preferred as well because the weights highly depend on their prior distributions.

Hence, the suggested candidate models for treatment effect estimation in an snSMART are the BJSM and the power prior models with BOM or FET when considering both the performance of treatment effect estimation and reasonable values of $\boldsymbol{\delta}$.

For the FET, we acknowledge that the stage 2 outcomes from stage 1 responders and their stage 1 outcomes may not be independent, which is an assumption of the Fisher's exact test, but we believe that the smaller p-values are reasonable because less weight should be put on second stage data when the within-individual correlation of first and second stage responses can affect the determination of dependency between stages and responses to treatment. Moreover, the number of correlated observations is likely to be small especially in rare disease trials. For the BOM, we can see that the assigned weights to "historical" data tend to be larger than the weights from other methods. For the BJSM, we need to assume that the relationship between first

and second stage response rates can be described proportionally through the linkage parameters. This assumption may be difficult to justify.

When selecting a primary method of analysis, some background information about the treatments of interest in an snSMART may influence model choice in the estimation of treatment effects. If investigators believe that the second stage response rates are proportional to the first stage response rates and the proportionality (linkage) parameters do not depend on first and second stage treatments, then the BJSM may be preferred since it is most efficient when its assumptions are met. For example, the BJSM can be used if we believe that the response rates of all treatments will double in the second stage for all first stage responders. However, if this assumption is violated, which may be very likely, then power prior models with BOM or FET may be considered. The BOM is preferred if the data from two stages are more compatible, while the FET is preferred in the cases of less compatibility between data from two stages. If prior information about possible first and second stage response rates of all treatments exists, simulation studies can be conducted to help decide the prior distribution of $\boldsymbol{\delta}$ and $\boldsymbol{\pi}$.

An extension of the SMART is the proposal by *Liu et al.* (2017) that the design be enriched at later stages of the trial by the inclusion of subjects that received previous stage treatments outside of the trial. They used the term, SMARTER, for a SMART with enrichment. While this design assumed larger sample sizes, the same idea can apply to an snSMART. In an snSMART, it is not clear how a subject's information outside of the trial should be incorporated by the BJSM. However, this enrichment is not a problem for the power prior model methods since these methods do not link an individual subject's responses between stages. Thus, our power prior models might be more appropriate for SMARTER designs.

Moreover, a different number of subgroups in stage 2 of an snSMART can be pre-

specified instead of $J = 2$ in our study. In simulations, we have tried $J = 6$, where the δ can differ depending on the individuals' first stage responses and their stage 1 treatment assignments. However, due to the resulting small sample sizes in each subgroup, the extra power parameters did not improve the bias and efficiency of the estimation (results not shown). The application of a Bhattacharyya's overlap measure or Fisher's exact test in power prior models is not limited to our snSMART settings, but also can be used in more general cases when data from historical trials are used to facilitate the data analysis of a current clinical trial. In this setting, the potential issue of independence between samples no longer exists because patients from different trials should be uncorrelated.

CHAPTER V

Summary and Future Work

Motivated by ARAMIS, an snSMART for skin vasculitis, this dissertation has focused on different models that can be applied to estimate the response rates of first stage treatments or DTRs in small samples. In addition, we modified the standard snSMART by incorporating a group sequential design to allow dropping of an inferior treatment arm at an interim analysis.

In Chapter II, we demonstrated how the Bayesian joint stage model (BJSM) and the joint stage regression model (JSRM) proposed by *Wei et al.* (2018) can be used to estimate DTRs in an snSMART and compared them with the existing weighted and replicated regression model by *Nahum-Shani et al.* (2012). The BJSM and JSRM perform better in terms of efficiency because the data from both stages are used. We also proposed a simulation-based sample size calculation method using the JSRM for an snSMART when the goal of the trial is the estimation of first stage treatment response rates. This approach involves Dunnett's correction method for multiple comparisons under the GEE model.

In Chapter III, we proposed a group sequential snSMART where a decision of whether an arm should be removed can be made at each interim look. Compared to a standard snSMART, more participants were expected to be assigned to the better performing treatments in our group sequential snSMART, which is an attractive property of this

design. Moreover, the probability of incorrectly removing an arm during all interim analyses can be controlled by the pre-specified cutoff values for the decision rules.

In Chapter IV, we introduced a new application of power prior models to the estimation of treatment effect in an snSMART by assuming that first stage outcomes are "current data" and second stage outcomes are "historical data". Compared to the BJSM, the power prior model performs better when second stage treatment response rates change with respect to both first and second stage treatments, which violates an assumption of the BJSM.

In addition to the group sequential snSMART design that we introduced in Chapter III, we examined some other adaptive designs that change the treatment allocation rule. For example, we explored Bayesian adaptive randomization, where the randomization probabilities to each treatment are altered based on the posterior probabilities from interim outcomes (*Thall and Wathen*, 2007). However, we found some potential issues with this approach. First, it is hard to decide an appropriate mapping from interim outcomes to randomization probabilities. In our group sequential snSMART, we can pre-specify a desired probability of incorrectly removing an arm through thresholds for the posterior probabilities in the two-step decision rule. However, in Bayesian adaptive randomization, we do not have an objective criterion to choose an appropriate power term c if the randomization probability of $j = J$ is determined by $r_J = \frac{P(\pi_J \text{ is the largest})^c}{\sum_{\text{all}j} P(\pi_j \text{ is the largest})^c}$, which is similar to *Thall and Wathen* (2007). Second, since the sample size for an snSMART is small, the effect of changing randomization probabilities is also small in terms of treatment allocation compared to that of a standard snSMART. Based on our simulation results (not shown), if our group sequential snSMART is applied instead of other adaptive randomization approaches, more patients are expected be allocated to better performing treatments.

In Chapter IV, we introduced a new application of measures of closeness to the

estimation of power parameters in the power prior models. However, we only have tried this approach in our snSMART with binary outcomes. In the future, since more and more data will be generated or collected from different clinical studies, drawing inference from a combination of data from several sources would be of more importance. Thus, it would be interesting to investigate if this application can be generalized to the incorporation, or integration, of historical and current data of clinical trials with different endpoints.

APPENDICES

APPENDIX A

Chapter II: Additional Simulation Results

Table A.1: The bias and root mean squared error (rMSE) of the dynamic treatment regimen (DTR) response rate estimates using Bayesian Joint Stage Model (BJSM), Joint Stage Regression Model (JSRM), and Weighted and Replicated Regression Model (WRRM). The sample sizes for scenarios 1a-c, 2a-c, 3a-c, and 4a-c (see Table 2.1 in the main text), are 135, 90, 120 and 120, respectively.

Begin of Table							
Scenario		BJSM		JSRM		WRRM	
		Bias	rMSE	Bias	rMSE	Bias	rMSE
1a	AAB	-0.018	0.056	-0.009	0.080	-0.007	0.087
	AAC	-0.019	0.052	-0.009	0.062	-0.009	0.076
	BBA	0.013	0.056	0.002	0.075	-0.001	0.082
	BBC	0.003	0.050	0.001	0.060	0.004	0.073
	CCA	0.033	0.063	0.000	0.057	0.000	0.071
	CCB	0.037	0.066	0.000	0.057	0.000	0.071
1b	AAB	-0.024	0.066	0.000	0.082	0.001	0.090
	AAC	-0.027	0.063	0.001	0.067	0.001	0.083
	BBA	0.017	0.057	0.002	0.075	-0.001	0.082
	BBC	0.000	0.050	0.001	0.060	0.004	0.073
	CCA	0.045	0.070	0.001	0.054	0.001	0.068
	CCB	0.045	0.070	0.001	0.054	0.001	0.069
1c	AAB	-0.001	0.062	0.012	0.080	0.000	0.088
	AAC	-0.017	0.060	0.006	0.068	0.001	0.082
	BBA	0.006	0.055	-0.006	0.078	-0.001	0.085

Continuation of Table A.1

Scenario		BJSM		JSRM		WRRM	
		Bias	rMSE	Bias	rMSE	Bias	rMSE
	BBC	-0.003	0.049	0.000	0.059	0.004	0.073
	CCA	0.002	0.057	-0.036	0.076	0.002	0.087
	CCB	0.073	0.089	0.039	0.072	0.000	0.072
2a	AAB	-0.029	0.070	-0.014	0.099	-0.011	0.106
	AAC	-0.030	0.067	-0.012	0.078	-0.012	0.094
	BBA	0.009	0.066	0.003	0.095	0.000	0.102
	BBC	-0.004	0.063	0.002	0.077	0.006	0.092
	CCA	0.044	0.077	0.000	0.071	0.000	0.088
	CCB	0.049	0.082	0.000	0.071	-0.001	0.088
2b	AAB	-0.038	0.083	0.000	0.100	0.000	0.109
	AAC	-0.040	0.081	0.001	0.084	0.003	0.104
	BBA	0.014	0.067	0.003	0.095	0.000	0.102
	BBC	-0.007	0.063	0.002	0.077	0.006	0.092
	CCA	0.060	0.088	0.001	0.068	0.002	0.086
	CCB	0.058	0.087	0.002	0.068	0.000	0.085
2c	AAB	-0.010	0.074	0.015	0.099	0.000	0.108
	AAC	-0.027	0.076	0.010	0.086	0.004	0.104
	BBA	0.001	0.064	-0.007	0.098	-0.002	0.104
	BBC	-0.012	0.063	-0.002	0.075	0.006	0.092
	CCA	0.005	0.067	-0.039	0.093	0.004	0.108
	CCB	0.086	0.104	0.046	0.089	0.001	0.090
3a	AAB	-0.029	0.060	-0.009	0.073	-0.009	0.082
	AAC	-0.029	0.061	-0.010	0.073	-0.009	0.083
	BBA	0.027	0.062	0.002	0.081	-0.001	0.089
	BBC	0.017	0.041	0.001	0.049	0.004	0.068
	CCA	0.047	0.077	0.000	0.070	0.000	0.078
	CCB	0.024	0.045	0.000	0.043	0.000	0.056
3b	AAB	-0.036	0.074	0.001	0.078	0.000	0.090
	AAC	-0.038	0.073	0.000	0.078	0.003	0.091
	BBA	0.033	0.065	0.002	0.081	-0.001	0.089
	BBC	0.013	0.039	0.001	0.049	0.004	0.068
	CCA	0.060	0.085	0.001	0.067	0.001	0.076

Continuation of Table A.1

Scenario		BJSM		JSRM		WRRM	
		Bias	rMSE	Bias	rMSE	Bias	rMSE
	CCB	0.033	0.048	0.001	0.038	0.001	0.052
3c	AAB	-0.015	0.066	0.013	0.078	0.000	0.088
	AAC	-0.025	0.068	0.006	0.078	0.003	0.090
	BBA	0.021	0.061	-0.005	0.086	-0.002	0.094
	BBC	0.015	0.040	0.007	0.050	0.004	0.068
	CCA	0.014	0.061	-0.022	0.087	0.003	0.095
	CCB	0.048	0.060	0.027	0.051	0.001	0.055
4a	AAB	-0.027	0.061	-0.010	0.079	-0.009	0.087
	AAC	-0.027	0.060	-0.010	0.071	-0.009	0.083
	BBA	0.022	0.060	0.002	0.081	-0.001	0.088
	BBC	0.010	0.045	0.000	0.056	0.004	0.072
	CCA	0.042	0.074	0.000	0.066	0.000	0.078
	CCB	0.032	0.057	0.000	0.051	-0.001	0.066
4b	AAB	-0.034	0.073	0.001	0.082	0.000	0.093
	AAC	-0.036	0.073	0.000	0.076	0.003	0.091
	BBA	0.027	0.063	0.002	0.081	-0.001	0.088
	BBC	0.007	0.044	0.000	0.056	0.004	0.072
	CCA	0.056	0.082	0.001	0.064	0.001	0.076
	CCB	0.041	0.061	0.001	0.047	0.000	0.063
4c	AAB	-0.011	0.066	0.014	0.082	0.000	0.092
	AAC	-0.023	0.067	0.007	0.076	0.003	0.091
	BBA	0.016	0.058	-0.003	0.085	0.000	0.092
	BBC	0.006	0.043	0.004	0.055	0.004	0.072
	CCA	0.009	0.062	-0.031	0.085	0.003	0.095
	CCB	0.064	0.077	0.035	0.064	0.001	0.066

End of Table

APPENDIX B

Chapter IV: Additional Simulation Results

Scenario	$N = 90$									
	MPP		PLC		MLC		BOM		FET	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
5	0.42(0.10)	0.47(0.13)	0.28(0.03)	0.22(0.02)	0.29(0.37)	0.42(0.37)	0.51(0.15)	0.61(0.17)	0.30(0.18)	0.35(0.19)
6	0.51(0.06)	0.54(0.07)	0.32(0.04)	0.23(0.02)	0.65(0.42)	0.75(0.35)	0.76(0.11)	0.81(0.11)	0.62(0.19)	0.59(0.18)
7	0.49(0.07)	0.50(0.10)	0.32(0.04)	0.23(0.02)	0.56(0.44)	0.62(0.40)	0.74(0.11)	0.76(0.13)	0.59(0.20)	0.52(0.19)
Scenario	$N = 300$									
	MPP		PLC		MLC		BOM		FET	
	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
5	0.21(0.10)	0.27(0.14)	0.17(0.01)	0.14(0.01)	0.04(0.08)	0.10(0.11)	0.21(0.11)	0.33(0.15)	0.06(0.08)	0.13(0.12)
6	0.52(0.06)	0.55(0.07)	0.24(0.11)	0.13(0.04)	0.67(0.41)	0.77(0.34)	0.75(0.11)	0.81(0.12)	0.57(0.18)	0.55(0.18)
7	0.46(0.10)	0.44(0.12)	0.24(0.11)	0.13(0.04)	0.44(0.43)	0.43(0.40)	0.67(0.13)	0.66(0.15)	0.46(0.19)	0.38(0.18)

Table B.1: The means and standard errors (in parentheses) of δ_1 and δ_2 obtained from each of the three power prior approaches, which are modified power prior model (MPP), power prior model with penalized likelihood-type criterion (PLC), and power prior model with marginal likelihood criterion (MLC). Scenarios 5-7 in Table 4.1 are used to evaluate how these δs change with different levels of compatibility between first and second stage data. All simulation studies are done at $N = 90$ or 300.

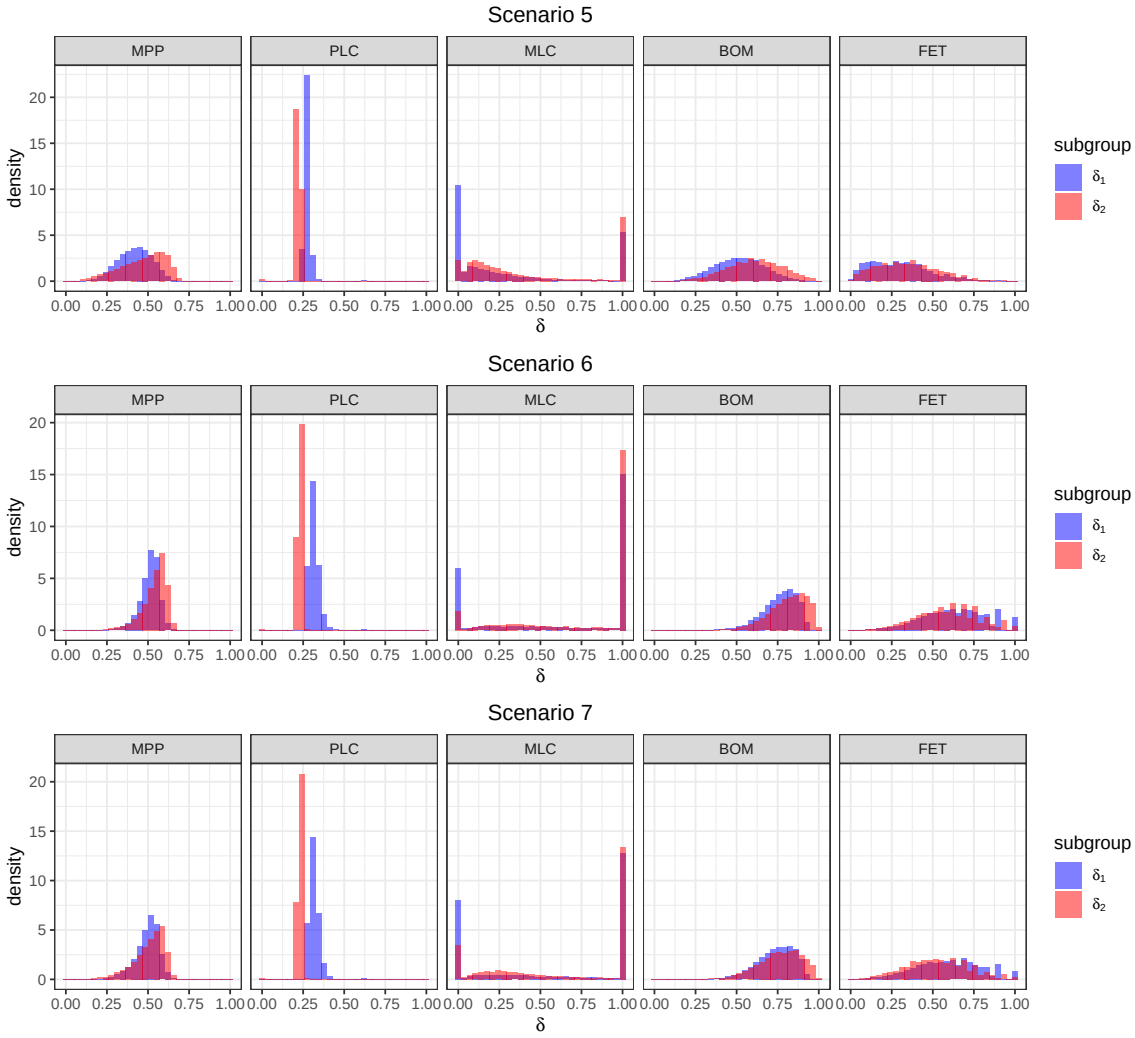

Figure B.1: The distributions of δ_1 and δ_2 from modified power prior (MPP), power prior with penalized likelihood-type criterion (PLC), marginal likelihood criterion (MLC), Bhattacharyya's overlap measure (BOM) and measure from Fisher's exact test (FET) under scenarios 5-7. $N = 90$

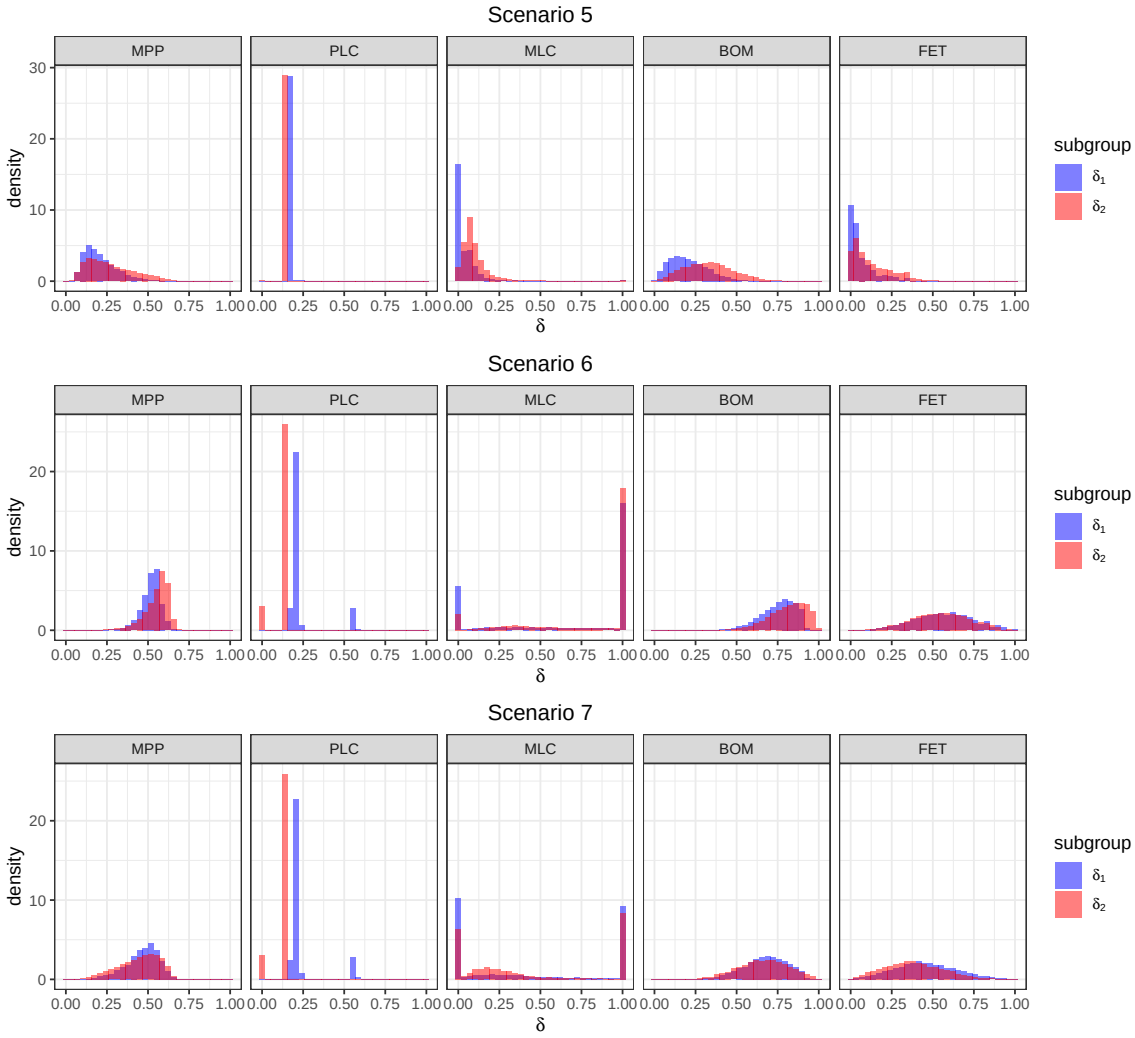

Figure B.2: The distributions of δ_1 and δ_2 from modified power prior (MPP), power prior with penalized likelihood-type criterion (PLC), marginal likelihood criterion (MLC), Bhattacharyya's overlap measure (BOM) and measure from Fisher's exact test (FET) under scenarios 5-7. $N = 300$

Scenario	Treatment	MPP	PLC	MLC	BOM	FET	BJSM	$\delta = 1$	$\delta = 0$
1	A	0.012	0.015	0.011	0.010	0.011	0.015	0.010	0.020
	B	0.011	0.012	0.010	0.008	0.009	0.016	0.008	0.017
	C	-0.001	0.005	0.003	0.003	0.004	0.013	0.002	0.008
2	A	0.023	0.024	0.018	0.021	0.018	0.019	0.030	0.020
	B	0.032	0.032	0.024	0.031	0.024	0.025	0.051	0.017
	C	0.036	0.038	0.027	0.044	0.032	0.032	0.075	0.008
3	A	-0.005	0.000	0.000	-0.013	-0.002	0.011	-0.025	0.020
	B	-0.015	-0.010	-0.007	-0.027	-0.013	0.009	-0.043	0.017
	C	-0.037	-0.026	-0.020	-0.044	-0.027	0.004	-0.064	0.008
4	A	0.039	0.036	0.030	0.046	0.034	0.022	0.065	0.020
	B	0.057	0.050	0.037	0.070	0.049	0.031	0.102	0.017
	C	0.068	0.062	0.040	0.097	0.068	0.039	0.141	0.008
5	A	0.070	0.056	0.051	0.083	0.058	0.099	0.120	0.020
	B	0.054	0.045	0.038	0.061	0.042	0.051	0.089	0.017
	C	-0.018	-0.003	-0.011	-0.022	-0.011	-0.066	-0.028	0.008
6	A	0.007	0.009	0.006	0.006	0.007	0.014	0.005	0.013
	B	0.010	0.011	0.009	0.007	0.008	0.015	0.007	0.016
	C	0.004	0.010	0.007	0.007	0.008	0.014	0.006	0.014
7	A	-0.022	-0.010	-0.016	-0.033	-0.021	-0.035	-0.043	0.013
	B	0.010	0.011	0.010	0.008	0.009	0.014	0.007	0.016
	C	0.032	0.029	0.030	0.045	0.036	0.059	0.054	0.014

Table B.2: The bias of the estimates of treatment response rates under different methods. MPP, PLC, MLC, BOM, FET and BJSM stand for modified power prior model, power prior model with penalized likelihood-type criterion, power prior model with marginal likelihood criterion, power prior model with Bhattacharyya's overlap measure and power prior models with Fisher's exact test and Bayesian joint stage model, respectively. Power prior model is also applied with all δ fixed at 1 (or 0), meaning that the second stage data are completely used (or ignored). $N = 90$.

Scenario	Treatment	MPP	PLC	MLC	BOM	FET	BJSM	$\delta = 1$	$\delta = 0$
1	A	0.055	0.060	0.058	0.053	0.056	0.058	0.052	0.071
	B	0.062	0.067	0.065	0.059	0.062	0.066	0.058	0.080
	C	0.064	0.069	0.069	0.060	0.063	0.072	0.059	0.085
2	A	0.061	0.066	0.061	0.061	0.061	0.057	0.065	0.071
	B	0.073	0.077	0.072	0.072	0.072	0.065	0.083	0.080
	C	0.080	0.085	0.082	0.082	0.078	0.074	0.102	0.085
3	A	0.056	0.058	0.061	0.055	0.059	0.060	0.054	0.071
	B	0.065	0.067	0.072	0.066	0.066	0.069	0.069	0.080
	C	0.076	0.076	0.080	0.077	0.073	0.075	0.087	0.085
4	A	0.072	0.071	0.071	0.075	0.071	0.055	0.087	0.071
	B	0.090	0.086	0.085	0.097	0.087	0.064	0.121	0.080
	C	0.103	0.099	0.095	0.120	0.102	0.075	0.156	0.085
5	A	0.096	0.085	0.090	0.105	0.088	0.114	0.136	0.071
	B	0.088	0.084	0.086	0.091	0.084	0.079	0.110	0.080
	C	0.076	0.074	0.085	0.074	0.078	0.094	0.070	0.085
6	A	0.061	0.066	0.065	0.058	0.062	0.066	0.057	0.079
	B	0.062	0.067	0.065	0.059	0.062	0.067	0.058	0.080
	C	0.061	0.067	0.066	0.058	0.061	0.066	0.058	0.081
7	A	0.063	0.064	0.069	0.064	0.064	0.071	0.067	0.079
	B	0.063	0.067	0.068	0.059	0.063	0.067	0.058	0.080
	C	0.072	0.074	0.077	0.076	0.074	0.088	0.081	0.081

Table B.3: The root mean square error (rMSE) of the estimates of treatment response rates under different methods. MPP, PLC, MLC, BOM, FET and BJSM stand for modified power prior model, power prior model with penalized likelihood-type criterion, power prior model with marginal likelihood criterion, power prior model with Bhattacharyya's overlap measure, power prior models with Fisher's exact test and Bayesian joint stage model, respectively. Power prior model is also applied with all δ fixed at 1 (or 0), meaning that the second stage data are completely used (or ignored). $N = 90$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Banbeta, A., J. van Rosmalen, D. Dejardin, and E. Lesaffre (2019), Modified power prior with multiple historical trials for binary endpoints, *Statistics in Medicine*, *38*(7), 1147–1169.

Bhattacharyya, A. (1946), On a measure of divergence between two multinomial populations, *Sankhyā: the Indian Journal of Statistics*, *7*(4), 401–406.

Chao, Y.-C., H. Trachtman, D. S. Gipson, C. Spino, T. M. Braun, and K. M. Kidwell (2020), Dynamic treatment regimens in small n, sequential, multiple assignment, randomized trials: An application in focal segmental glomerulosclerosis, *Contemporary Clinical Trials*, p. 105989.

Chen, M.-H., J. G. Ibrahim, et al. (2006), The relationship between the power prior and hierarchical models, *Bayesian Analysis*, *1*(3), 551–574.

D'Agati, V. D., F. J. Kaskel, and R. J. Falk (2011), Focal segmental glomerulosclerosis, *New England Journal of Medicine*, *365*(25), 2398–2411.

Duan, Y., K. Ye, and E. P. Smith (2006), Evaluating water quality using power priors to incorporate historical information, *Environmetrics: The Official Journal of the International Environmetrics Society*, *17*(1), 95–106.

Estey, E. H., P. F. Thall, S. Pierce, J. Cortes, M. Beran, H. Kantarjian, M. J. Keating, M. Andreeff, and E. Freireich (1999), Randomized phase ii study of fludarabine+ cytosine arabinoside+ idarubicin±all-trans retinoic acid±granulocyte colony-stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome, *Blood, The Journal of the American Society of Hematology*, *93*(8), 2478–2484.

Gravestock, I., and L. Held (2017), Adaptive power priors with empirical bayes for clinical trials, *Pharmaceutical Statistics*, *16*(5), 349–360.

Henderson, R., P. Ansell, and D. Alshibani (2010), Regret-regression for optimal dynamic treatment regimes, *Biometrics*, *66*, 1192–1201.

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

Hsu, J. C. (1992), The factor analytic approach to simultaneous inference in the general linear model, *Journal of Computational and Graphical Statistics*, *1*(2), 151–168, doi:10.1080/10618600.1992.10477011.

Ibrahim, J. G., M.-H. Chen, et al. (2000), Power prior distributions for regression models, *Statistical Science*, *15*(1), 46–60.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2003), On optimality properties of the power prior, *Journal of the American Statistical Association*, *98*(461), 204–213.

Ibrahim, J. G., M.-H. Chen, Y. Gwon, and F. Chen (2015), The power prior: theory and applications, *Statistics in medicine*, *34*(28), 3724–3749.

Jennison, C., and B. W. Turnbull (1999), *Group sequential methods with applications to clinical trials*, Chapman and Hall/CRC.

Kelleher, S. A., et al. (2017), Optimizing delivery of a behavioral pain intervention in cancer patients using a sequential multiple assignment randomized trial smart, *Contemporary clinical trials*, *57*, 51–57.

Kidwell, K. M., N. J. Seewald, Q. Tran, C. Kasari, and D. Almirall (2017), Design and analysis considerations for comparing dynamic treatment regimens with binary outcomes from sequential multiple assignment randomized trials, *Journal of Applied Statistics*, *0*, 1–24.

Lavori, P. W., and R. Dawson (2000), A design for testing clinical strategies: biased adaptive within-subject randomization, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*(1), 29–38.

Lavori, P. W., and R. Dawson (2004), Dynamic treatment regimes: practical design considerations, *Clinical trials*, *1*(1), 9–20.

Lavori, P. W., R. Dawson, and A. J. Rush (2000), Flexible treatment strategies in chronic disease: clinical and research implications, *Biological psychiatry*, *48*(6), 605–614.

Lei, H., I. Nahum-Shani, K. Lynch, D. Oslin, and S. Murphy (2012), A SMART design for building individualized treatment sequences, *Annual Review of Clinical Psychology*, *8*, 21–48.

Li, Y., and Y. Yuan (2020), PA-CRM: A continuous reassessment method for pediatric phase I oncology trials with concurrent adult trials, *Biometrics*.

Liu, Y., Y. Wang, and D. Zeng (2017), Sequential multiple assignment randomization trials with enrichment design, *Biometrics*, *73*(2), 378–390.

Magirr, D., T. Jaki, and J. Whitehead (2012), A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection, *Biometrika*, *99*(2), 494–501.

Mancl, L. A., and T. A. DeRouen (2001), A covariance estimator for GEE with improved small-sample properties, *Biometrics*, *57*, 126–134.

Micheletti, R. G., C. Pagnoux, R. N. Tamura, P. C. Grayson, C. A. McAlear, R. Borchin, J. P. Krischer, P. A. Merkel, and V. C. R. Consortium (2020), Protocol for a randomized multicenter study for isolated skin vasculitis (ARAMIS) comparing the efficacy of three drugs: azathioprine, colchicine, and dapsone, *Trials*, *21*, 1–9.

Murphy, S. A. (2003), Optimal dynamic treatment regimes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 331–355.

Murphy, S. A. (2005a), An experimental design for the development of adaptive treatment strategies, *Statistics in Medicine*, *24*(10), 1455–1481.

Murphy, S. A. (2005b), A generalization error for Q-learning, *Journal of Machine Learning Research*, *6*, 1073–1097.

Nahum-Shani, I., M. Qian, D. Almirall, W. E. Pelham, B. Gnagy, G. A. Fabiano, J. G. Waxmonsky, J. Yu, and S. A. Murphy (2012), Experimental design and primary data analysis methods for comparing adaptive interventions., *Psychological Methods*, *17*, 457.

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

Nikolakopoulos, S., I. van der Tweel, and K. C. Roes (2018), Dynamic borrowing through empirical power priors that control type I error, *Biometrics*, *74*(3), 874–880.

Orelien, J. G., J. Zhai, R. Morris, and R. Cohn (2002), An approach to performing multiple comparisons with a control in gee models, *Communications in Statistics - Theory and Methods*, *31*(1), 87–105, doi:10.1081/STA-120002436.

Pan, H., Y. Yuan, and J. Xia (2017), A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *66*(5), 979–996.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, in *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326, Springer.

Rosenberg, A. Z., and J. B. Kopp (2017), Focal segmental glomerulosclerosis, *Clinical Journal of the American Society of Nephrology*, *12*(3), 502–517.

Rosner, G. L., and D. A. Berry (1995), A Bayesian group sequential design for a multiple arm randomized clinical trial, *Statistics in Medicine*, *14*(4), 381–394.

Ruppert, A., J. Yin, M. Davidian, A. Tsiatis, J. Byrd, J. Woyach, and S. J. Mandrekar (2019), Application of a sequential multiple assignment randomized trial (smart) design in older patients with chronic lymphocytic leukemia, *Annals of Oncology*, *30*(4), 542–550.

Rush, A. J., et al. (2004), Sequenced treatment alternatives to relieve depression (star* d): rationale and design, *Controlled clinical trials*, *25*(1), 119–142.

Saarela, O., E. Arjas, D. A. Stephens, and E. E. Moodie (2015), Predictive Bayesian inference and dynamic treatment regimes, *Biometrical Journal*, *57*, 941–958.

Shi, H., and G. Yin (2019), Control of Type I error rates in Bayesian sequential designs, *Bayesian Analysis*, *14*(2), 399–425.

Shih, M.-C., and P. W. Lavori (2013), Sequential methods for comparative effectiveness experiments: Point of care clinical trials, *Statistica Sinica*, *23*(4), 1775–1791.

Stallard, N., and T. Friede (2008), A group-sequential design for clinical trials with treatment selection, *Statistics in Medicine*, *27*(29), 6209–6227.

Stallard, N., and S. Todd (2003), Sequential designs for phase III clinical trials incorporating treatment selection, *Statistics in Medicine*, *22*(5), 689–703.

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

Thall, P. F., and J. K. Wathen (2007), Practical bayesian adaptive randomisation in clinical trials, *European Journal of Cancer*, *43*(5), 859–866.

Thall, P. F., L. H. Wooten, C. J. Logothetis, R. E. Millikan, and N. M. Tannir (2007), Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring, *Statistics in Medicine*, *26*, 4687–4702.

van Rosmalen, J., D. Dejardin, Y. van Norden, B. Löwenberg, and E. Lesaffre (2018), Including historical data in the analysis of clinical trials: Is it worth the effort?, *Statistical Methods in Medical Research*, *27*(10), 3167–3182.

Wei, B., T. M. Braun, R. N. Tamura, and K. M. Kidwell (2018), A Bayesian analysis of small n sequential multiple assignment randomized trials (snSMARTs), *Statistics in Medicine*, *37*, 3723–3732.

Williamson, T., M. Eliasziw, and G. H. Fick (2013), Log-binomial models: exploring failed convergence, *Emerging Themes in Epidemiology*, *10*, 14.

Xu, Y., P. Müller, A. S. Wahed, and P. F. Thall (2016), Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times, *Journal of the American Statistical Association*, *111*, 921–950.

Yin, G., N. Chen, and J. Jack Lee (2012), Phase II trial design with Bayesian adaptive randomization and predictive probability, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(2), 219–235.

Zajonc, T. (2012), Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking, *Journal of the American Statistical Association*, *107*, 80–92.

Zhu, H., Q. Yu, and D. E. Mercante (2017), A Bayesian sequential design with binary outcome, *Pharmaceutical Statistics*, *16*(3), 192–200.