Bayesian Methods in Phase I Trials and Small n Sequential Multiple Assignment Randomized Trials

by

Boxian Wei

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

Doctoral Committee:

Associate Professor Kelley M. Kidwell, Chair
Professor Thomas M. Braun
Associate Professor Roy N. Tamura, University of South Florida
Associate Professor Lu Wang

Boxian Wei

boxian@umich.edu

ORCID iD: 0000-0002-6941-0812

ACKNOWLEDGEMENTS

I would like to express my great appreciation and thanks to my committee members Dr. Kelley M. Kidwell, Dr. Thomas M. Braun, Dr. Roy Tamura, and Dr. Lu Wang, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a statistician. Your advice on both research as well as on my career have been priceless. I also want to thank you for letting my defense be an impressive moment, and for your brilliant comments and suggestions, thanks to you. I am deeply grateful for the support from a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1507-31108) for Chapter II and III in this dissertation. Chapter II in its entirety first appeared as *Wei et al.* (2018).

My sincere thanks also goes to Dr. Peter X.K. Song and Dr. Min Zhang, for offering me the research assistant opportunities in their groups and leading me working on diverse exciting projects.

I would also like to thank all of my friends who supported me and encouraged me to strive towards my goal. To the very best of times we had in the last six years at the University of Michigan.

A special thanks to my families. I cannot express how grateful I am to my mother Xiaohong Yang and my father Xiang Wei for all the love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

In this dissertation, we present new methods for Phase I trials and Small n Sequential Multiple Assignment Randomized Trials (snSMARTs), both in a Bayesian framework.

The Bayesian formulation of the Continual Reassessment Method (CRM) is implemented with a one-parameter model describing the association of dose with the probability of dose-limiting toxicity (DLT). Implementation of the CRM requires the user to select two "tuning parameters": (1) the "skeleton," or vector of *a priori* probabilities of DLT for each dose, and (2) the prior standard deviation for the model parameter. Existing methods search the values for each from a range of plausible values through simulation, which is time consuming. Therefore, in the first project, we propose a systematic way of recommending the skeleton and prior standard deviation that avoids simulations. We compare the percentage that each dose level is recommended as the MTD using the proposed approach and existing approaches as comparators. We demonstrate that our approach is computationally faster and maintains a good precision of selecting the MTD in various scenarios.

In the second chapter, we continue trial design in small samples, but change gears to the small-n Sequential Multiple Assignment Randomized Trial (snSMART). In an snSMART, patients are first randomized to one of multiple treatments (stage 1) and patients who respond to their initial treatment continue the same treatment for another stage, while those who fail to respond are re-randomized to one of the remaining treatments (stage 2). The data from both stages are used to estimate the efficacy of three active treatments in the setting of rare disease. Analysis approaches for snSMARTs are limited. Therefore, in the second project, we propose a Bayesian approach that allows for borrowing of information across both stages. Through simulation, we compare the bias, root mean-square error (rMSE),

width and coverage rate of 95% confidence/credible interval (CI) of estimators from of our approach to estimators produced from (a) standard approaches that only use the data from stage 1, and (b) a log-Poisson model using data from both stages whose parameters are estimated via generalized estimating equations. We demonstrate the rMSE and width of 95% CIs of our estimators are smaller than the other approaches in realistic settings, so that the collection and use of stage 2 data in snSMARTs provide improved inference for treatments of rare diseases.

In the previous project, a Bayesian method for estimating the response rate of each individual treatment in a three-arm snSMART demonstrated efficiency gains for a given sample size relative to other existing frequentist approaches. However, these efficiency gains are dependent upon knowing the sample size. Because few sample size calculation methods for snSMARTs exist, in the third project, we propose a Bayesian sample size calculation for an snSMART designed to distinguish the best treatment from the second-best treatment. Although our methods are based on asymptotic approximations, we demonstrate via simulations that our proposed sample size calculation approach produces the desired statistical power, even in small samples. Moreover, our methods produce sample sizes quickly, thereby saving time relative to using simulations to determine the appropriate sample size. We compare our proposed sample size to an existing frequentist method based upon a weighted Z-statistic and demonstrate that the Bayesian method requires far fewer patients than the frequentist method for a study with the same design parameters.

CHAPTER I

Introduction

In the first project, we propose new methods for Phase I trials of a single agent in a Bayesian framework. Phase I clinical trials are first-in-human studies for a new treatment. In Phase I oncology trials, investigators are interested in determining a treatment regimen that is not only safe but also likely to be efficacious in a small group of patients, typically ranging from 20 to 40 patients. A reasonable approach is to specify a target toxicity rate η that should be sufficiently low to indicate safety and sufficiently high to indicate efficacy, where η usually falls in the interval $[0.2, 0.3]$. Toxicity here does not include all types of adverse events but is a dose-limiting toxicity (DLT), which, even though varying among trials, often includes Grade 3 or higher toxicity according to the National Cancer Institute. The largest dose among those examined that leads to DLTs closest to η is usually defined as the maximum tolerated dose (MTD). One of the main goals of Phase I clinical trials in oncology is to establish the MTD from a pre-specified set of dose levels that will be examined further in a Phase II trial for efficacy.

Although numerous approaches for MTD identification in Phase I exist (*Braun*, 2014), our work focuses upon the continual reassessment method (CRM) (*O'Quigley et al.*, 1990). The CRM requires the user to specify *a priori* probabilities of dose-limiting toxicity (DLT) for each dose level, known as the skeleton, and the prior standard deviation for the dose-toxicity model. The skeleton and the prior standard deviation impact the ability of the continual

reassessment method (CRM) to correctly identify the MTD. Thus, the values assigned to the skeleton and prior standard deviation need to be carefully considered, and the process of selecting appropriate skeleton and prior standard deviation ideally occurs with clinical input. Unfortunately, often there is little existing prior data to provide input for appropriate values.

To address the sensitivity of the CRM to the skeleton, *Yin and Yuan* (2009) suggested using Bayesian model averaging techniques. *Lee and Cheung* (2009) proposed developing a single skeleton through prior ideas developed in *Shen and O'Quigley* (1996) and *Cheung and Chappell* (2002) describing consistency properties of the CRM. *Lee and Cheung* (2011) generalized the skeleton developing to the skeleton and prior standard deviation jointly. in *Lee and Cheung* (2011), to figure out a pair of values of skeleton and prior standard deviation for any setting of Phase I trial requires thousands of simulated trials, which is time consuming. Moreover, the work of *Shen and O'Quigley* (1996) and *Cheung and Chappell* (2002) were motivated by one of the scenarios in the simulation study of *O'Quigley et al.* (1990) where the CRM did a poor job of correctly identifying the MTD, even in large samples. Therefore, we propose a systematic approach to determine values of skeleton and prior standard deviation that will be more time efficient, improve the ability of the CRM to identify the MTD in this setting, and maintain good MTD identification in other settings.

In the second and third projects, we present a Bayesian analysis of Small n Sequential Multiple Assignment Randomized Trials (snSMARTs) and provide a sample size calculation. snSMARTs are motivated by ARAMIS (A RAndomized Multi-center study for Isolated Skin vasculitis trial), the design of which mimics the SMART design in metastatic renal cancer *Thall et al.* (2007); *Thall* (2016). ARAMIS (NCT02939573) is a multi-national trial to find the best treatment for patients with skin vasculitis.

We propose methodological improvements for ARAMIS in the setting of small sample sizes rare diseases. Standard SMARTs have not been previously implemented in rare diseases and

estimating or comparison of DTRs with large sample sizes. Due to the small sample size and treatments considered, a snSMART differs from a standard SMART in terms of the research questions of interest and analytic methods necessary to address these questions. Thus, our proposed methods differ from those for standard SMARTs because we propose to use the multi-stage randomization feature to borrow information across different stages to make conclusions about individual treatments. The snSMART design may improve recruitment due to continuing treatment for those who respond and switching treatment for those who do not, and may be more efficient than current trial designs for rare diseases.

Our overall goal of chapters 2 and 3 is to develop an analytical method that efficiently estimates the individual treatment efficacies for 3 treatments and calculations for sample size that compares the best 2 treatments in a snSMART with 3 or more treatment arms with sufficient statistical power and coverage rate. By using the efficient proposed methods, smaller clinical trials can provide rigorous evidence of treatment effectiveness to expand the design options for rare disease comparative effectiveness trials.

CHAPTER II

A Systematic Approach to Skeleton and Prior Variance Selection in the Continual Reassessment Method

2.1 Introduction

In Phase I oncology clinical trials, researchers seek to model the association between doses of a drug and the probability of a dose limiting toxicity (DLT), which is a toxicity that prevents further administration of the agent at that dose level. The end goal of the trial is to identify the maximum tolerated dose (MTD), which is defined as the largest dose among those examined that leads to DLTs in an acceptable proportion of subjects. This acceptable proportion is referred to as the target DLT rate and is usually in the interval [0.20, 0.30]. Although numerous approaches for Phase I designs exist (*Braun*, 2014), our work focuses upon the continual reassessment method (CRM) (*O'Quigley et al.*, 1990). An excellent tutorial of the CRM was published by *Garrett-Mayer* (2006).

The CRM design begins by assigning a cohort of subject(s) to a dose and following each member of the cohort for the occurrence of DLT. Although the original formulation of the CRM recommended the dose for the first cohort should be the one believed *a priori* to be the MTD, later modifications to the CRM suggested for subject safety reasons that the starting dose should be the lowest dose (*Goodman et al.*, 1995; *Faries*, 1994; *Moller*, 1995). This latter convention has since been adopted in most applications of the CRM.

The CRM quantifies the association of dose and DLT rate through a one-parameter model. A prior distribution is placed on the parameter and the observed proportion of DLTs at each dose is used to compute the posterior mean of the parameter. The posterior mean is then inserted in the CRM model to compute the posterior estimate of DLT probability for each dose. The next cohort of subjects is then assigned to the dose with the posterior probability of DLT closest to the target DLT rate. The posterior estimates of the DLT probabilities are continually updated with each successive cohort and enrollment stops once the desired sample size is reached.

One important aspect for the CRM is determining the magnitude of the prior standard deviation for the model parameter, as this will determine how informative the prior distribution will be. We cannot use a traditionally non-informative variance because of the small sample sizes seen in most Phase I trials. Instead, the prior variance must be selected such that the prior provides direction for the posterior early in the trial when little data has been collected, yet allows the data to drive decisions later in the trial once sufficient data have been collected. *Zhang et al.* (2013) developed approaches for adaptively changing the prior variance during a Phase I trial that include increasing the prior standard deviation gradually during the trial, or through formal hypothesis testing. Although *Chevret* (1993) suggested that the CRM is robust to the choice of prior as long as the prior is uninformative, *Lee and Cheung* (2011) suggests that a larger value of the prior standard deviation is not necessarily uninformative and an uninformative prior for a given model may not be uninformative for another model. Thus, *Lee and Cheung* (2011) proposed methods to jointly select the prior standard deviation and a set of *a priori* DLT rates, known as the skeleton, which also plays a role in the performance of the CRM.

The one-parameter model in the CRM is often some variant of a regression model that includes a linear combination of the model parameter and a numeric value for each dose. However, these dose values are typically not the actual clinical values of the doses, which can often vary by factors of 100. Again, due to the small sample sizes used in Phase I trials, we

need to use dose values that are less variable than the clinical values, ones that will improve the fit of the model. The skeleton allows us to determine the dose values, as the skeleton contains the *a priori* estimate of the DLT rate for each dose. Replacing the model parameter by its prior mean and equating the model to each value in the skeleton allows us to solve for a dose value that we can use with each dose. The skeleton will have an impact on the performance of the CRM because the fit of the model will diminish as the true unknown DLT rates deviate from the skeleton. Thus, the skeleton needs to be carefully selected, and the process of selecting an appropriate skeleton ideally occurs with clinical input. Unfortunately, often there is little existing prior data to provide input for an appropriate skeleton.

To address the sensitivity of the CRM to the skeleton, *Yin and Yuan* (2009) suggested using Bayesian model averaging techniques. In this approach, several skeletons are proposed at the beginning of the trial, with equal prior preference (weight) given to each skeleton. As the clinical trial continually accrues more data, the posterior weights given to each skeleton change to reflect how they fit the accrued data. These updated weights are then incorporated into the posterior estimates of DLT for each dose so that the impact of skeletons that deviate far from the true DLT rates have little influence on estimation of the MTD. However, the approaches that determine the number of skeletons and the values contained in each skeleton are currently unknown.

As an alternative to Bayesian model averaging, we could conceive of an approach where the performance of the CRM is examined via simulation using every possible configuration of true DLT rates with every possible configuration of skeleton values. However, there is an uncountable number of such combinations, and even examining an exhaustive finite grid of combinations would still prove to be computationally time-consuming. To reduce the dimensionality of this problem, *Lee and Cheung* (2009) proposed developing a single skeleton through prior ideas developed in *Shen and O'Quigley* (1996) and *Cheung and Chappell* (2002), which was then extended in *Lee and Cheung* (2011) to also incorporate selection of the prior variance.

However, the methods of *Lee and Cheung* (2011) are based on intensive simulations, which is time-consuming because dose assignments and DLT outcomes are simulated subject-by-subject due to the adaptive process of the CRM. We propose an alternative approach for calibrating the skeleton and prior variance. Our methods are motivated by a setting examined in *O'Quigley et al.* (1990), where the CRM did a poor job of correctly identifying the MTD, even in large samples. Using the weighted *expected* outcome of DLT, our process of calibration is less time-consuming than the simulation based calibration approaches and the CRM using our skeleton and prior standard deviation performs as other approaches. We describe the details of our approach in Section 2.2 and examine the performance of our methods via simulation in Section 2.3. We close with a discussion in Section 3.4.

2.2 Methods

2.2.1 CRM Background

The design of a Phase I trial has three components: (i) J dose levels of an investigational treatment, (ii) N, the total number of subjects to enroll, and (iii) η, the target DLT rate. A one-parameter model, denoted as $p_j = f(d_j; \beta)$, is used to associate dose, $j = 1, 2, \ldots, J$, with its corresponding DLT rate, p_j, in the CRM, in which d_j is a numerical value given to dose j. The two most-commonly used models in the CRM are the power (or empiric) model, $p_j = d_j^{exp(\beta)}$, and the logistic model, $log(p_j/[1-p_j]) = 3 + exp(\beta)d_j$, in which β is the unknown model parameter.

We let π_j denote an *a priori* DLT rate for dose j. Collectively the vector, $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_J\}$, is called the skeleton for the CRM. As suggested earlier, the value of d_j is not the actual clinical dose, but is a value solved using the assumed model, $f(d_j; \beta)$, once $\boldsymbol{\pi}$ is specified. Setting β equal to its prior mean, denoted as μ, we use π_j and $f(d_j; \beta)$ to solve for d_j, i.e. $d_j = \pi_j/exp(\mu)$ for the power model and $d_j = [log(\pi_j/[1-\pi_j]) - 3]/exp(\mu)$ for the logistic

model. In the Bayesian formulation of the CRM, a prior distribution, $g(\beta)$ is assumed for β, such that β has support on the real line. We emphasize that in our model we exponentiate the value of β because we assume that the probability of DLT must increase with the dose. If the support of $g(\beta)$ is already constrained to the positive line, then β can be used directly in the model rather than $exp(\beta)$.

The dose assignment for each cohort is determined adaptively based upon the dose assignments and DLT outcomes observed on previously enrolled subjects. Specifically, if we have enrolled k subjects, in which subject $i = 1, 2, \ldots, k$ has dose assignment $d_{[i]}$, which is one of the values in $\{d_1, d_2, ...d_J\}$, and a binary indicator of DLT Y_i, we can compute the posterior mean DLT rate for dose j, $\hat{p}_j$, as:

$$\hat{p}_j = \int_{-\infty}^{\infty} f(d_j; \beta) \frac{L(\beta|\mathbf{Y})g(\beta)}{\int_{-\infty}^{\infty} L(\beta|\mathbf{Y})g(\beta)d\beta} d\beta, \tag{2.1}$$

where

$$L(\beta|\mathbf{Y}) = \prod_{i=1}^{k} f(d_{[i]}; \beta)^{Y_i}[1 - f(d_{[i]}; \beta)]^{1-Y_i} \tag{2.2}$$

is the likelihood for β and $\mathbf{Y} = \{Y_1, Y_2, ...Y_k\}$.

The dose level, j^*, recommended for the next enrolled cohort is the dose with the estimated DLT rate closest to η. After observing the DLT outcomes for this most-recently enrolled cohort, we use Equation 2.1 to update $\hat{p}_j$ and update which dose has a DLT rate closest to the targeted DLT rate. At the end of the study, we determine the MTD as the index j^* based upon the data from all N subjects.

2.2.2 Existing Methods for Calibrating π and σ

For a given vector of true DLT rates $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$, in which α_k is the true DLT rate for dose k, the performance of the CRM is usually quantified by how frequently in repeated simulations that the CRM correctly identifies the MTD at the end of the study.

This quantity is often referred to as the probability of correct selection (PCS). Of course, the PCS is also a function of the assumed model, $f(d_j;\beta)$, the skeleton, $\boldsymbol{\pi}$, and the prior standard deviation of the model parameter, σ. Thus, in order to identify a "best" $\boldsymbol{\pi}$, one could envision identifying a candidate set of values for $\boldsymbol{\pi}$ with a fixed value of σ, repeating simulations over several values of $\boldsymbol{\alpha}$, and choosing the $\boldsymbol{\pi}$ that has the highest average PCS among all the possible values of $\boldsymbol{\alpha}$.

Lee and Cheung (2009) provided a systematic process for selecting both the set of skeleton values, $\boldsymbol{\pi}$, and the set of true DLT rates, $\boldsymbol{\alpha}$. The set of skeletons is selected through the concept of an indifference interval, which was first proposed by *Cheung and Chappell* (2002). For a given target DLT rate η, the indifference interval is an interval of DLT rates such that any dose with a true DLT rate in that interval could be selected as the MTD at the end of the study. *Cheung and Chappell* (2002) showed that there is a simple correspondence between the values in $\boldsymbol{\pi}$, and δ, the half-width of the desired indifference interval. Thus, *Lee and Cheung* (2009) proposed defining a sequence of values for δ that start at 0.01 and increase in increments of 0.01 up to a pre-defined upper bound; each of those values of δ would produce a candidate value for $\boldsymbol{\pi}$ to examine.

The set of true DLT rates is based upon what *Lee and Cheung* (2009) call a "plateau configuration" consisting of J vectors. Each vector $j = 1, 2, \ldots J$ assumes dose j is the MTD and has a DLT rate equal to the target DLT rate η. All other doses have DLT rates that vary around dose j through an odds ratio Ψ, such that all doses below the MTD have a DLT rate equal to $\eta/[\Psi + \eta(1-\Psi)]$ and all doses above the MTD have a DLT rate equal to $\eta\Psi/[1-\eta(1-\Psi)]$. Thus, there is a change in the DLT rate directly above and below the MTD, but then the DLT rates plateau in either direction with the remaining doses.

With σ set equal to $\sqrt{1.34}$, which is a value used in other publications (*Lee and Cheung*, 2009; *O'Quigley and Shen*, 1996), each skeleton is examined with each of the J vectors of true DLT rates in a large number, i.e. perhaps 1,000, of simulated trials. Each simulation

is summarized by the resulting PCS, with a final average PCS computed for each skeleton over the J vectors of true DLT rates. The "best" skeleton is defined as the one with largest average PCS. *Lee and Cheung* (2011) then expanded this concept to also consider a set of potential values for σ, thereby developing a single approach for identifying both $\boldsymbol{\pi}$ and σ.

Note that the calibration of $\boldsymbol{\pi}$ and σ depends on the choice of the odds ratio Ψ, which quantifies the spread of DLT rates near the MTD. Values of $\Psi = 2, 3$, and 5 were considered in *Lee and Cheung* (2009) as values that produced DLT rates most likely to be seen in practical applications. However, in *Lee and Cheung* (2011), where both $\boldsymbol{\pi}$ and σ were calibrated, only the value of $\Psi = 2$ was considered due to the amount of computing time needed for simulations. We propose an alternative approach to jointly calibrate $\boldsymbol{\pi}$ and σ that is not based on simulations, and hence more time efficient compared to *Lee and Cheung* (2011). Thus, in our proposed methods, we will continue to define our true DLT rates through the plateau configuration, but expanded to use odds ratios of $\Psi = 2, 3$, and 5.

2.2.3 Proposed Methods for Calibrating $\boldsymbol{\pi}$ and σ

To identify the calibration set of $\boldsymbol{\pi}$ and σ, we consider a class of candidate values for $\boldsymbol{\pi}$ that are (i) equally spaced by a positive value denoted as Δ, where Δ takes values in $\{0.01, 0.02, ...\}$, and (ii) the *a priori* MTD has a DLT rate equal to η. Naturally, the resulting skeleton must also contain sensible values of DLT rates; we only consider skeletons in which the DLT rate of the lowest dose is no more than 0.10 and the DLT rate of the highest dose is no more than 0.90. The candidate values of σ are a sequence from 1 to 4 with an increments of 0.20. As stated earlier, *Lee and Cheung* (2009) and *Lee and Cheung* (2011) use simulations to estimate the average PCS value for every combination of $\boldsymbol{\pi}$ and σ and select the combination with the largest average PCS. Given that our calibration set of $\boldsymbol{\pi}$ and σ is larger than *Lee and Cheung* (2011), the number of simulations required for every combination of $\boldsymbol{\pi}$ and σ is prohibitively time-consuming. Thus, our goal is to develop a

metric to find the best combination of $\boldsymbol{\pi}$ and σ without using simulations.

As a first step, we adopt the methods of *Cheung and Chappell* (2002), who demonstrated that the support of the model parameter β can be divided into J non-overlapping intervals $H_1, H_2, \ldots H_J$. Assuming that β has support on the real line, the interval H_j is equal to (b_j, b_{j+1}), where $b_1 = -\infty$, $b_{J+1} = \infty$, and $b_j, j = 2, 3, \ldots J$ solves $f(d_{j-1}; b_j) + f(d_j; b_j) = 2\eta$. Essentially, interval H_j contains the values of β that lead to dose j having a modeled DLT rate closest to the targeted DLT rate η.

When we have collected data on k subjects, we will have a posterior distribution for β and we define $\omega_{j,k}$ to be the amount of posterior mass in H_j, such that $\sum_j \omega_{j,k} = 1$. Thus, ω_{j0} is the amount of *a priori* mass contained in H_j and ω_{jN} is the amount of posterior mass at the end of the study and we have data for all N subjects.

However, the posterior distribution of β is dependent upon the data collected during the study, which again, because of the adaptive nature of the CRM, would have to be determined via simulation. To avoid simulations, we will choose one outcome at every dose for each subject, similar to the idea of the non-parametric optimal design (NPOD) methodology of *O'Quigley et al.* (2002). In the NPOD, each subject is given a binary indicator of DLT at each dose, so that the data now contain J times more information than an actual trial, and the NPOD supplies an upper bound for the performance of the CRM.

However, in our methods, each subject does not have a binary indicator of DLT for each dose. Instead, each subject is given the *expected* outcome of DLT for each dose, which is simply the vector of true DLT rates, $\boldsymbol{\alpha}$. To account for the fact that each subject has J outcomes rather than one outcome, we incorporate weights into the likelihood used by the CRM. When the subject $k+1$ enters the trial and we have data for k subjects, the weights for subject $k+1$ are $\omega_{1,k}, \omega_{2,k}, \ldots \omega_{J,k}$, which are the amount of posterior mass of β in H_j

described earlier. Explicitly, the likelihood in Equation (2.2) changes to

$$\begin{aligned} L(\beta|\boldsymbol{\alpha}) &= \prod_{i=1}^{k+1}\prod_{j=1}^{J} f(d_j;\beta)^{\omega_{j,(i-1)}\alpha_j}[1-f(d_j;\beta)]^{\omega_{j,(i-1)}(1-\alpha_j)} \\ &= \prod_{i=1}^{k+1}\prod_{j=1}^{J} \left\{f(d_j;\beta)^{\alpha_j}[1-f(d_j;\beta)]^{(1-\alpha_j)}\right\}^{\omega_{j,(i-1)}} \end{aligned} \tag{2.3}$$

Thus, in this likelihood, every subject has the same vector of outcomes, quantified by $\boldsymbol{\alpha}$, but each subject has a different set of weights.

The posterior distribution of β is recursively updated with each new subject, and after the data collection of all N subjects, the values $\omega_{1N}, \omega_{2N}, \ldots, \omega_{JN}$ quantify the posterior probability that each dose is the MTD. Better performance of the CRM is indicated by having large ω_{jN} at the true MTD and smaller values of ω_{jN} at the other doses. Thus, if we then define $q_j = 1$ if dose j is the true MTD and $q_j = 0$ otherwise, we can compute the deviance

$$\mathcal{D} = \frac{1}{J}\sum_{j=1}^{J}(\omega_{jN} - q_j)^2. \tag{2.4}$$

A smaller deviance is associated with a higher precision of correctly identifying the MTD. Thus, for a given skeleton $\boldsymbol{\pi}$ and a given prior standard deviation σ, we have a deviance for every vector of true DLT rates, $\boldsymbol{\alpha}$. We recommend the pair of $\boldsymbol{\pi}$ and σ that has the smallest median deviance over all the possible values of $\boldsymbol{\alpha}$ examined.

Hence, the skeleton and prior standard deviation selection process is as follows:

1. Determine J, the number of doses for investigation, N, the number of subjects to enroll, and η, the target DLT rate.
2. Generate, $\boldsymbol{\alpha}$, the set of vectors of true DLT rates, under the plateau configuration with $\Psi = 2, 3$, and 5.
3. Generate the candidate values of skeleton, $\boldsymbol{\pi}$, and prior standard deviation, σ, using the criteria defined in the first paragraph of Section 2.2.3.

4. Over all true DLT rates in $\boldsymbol{\alpha}$, calculate the median deviance for each pair of candidate values of $\boldsymbol{\pi}$ and σ using the likelihood proposed in Equation (2.3) and deviance measure defined in Equation (2.4).

5. The pair of $\boldsymbol{\pi}$ and σ associated with the smallest median deviance are the values of $\boldsymbol{\pi}$ and σ recommended for use in the actual Phase I trial.

To illustrate this process, we consider a Phase I clinical trial in which six doses are being studied, 25 subjects will be enrolled and η is equal to 0.20. We use a power model with parameter, β, which has a prior normal distribution with mean 0 and standard deviation σ. In this setting, we will have 18 vectors of true DLT rates, 35 skeletons under consideration, and 20 values of $\sigma \in \{1.0, 1.2, \ldots, 4.0\}$. Let us now consider one specific combination in which the true DLT rates $\boldsymbol{\alpha} = \{0.20, 0.33, 0.33, 0.33, 0.33, \text{ and } 0.33\}$, the skeleton $\boldsymbol{\pi} = \{0.20, 0.21, 0.22, 0.23, 0.24, \text{ and } 0.25\}$, and, $\sigma = 1$.

This information then allows us to divide the support of β into the six intervals $H_1 = (-\infty, 0.02)$, $H_2 = (0.02, 0.05)$, $H_3 = (0.05, 0.08)$, $H_4 = (0.08, 0.11)$, $H_5 = (0.11, 0.13)$, and $H_6 = (0.13, \infty)$, which have respective prior masses of 0.51, 0.01, 0.01, 0.01, 0.01, and 0.45. These prior masses are then used as weights for the first subject in Equation (2.3), which is then used to compute the respective posterior mass of each H_j as 0.56, 0.02, 0.02, 0.01, 0.01, and 0.38. The posterior masses are the weights given to the second subject in Equation (2.3), which is used to compute weights for the third subject. We repeat this procedure until all 25 subjects are incorporate and produce final posterior probabilities $\{\omega_{25,1}, \ldots, \omega_{25,6}\} = \{0.69, 0.04, 0.04, 0.04, 0.03, 0.16\}$. Given that the vector of true DLT rates sets the first dose as the MTD, we have $q_1 = 1$ and $q_2, \ldots, q_6$ are each equal to 0. According to Equation 2.4, the deviance is equal to 0.14.

For this given pair of $\boldsymbol{\pi}$ and σ, we repeat this process for all 18 vectors of true DLT rates and compute the median deviance. We then recursively compute a median deviance for every possible pair of $\boldsymbol{\pi}$ and σ. The smallest median deviance occurs when $\boldsymbol{\pi} = \{0.20, 0.33, 0.46,$

0.59, 0.72, 0.85} and $\sigma = 3$, which are the values we recommend being used in the actual Phase I clinical trial.

2.3 Simulations

We compare our proposed skeleton and prior standard deviation calibration, denoted as Wei, to the method proposed in *Lee and Cheung* (2011), in settings in which the CRM uses the power model, there are five scenarios of true DLT rates for $J = 6$ doses, the target DLT $\eta = 0.20$, and total sample size $N = 25$. For $J = 6$, $\eta = 0.20$, and $N = 25$, our approach suggests the optimal skeleton $\boldsymbol{\pi}$ being equal to $\{0.20, 0.33, 0.46, 0.59, 0.72, 0.85\}$ and prior standard deviation σ being equal to 3.

Lee and Cheung (2011) proposes two algorithms for the skeleton and prior standard deviation calibration, denoted as LC11A and LC11B respectively. LC11A suggests $\boldsymbol{\pi}$ being equal to $\{0.05, 0.11, 0.20, 0.31, 0.42, 0.53\}$ and σ being equal to 0.69, while LC11B suggests $\boldsymbol{\pi}$ being equal to $\{0.07, 0.13, 0.20, 0.29, 0.38, 0.47\}$ and σ being equal to 0.55. *Lee and Cheung* (2011) compares LC11A and LC11B with other two approaches proposed in *Lee and Cheung* (2009) and *O'Quigley et al.* (1990), respectively, denoted as LC09 and OQ90, respectively. LC09 suggests $\boldsymbol{\pi}$ being equal to $\{0.01, 0.07, 0.20, 0.38, 0.56, 0.71\}$ and σ being equal to 1.16. OQ90 suggests $\boldsymbol{\pi}$ being equal to $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.70\}$ and uses a different prior settings, where β has an exponential prior with rate 1. Using simulations with $2,000$ realizations, we compare the percentage with which each dose level is recommended as the MTD (the larger the better) using our proposed approach, and all other four approaches as comparators. We also examine our approach without weighting, denoted as Wei2. We change the candidate values for σ in Wei2 from the original values into $\sigma \in \{1.0, 1.2, \ldots, 10\}$ because the selected σ is equal to the upper bound, 4, in original settings. In Wei2, the optimal skeleton is $\{0.20, 0.26, 0.32, 0.38, 0.44, 0.50\}$ and prior standard deviation σ is 7.5.The results are shown in Table 2.1.

Table 2.1: The simulated percentage with which each dose level is recommended as the MTD, where the number of doses for investigation is $J = 6$, the target DLT rate is $\eta = 0.20$, and the number of subjects $n = 25$. The number of simulation is $2,000$. The true MTD in each scenario is in italic. The percentage that the true MTD is recommended as the MTD is in bold. OQ90 represents the original setting of the CRM in *O'Quigley et al.* (1990); LC09 represents the CRM using the calibration of $\boldsymbol{\pi}$ with $\sigma = \sqrt{1.34}$ in *Lee and Cheung* (2009); LC11A and LC11B represent the CRM using two different algorithms for the calibration of $\boldsymbol{\pi}$ and σ proposed in *Lee and Cheung* (2011), respectively. Wei2 represents the CRM using proposed algorithm without weighting the expected outcome.

		percentage of selection as MTD					
Scenario	Method	1	2	3	4	5	6
1	DLT rate	0.05	0.10	*0.20*	0.30	0.50	0.70
	OQ90	0.01	0.18	**0.50**	0.29	0.01	0.00
	LC09	0.02	0.21	**0.53**	0.24	0.01	0.00
	LC11A	0.02	0.23	**0.52**	0.23	0.01	0.00
	LC11B	0.02	0.23	**0.51**	0.23	0.01	0.00
	Wei	0.03	0.25	**0.47**	0.24	0.01	0.00
	Wei2	0.06	0.27	**0.42**	0.21	0.03	0.00
2	DLT rate	*0.30*	0.40	0.52	0.61	0.76	0.87
	OQ90	**0.91**	0.08	0.01	0.00	0.00	0.00
	LC09	**0.89**	0.11	0.01	0.00	0.00	0.00
	LC11A	**0.92**	0.08	0.00	0.00	0.00	0.00
	LC11B	**0.92**	0.08	0.00	0.00	0.00	0.00
	Wei	**0.95**	0.05	0.00	0.00	0.00	0.00
	Wei2	**0.97**	0.03	0.00	0.00	0.00	0.00
3	DLT rate	0.05	0.06	0.08	0.11	*0.19*	0.34
	OQ90	0.00	0.01	0.05	0.32	**0.57**	0.06
	LC09	0.00	0.01	0.08	0.29	**0.49**	0.13
	LC11A	0.00	0.01	0.08	0.29	**0.48**	0.14
	LC11B	0.00	0.01	0.09	0.32	**0.44**	0.14
	Wei	0.01	0.02	0.07	0.24	**0.51**	0.16
	Wei2	0.00	0.03	0.09	0.25	**0.37**	0.26
4	DLT rate	0.06	0.08	0.12	*0.18*	0.40	0.71
	OQ90	0.00	0.04	0.22	**0.59**	0.14	0.00
	LC09	0.01	0.06	0.26	**0.57**	0.11	0.00
	LC11A	0.01	0.06	0.27	**0.54**	0.13	0.00
	LC11B	0.01	0.06	0.27	**0.55**	0.12	0.00
	Wei	0.02	0.08	0.24	**0.54**	0.12	0.00
	Wei2	0.03	0.11	0.28	**0.42**	0.16	0.00
5	DLT rate	0.00	0.00	0.03	0.05	0.11	*0.22*
	OQ90	0.00	0.00	0.00	0.07	0.64	**0.29**
	LC09	0.00	0.00	0.01	0.09	0.42	**0.49**
	LC11A	0.00	0.00	0.00	0.07	0.42	**0.52**
	LC11B	0.00	0.00	0.00	0.08	0.42	**0.49**
	Wei	0.00	0.00	0.00	0.05	0.37	**0.57**
	Wei2	0.00	0.00	0.01	0.06	0.24	**0.70**

In the first scenario, where the MTD is the third dose, our proposed $\boldsymbol{\pi}$ and σ leads to the proportion for selecting the third dose (the MTD) of 47%, which is close to the MTD selections from other four approaches. In the second scenario, where the MTD is the first dose, all approaches provide high chances for selecting the MTD. Our approach shares the highest proportion for selecting the MTD of 95% with both LC11A and LC11B. In the third scenario, where the MTD is the fifth dose, our approach yields the second highest proportion for selecting the MTD of 51%. In the fourth scenario, where the MTD is the fourth dose, our proposed approach leads to the proportion for selecting the MTD of 54% which is close to the MTD selections from LC11A and LC11B. The last scenario is originally from *O'Quigley et al.* (1990) which is the scenario that motivates our research. The MTD is the last dose. With the original setting in *O'Quigley et al.* (1990), the proportion of selecting the sixth dose (the MTD) was just 29%, while the proportion of selecting one dose lower than the MTD is 53%. This means the CRM will be more likely to incorrectly identify the MTD if the DLT rates for the six doses are as those in the last scenario. Our method and the LC11A increase the proportion of MTD selection to 57% and 52%, respectively. Overall, the averaged proportion of the MTD selection is 61% of the five scenarios for our proposed approach, which is the highest compared to all other four comparator approaches. Our approach without weighting only outperforms other four comparator approaches in Scenarios 2 and 5, where the MTDs are at the edges. The recommended σ is 7.5, which is a large value for prior standard deviance, that leads a U-shaped *a priori* distribution for the model-based MTD (*Lee and Cheung*, 2011) and hence the CRM is more likely to select the doses at the edges as the MTDs.

2.4 Concluding Remarks

Our method is motivated by the last scenario in *O'Quigley et al.* (1990) where the CRM as originally proposed does not perform well. We simultaneously consider different scenar-

ios of true DLT rates and use them as data after properly weighting for the skeleton and prior standard deviation calibration. The proposed method yields better MTD selection in this scenario and similarly good MTD selection in other scenarios, but with much less computation time.

Our idea of recommending an optimal skeleton and prior standard deviation given the clinical trial design (number of doses involved, sample size, and target DLT rates) is selecting the skeleton and prior standard deviation that lead to the highest median deviance of MTD selection on a set of systematically created hypothetical true DLT rates when applied in the CRM. The simulation results suggests that the proposed recommendation of the skeleton and prior standard deviation enhance the performance of CRM in handling the motivating scenario and also maintain a good performance on MTD selection in different scenarios.

In estimating the chance for each dose to be the MTD, we apply a weighted expected outcome as a Bernoulli outcome to the Bernoulli likelihood, which results Equation 2.3. Though the expected outcome of having a DLT after weighting is a proportion instead of a Bernoulli outcome, the rationality of Equation 2.3 is explained via a quasi-Bayesian likelihood (*Chernozhukov*, 2007) approach. Weighting the expected outcome of DLT for investigational doses with adaptive weights is in fact reversely applying a propensity score (*Rosenbaum*, 1987) for the adaptive dose-assignment for each subject using the CRM. In Phase I clinical trials using the CRM, each subject is adaptively assigned to one dose that is most likely to be the MTD. Our goal is to calibrating the parameters of CRM using data from this adaptive process. We can not directly apply the population level expected DLT outcomes for all doses from one subject because this is not consistent with the dose assignment process and hence leads to potentially less optimal operating characteristics. Further simulation studies support the expectations that the calibrations are less optimal without using weights.

Though we have focused our simulations using the power model for the CRM, the proposed

method for the skeleton and prior standard deviation calibration is general. For the logistic dose-toxicity model, the settings for the true DLT rates, the construction of the candidate value of the skeleton and prior standard deviance are the same as the power model. The only difference is to use the logistic dose-toxicity model, instead of the power model, in the proposed likelihood derivation for the model parameter and in the computation of the weights. We also assume that the first subject entering the trial is assigned to the middle dose for the convenience of the comparison to other methods in our simulations. In real applications, for safety considerations, a trial may start with the lowest dose. Our approach applies for any starting dose. An R package to implement the skeleton and standard deviation recommendations for any user specified design of Phase I clinical trials is under development.

CHAPTER III

A Bayesian Analysis of Small n Sequential Multiple Assignment Randomized Trials (snSMARTs)

3.1 Introduction

The Orphan Drug Act defines rare diseases as disorders affecting fewer than 200,000 individuals in the United States *107th Congress* (2002). More than 8,000 recognized rare diseases affect almost 30 million individuals and their families in the United States *Griggs et al.* (2009). Identifying optimal treatment options for patients living with rare diseases is challenging due to the low number of individuals affected. Randomized clinical trials (RCTs) are generally regarded as providing the strongest scientific evidence for the efficacy of a treatment. RCTs attempt to minimize bias and balance confounders across treatments by employing randomization *Levin* (2007). However, confirmatory RCTs often require a large number of subjects, which is difficult to attain in rare disease trials. Thus, RCTs studying treatment for rare diseases commonly have reduced power compared to studies of non-rare diseases. As a result, rare disease trials are more likely to be single arm (63.0% vs. 29.6% for non-rare disease trials) and non-randomized (64.5% vs. 36.1% for non-rare disease trials) *Bell and Smith* (2014). Small sample trials of rare diseases that are randomized and multi-arm are most likely crossover, n-of-1, or adaptive designs *Gupta et al.* (2011).

There are disadvantages of the trial designs currently used in the rare disease landscape.

For example, single arm studies are employed when the objective of the trial is to obtain preliminary evidence of the treatment efficacy and to collect additional safety data. As a result, single arm trials are not generally used to confirm efficacy *Evans* (2010). In a crossover study, all subjects receive all experimental treatments. By design, each subject is their own control so that confounding and between-subject variance is reduced, leading to the need of fewer subjects than a standard RCT. However, a treatment effect may extend beyond the phase of the study and affect further treatment (i.e., a carryover effect). To ensure carryover effects are not a problem, a long time period may be needed between treatments, which could increase the trial duration and patient dropout. Moreover, since all participants receive a sequence of different treatments, the treatments may expose participants to additional toxicities than a standard RCT or switch participants from an efficacious treatment to a non-efficacious treatment. These challenges have inspired the design of alternative crossover trials so that participants are less likely to drop out of the study *Makubate and Senn* (2010).

An n-of-1 trial is conducted in a single participant with multiple crossover treatment assignments. In an n-of-1 trial, a participant is their own control so that confounding is reduced and the data can suggest which treatment is optimal for the participant. However, an n-of-1 trial usually requires multiple crossover treatment assignments to defend against the effect of treatment across time, measurement error, and error from the participant's condition differing across time. The multiple crossover treatment assignments potentially prolong the duration of the trial which may be burdensome for the participant and requires a well developed trial protocol to keep the participant engaged.

Adaptive designs allow for design parameters, such as the sample size, randomization fraction, population recruited, or doses, to be altered during the trial after interim data evaluation *Evans* (2010). The adaptiveness may reduce the number of subjects recruited to inferior treatment, increase efficiency, improve recruitment and take advantage of accumulating data to enable early stopping of the trial. Alternatively, an adaptive trial is often more complex to design and analyze than other standard clinical trial designs and is susceptible to bias

due to temporal drift in participant characteristics.

Although the designs described above may be useful to study treatments in rare diseases, many have called for more innovative trial designs *Gupta et al.* (2011). Here, we propose design and methodological improvements for a small n sequential multiple assignment randomized trial, the snSMART *Tamura et al.* (2016). An snSMART is an application of a SMART design *Lavori and Dawson* (2000); *Murphy* (2005); *Dawson and Lavori* (2011) in small samples, particularly for rare disease research. In a two-stage snSMART design, patients are randomized at baseline between a number of treatments and re-randomized at some timepoint according to their response to their initial treatment. Compared to a traditional multi-stage design, such as a crossover design, the snSMART is attractive because it allows participants who benefit from their treatment to continue to receive that treatment and who do not benefit from the treatment to switch to another treatment. Hence, an snSMART design may help to improve participant recuitment and retention. However, analytic methods for an snSMART are not fully established, so that the efficiency gains of an snSMART design compared to other designs in rare disease research have not yet been confirmed.

We want to emphasize the similarities and differences between an snSMART and SMART. In both snSMARTs and SMARTs, patients may be sequentially randomized to treatments where second-stage treatment may depend on response to first-stage treatment. However, the primary aim of an snSMART and a SMART differ. The primary aim of an snSMART is to compare treatments by pooling data from all stages to find one superior treatment. In contrast, the primary aim of a SMART is to develop effective dynamic treatment regimes *Robins* (1986); *Murphy* (2003) that define the personalized treatment guidelines consisting of a first stage treatment followed by a second stage treatment for patients.

Our methods are motivated by the ARAMIS (A RAndomized Multi-center study for Isolated Skin vasculitis trial), the design of which mimics the SMART design in metastatic renal

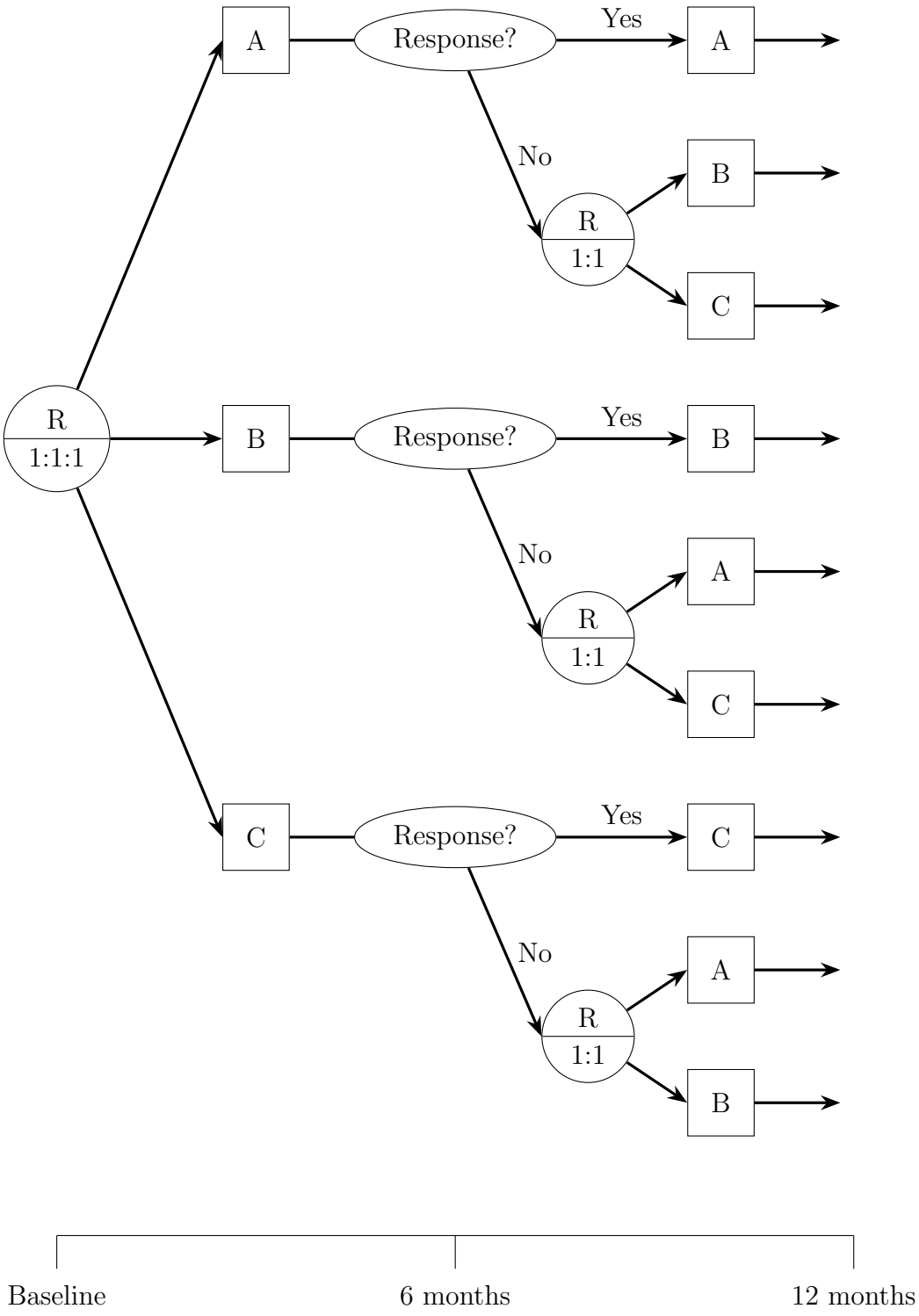
cancer *Thall et al.* (2007); *Thall* (2016). ARAMIS (NCT02939573) is a multi-national trial to evaluate different treatment options for patients with skin vasculitis. Vasculitides are uncommon diseases which can affect almost any organ, although vasculitis frequently involves the skin as an isolated process or as part of systemic vasculitis. Without high quality studies to guide the management of skin vasculitis, treatment decisions are made based on anecdotal experience and expert opinions. This uncertainty is reflected in variation between providers, leading to patients being treated with agents of uncertain efficacy and unknown relative merit. ARAMIS compares the efficacy of three of the most commonly used treatments for the treatment of skin vasculitis: colchicine, dapsone and azathioprine (Figure 3.1). Eligible patients are randomized with equal chance of receiving one of the three treatments under investigation for six months. Those who do not respond after the first stage (i.e., six months) are re-randomized equally between the other two treatments. Responders in stage 1 remain on their treatment in stage 2. The outcome of interest is response to treatment at six months as defined by a combination of participant and physician measures.

In Section 4.2, we present a method to analyze data from an snSMART by sharing information across stages to evaluate the overall treatment efficacy. The efficacy of a treatment is defined as the response rate at 6 months after initiating that treatment. In Section 4.3, we present simulation studies to illustrate our model's properties under various scenarios. Our manuscript concludes with a discussion in Section 3.4.

3.2 Method

The outcome of interest after each stage is a binary variable, where 1 denotes response and 0 denotes non-response to the assigned treatment. We propose a Bayesian approach that borrows information across both stages to estimate the response rate of each treatment. We model the first stage outcome as the probability of having a response to the first stage treatment. The second stage outcome is modeled conditionally on the first stage outcome

Figure 3.1: Study design of snSMART. Patients are randomized (R) to one of the treatment arms, A, B or C equally and followed up for 6 months. The responders keep the same treatment for another six months, while the non-responders are re-randomized to one of the remaining treatments and followed up for another six months



linking the first and second stage response probabilities through linkage parameters.

We use different linkage parameters for the first stage responders and first stage non-responders. In this way, a patient's second stage response rate is: (1) at least as high as one's first stage response rate if she/he is a responder and, (2) is at most as high as one's first stage response rate if he/she is a non-responder after the first stage treatment. We compare the estimator of the response rate using the proposed method to estimators produced from three other methods: (a) a log-Poisson model using data from both stages whose parameters are estimated via generalized estimating equations (GEE), (b) a Bayesian method using only the first stage data, and (c) a maximum likelihood method (MLE) using only the first stage data. The details of our proposed model and the log-Poisson model will be discussed next and simulation results for the comparison of estimators produced from the four methods will be shown in the Section 4.3.

3.2.1 Bayesian Joint Stage Modeling

For each subject $i = 1, \ldots, N$, stage $j = 1, 2$, and treatment $k = A, B, \ldots$ K, where N denotes the total sample size and K denotes the number of arms, let Y_{ijk} denote the observed response outcome. We model the first stage outcome and the second stage outcome given the first stage outcome each as a Bernoulli random variable. The first stage response rate is denoted as π_k for treatment k. The second stage response rate for first stage responders is equal to $\beta_1\pi_k$. For non-responders to treatment k in the first stage who receive treatment k' in the second stage, the second stage response rate in the second stage is equal to $\beta_0\pi_{k'}$. In practice, we assume: (i) The linkage parameters (β_0, β_1) do not depend on the initial treatment k. (ii) The linkage parameter for non-responders is smaller than 1, i.e., $\beta_0 < 1$. (iii) The linkage parameter for responders is greater than 1, i.e, $\beta_1 > 1$. Via simulation, we examine the violations of the assumptions in section 4.3.

Our proposed Bayesian joint stage model (BJSM) is as follows:

$$Y_{i1k}|\pi_k \sim Bernoulli(\pi_k)$$

$$Y_{i2k'}|Y_{i1k}, \pi_k \sim Bernoulli((\beta_1\pi_k)^{Y_{i1k}}(\beta_0\pi_{k'})^{1-Y_{i1k}})$$

Prior distributions on the first stage response rates and the linkage parameters are used to incorporate physician beliefs about the treatments. For the ARAMIS trial, we specify priors for the parameters involved in the model: $\pi_k \sim Beta(\zeta_k, \eta_k)$, $\beta_0 \sim Beta(\zeta_0, \eta_0)$, $\beta_1 \sim Pareto(1, \phi)$. For π_k, we have chosen to use the hyperparameter values $\zeta_k = 0.4$ and $\eta_k = 1.6$ for two reasons. First, these parameters lead to a prior mean of $\zeta_k/(\zeta_k + \eta_k) = 0.2$ for each of the arms, which was a reasonable *a priori* setting for the ARAMIS study. Second, the sum of the two parameters of a Beta distribution can be viewed as a prior sample size because the prior variance is inversely proportional to that sum. Thus, we assume our prior information is based upon a sample size of $\zeta_k + \eta_k = 2$ patients. For β_0, we have chosen hyperparameter values $\zeta_0 = 1$ and $\eta_0 = 1$, which lead to a uniform distribution over the interval $[0, 1]$. For β_1, we have assumed a hyperparameter value of $\phi = 3$, so that, on average, the second stage response rate is $\phi/(\phi - 1) = 1.5$ times as large as the first stage response rate.

3.2.2 Log-Poisson Joint Stage Modeling

The log-Poisson joint stage model, which we refer to LPJSM, is a frequentist way of modeling data from two stages, where we use a log link to model the mean and the Poisson family to model the variance of the outcome. We model the the log of each response rate instead of the logit of each response rate mainly for interpretability. The model is as follows:

$$log(E(Y_{i1k})) = log(\mu_{i1k}) = \alpha_A \mathbb{1}\{k = A\} + \alpha_B \mathbb{1}\{k = B\} + \alpha_C \mathbb{1}\{k = C\} \qquad (3.1)$$

$$log(E(Y_{i2k'})) = log(\mu_{i2k'}) = \alpha_A \mathbb{1}\{k' = A\} + \alpha_B \mathbb{1}\{k' = B\} + \alpha_C \mathbb{1}\{k' = C\} + \gamma_1 Y_{i1k} + \gamma_0 (1 - Y_{i1k}) \tag{3.2}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function. The response rates, π_k, and the linkage parameters β_1 and β_0 from the BJSM are equivalent to the exponentiated values of α_k, γ_1 and γ_0, respectively.

We estimate the parameters via GEE *Zeger and Liang* (1986):

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\theta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{3.3}$$

where $\boldsymbol{Y}_i = (Y_{i1k}, Y_{i2k'})^T$, $\boldsymbol{\mu}_i = (\mu_{i1k}; \mu_{i2k'})^T$, $\boldsymbol{\theta}$ is the parameter vector with $\boldsymbol{\theta} = (\alpha_A, \alpha_B, \alpha_C, \gamma_1, \gamma_0)^T$; $\boldsymbol{V}_i$ is the working covariance matrix of $\boldsymbol{Y}_i$ with $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\alpha) \boldsymbol{A}_i^{1/2}$ *Pan and Connett* (2002), where $\boldsymbol{A}_i^{1/2}$ is a diagonal matrix with elements being the square root of $Var(Y_{ijk})$, the variance of the outcome of the i^{th} patient at the j^{th} stage under treatment k. The variance of the outcome of the i^{th} patient at the j^{th} stage is modeled with a Poisson family variance structure, $Var(Y_{ijk}) = \mu_{ijk}$. We use the Poisson family variance structure to construct V_i in equation (3.3) to find the estimator of $\boldsymbol{\theta}$ as opposed to the binomial family variance structure, because others have reported that estimation sometimes fails to converge when attempting to fit log-binomial models with a small sample size *Williamson et al.* (2013). In addition, we use an independence working correlation structure $\boldsymbol{R}(\alpha) = \mathbf{I}_{2\times 2}$ in the estimating equation because the independence working correlation structure is recommended when binary responses have less than binomial variation over clusters *Hanley et al.* (2000).

In estimating the variance of $\hat{\boldsymbol{\theta}}$, we use the robust 'sandwich' covariance estimator, $\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_0^{-1}$, where $\boldsymbol{\Sigma}_0 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\theta}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_1 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\theta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\theta}}$. We use the binomial family variance structure, $Var(Y_{ijk}) = \mu_{ijk}(1 - \mu_{ijk})$, to construct $\boldsymbol{V}_i$ in $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ to estimate variance of $\hat{\boldsymbol{\theta}}$, because the 'sandwich' estimator is consistent, when $\boldsymbol{V}_i$ is correctly specified and even if $\boldsymbol{R}(\alpha)$ misspecified *Halekoh et al.* (2006).

3.3 Simulations

We set up the scenarios for our simulation studies in two subsections: simulations when the assumptions for BJSM are satisfied and simulations when our assumptions are violated. We compare the bias, root mean-square error (rMSE), coverage rate and width of the 95% credible/confidence interval (CI) of the estimated parameters. When the BJSM assumptions are true, we compare the estimators from our proposed method the BJSM to estimators produced from other three methods: the log-Poisson joint stage model (LPJSM), described in Section 3.2.2, Bayesian first stage model (BFSM), and a first stage maximum likelihood estimates (FSMLE). For both the BFSM and FSMLE, we only use the first stage data for estimation and for the BFSM, we assume the same prior distribution for π_k as the BJSM. When assumptions are violated, we compare the BJSM to an extension with multiple linkage parameters (BJSMM), and the LPJSM. The BJSMM is the same as that of the BJSM except that we allow the linkage parameters to depend on the initial treatment, i.e., β_0 is now replaced by β_{0k} and β_1 is now replaced by β_{1k}. We use exactly the same values for the hyperparameters as the BJSM for the prior densities of the BJSMM so that we allow for estimating β_{0k} and β_{1k} values that differ among different treatments k, but we give each the same prior distribution.

Bias is defined as the average of the differences between the true value of π_k and the estimated π_k in all simulations. The rMSE is calculated by taking the square root of the mean-square error of the estimators in all simulations. The simulated coverage rate is the frequency that the true value of the response parameter falls in the 95% CI for all simulations. The 95% CIs for the BJSM, BFSM and BJSMM are the highest posterior density (HPD) credible intervals, which is the narrowest interval that covers the 95% of the posterior distribution of π_k. The 95% CIs for the LPJSM and FSMLE are derived based on the asymptotic normality of the estimator of π_k in these two methods, and calculated by the estimator plus or minus 1.96 times of the standard error of the estimator. The parameters are estimated via the R

function jags and gee in the R package rjags and gee respectively. The computer programs used to derive estimates and CIs are available upon request from the primary author.

3.3.1 Simulation Scenarios

We simulate 2000 realizations per scenario; each scenario is a three-arm snSMART. The true values of the response rates in each arm and the linkage parameters in each stage vary in different scenarios; details are presented in Table 3.1. We focus on simulation results where the total snSMART sample size is 90 (30 patients per treatment), but we provide results for total sample sizes of 45 and 180 in Section 3.5.

Scenarios 1, 2 and 3 represent three ideal settings. In these scenarios, the linkage parameters for non-responders and for responders are the same for all three treatments, which means the model specification is the same as the data generating process. In scenarios 4 to 12 we vary the values of the linkage parameters to investigate model properties when assumptions are violated (the assumptions are enumerated in Section 3.2.1). Assumption (i) is violated in scenarios 4-7 and 10-12. Assumption (ii) is violated in scenarios 8, 10 and 12 and assumption (iii) is violated in scenarios 9, 11 and 12.

3.3.2 Simulation Results When the BJSM Assumptions are True

For scenarios 1, 2 and 3, the bias and rMSE for estimators of the response rates are shown in Table 3.2. The response rate estimators of the BJSM have the smallest rMSEs among all four methods. The rMSEs of the estimators from the BJSM and LPJSM are smaller than the rMSEs provided by the BFSM and FSMLE, which only use data from the first stage. In scenario 2, the BJSM provides the estimators with smallest bias compared to other three methods. In scenarios 1 and 3, the bias of the estimators for the BJSM is still small but slightly higher than the bias for the other three methods. This may be because the prior mean for the linkage parameter for non-responders is 0.5 which is closer to 0.6, the setting

Table 3.1: Simulation scenarios. π_k is the response rate at six months for the treatment $k = A, B, C$. β_{0k} is the linkage parameter for the first stage non-responders treated with treatment k. β_{1k} is the linkage parameter for the first stage responders treated with treatment k. The linkage parameters link the second stage response rates with the first stage response rate in our proposed model. The three assumptions are: (i) The linkage parameters do not depend on the initial treatment k, i.e., $\beta_{1k} = \beta_1$ and $\beta_{0k} = \beta_0$. (ii) The linkage parameter for non-responders is smaller than 1, i.e., $\beta_0 < 1$. (iii) The linkage parameter for responders is greater than 1, i.e, $\beta_1 > 1$.

Scenarios	π_A	π_B	π_C	β_{0A}	β_{0B}	β_{0C}	β_{1A}	β_{1B}	β_{1C}	Assumptions Violated
1	0.3	0.3	0.3	0.8	0.8	0.8	1.5	1.5	1.5	none
2	0.2	0.3	0.4	0.6	0.6	0.6	1.5	1.5	1.5	none
3	0.2	0.3	0.4	0.8	0.8	0.8	1.5	1.5	1.5	none
4	0.2	0.3	0.4	0.3	0.6	0.9	1.5	1.5	1.5	(i)
5	0.2	0.3	0.4	0.6	0.6	0.6	1.2	1.5	1.8	(i)
6	0.2	0.3	0.4	0.3	0.6	0.9	1.2	1.5	1.8	(i)
7	0.2	0.3	0.4	0.9	0.6	0.3	1.2	1.5	1.8	(i)
8	0.2	0.3	0.4	1.2	1.2	1.2	1.5	1.5	1.5	(ii)
9	0.2	0.3	0.4	0.6	0.6	0.6	0.8	0.8	0.8	(iii)
10	0.2	0.3	0.4	0.3	0.6	1.2	1.2	1.5	1.8	(i), (ii)
11	0.2	0.3	0.4	1.2	1.2	1.2	0.8	0.8	0.8	(ii), (iii)
12	0.2	0.3	0.4	0.3	0.6	1.2	0.8	1.5	1.8	(i), (ii), (iii)

in scenario 2, than 0.8, the setting in scenarios 1 and 3. These observations suggest that, in settings where the assumptions are satisfied, jointly modeling data from two stages provides improved estimators for treatment due to smaller rMSEs. In particular, the biggest gain in rMSE is given by the BJSM which also produces small to negligible bias. Table 3.3 presents the 95% CI width and coverage rates. Here we see the average width of the 95% CI of the BJSM is smaller than the other approaches and the coverage rate is around the target 95%.

3.3.3 Simulation Results When the BJSM Assumptions are Violated

Simulation results when the assumptions for the BJSM are violated are shown in Tables 3.4 and 3.5. When only assumption (i) is violated (scenarios 4-7), we see that the bias for the response rate estimators is small for all three methods. The estimators of the BJSMM has the smallest bias in scenario 4-6 and BJSM has the smallest bias in scenario 7. The estimators of the BJSM and BJSMM have smaller rMSEs than the LPJSM approach. When only assumption (ii) is violated, as in scenario 8, we see that the response rates are overestimated for all treatment arms by the BJSM and BJSMM. The response rates are overestimated to balance the effect of the underestimated β_0 on the stage 2 response rates of non-responders. The bias for the estimators is higher for BJSM and BJSMM but still small. The estimators from the LPJSM have the smallest bias and rMSEs. When only assumption (iii) is violated as in scenario 9, the response rates are underestimated for all arms by the BJSM and BJSMM. The response rates are underestimated to balance the effect of the overestimated β_1 on the stage 2 response rates of responders. The estimators from the LPJSM have smallest bias but the rMSEs are higher compared with those from the BJSM and BJSMM.

When more than one assumption is violated (scenarios 10-12), the bias for the estimators of the response rates is lower in scenario 11, where assumption (i) holds and assumptions (ii) and (iii) are violated. This finding occurs because when assumption (ii) is violated, the BJSM and BJSMM tend to overestimate the response rates, and when assumption (iii) is

Table 3.2: Simulated bias and root mean-square error (rMSE) for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 30. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE	Bias	rMSE
1	π_A	0.008	0.062	-0.001	0.069	-0.008	0.079	-0.002	0.084
	π_B	0.008	0.062	0.002	0.069	-0.006	0.078	0.001	0.083
	π_C	0.008	0.061	-0.002	0.068	-0.008	0.078	-0.001	0.083
2	π_A	-0.001	0.056	-0.001	0.059	-0.002	0.069	-0.002	0.074
	π_B	0.001	0.063	0.000	0.070	-0.006	0.078	0.001	0.083
	π_C	0.000	0.067	0.002	0.077	-0.014	0.085	-0.002	0.089
3	π_A	0.005	0.056	-0.001	0.057	-0.002	0.069	-0.002	0.074
	π_B	0.008	0.062	0.000	0.069	-0.006	0.078	0.001	0.083
	π_C	0.011	0.064	0.002	0.076	-0.014	0.085	-0.002	0.089

Table 3.3: Simulated width and coverage of 95% CI for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 30. π_k is the true response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Width	CR	Width	CR	Width	CR	Width	CR
1	π_A	0.240	0.944	0.265	0.931	0.299	0.903	0.321	0.950
	π_B	0.240	0.948	0.266	0.936	0.300	0.908	0.322	0.949
	π_C	0.240	0.944	0.265	0.934	0.299	0.908	0.321	0.950
2	π_A	0.213	0.929	0.228	0.932	0.256	0.945	0.277	0.945
	π_B	0.245	0.940	0.269	0.936	0.300	0.908	0.322	0.949
	π_C	0.265	0.948	0.305	0.937	0.323	0.930	0.344	0.930
3	π_A	0.210	0.936	0.222	0.936	0.256	0.945	0.277	0.945
	π_B	0.240	0.942	0.263	0.936	0.300	0.908	0.322	0.949
	π_C	0.258	0.956	0.300	0.937	0.323	0.930	0.344	0.930

violated, the BJSM and BJSMM tend to underestimate the response rates. and when both assumptions (ii) and (iii) are violated, two errors cancel each other. The BJSM and BJSMM have smaller rMSEs in all three scenarios.

In general, when the assumptions are violated, the response rate estimators of the BJSM have smaller bias and rMSEs than the LPJSM in most of the settings, and smaller rMSEs than the standard approaches that only use the data in stage 1. When multiple linkage parameters are considered in the BJSMM, we do not see a large reduction of bias or rMSEs.

In Table 3.5, we can see the average width of the 95% CI of the BJSM is smaller than the other approaches. When only assumption (i) is violated (scenarios 4-7), the coverage rates of 95% CIs for treatments B and C are around the target 95%. The coverage rate is readily below the target for all three approaches in scenarios 4 and 6. When only assumption (ii) (scenario 8) or assumption (iii) is violated (scenario 9), the coverage rate for the treatment C is below the target. This can be explained by the larger bias in the response rate estimator for treatment C in scenarios 8 and 9. In scenario 10, (when assumption (i) and (ii) are violated), and Scenario 12, when all the assumptions are violated, the coverage rates of 95% CIs are below the target 95% for treatment A and C because the response rate estimators have higher bias compared to the estimators of the treatment B. When assumptions (ii) and (iii) are violated at the same time, the coverage rates are greater than the target 95% for the BJSM and BJSMM and below the target 95% for the LPJSM. Similar trends were observed for sample sizes of n=45 and 180; details are given in Table 3.6-3.13 in the Section 3.5.

3.4 Concluding Remarks

In this manuscript, we present a Bayesian method (BJSM) to estimate the response rates of multiple treatments from snSMART with two stages. The BJSM is a novel method that links the response rates from two stages of one clinical trial via linkage parameters. The BJSM provides accurate estimators and straight forward clinical interpretations for the parameters.

Table 3.4: Simulated bias and root mean-square error (rMSE) for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 30. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		BJSMM		LPJSM	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE
4	π_A	-0.024	0.060	-0.021	0.059	-0.029	0.064
	π_B	-0.004	0.063	-0.001	0.062	-0.007	0.070
	π_C	0.027	0.068	0.032	0.071	0.037	0.086
5	π_A	-0.010	0.054	-0.005	0.055	-0.011	0.056
	π_B	-0.003	0.062	0.003	0.063	-0.005	0.068
	π_C	0.022	0.072	0.017	0.070	0.017	0.080
6	π_A	-0.033	0.061	-0.027	0.060	-0.038	0.066
	π_B	-0.008	0.062	0.000	0.062	-0.012	0.069
	π_C	0.048	0.080	0.048	0.081	0.050	0.093
7	π_A	0.014	0.055	0.017	0.058	0.017	0.059
	π_B	0.001	0.062	0.007	0.063	0.001	0.069
	π_C	-0.004	0.073	-0.013	0.073	-0.018	0.080
8	π_A	0.023	0.060	0.029	0.063	-0.001	0.054
	π_B	0.036	0.069	0.042	0.073	0.000	0.065
	π_C	0.047	0.076	0.054	0.081	0.001	0.073
9	π_A	-0.015	0.054	-0.016	0.054	-0.001	0.061
	π_B	-0.029	0.064	-0.030	0.063	0.000	0.073
	π_C	-0.047	0.075	-0.047	0.075	0.002	0.081
10	π_A	-0.038	0.063	-0.030	0.060	-0.045	0.069
	π_B	-0.015	0.063	-0.005	0.062	-0.024	0.070
	π_C	0.073	0.095	0.078	0.099	0.071	0.105
11	π_A	0.011	0.053	0.014	0.054	0.000	0.056
	π_B	0.010	0.055	0.013	0.056	0.000	0.067
	π_C	0.006	0.055	0.010	0.055	0.001	0.076
12	π_A	-0.047	0.066	-0.039	0.063	-0.053	0.073
	π_B	-0.014	0.063	-0.005	0.062	-0.022	0.070
	π_C	0.075	0.097	0.078	0.099	0.076	0.109

Table 3.5: Simulated width and coverage of 95% CI for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 30. π_k is the true response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		BJSMM		LPJSM	
Scenario		Width	CR	Width	CR	Width	CR
4	π_A	0.200	0.868	0.206	0.890	0.220	0.854
	π_B	0.240	0.933	0.247	0.948	0.266	0.929
	π_C	0.262	0.950	0.264	0.938	0.303	0.931
5	π_A	0.208	0.932	0.213	0.936	0.215	0.915
	π_B	0.243	0.942	0.250	0.950	0.262	0.930
	π_C	0.270	0.942	0.278	0.951	0.311	0.942
6	π_A	0.194	0.854	0.203	0.880	0.207	0.824
	π_B	0.238	0.936	0.248	0.946	0.259	0.923
	π_C	0.266	0.912	0.271	0.912	0.310	0.917
7	π_A	0.220	0.957	0.223	0.949	0.222	0.948
	π_B	0.247	0.948	0.253	0.950	0.264	0.936
	π_C	0.272	0.928	0.284	0.935	0.313	0.930
8	π_A	0.211	0.926	0.219	0.927	0.210	0.930
	π_B	0.235	0.922	0.245	0.922	0.252	0.938
	π_C	0.246	0.899	0.255	0.882	0.291	0.944
9	π_A	0.206	0.926	0.207	0.933	0.238	0.934
	π_B	0.235	0.916	0.236	0.918	0.282	0.927
	π_C	0.250	0.892	0.251	0.900	0.316	0.932
10	π_A	0.188	0.834	0.199	0.868	0.202	0.787
	π_B	0.232	0.922	0.244	0.942	0.255	0.902
	π_C	0.260	0.825	0.264	0.811	0.308	0.876
11	π_A	0.207	0.949	0.212	0.950	0.217	0.928
	π_B	0.230	0.960	0.235	0.968	0.260	0.932
	π_C	0.238	0.972	0.243	0.974	0.298	0.944
12	π_A	0.183	0.792	0.194	0.838	0.192	0.735
	π_B	0.232	0.925	0.244	0.942	0.258	0.907
	π_C	0.259	0.814	0.264	0.810	0.312	0.863

We compared the proposed method to three other methods via simulation and found that the BJSM provides the most accurate estimators among all four methods in small samples.

The BJSM relies on three key assumptions. These assumptions simplify our model and make our model easier to interpret. However, there might be situations where the assumptions are violated. Simulation results suggest that our method works well even when the linkage parameters vary among the treatment arm. We can make our model more flexible by including different linkage parameters for different treatments (i.e., BJSMM). When we introduce the additional parameters to analyze data from snSMARTs with total sample sizes of 45, 90, or 180, the gain in bias and efficiency are small to negligible and not uniform across the different treatment response rates. In further simulations, the BJSMM was uniformly superior to the BJSM when sample sizes increase above 120 per arm (total sample size of 360). The model as proposed with homogeneous linkage parameters (i.e., BJSM) has sufficiently low bias and high efficiency in smaller sample sizes realistic for rare disease research not warranting estimation of the additional linkage parameters.

Assumptions (ii) and (iii) that constrain the values of the linkage parameters generally hold when the response rates are similar and low (i.e., less than 50%) for all treatments in the trial. For many rare diseases, these assumptions are realistic and, thus, violations generally do not pose problems. If the assumptions are violated, the BJSM provides more biased estimates than the LPJSM, but is more efficient. Sensitivity analyses using the LPJSM can always be done to compare results for the two methods.

A limitation that develops from our assumptions and corresponding priors is that the posterior distributions of β_1 and π_k can have positive probability for $\beta_1\pi_k > 1$. In reality, we can not have a response rate greater than 1, but our models allow for this. To circumvent this potential problem we considered a power model formulation of the BJSM. In the power model formulation, the second stage response rates are defined as $\pi_{k'}^{\beta_0}$ and $\pi_k^{\beta_1}$ for non-responders and responders, respectively. This allows β_0 and β_1 to vary on the positive

real line. Ultimately, we decided against the power formulation as the linkage parameters are not clinically interpretable and our simulations for the proposed version of BJSM did not draw any samples such that $\beta_1\pi_k > 1$, making it unlikely for this limitation to be a problem in practice in similar settings.

Future work includes extending the BJSM to non-binary outcomes (i.e. continuous and survival outcomes) and establishing sample size calculations based on the analysis of snSMARTs using the BJSM. We aim to develop sample size calculations and an easy-to-use corresponding applet that can target specific differences between treatment arms. The sample size calculations will lead us to consider alternative designs of snSMARTs where more than three treatments are involved or there is an imbalance in second-stage randomization.

3.5 Supplementary Materials

Simulation results for sample sizes of 45 and 180 (15 and 60 per arm) are in Table 3.6-3.13.

Table 3.6: Simulated bias and root mean-square error (rMSE) for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 15. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE	Bias	rMSE
1	π_A	0.004	0.084	0.001	0.099	-0.014	0.106	-0.003	0.119
	π_B	0.007	0.084	0.002	0.099	-0.011	0.104	0.001	0.118
	π_C	0.006	0.082	0.001	0.097	-0.013	0.104	-0.002	0.117
2	π_A	-0.002	0.076	-0.001	0.084	-0.002	0.092	-0.002	0.104
	π_B	-0.003	0.085	0.000	0.101	-0.011	0.104	0.001	0.118
	π_C	-0.007	0.088	0.005	0.111	-0.025	0.113	-0.002	0.125
3	π_A	0.005	0.076	-0.002	0.083	-0.002	0.092	-0.002	0.104
	π_B	0.006	0.084	0.000	0.098	-0.011	0.104	0.001	0.118
	π_C	0.006	0.085	0.005	0.110	-0.025	0.113	-0.002	0.125

Table 3.7: Simulated width and coverage of 95% CI for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 15. π_k is the true response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Width	CR	Width	CR	Width	CR	Width	CR
1	π_A	0.319	0.936	0.372	0.913	0.386	0.930	0.441	0.947
	π_B	0.320	0.928	0.374	0.917	0.389	0.934	0.444	0.947
	π_C	0.320	0.937	0.373	0.920	0.388	0.934	0.443	0.948
2	π_A	0.278	0.920	0.317	0.898	0.327	0.943	0.376	0.801
	π_B	0.323	0.922	0.376	0.912	0.389	0.934	0.444	0.947
	π_C	0.346	0.938	0.428	0.928	0.423	0.899	0.478	0.943
3	π_A	0.279	0.926	0.308	0.892	0.327	0.943	0.376	0.801
	π_B	0.320	0.926	0.368	0.918	0.389	0.934	0.444	0.947
	π_C	0.341	0.947	0.422	0.926	0.423	0.899	0.478	0.943

Table 3.8: Simulated bias and root mean-square error (rMSE) for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 60. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE	Bias	rMSE
1	π_A	0.006	0.046	0.000	0.048	-0.004	0.058	-0.001	0.060
	π_B	0.007	0.045	-0.001	0.048	-0.003	0.057	0.000	0.059
	π_C	0.007	0.044	-0.002	0.048	-0.004	0.057	-0.001	0.059
2	π_A	0.000	0.041	-0.003	0.043	-0.001	0.050	-0.001	0.052
	π_B	0.002	0.047	0.000	0.050	-0.003	0.057	0.000	0.059
	π_C	0.002	0.050	-0.001	0.055	-0.008	0.061	-0.001	0.063
3	π_A	0.004	0.040	-0.002	0.042	-0.001	0.050	-0.001	0.052
	π_B	0.007	0.045	0.000	0.048	-0.003	0.057	0.000	0.059
	π_C	0.010	0.048	-0.002	0.055	-0.008	0.061	-0.001	0.063

Table 3.9: Simulated width and coverage of 95% CI for the estimators of π_k where the BJSM assumptions are satisfied. Four modeling approaches: Bayesian joint stage modeling (BJSM), log-Poisson joint stage modeling (LPJSM), Bayesian first stage modeling (BFSM) and first stage maximum likelihood estimation (FSMLE) are compared. The sample size per treatment arm is 60. π_k is the true response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		LPJSM		BFSM		FSMLE	
Scenario		Width	CR	Width	CR	Width	CR	Width	CR
1	π_A	0.176	0.942	0.188	0.946	0.221	0.931	0.229	0.947
	π_B	0.176	0.951	0.188	0.942	0.222	0.933	0.230	0.948
	π_C	0.176	0.952	0.187	0.938	0.221	0.935	0.229	0.948
2	π_A	0.157	0.932	0.162	0.931	0.191	0.935	0.199	0.922
	π_B	0.181	0.947	0.190	0.945	0.222	0.932	0.230	0.948
	π_C	0.200	0.955	0.215	0.936	0.238	0.938	0.246	0.931
3	π_A	0.154	0.940	0.158	0.935	0.191	0.935	0.199	0.922
	π_B	0.176	0.949	0.186	0.940	0.222	0.932	0.230	0.948
	π_C	0.192	0.960	0.212	0.938	0.238	0.938	0.246	0.931

Table 3.10: Simulated bias and root mean-square error (rMSE) for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 15. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		BJSMM		LPJSM	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE
4	π_A	-0.023	0.078	-0.020	0.077	-0.027	0.085
	π_B	-0.006	0.085	-0.004	0.084	-0.007	0.099
	π_C	0.020	0.086	0.024	0.087	0.038	0.116
5	π_A	-0.010	0.073	-0.006	0.074	-0.010	0.080
	π_B	-0.005	0.085	-0.001	0.085	-0.006	0.099
	π_C	0.016	0.093	0.015	0.091	0.020	0.115
6	π_A	-0.031	0.077	-0.026	0.076	-0.036	0.084
	π_B	-0.009	0.085	-0.003	0.084	-0.012	0.097
	π_C	0.043	0.096	0.044	0.097	0.052	0.123
7	π_A	0.012	0.073	0.014	0.075	0.016	0.085
	π_B	-0.002	0.084	0.002	0.085	0.001	0.099
	π_C	-0.011	0.098	-0.015	0.094	-0.015	0.114
8	π_A	0.024	0.079	0.030	0.081	-0.003	0.079
	π_B	0.034	0.088	0.040	0.091	0.001	0.093
	π_C	0.041	0.091	0.047	0.094	0.004	0.106
9	π_A	-0.016	0.071	-0.016	0.070	0.001	0.088
	π_B	-0.034	0.083	-0.034	0.082	0.002	0.105
	π_C	-0.056	0.095	-0.055	0.094	0.005	0.114
10	π_A	-0.034	0.077	-0.028	0.076	-0.043	0.086
	π_B	-0.013	0.085	-0.007	0.084	-0.025	0.098
	π_C	0.069	0.107	0.072	0.110	0.072	0.132
11	π_A	0.012	0.071	0.015	0.071	-0.001	0.082
	π_B	0.007	0.075	0.010	0.074	0.002	0.096
	π_C	-0.001	0.073	0.002	0.072	0.005	0.106
12	π_A	-0.043	0.077	-0.037	0.075	-0.051	0.086
	π_B	-0.013	0.085	-0.007	0.084	-0.022	0.098
	π_C	0.070	0.108	0.072	0.110	0.078	0.137

Table 3.11: Simulated width and coverage of 95% CI for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 15. π_k is the true response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		BJSMM		LPJSM	
Scenario		Width	CR	Width	CR	Width	CR
4	π_A	0.262	0.876	0.268	0.898	0.306	0.844
	π_B	0.319	0.919	0.325	0.926	0.372	0.909
	π_C	0.345	0.952	0.348	0.950	0.426	0.924
5	π_A	0.272	0.921	0.278	0.926	0.300	0.889
	π_B	0.321	0.924	0.327	0.929	0.366	0.908
	π_C	0.352	0.941	0.358	0.947	0.438	0.929
6	π_A	0.256	0.872	0.264	0.894	0.289	0.822
	π_B	0.317	0.920	0.325	0.928	0.363	0.903
	π_C	0.349	0.933	0.354	0.932	0.435	0.921
7	π_A	0.288	0.945	0.291	0.942	0.311	0.926
	π_B	0.325	0.932	0.329	0.930	0.371	0.913
	π_C	0.353	0.912	0.361	0.926	0.439	0.922
8	π_A	0.283	0.928	0.293	0.937	0.292	0.884
	π_B	0.319	0.930	0.329	0.932	0.353	0.912
	π_C	0.332	0.936	0.341	0.928	0.409	0.926
9	π_A	0.270	0.920	0.271	0.925	0.332	0.902
	π_B	0.310	0.912	0.311	0.918	0.394	0.914
	π_C	0.331	0.914	0.332	0.918	0.443	0.928
10	π_A	0.250	0.856	0.261	0.890	0.282	0.794
	π_B	0.311	0.912	0.322	0.927	0.356	0.890
	π_C	0.343	0.892	0.348	0.883	0.433	0.908
11	π_A	0.278	0.938	0.284	0.948	0.303	0.889
	π_B	0.311	0.954	0.318	0.962	0.366	0.913
	π_C	0.324	0.966	0.330	0.970	0.419	0.927
12	π_A	0.244	0.845	0.254	0.879	0.268	0.767
	π_B	0.312	0.912	0.322	0.924	0.360	0.891
	π_C	0.344	0.888	0.348	0.887	0.439	0.903

Table 3.12: Simulated bias and root mean-square error (rMSE) for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 60. π_k is the response rate at six months for treatment k, $k = A, B, C$.

		BJSM		BJSMM		LPJSM	
Scenario		Bias	rMSE	Bias	rMSE	Bias	rMSE
4	π_A	-0.025	0.048	-0.022	0.046	-0.030	0.051
	π_B	-0.005	0.046	-0.001	0.046	-0.007	0.049
	π_C	0.029	0.055	0.034	0.057	0.033	0.064
5	π_A	-0.011	0.040	-0.005	0.041	-0.012	0.042
	π_B	-0.005	0.046	0.003	0.046	-0.005	0.049
	π_C	0.022	0.055	0.012	0.053	0.014	0.058
6	π_A	-0.035	0.051	-0.028	0.048	-0.039	0.055
	π_B	-0.011	0.046	-0.001	0.046	-0.012	0.049
	π_C	0.046	0.066	0.044	0.065	0.046	0.073
7	π_A	0.014	0.041	0.018	0.044	0.016	0.044
	π_B	0.001	0.046	0.009	0.047	0.002	0.049
	π_C	-0.002	0.054	-0.018	0.058	-0.021	0.061
8	π_A	0.022	0.045	0.027	0.049	-0.003	0.039
	π_B	0.035	0.055	0.041	0.060	0.000	0.046
	π_C	0.047	0.063	0.054	0.070	-0.001	0.053
9	π_A	-0.014	0.041	-0.016	0.041	-0.002	0.044
	π_B	-0.025	0.049	-0.028	0.050	0.001	0.052
	π_C	-0.040	0.059	-0.042	0.060	-0.002	0.058
10	π_A	-0.041	0.055	-0.032	0.050	-0.046	0.060
	π_B	-0.019	0.047	-0.007	0.045	-0.025	0.053
	π_C	0.070	0.083	0.075	0.087	0.067	0.087
11	π_A	0.011	0.039	0.013	0.040	-0.002	0.041
	π_B	0.011	0.041	0.014	0.042	0.001	0.048
	π_C	0.010	0.040	0.013	0.041	-0.002	0.055
12	π_A	-0.050	0.060	-0.040	0.054	-0.054	0.065
	π_B	-0.018	0.047	-0.007	0.045	-0.022	0.052
	π_C	0.073	0.085	0.074	0.087	0.072	0.092

Table 3.13: Simulated width and coverage of 95% CI for the estimators of π_k when assumptions are violated. Three modeling approaches: Bayesian joint stage modeling (BJSM), Bayesian joint stage modeling with multiple linkage parameters (BJSMM), and log-Poisson joint stage modeling (LPJSM) are compared. The sample size per treatment arm is 60. π_k is the response rate at six months for the treatment k, $k = A, B, C$. CR=Coverage Rate.

		BJSM		BJSMM		LPJSM	
Scenario		Width	CR	Width	CR	Width	CR
4	π_A	0.147	0.857	0.153	0.884	0.157	0.838
	π_B	0.177	0.938	0.185	0.950	0.188	0.938
	π_C	0.197	0.938	0.198	0.920	0.214	0.918
5	π_A	0.152	0.928	0.158	0.935	0.153	0.907
	π_B	0.178	0.944	0.187	0.951	0.186	0.937
	π_C	0.202	0.939	0.212	0.955	0.220	0.94
6	π_A	0.142	0.800	0.150	0.858	0.148	0.780
	π_B	0.174	0.934	0.185	0.950	0.184	0.928
	π_C	0.200	0.878	0.205	0.894	0.219	0.884
7	π_A	0.161	0.958	0.165	0.944	0.158	0.934
	π_B	0.182	0.952	0.189	0.952	0.187	0.940
	π_C	0.203	0.932	0.218	0.924	0.221	0.919
8	π_A	0.152	0.917	0.159	0.911	0.149	0.934
	π_B	0.169	0.895	0.178	0.882	0.178	0.938
	π_C	0.179	0.835	0.187	0.810	0.206	0.938
9	π_A	0.153	0.924	0.153	0.923	0.170	0.930
	π_B	0.173	0.903	0.174	0.906	0.200	0.943
	π_C	0.186	0.882	0.186	0.877	0.223	0.938
10	π_A	0.136	0.757	0.147	0.836	0.144	0.713
	π_B	0.168	0.909	0.181	0.943	0.180	0.894
	π_C	0.193	0.723	0.197	0.703	0.218	0.789
11	π_A	0.149	0.948	0.153	0.950	0.154	0.933
	π_B	0.165	0.957	0.170	0.964	0.185	0.940
	π_C	0.171	0.970	0.175	0.968	0.211	0.941
12	π_A	0.133	0.679	0.143	0.786	0.137	0.621
	π_B	0.168	0.912	0.181	0.944	0.183	0.900
	π_C	0.193	0.699	0.197	0.705	0.220	0.766

CHAPTER IV

Sample Size Determination for Bayesian Analysis of small n Sequential, Multiple Assignment, Randomized Trials (snSMARTs) with Three Agents

4.1 Introduction

A rare disease is defined as any disorder that affects fewer than 200,000 individuals in the United States (*107th Congress*, 2002); collectively almost 30 million individuals and their families in the United States are impacted by rare diseases (*Griggs et al.*, 2009). With low numbers of individuals affected for any single rare disease, identifying optimal treatments for a rare disease via clinical trials is challenging because of limited allocation of individuals, insufficient funding, and challenges with individual recruitment. To meet the challenges posed by traditional study designs, a suggested clinical trial design for rare diseases is known as the small n, Sequential, Multiple Assignment, Randomized Trial, snSMART (*Tamura et al.*, 2016). In a two-stage snSMART (Figure 4.1), individuals are randomized in the first stage to one of three treatments. In the second stage, individuals who respond to their initial treatment continue the same treatment, while those who fail to respond are re-randomized to one of the two remaining treatments. The snSMART design is attractive because it allows participants who benefit from their treatment to continue to receive that treatment and those

who do not benefit from the treatment to switch to another treatment. Hence, an snSMART design may be more attractive to potential participants than a crossover design.

Tamura et al. (2016) proposed a weighted Z-statistic to identify the best treatment via hypothesis testing and provided a way to calculate the sample size for an snSMART based on this weighted Z-statistic. Specifically, the statistic is a weighted average of: (1) the observed difference in the response rates of the two best treatments in the first stage, and (2) the observed difference in the response rates of the same two treatments in the second stage, restricted to non-responders of the treatment with the lowest response rate in stage 1. Via simulation, one can calculate the sample size that achieves a desired statistical power with a pre-specified type I error rate, weights for the weighted Z-statistic and the hypothesized response rates. Although the weighted Z-statistic identifies the best treatment using information from both stages, much of the data from the second stage are ignored.

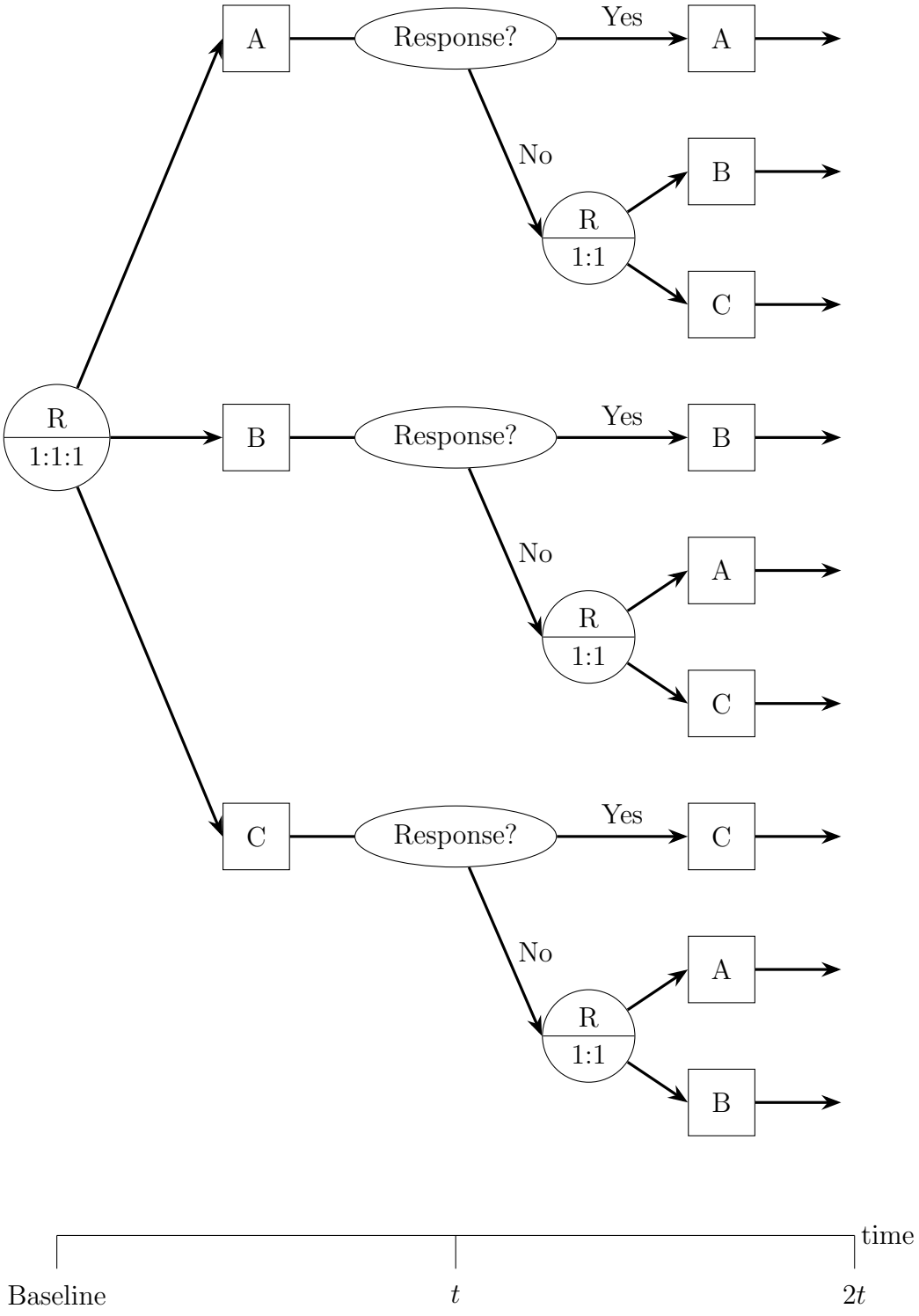
Wei et al. (2018) presented a more efficient Bayesian joint stage model (BJSM) that uses all of the data collected from an snSMART. Because of the efficiency of the BJSM relative to the weighted Z-statistic approach of *Tamura et al.* (2016), we would expect that the sample size calculation corresponding to *Tamura et al.* (2016) would be larger than necessary for a given power when using the BJSM. As a result, our goal was to develop a sample size calculation specific to the BJSM.

There are several methods for calculating sample sizes using Bayesian methods (*Adcock*, 1988; *Pham-Gia and Turkkan*, 1992; *Joseph and Bélisle*, 1997; *Cao et al.*, 2009).

These methods use a variety of metrics for determining sample size, including the variance of the posterior distribution, Bayes risk, length of the posterior credible interval, and coverage rate of the posterior credible interval.

Sample size calculations based upon the length and coverage rate of posterior intervals further vary depending on the use of the average coverage criterion (ACC), the average length criterion (ALC), or the worst outcome criterion (WOC). For a fixed length of the posterior

Figure 4.1: The study design of an snSMART. Patients are randomized (R) to one of the treatment arms, A, B or C equally (1:1:1) and followed for time t. The responders continue the same treatment for another time t, while the non-responders are re-randomized to one of the remaining treatments for an additional time t.



credible interval of the test statistic, the ACC (*Adcock*, 1988) provides an estimated sample size so that the average posterior coverage rate over the statistic's predictive distribution achieves a desired coverage rate. In contrast, the ALC (*Joseph and Bélisle*, 1997) determines an estimated sample size so that the average length of the posterior credible interval over the test statistic's predictive distribution achieves a desired length for a fixed coverage rate. For a fixed length of the posterior credible interval of the test statistic, the WOC provides an estimated sample size so that the minimum coverage rate over the statistic's predictive distribution achieves the desired coverage rate. *Cao et al.* (2009) provides a detailed comparison of results using these three criteria and found that the WOC (*Joseph and Bélisle*, 1997) produces the largest sample size among the three criteria in most settings. The size using ACC or ALC depends on the desired coverage rate rather than the desired length of the posterior credible interval.

Our work proposes to use the ACC as a way of determining a sample size for an snSMART analyzed with the BJSM. In Section 4.2 of this manuscript, we describe a sample size calculation for the snSMART that analyzes data using the BJSM. In Section 4.3, we present results from simulations to illustrate the properties of our calculated sample size under various scenarios and compare the proposed sample size to that given by the weighted Z-statistic. We conclude with a discussion in Section 4.4.

4.2 Methods

4.2.1 Review of the BJSM

We assume the same setting as in *Wei et al.* (2018). We have an snSMART that randomizes N subjects over two stages in order to compare the response rates of three treatments. We assume N subjects are equally allocated to three treatments in the first stage, and denote the number of subjects in each arm as $n = N/3$. For subject $i = 1, \ldots, N$, stage $j = 1, 2$,

and treatment $k = A, B, C$, let Y_{ijk} equal 1 for response and 0 for non-response. In the first stage, let the response rate for treatment k be $\pi_k = Pr(Y_{i1k} = 1)$. In the second stage, we assume the response rate for first stage responders who receive treatment k again in the second stage is equal to $\beta_1\pi_k$, while non-responders to treatments k' ($k' \neq k$) in the first stage who receive treatment k in the second stage have a response rate equal to $\beta_0\pi_k$. Our primary goal is to estimate the first stage response rates π_k, $k = A, B, C$, using data from both stages.

We refer to β_1 and β_0 as linkage parameters since they link the first stage response to the second stage response. The BJSM has three assumptions for the linkage parameters: (i) they do not depend on the initial treatment k; (ii) non-responders are less likely to respond to other treatments in Stage 2 ($\beta_0 < 1$); (iii) responders are more likely to respond in stage 2 when given the same treatment again ($\beta_1 > 1$).

In order to estimate the parameters of the BJSM, we specify the following priors for each parameter:

$$\pi_k \sim Beta(a_k, b_k),$$
$$\beta_0 \sim Beta(\gamma, \kappa),$$
$$\beta_1 \sim Pareto(1, \psi),$$

and use the posterior means of each parameter to estimate the response rate of each treatment. The suggested values for hyperparameters a_k, b_k, γ, κ, and ψ are discussed in *Wei et al.* (2018).

4.2.2 Approximate Prior Distribution

Our goal is to determine the minimum number of individuals to enroll in an snSMART so that the BJSM can be used to distinguish between the best treatment and the second best

treatment with sufficient confidence using posterior credible intervals. Unfortunately, the BJSM does not lead to closed-form expressions for the posterior distributions of the parameters, making direct calculation of the sample size intractable. To facilitate the calculation, we require several approximations for the posterior distribution of the best and second best treatments.

The BJSM states that π_k has a prior $Beta(a_k, b_k)$ distribution. Thus, π_k has a prior mean response rate of $a_k/(a_k+b_k)$ and a prior sample size of $m_k = a_k+b_k$. For the second stage, we let $\theta_k = \beta_1\pi_k$ for subjects who receive treatment k in Stage 2 after responding to treatment k in Stage 1, and we let $\phi_k = \beta_0\pi_k$ for subjects who receive treatment k in Stage 2 after not responding to treatments k' in Stage 1. We can see that neither θ_k nor ϕ_k will have Beta distributions, regardless of whether β_1 and β_0 are fixed or random.

In a sample size computation for Bayesian adaptive Phase I trials, *Braun* (2018) used a Beta distribution to approximate the prior distribution of a proportion that lacked a closed-form expression. We adopt the same idea for our methods to simplify calculating the posterior distribution of the proportions. For given values of β_1 and β_0, we can use the method-of-moments to approximate the prior distributions of θ_k and ϕ_k with respective Beta distributions, $Beta(c_k, d_k)$ and $Beta(e_k, f_k)$, where

$$\frac{c_k}{c_k + d_k} = \beta_1 \frac{a_k}{a_k + b_k}, \tag{4.1}$$

$$\frac{e_k}{e_k + f_k} = \beta_0 \frac{a_k}{a_k + b_k}. \tag{4.2}$$

We also assume that the prior sample size for θ_k, conditional on β_1, is equal to the expected prior number of responders to treatment k in the first stage. Similarly, the prior sample size for ϕ_k, conditional on β_0, is equal to the prior expected number of non-responders to all treatments $k' \neq k$ in the first stage who then receive treatment k in the second stage. These assumptions lead to a second pair of equations for the denominators in Equations (4.1) and

(4.2):

$$\begin{aligned} c_k + d_k &= a_k, \quad (4.3) \\ e_k + f_k &= \frac{1}{2}\sum_{k' \neq k} b_{k'}. \quad (4.4) \end{aligned}$$

Through Equations (4.1) and (4.3), we have

$$\begin{aligned} c_k &= \frac{\beta_1 a_k^2}{a_k + b_k}, \\ d_k &= \frac{a_k^2 + a_k b_k - \beta_1 a_k^2}{a_k + b_k}, \end{aligned} \quad (4.5)$$

and through Equations (4.2) and (4.4), we have

$$\begin{aligned} e_k &= \frac{\beta_0 a_k}{a_k + b_k} \sum_{k' \neq k} \frac{b_{k'}}{2}, \\ f_k &= \left(1 - \frac{\beta_0 a_k}{a_k + b_k}\right) \sum_{k' \neq k} \frac{b_{k'}}{2}. \end{aligned} \quad (4.6)$$

As a result, the prior distribution of π_k, given β_1 and β_0, is proportional to the product of three Beta distributions, i.e. $f(\pi_k|\beta_1, \beta_0) \propto Beta(a_k, b_k) \times Beta(c_k, d_k) \times Beta(e_k, f_k)$, where all parameters are functions of a_k, b_k, β_0, and β_1, which are fixed.

4.2.3 Approximate Likelihood and Posterior Distribution of Response Rates

To develop a likelihood for the response rates, we first define summary statistics based upon the numbers of responders for each arm in each stage. We let R_k denote the number of individuals who responded to treatment k in the first stage, $\tilde{R}_{k'k}$ denote the total number of individuals who failed to respond to treatments $k' \neq k$ in stage 1 and then switched to treatment k in stage 2, S_{kk} denote the number of individuals who responded to treatment k in both stages, and $T_{k'k}$ denote the total number of individuals who failed to respond to

treatments $k' \neq k$ in stage 1 and then responded to treatment k in stage 2.

Each of these statistics has a conditional Binomial distribution written explicitly as:

$$\begin{aligned} R_k|n,\pi_k &\sim Binomial(\pi_k, n) && (4.7)\\ S_{kk}|R_k,\pi_k,\beta_1 &\sim Binomial(\beta_1\pi_k, R_k) && (4.8)\\ T_{k'k}|\tilde{R}_{k'k},\pi_k,\beta_0 &\sim Binomial(\beta_0\pi_k, \tilde{R}_{k'k}). && (4.9) \end{aligned}$$

Hence, for each arm k, the joint distribution of the numbers of responders and non-responders across both stages is

$$\begin{aligned} f(R_k,\tilde{R}_{k'k},S_{kk},T_{k'k}|\pi_k,\beta_1,\beta_0) &= f(R_k,\tilde{R}_{k'k}|\pi_k,\beta_1,\beta_0)\times f(S_{kk},T_{k'k}|\pi_k,\beta_1,\beta_0,R_k,\tilde{R}_{k'k})\\ &\propto f(R_k|\pi_k)\times f(S_{kk}|R_k,\pi_k,\beta_1)\times f(T_{k'k}|\tilde{R}_{k'k},\pi_k,\beta_0) \quad (4.10) \end{aligned}$$

which is a product of the three conditional Binomial distributions given in Equations (4.7) - (4.9).

We now have a prior distribution for the response rates expressed as a product of three Beta distributions and a likelihood that is proportional to a product of three Binomial distributions. The conjugacy of Beta and Binomial distributions leads to a posterior distribution of π_k given β_1, β_0, and the data that is proportional to the product of three Beta distributions

$$\begin{aligned} f(\pi_k|R_k,\tilde{R}_{k'k},S_{kk},T_{k'k},\beta_1,\beta_0) &\propto f(R_k,\tilde{R}_{k'k},S_{kk},T_{k'k}|\pi_k,\beta_1,\beta_0)\times f(\pi_k|\beta_1,\beta_0)\\ &\propto Binomial(\pi_k,n)\times Beta(a_k,b_k)\\ &\quad \times Binomial(\beta_1\pi_k,R_k)\times Beta(c_k,d_k)\\ &\quad \times Binomial(\beta_0\pi_k,\tilde{R}_{k'k})\times Beta(e_k,f_k)\\ &\propto Beta(a_k+R_k,n-R_k+b_k)\times Beta(S_{kk}+c_k,R_k-S_{kk}+d_k)\\ &\quad \times Beta(T_{k'k}+e_k,\tilde{R}_{k'k}-T_{k'k}+f_k). \quad (4.11) \end{aligned}$$

However, due to the mathematical intractability of using a product of three Beta distributions, we further propose a normal approximation for each of the Beta distributions in Equation (4.11). Using the method-of-moments again, $Beta(a_k + R_k, n - R_k + b_k)$ is approximated by a normal distribution with mean ι_k and variance ω_k^2, denoted as $N(\iota_k, \omega_k^2)$, with its parameters found via

$$\begin{aligned} \iota_k &= \frac{a_k + R_k}{a_k + b_k + n}, \\ \omega_k^2 &= \frac{(a_k + R_k)(b_k + n - R_k)}{(a_k + b_k + n)^2(a_k + b_k + n + 1)}. \end{aligned} \tag{4.12}$$

Similarly, $Beta(S_{kk}+c_k, R_k - S_{kk}+d_k)$ is approximated by $N(\nu_k, \tau_k^2)$ and $Beta(T_{k'k}+e_k, \tilde{R}_{k'k} - T_{k'k} + f_k)$ by $N(\zeta_k, \lambda_k^2)$, where the parameters of $N(\nu_k, \tau_k^2)$ and $N(\zeta_k, \lambda_k^2)$ are determined by equating moments in the same way, i.e.,

$$\begin{aligned} \nu_k &= \frac{S_{kk} + c_k}{c_k + R_k + d_k}, \\ \tau_k^2 &= \frac{(S_{kk} + c_k)(R_k - S_{kk} + d_k)}{(c_k + R_k + d_k)^2(c_k + R_k + d_k + 1)}, \\ \zeta_k &= \frac{T_{k'k} + e_k}{e_k + \tilde{R}_{k'k} + f_k}, \\ \lambda_k^2 &= \frac{(T_{k'k} + e_k)(\tilde{R}_{k'k} - T_{k'k} + f_k)}{(e_k + \tilde{R}_{k'k} + f_k)^2(e_k + \tilde{R}_{k'k} + f_k + 1)}. \end{aligned} \tag{4.13}$$

From the above derivations, we can approximate the posterior distribution of π_k as proportional to the product of $N(\iota_k, \omega_k^2)$, $N(\nu_k, \tau_k^2)$, and $N(\zeta_k, \lambda_k^2)$. Our final approximation is based on the work of *Bromiley* (2003) who found that the product of three normal densities can be written as a single normal kernel. Although Bromiley's work focused on multiple normal kernels with the same argument, the concepts apply in our situation because θ_k and ϕ_k are scalar multiples of π_k. Thus, we write Equation (4.11) as proportional to a normal kernel equal to $exp(-(\pi_k - \mu_k)^2/(2\sigma_k^2))$, where

$$\begin{aligned}\mu_k &= \frac{\iota_k/\omega_k^2 + \beta_0\zeta_k/\lambda_k^2 + \beta_1\nu_k/\tau_k^2}{1/\omega_k^2 + \beta_0^2/\lambda_k^2 + \beta_1^2/\tau_k^2}, \\ \sigma_k^2 &= \frac{1}{1/\omega_k^2 + \beta_0^2/\lambda_k^2 + \beta_1^2/\tau_k^2}. \end{aligned} \tag{4.14}$$

Hence, we approximate the posterior distribution of π_k by $N(\mu_k, \sigma_k^2)$.

One remaining challenge for all Bayesian sample size calculations is that the posterior distribution is a function of yet to be collected data. To remedy this problem, we use Equations (4.7) - (4.9) to generate each of R_k, S_{kk}, and $T_{k'k}$ that are equal to their respective expected values. This approach is similar to the idea of an exemplary dataset proposed by *O'Brien and Shieh* (1998). The exemplary dataset is a function of both the sample size and the values of β_0 and β_1; we set β_0 and β_1 equal to their prior means, with β_1 truncated so that the response rate of the second stage responders is never greater than one.

4.2.4 Approximating the Posterior Mean of Difference of Two Best Arms

Unlike many trial designs in which there is a control arm or a treatment arm to which all other arms are compared, there is no such reference arm in the proposed snSMART design. Instead, once the posterior response rate of each arm is determined, inference is based upon the difference of the two largest posterior response rates. Thus, our sample size calculation must reflect the ordering that is implicit in inference (*Kim*, 1988). If we suppose the ordering of π_A, π_B, and π_C such that $\pi_{(3)} < \pi_{(2)} < \pi_{(1)}$, our goal is to determine the posterior mean and variance of $\mathcal{D} = \pi_{(1)} - \pi_{(2)}$, although we do not know the true ordering of π_A, π_B, and π_C.

We let $f_{(1)}(x)$ and $f_{(2)}(x)$ denote the posterior probability density functions (pdfs) of $\pi_{(1)}$ and $\pi_{(2)}$, respectively. Because our approximations for the posterior distributions of π_A, π_B, and π_C are conditionally independent, we have $F_{(1)}(x) = Pr(\pi_{(1)} \leq x) = Pr(\pi_A \leq x, \pi_B \leq$

$x, \pi_C \leq x) = F_A(x)F_B(x)F_C(x)$, where $F_k(x)$ is the cumulative distribution function (cdf) for π_k. This leads to the pdf

$$f_{(1)}(x) = f_A(x)F_B(x)F_C(x) + f_B(x)F_A(x)F_C(x) + f_C(x)F_A(x)F_B(x), \quad (4.15)$$

in which $f_k(x)$ is the pdf for π_k. The cdf and pdf of π_k are derived based on the posterior distribution for π_k in Equation 4.14. Similarly, $f_{(2)}(x)$ is a summation of all combinations for the probability that one of π_A, π_B, and π_C is greater than x, one is smaller than x, and one is from $[x, x+\epsilon)$, where ϵ goes to 0. The explicit expression of $f_{(2)}(x)$ is as follows:

$$\begin{aligned} f_{(2)}(x) = {} & [f_A(x)F_B(x) + f_B(x)F_A(x)][1 - F_C(x)] + \\ & [f_A(x)F_C(x) + f_C(x)F_A(x)][1 - F_B(x)] + \\ & [f_C(x)F_B(x) + f_B(x)F_C(x)][1 - F_A(x)]. \end{aligned} \quad (4.16)$$

Thus, the pdf of $\mathcal{D} = \pi_{(1)} - \pi_{(2)}$, denoted as $f_{\mathcal{D}}(D)$, is a simple convolution

$$f_{\mathcal{D}}(D) = \int_{-\infty}^{\infty} f_{(1)}(D+x) f_{(2)}(D) dx \quad (4.17)$$

and we can compute the mean of $\mathcal{D}$ as

$$\begin{aligned} E(\mathcal{D}) &= \int_0^{\infty} D f_{\mathcal{D}}(D) dD \\ &= \int_0^{\infty} D \int_{-\infty}^{\infty} f_{(1)}(D+x) f_{(2)}(x) dx dD \\ &= \int_{-\infty}^{\infty} f_{(2)}(x) \int_0^{\infty} D f_{(1)}(D+x) dD dx. \end{aligned} \quad (4.18)$$

with an analogous computation for the expected value of $\mathcal{D}^2$, from which we can derive the variance of $\mathcal{D}$, denoted as $var(\mathcal{D})$. *Bayarri and Berger* (2004) have advocated using the posterior distribution as an approximation to the sampling distribution when the prior information is reasonably objective. Hence, we approximate the posterior distribution of $\mathcal{D}$

as a normal distribution with mean being equal to $E(\mathcal{D})$ and variance being equal to $var(\mathcal{D})$, and from which we can simplify the sample size calculation.

4.2.5 ACC Sample Size Calculation

In order to compute a sample size based on the ACC, we need a fixed length, ℓ, for the posterior credible interval for $\mathcal{D} = \pi_{(1)} - \pi_{(2)}$ that achieves a desired average coverage rate, $1-\alpha$ over the posterior distribution. However, because the value of ℓ is often hard to specify, we instead focus on statistical power which has a better clinical interpretation and has a one-to-one relationship with the value of ℓ. We define the statistical power of an snSMART as the probability that we claim the best treatment and the second best treatment are significantly different when $\pi_{(1)} > \pi_{(2)}$. (i.e. when the credible interval for $\mathcal{D}$ excludes zero). The desired level of power is denoted as $1 - \xi$. From the statistical power and coverage rate, ℓ can be determined via a grid search.

Hence, under the ACC, the sample size in each arm, n, is the solution to the following equation,

$$1 - 2\Phi(-\frac{\ell}{2\sqrt{Var(\mathcal{D})}}) = 1 - \alpha, \tag{4.19}$$

where $\Phi(\cdot)$ is the pdf of a standard normal distribution.

4.2.6 Synopsis of Sample Size Algorithm

The results of Sections 4.2.1-4.2.5 can be summarized to the following algorithm to implement the sample size calculation for a three arm snSMART:

1. Specify the value for π_A, π_B and π_C and the hyper-parameters, a_k and b_k, $k = A, B, C$, for the priors of π_A, π_B and π_C.
2. Specify the expected coverage rate, $1 - \alpha$, and desired power $1 - \xi$.

3. Specify the hyper parameters, γ, κ, and ψ, for the priors of the linkage parameters, β_1 and β_0.

4. Let the value of β_1 be equal to the prior mean of $Pareto(1, \psi)$ right-truncated at $1/\pi_C$ and the value of β_0 be equal to the prior mean of $Beta(\gamma, \kappa)$.

5. For each value of $\ell \in \{2 \times (\pi_{(1)} - \pi_{(2)}), \ldots, 0.02, 0.01\}$, calculate the sample size using Equation (4.19) and the resulting power, until the power is at least $1 - \xi$.

6. The last calculated sample size in step (5) is the final recommended sample size.

4.2.7 An Example of Sample Size Calculation

To illustrate, consider an snSMART in which $\pi_A = 0.25$, $\pi_B = 0.25$, and $\pi_C = 0.5$. For prior distributions, we use $\pi_A \sim Beta(0.5, 1.5)$, $\pi_B \sim Beta(0.5, 1.5)$, $\pi_C \sim Beta(1, 1)$, $\beta_0 \sim Beta(1, 1)$, and $\beta_1 \sim Pareto(1, 3)$. We have a desired power of $1 - \xi = 0.8$ and coverage rate of $1 - \alpha = 0.9$. In this example, the prior mean of $Pareto(1, 3)$ with a truncation of the upper tail at $1/\pi_C$ is equal to 1.29 and the prior mean of $Beta(1, 1)$ is equal to 0.5. Hence, we let β_1 and β_0 be equal to 1.29 and 0.5 respectively. Following the algorithm in Section 4.2.6, we require 27 subjects per arm to compare the best and second best treatment.

4.3 Simulations

We present results from eight scenarios in which one treatment is superior to the others. The response rates can be found in Table 4.1. In scenarios 1-6, the response rate of the second best treatment is not unique and in scenarios 7-8 all three response rates differ. The goal of simulations is to compare the simulated power to the desired statistical power given the coverage rate (i.e., type I error rate). We set the desired statistical power at 80% and the coverage rate at 90% (i.e., a 5% one-sided type I error rate). We ran 2000 simulations for every scenario.

Table 4.1 presents results from ideal settings in which the prior means for each π_k are equal to the true response rates. In each scenario, we calculated the sample size and simulated power using both the proposed BJSM, and the weighted Z-statistic approach from *Tamura et al.* (2016). The weighted Z-statistic approach is based on simulations over a grid search of potential sample sizes until the sample size that provides a simulated power closest to the desired power is identified.

The BJSM approach suggests a lower required sample size than the weighted Z-statistic approach in all of the scenarios. Even with lower suggested sample sizes, the BJSM approach is often slightly conservative, since in scenarios 2-8, the simulated power is slightly higher than the desired power. In scenarios 1 and 2, the BJSM approach suggests 27 and 26 subjects per arm (for total sample sizes of 81 and 78), respectively. In these scenarios, the BJSM approach requires 46% and 37% less subjects per arm than the weighted Z-statistic approach, respectively.

Compared to scenarios 1 and 2, the difference in response rates between the best and second best treatments in scenarios 3 and 4 decreases from 0.25 to 0.20, therefore requiring 20 more subjects per arm to achieve the same desired power. The BJSM approach suggests 43% and 33% less subjects per arm, respectively, than the weighted Z-statistic approach.

Compared to scenarios 1 and 2, the difference in response rates between the best and second best treatments decreases from 0.25 to 0.15 in scenarios 5 and 6, therefore requiring around 70 more subjects per arm to achieve the same desired power. In scenarios 5 and 6, the BJSM approach saves 39% and 28% of subjects per arm, respectively, than the weighted Z-statistic approach. Results are similar for scenarios 7 and 8.

The power of significantly recognizing the best and second best treatments using the BJSM varies depending upon the prior distribution that we use for each π_k. In Table 4.2, we calculated the sample size and simulated power under less ideal settings than the settings in Table 4.1, when the prior means for each π_k were set to either the highest response rate

or the lowest response rate. Comparing Table 4.2 to Table 4.1, our sample size calculation suggests slightly larger sample sizes in scenarios 1, 3 and 4.

In scenarios 1 and 2, the BJSM yields 0.023 and 0.038 lower power comparing Table 4.2 to Table 4.1. The slightly lower power in Table 4.2 for these scenarios is due to the prior specification leads to closer treatment estimates compared to the truth. The specification of priors is influential when the calculated sample size is small. In scenarios 3-8, the BJSM yields similar power comparing Table 4.2 to Table 4.1, in these scenarios, the sample size is larger and thus the specification of priors is less influential on the statistical power.

Table 4.1: The estimated sample size and simulated power to identify the best from the second best treatment, when the prior mean of each π is equal to the true response rate. BJSM denotes the Bayesian joint stage model. We assume β_1 and β_0 are fixed at prior means of $Pareto(1,3)$ with a truncation of the upper tail at $1/max(\pi_A, \pi_B, \pi_C)$ and $Beta(1,1)$, respectively. The sample size per treatment arm is denoted as n. The power is the proportion of simulations that the difference between the best and second best treatments is significantly recognized. We set desired power to be 0.80 and coverage rate to be 0.90.

	Response rates			BJSM		Weighted Z-statistic
scenario	π_A	π_B	π_C	n	power	n
1	0.25	0.25	0.50	27	0.796	50
2	0.15	0.15	0.40	26	0.815	41
3	0.30	0.30	0.50	47	0.820	82
4	0.20	0.20	0.40	46	0.817	69
5	0.35	0.35	0.50	94	0.859	153
6	0.25	0.25	0.40	94	0.849	131
7	0.30	0.40	0.50	171	0.843	284
8	0.20	0.30	0.40	174	0.860	248

Table 4.2: The estimated sample size and simulated power to identify the best from the second best treatment, when the prior mean of each π is equal to the highest response rate or lowest response rate. The sample size calculation approach is the Bayesian joint stage model (BJSM). We assume β_1 and β_0 are fixed at prior means of $Pareto(1,3)$ with a truncation of the upper tail at $1/max(\pi_A, \pi_B, \pi_C)$ and $Beta(1,1)$, respectively. The sample size per treatment arm is denoted as n. The power is the proportion of simulations that the difference between the best and second best treatments is significantly recognized. We set desired power to be 0.80 and coverage rate to be 0.90.

	Response rates			Prior set at highest		Prior set at lowest	
scenario	π_A	π_B	π_C	n	power	n	power
1	0.25	0.25	0.50	28	0.773	28	0.778
2	0.15	0.15	0.40	26	0.777	26	0.786
3	0.30	0.30	0.50	48	0.818	48	0.817
4	0.20	0.20	0.40	47	0.822	46	0.807
5	0.35	0.35	0.50	94	0.852	94	0.848
6	0.25	0.25	0.40	94	0.843	94	0.837
7	0.30	0.40	0.50	171	0.842	171	0.845
8	0.20	0.30	0.40	174	0.854	174	0.855

4.4 Concluding Remarks

In this manuscript, we present a Bayesian method to calculate the sample size required for a two-stage snSMART with three active treatments. We compared the BJSM approach to a simulation-based weighted Z-statistic approach and found that the BJSM approach requires fewer subjects to achieve the same level of power.

We made two key assumptions to simplify our calculations: (i) the response rates of the three active treatments, given β_1 and β_0, are independent; and (ii) the response rate of the best treatment is unique. Although we assume independence of the response rates of the three treatments, simulation results show that the correlation between the best and second best treatment ($\pi_{(1)}$ and $\pi_{(2)}$) is positive, but weak. Thus, this violation leads to a mildly conservative sample size calculation.

For assumption (ii), it is possible to have settings where the best treatment response rate is not unique. For example, investigators may be interested in comparing two novel treatments to a standard of care. We can generalize the proposed sample size calculations to this objective by directly applying the same sample size calculation procedure to the largest and the smallest posterior response rates. The total sample size for an snSMART to compare two novel treatments to the standard treatment will thus be the calculated sample size times three.

The proposed sample size calculation uses method-of-moments to approximate Beta distributions with normal distributions. Simulations and Q-Q plots show the approximation works well when the sample size is at least 10 per arm, which is valid in realistic settings for the snSMART. Finally, the proposed sample size calculation approximates the posterior distribution of the difference between the best and second best treatment from the BJSM with a normal distribution. Simulation results suggest that the approximation is appropriate even in small sample settings.

The proposed sample size calculation is based upon a three-arm snSMART, but the ap-

proach can be generalized to snSMARTs with more than three arms. The generalization is a natural extension such that the prior distributions, likelihood, and posterior distributions for all added arms are first approximated. The sample size is then calculated based on the comparison of the best and second best treatments with the same procedure as described in Section 4.2. Software to implement the sample size calculations for a three arm snSMART is available from the authors.

CHAPTER V

Summary and Future Work

In this dissertation, we explored Bayesian methods for Phase I Trials and small n Sequential Multiple Assignment Randomized Trials (snSMARTs). We proposed a systematic approach in selecting the skeleton and prior standard deviation that saves computation time and provides good precision of the MTD selection. We also proposed a data analysis method and sample size calculation for snSMART designs. The proposed joint stage model is efficient in estimating the response rates from three active treatments.

In the first project, the proposed method of selecting the skeleton and prior standard deviation for the CRM in Phase I trials of a single agent provides a simulation free way that is more time efficient and maintains good MTD identification. Specifically, the proposed method indicates that using weights and the expected outcome of DLT accommodates the information of DLT rates for all doses and the adaptive treatment assignments towards the end of study. The weights and expected outcome of DLT are determined when the design parameters of a trial, the true DLT rates and the values of skeleton and prior standard deviation are given, which allows us to extend our research in studying the typical characteristics of skeleton and prior standard deviation that leads to high precision in MTD identification. *Lee and Cheung* (2011) have already shown that a smaller value of prior standard deviation will lead to a better selection when the MTD is in the middle. In a future study, we will investigate how the space of skeleton and the *a priori* MTD in the skeleton, jointly with

prior standard deviation, are associated with the MTD identification.

In the second project, the Bayesian joint stage model (BJSM) jointly analyzes the data from the first and second stages of an snSMART design. Via modeling the correlation between response rates at the first and second stages using the linkage parameter between the data from two stages due to the multiple treatment assignment, the BJSM accommodates information from two stages and is more efficient than one stage estimators. The method in the second project could be expanded to include covariates. By including values that may be associated with outcome of response, we could potentially improve the estimation efficiency. Additionally, other snSMART designs that consider not only binary outcome of response but also continuous or time to event outcomes can be analyzed by modifying the BJSM, which provides extra flexibility in the design and analysis of snSMARTs. We did not equip the BJSM to handle carry-over effects because the carry-over effects for three specific treatments for investigations in our application of ARAMIS were negligible. The BJSM can be improved to capture the carry-over effect by calibrating the assumptions on the linkage parameters.

In the third project, we propose an explicit sample size calculation equation that provides reasonable and fast sample size estimations for a desired statistical power and coverage rate based on the BJSM. We have developed a corresponding R shiny applet that enable the researchers to calculate the sample size required for an snSMART. In the sample size estimations, we propose a Normal approximation method for the posterior difference of the response rates between the top two efficacious treatments that lack an explicit distribution and found our approximation works well in realistic settings for rare disease. The sample size calculations in project three could be expanded to calculate the sample size for snSMARTs with more than three arms or non-binary outcomes. Though the sample size calculation is developed to identify the most efficacious treatment from three active treatments in an snSMART, we can also expand our method to calculate sample size for snSMARTs when the goal is instead comparing treatments for investigations with placebo or reference treatment

by taking contrasts of the most and least efficacious treatments instead of the top two efficacious treatments.

BIBLIOGRAPHY

107th Congress (2002), Rare diseases act of 2002, *Public Law 107–280*.

Adcock, C. (1988), A Bayesian approach to calculating sample sizes, *The Statistician*, *37*, 433–439.

Bayarri, M. J., and J. O. Berger (2004), The interplay of Bayesian and frequentist analysis, *Statistical Science*, *19*(1), 58–80.

Bell, S. A., and C. T. Smith (2014), A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of clinicaltrials.gov, *Orphanet Journal of Rare Diseases*, *9*(170), 1–11.

Braun, T. M. (2014), The current design of oncology phase I clinical trials: progressing from algorithms to statistical models, *Chinese Clinical Oncology*, *3*, 1–11.

Braun, T. M. (2018), Motivating sample sizes in adaptive phase I trials via Bayesian posterior credible intervals, *Biometrics*, *74*(3), 1065–1071.

Bromiley, P. (2003), Products and convolutions of Gaussian probability density functions, *Tina Vision Memo*, *3*(4), 1.

Cao, J., J. J. Lee, and S. Alber (2009), Comparison of Bayesian sample size criteria: ACC, ALC, and WOC, *Journal of Statistical Planning and Inference*, *139*(12), 4111–4122.

Chernozhukov, V. (2007), course materials for 14.385 nonlinear econometric analysis, *MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology*, p. Downloaded on 25 June 2019.

Cheung, Y. K., and R. Chappell (2002), A simple technique to evaluate model sensitivity in the continual reassessment method, *Biometrics*, *58*, 671–674.

Chevret, S. (1993), The continual reassessment method in cancer phase i clinical trials: a simulation study, *Statistics in Medicine*, *12*(12), 1093–1108.

Dawson, R., and P. W. Lavori (2011), Efficient design and inference for multistage randomized trials of individualized treatment policies, *Biostatistics*, *13*(1), 142–152.

Evans, S. R. (2010), Clinical trial structures, *Journal of Experimental Stroke & Translational Medicine*, *3*(1), 8–18.

Faries, D. (1994), Practical modifications of the continual reassessment method for phase I cancer clinical trials, *Journal of Biopharmaceutical Statistics*, *4*, 147–164.

Garrett-Mayer, E. (2006), The continual reassessment method for dose-finding studies: a tutorial, *Clinical Trials*, *3*, 57–71.

Goodman, S., M. Zahurak, and S. Piantadosi (1995), Some practical improvements in the continual reassessment method for phase I studies, *Statistics in Medicine*, *14*, 1149–1161.

Griggs, R. C., et al. (2009), Clinical research for rare disease: Opportunities, challenges, and solutions, *Molecular Genetics and Metabolism*, *96*(1), 20–26.

Gupta, S., M. E. Faughnan, G. A. Tomlinson, and A. M. Bayoumi (2011), A framework for applying unfamiliar trial designs in studies of rare diseases, *Journal of Clinical Epidemiology*, *64*(10), 1085–1094.

Halekoh, U., S. Højsgaard, J. Yan, et al. (2006), The R package geepack for generalized estimating equations, *Journal of Statistical Software*, *15*(2), 1–11.

Hanley, J. A., A. Negassa, et al. (2000), GEE analysis of negatively correlated binary responses: a caution, *Statistics in Medicine*, *19*(5), 715–722.

Joseph, L., and P. Bélisle (1997), Bayesian sample size determination for normal means and differences between normal means, *The Statistician*, *46*(2), 209–226.

Kim, W.-C. (1988), On detecting the best treatment, *Journal of the Korean Statistical Society*, *17*(2), 82–92.

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.

Lee, S., and Y. Cheung (2009), Model calibration in the continual reassessment method, *Clinical Trials*, *6*, 227–238.

Lee, S. M., and Y. K. Cheung (2011), Calibration of prior variance in the Bayesian continual reassessment method, *Statistics in Medicine*, *30*(17), 2081–2089.

Levin, K. A. (2007), Study design VII. randomised controlled trials, *Evidence-Based Dentistry*, *8*(1), 22–23.

Makubate, B., and S. Senn (2010), Planning and analysis of cross-over trials in infertility, *Statistics in Medicine*, *29*(30), 3203–3210.

Moller, S. (1995), An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients in order to investigate a greater range of doses, *Statistics in Medicine*, *14*, 911–922.

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

O'Brien, R. G., and G. Shieh (1998), A simpler method to compute power for likelihood ratio tests in generalized linear models, in *Annual Joint Statistical Meetings of the American Statistical Association of 1998, Dallas, Texas*.

O'Quigley, J., and L. Z. Shen (1996), Continual reassessment method: a likelihood approach, *Biometrics*, pp. 673–684.

O'Quigley, J., M. Pepe, and L. Fisher (1990), Continual reassessment method: A practical design for phase I clinical trials in cancer, *Biometrics*, *46*, 33–48.

O'Quigley, J., X. Paoletti, and J. Maccario (2002), Non-parametric optimal design in dose finding studies, *Biostatistics*, *3*(1), 51–56.

Pan, W., and J. E. Connett (2002), Selecting the working correlation structure in generalized estimating equations with application to the lung health study, *Statistica Sinica*, *12*(2), 475–490.

Pham-Gia, T., and N. Turkkan (1992), Sample size determination in Bayesian analysis, *The Statistician*, *41*, 389–397.

Robins, J. (1986), A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect, *Mathematical Modelling*, *7*(9–12), 1393–1512.

Rosenbaum, P. R. (1987), Model-based direct adjustment, *Journal of the American Statistical Association*, *82*(398), 387–394.

Shen, L. Z., and J. O'Quigley (1996), Consistency of continual reassessment method under model misspecification, *Biometrika*, *83*, 395–406.

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. (2016), Smart design, conduct, and analysis in oncology, in *Adaptive Treatment Strategies in Practice*, edited by M. R. Kosorok, E. E. M. Moodie, and P. F. Thall, chap. 4, pp. 41–54, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

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*(26), 4687–4702.

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*(26), 3723–3732.

Williamson, T., M. Eliasziw, and G. H. Fick (2013), Log-binomial models: exploring failed convergence, *Emerging Themes in Epidemiology*, *10*(14), 1–10.

Yin, G., and Y. Yuan (2009), Bayesian model averaging continual reassessment method in phase I clinical trials, *Journal of the American Statistical Association*, *104*, 954–968.

Zeger, S. L., and K.-Y. Liang (1986), Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, *42*(1), 121–130.

Zhang, J., T. M. Braun, and J. M. Taylor (2013), Adaptive prior variance calibration in the Bayesian continual reassessment method, *Statistics in Medicine*, *32*(13), 2221–2234.