2019

Site- and Location-Adjusted Approaches to Adaptive Allocation Clinical Trial Designs

Brian S. Di Pace
Virginia Commonwealth University

Site- and Location-Adjusted Approaches to Adaptive Allocation Clinical Trial Designs

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

Brian S. Di Pace, M.P.H.

Committee Members:
Roy T. Sabo, Ph.D., Associate Professor, Department of Biostatistics (Director)
David C. Wheeler, Ph.D., M.P.H.,Department of Biostatistics
Le Kang, Ph.D., Department of Biostatistics
Amir Toor, M.D., Department of Internal Medicine
Qiong Zhang, Ph.D., M.A., School of Mathematical and Statistical Sciences, Clemson University

Department of Biostatistics
Virginia Commonwealth University
Richmond, Virginia

January 29, 2019

ACKNOWLEDGMENTS

The last few years have been some of the most challenging and humbling times of my life. This journey would not have been possible without the support of numerous people. I would first like to acknowledge my dissertation advisor, Dr. Roy Sabo. I am extremely thankful for his patience and his guidance throughout this process. He encouraged me to overcome all the challenges I have encountered in my research and has supported me along the way. I would also like to thank Dr. Tina D. Cunningham, my mentor from my master's program. Without her, I would not have discovered my interest in biostatistics or pursued this doctoral degree.

None of this would have been possible without the support and the team effort of my Fall 2014 cohort, BAMF – Yu Cao, Victoria Garcia, Camille Hochheimer, Alicia Johns and Anny-Claude Joseph. Thank you all for those late nights and long weekends working together to survive graduate school. I do not believe I would be here today without you. I would be honored to work with each of you in our future endeavors. I would also like to thank another colleague, Jordyn Wallenborn, for all the baked goods and encouraging words over the years.

I would also like to acknowledge all of the students and faculty in the Biostatistics department. Specifically, I would like to thank my committee members – Dr. Le Kang, Dr. Amir Toor, Dr. David Wheeler and Dr. Qiong Zhang – for their guidance with my research. I am grateful for all the help from Russ Boyle and Yvonne Hargrove, they always either knew the answer or knew exactly who to ask to find out. I would also like to thank my coworker, Viviana Rodriguez, and my boss, Dr. Adam Sima, for all their help and consulting advice in the last few years.

Finally, I would like to thank my family and friends for supporting me throughout the pursuit of my academic career. This journey has had many ups and downs, but you all have been there to keep pushing me forward. My mother has always believed in me and tells me how proud she is. She has accompanied me every step of the way, though my education has found me hundreds of miles away. I would also like to give a special thanks to WPC for all the support and love I have received. You have been a source of great motivation and happiness in the last two years.

Again, I offer my deepest gratitude to all these individuals for guiding me along this path. I look forward to beginning the next chapter of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

SITE- AND LOCATION-ADJUSTED APPROACHES TO ADAPTIVE ALLOCATION CLINICAL TRIAL DESIGNS

By Brian S. Di Pace, M.P.H.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2019

Director: Roy T. Sabo, Ph.D., Associate Professor, Department of Biostatistics

Response-Adaptive (RA) designs are used to adaptively allocate patients in clinical trials. These methods have been generalized to include Covariate-Adjusted Response-Adaptive (CARA) designs, which adjust treatment assignments for a set of covariates while maintaining features of the RA designs. Challenges may arise in multi-center trials if differential treatment responses and/or effects among sites exist. We propose Site-Adjusted Response-Adaptive (SARA) approaches to account for inter-center variability in treatment response and/or effectiveness, including either a fixed site effect or both random site and treatment-by-site interaction effects to calculate conditional probabilities. These success probabilities are used to update assignment probabilities for allocating patients between treatment groups as subjects accrue. Both frequentist and Bayesian models are considered. Treatment differences could also be attributed to differences in social determinants of health (SDH) that often manifest, especially if unmeasured, as spatial heterogeneity amongst the patient population. In these cases, patient residential location can be used as a proxy for these difficult to measure SDH. We propose the Location-Adjusted Response-Adaptive (LARA) approach to account for location-based variability in both treatment response and/or effectiveness. A Bayesian low-rank kriging model will interpolate spatially-varying joint treatment random effects to calculate the conditional probabilities of success, utilizing patient outcomes, treatment assignments and residential information. We compare the proposed methods with several existing allocation strategies that ignore site for a variety of scenarios where treatment success probabilities vary.

CHAPTER 1

INTRODUCTION

1.1 Background

Response-Adaptive (RA) allocation designs are used in clinical trials to achieve one of two aims: as an ethical approach to randomizing patients more effectively by minimizing treatment failures, or as an optimizational approach to maximize statistical power. This approach has been generalized to the Covariate-Adjusted Response-Adaptive (CARA) design with the goal to adjust the treatment assignment for a set of covariates or to guard against differential treatment effects, while maintaining the features of the adaptive design.

Clinical trials often recruit patients from multiple centers, where adaptive allocation algorithms, if used, estimate treatment effectiveness using all available data, irrespective of site[1]. This presents several challenges since treatment responses and/or effects can vary across sites; an extreme example may be where treatment effectiveness is reversed between sites. While the variability in treatment responses does not necessarily affect the estimated treatment differences across centers in a trial, variability in treatment effectiveness can result in treatment differences that cannot be applied universally to all centers. Consider a clinical trial where patients at one center have different success probabilities, and thus treatment effectiveness, than patients at all other sites where the treatment is superior. If we ignore site and conclude that overall there is a superior treatment effect, patients at the one site with different treatment effectiveness may be at risk [1].

Gallo (2008) refers to these differences in treatment effects as a treatment-by-center interaction that can be classified into two categories based on the direction and magnitude of the effects: quantitative and qualitative. Quantitative interactions result in treatment effects that differ in magnitude across centers yet have the same sign, while qualitative interactions result in both differences in magnitude and direction across centers [1]. A treatment-by-center interaction inhibits researchers from estimating a single treatment effect for all sites in the trial. However, knowledge of varying treatment effects can be advantageous because it provides researchers with information

[1]Site and center will be used interchangeably to define the location where patients receive medical treatment.

on the site and patient characteristics, such as where the treatment effect is largest, smallest or reversed [2].

Treatment differences could also be attributed to differences in social determinants of health (SDH), including variability in socioeconomic status (SES) among patients, population size/type, and quality of health care, for example, along with other lifestyle and landscape factors that drive heterogeneity between patient populations. There are some instances where SES has been associated with treatment adherence. For example, in a study of adjuvant systemic therapy guideline adherence in early breast cancer patients living in the Netherlands, women with medium and low SES were significantly less likely to be over-treated with adjuvant chemotherapy compared to those with high SES [3]. Treatment effectiveness could be affected by SES or other SDH although we may not have accurate measurements to capture them. Especially when unmeasured, these differences often manifest as spatial heterogeneity amongst the patient population. In these cases, trial site or patient residential location could be used as proxies for these difficult to measure SDH. This demonstrates the importance of incorporating location in adaptive allocation designs to account for variability in treatment responses and/or effectiveness.

A study by Longini Jr. et al. analyzed data from a placebo-controlled influenza vaccine multicenter trial (1996-1998) to estimate vaccine efficacy for susceptibility to culture-confirmed influenza (VE_S) in children. While adaptive randomization was not used, the goal was to account for inter-center variation along with additional covariates using a generalized linear mixed model (GLMM) approach, where the cities included were assumed to be a random sample from a larger population. The crude center-specific VE_S estimates for year 1 ranged from 0.72 in Baltimore to 1.00 in several cities, resulting in an inter-center standard deviation of 0.08. When including the vaccination status as a fixed effect and center as a random effect in the GLMM, the variability of the center effect is 0.08 and the overall estimated $VE_S = 0.94$ (95% CI : $0.89, 0.97$). Although the center effect did not reach statistical significance, it was included in the model to account for an additional source of variability in estimating the overall vaccine efficacy. The authors also considered models that incorporated vaccine-by-center interaction effects. If RA randomization was used instead of a 2:1 allocation ratio, the pooled VE_S would have resulted in more children being randomized to the vaccine group, regardless of the center-specific vaccine efficacy rates [4].

1.2 Current Methods

Traditionally, a balanced design has been used in binary response clinical trials because it often achieves high power for tests of treatment effects. Throughout a trial, this equal allocation corresponds to "clinical equipoise", or the stance that investigators have no preference for either treatment group [5, 6]. As a trial progresses, there are at times data suggesting that one treatment is more beneficial, though equal allocation ignores this information. This may benefit future patients that otherwise have a fifty percent chance of being allocated to the inferior treatment group [5].

The allocation rules discussed here are appropriate for trials where the response is ascertained shortly after treatment to allow for adaptation during the randomization process. An example of such a trial is the evaluation of crystalloid preload in hypotension prevention following spinal anesthesia for elective cesarean sections which utilized the 'play the winner' rule [7]. Adaptive allocation methods have been extended to handle delayed responses by calculating the joint distribution of the delayed response and an earlier correlated outcome and using this joint distribution to estimate allocation weights [8]; however, we will not cover these models here.

1.3 Proposed Methods

A potential method to account for inter-center variability in treatment effectiveness is to use separate adaptive allocation algorithms for each center. Instead of an overall treatment effect estimated using data from all sites, only local patient data would be used resulting in site-specific estimates of treatment effectiveness and assignment probabilities. The problem is that small samples sizes at centers would potentially result in convergence issues and unstable estimates, especially in trials with a large number of sites. The benefits of adapting would be negatively affected even with lead-in approaches to maintain balanced allocation until enough subjects are recruited to prevent estimator instability.

Alternatively, one could account for inter-center variability in response-adaptive randomization by extending the covariate-adjusted designs. In general, the CARA design estimates the predicted value of treatment success for each treatment group given a set of observed covariates. One method to account for inter-center variability in treatment effectiveness is to treat site as a fixed effect in

a generalized linear model (GLM) [9], specifically a marginal model [10]. Treatment estimates would be interpreted as a population-averaged treatment effect that applies only to the centers participating in the study. However, our focus is to adopt a method where the sites are assumed to come from a larger population, as in a previous study.

Another approach is to treat site as a random effect in a conditional model, where the treatment effect is estimated for individuals by averaging across centers, allowing for heterogeneity in the effects across centers. Here, sites are assumed to come from a larger population, as in several previous studies [4, 11, 10]. A GLMM allows inferences to be more generalizable than to just the centers chosen for the study, however, the assumption that centers are randomly selected from a larger population is debatable [12, 13]. A GLMM requires that the number of sites being sufficiently large for consistent estimation of the model parameters [11]. The consistency of the ordinary ML estimators for the difference of proportions, and other difference measures, breaks down under sparse large sample methods (asymptotics) for fixed effects models but not for random effects models [14]. Covariate-adjusted response-adaptive approaches could be extended to use random effects in a hierarchical framework. Due to the flexibility of the CARA approach, covariates can be added to the GLMM to adjust the model and ultimately the allocation probabilities.

1.4 Specific Aims

1.4.1 Aim 1: A Site-Adjusted Approach to Adaptive Allocation in Clinical Trials to account for site-based response variability.

Treatment responses can vary considerably across locations in a multi-center clinical trial due to a variety of reasons such as heterogeneous patient populations and differences in health care provider practices. We propose the Site-Adjusted Response-Adaptive (SARA) allocation approach to account for site-based variability in treatment responses in the presence of slight quantitative treatment-by-center interactions. A random site effect will be included in the generalized linear mixed model (GLMM) framework. SARA incorporates readily available information, namely site, about the subject so that conditional success probabilities are specific to each center. The allocation weights will be updated with the predicted probabilities of success conditioned on the site specific to the next patient enrolled.

1.4.2 Aim 2: A Bayesian Approach to Site-Adjusted Response-Adaptive Allocation in Clinical Trials to account for variability in treatment effectiveness.

This approach will extend the SARA design from Aim 1 by including a random treatment-by-site interaction effect to account for inter-site variability in treatment effectiveness. Quantitative and qualitative treatment-by-site interactions will be explored at both 'low' and 'high' magnitudes. A Bayesian approach will be used to overcome computational challenges from modeling both random site and treatment-by-site interaction effects.

1.4.3 Aim 3: A Bayesian Spatial Approach to Adaptive Allocation in Clinical Trials to account for location-based variability in treatment responses and/or effectiveness

This approach will focus on developing spatial statistical methods under the Bayesian framework to account for inter-site variability in both treatment response and effectiveness in the design phase. To adjust for the presence of a spatially heterogeneous treatment effect and/or response, spatially-varying joint treatment random effects will be estimated. The Location-Adjusted Response-Adaptive (LARA) approach will utilize treatment and residential information to predict a patient's response to each treatment and in turn update the allocation weights for subsequent patients.

CHAPTER 2

ACCOUNTING FOR SITE-BASED RESPONSE VARIABILITY

In this chapter, we introduce a site-adjusted approach to adaptive allocation in clinical trials to account for site-based response variability. Current response-adaptive methods ignore potential differences in treatment responses and/or effectiveness across sites, and as such are likely to translate these phenomena into biased or improper patient allocation. In Section 2.1, we summarized the existing adaptive methods and demonstrated that patient allocation and treatment effectiveness are adversely affected when multiple sites are not accounted for during randomization. In Section 2.2, we introduce two approaches to account for site-based variability in treatment responses in the presence of treatment-by-center interactions. A simulation study is conducted in Section 2.3 to compare the performance of our proposed methods to both balanced and RA randomization. The results are explored and a brief discussion of the implications follows.

2.1 Background and Existing Methods

2.1.1 Response-Adaptive Allocation

Response-adaptive (RA) randomization updates the allocation probabilities based on previous treatment assignments and patient responses to meet some aim, such as maximizing power or minimizing the number of patients assigned to an inferior treatment. Hu and Rosenberger (2006) describe the three features that define RA randomization, namely that they are 1) myopic, 2) fully randomized and 3) have a sample size fixed in advance. A myopic process is one which incorporates accrued data to make decisions on subsequent patients. The RA design is defined as a myopic process because both treatment assignments and responses of the previous subjects are used to iteratively update the allocation weights (w_1, w_2) for the next subject. The RA procedure is fully randomized since each patient is randomly assigned to a treatment group. While the allocation weights are dependent on data from the previous subjects, the assignment for each patient is random and thus independent. Finally, sample sizes are fixed before the trial begins, though early stopping rules may be implemented [15].

While much work has been done to generalize Efron's biased coin design (1971) [16] and Wei's urn design (1978) [17], we will focus on the RA procedure based on sequential maximum likelihood (ML) methods (doubly-adaptive biased coin design) [15]. In the RA design, allocation probabilities are skewed based on treatment successes with respect to the primary outcome. As such, it is necessary to have a single outcome that is ascertainable shortly after treatment so that the response of the current patient is available prior to the recruitment of the next individual. Here, this practically instantaneous response allows a sufficient number of patients to be recruited in a reasonable amount of time [18].

Consider two treatment groups, A and B, with a binary outcome ($Y = 1$ success, $Y = 0$ failure) and a success probability, $P(Y_i = 1) = p_i$, for the i^{th} patient, $i = 1, \ldots, n$. A success probability is the probability of observing patients responding favorably to the treatment, or treatment successes, given the existing data. The model for n patients is specified as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1 T_i, \ i = 1, \ldots, n, \tag{2.1}$$

where β_0 is the overall intercept, β_1 is the treatment effect and T_i is the treatment group indicator (fixed effect),

$$T_i = \begin{cases} 1 & \text{if treatment A,} \\ 0 & \text{otherwise.} \end{cases}$$

The estimated success probabilities $\hat{p}_t$ are defined for each treatment group, $t = 1, 2$, as follows:

$$\begin{aligned} \hat{p}_1 &= \text{logit}^{-1}\{\hat{\beta}_0 + \hat{\beta}_1\} \\ &= \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1\}}, \\ \hat{p}_2 &= \text{logit}^{-1}\{\hat{\beta}_0\} \\ &= \frac{\exp\{\hat{\beta}_0\}}{1 + \exp\{\hat{\beta}_0\}}, \end{aligned} \tag{2.2}$$

where $\hat{\beta}_0, \hat{\beta}_1$ are the ML estimates of the intercept and treatment effect from the GLM. Prior to implementing the RA design, a hard lead-in is initially used where equal allocation ($w_n = 0.5$) is used to randomize the first m patients between treatment groups until the regression parameters

in equation (2.1) become estimable. No conclusive research exists for selecting the lead-in length for balanced allocation, however, alternate treatment balancing methods [19], such as Pocock and Simon's procedure, or continual convergence checks [20] have been performed prior to adapting. Desirable properties of the lead-in length include a balance of being 1) long enough to overcome small sample estimation problems, but 2) short enough to allow for sufficient adaptation.

The RA design is used to allocate patients between treatment groups with various aims. For example, two goals an investigator could have are to maximize power or minimize the number of treatment failures. One way of achieving the latter goal is by allocating patients with higher probability to the best performing treatment arm [15]. These aims can be achieved by choosing different loss functions. When the simple difference measure ($f\{p_A, p_B\} = p_A - p_B$) is used to compare two binomial probabilities, the Optimal and Neyman allocation rules can be used. For a fixed variance of the test statistic, Optimal allocation minimizes the expected number of treatment failures. Let $\mathscr{F}_i$ be a function consisting of only treatment assignments and individual outcomes for the accrued $i = 1, \ldots, n$ patients. The Optimal allocation equations are often functions of the sample data as follows:

$$\begin{aligned} w_{1,O}(\mathscr{F}_i) &= \frac{\sqrt{\hat{p}_1}}{\sqrt{\hat{p}_1} + \sqrt{\hat{p}_2}}, \\ w_{2,O}(\mathscr{F}_i) &= 1 - w_{1,O}, \end{aligned} \tag{2.3}$$

where $\hat{p}_t$ is defined as the observed proportion of successes in treatment group t out of the total number of patients in that group, $t = 1, 2$. Rosenberger et al. (2001) note that by having less treatment failures, their optimal adaptive rule in equation (2.3) performs best with moderate success probabilities. However, this adaptive rule is protective of patients when the treatment effect is larger than expected at the beginning of the trial. Specifically, the adaptive design results in similar power but with less treatment failures compared to balanced allocation. When there is little to no treatment difference, there is no drawback to using this adaptive rule over equal allocation. Since the allocation weights will be close to 0.5, both methods will result in similar type I error rates [5].

Neyman Allocation refers to when patients are allocated proportional to the standard deviation of the parameter estimates with the goal to maximize power. For example, by allocating more subjects to the treatment group with a larger variability, the standard error of the proportion for

that group is reduced. Consequently, the overall variability of the test statistic is reduced, as it is a function of the standard errors for both proportions, increasing the magnitude of the test statistic and thus power of the design. The Neyman allocation equations are often functions of the sample data as follows:

$$\begin{aligned} w_{1,N}(\mathscr{F}_i) &= \frac{\sqrt{\hat{p}_1\hat{q}_1}}{\sqrt{\hat{p}_1\hat{q}_1}+\sqrt{\hat{p}_2\hat{q}_2}}, \\ w_{2,N}(\mathscr{F}_i) &= 1 - w_{1,N}, \end{aligned} \tag{2.4}$$

where $\hat{p}_t$ is defined as the observed proportion of successes in treatment group t out of the total number of patients in that group and $\hat{q}_t = 1 - \hat{p}_t$ is the proportion of treatment group t failures, $t = 1, 2$. Neyman allocation is unethical from a patient safety point of view when $\hat{p}_1 > (1-\hat{p}_2)$ since more subjects would be allocated to the inferior treatment group [5]. It is important to note that $w_{t,O}$ is proportional to $\hat{p}_t$ while $w_{t,N}$ is proportional to the standard deviation of $\hat{p}_t$. While other weight formulations could be used besides equations (2.3) and (2.4), their optimization criterion is not certain.

Adaptive allocation designs can cause an imbalance in treatment groups which sometimes results in lower power compared to balanced designs. However, there are exceptions where unbalanced designs result in better performance (e.g. power, number of treatment failures) than balanced allocation [19]. In cases where response rates are low (below 0.5), for example, the RA design may increase power. It has been shown that an allocation ratio as extreme as 7:3 is needed to severely impact power [18].

2.1.2 Covariate-Adjusted Response-Adaptive (CARA) Design

Adaptive randomization can result in skewed treatment groups which increases the likelihood of an imbalance in covariates. Consequently, adaptive designs that rely completely on patient responses fail to capture the complexity of heterogeneous patient populations. Patient characteristics or measurements may help to explain differential treatment efficacy or effectiveness. Some of these covariates include prognostic factors and time effects. Stratified or covariate-adaptive[1] randomization procedures can be performed to allocate patients within strata of an important characteristic

[1]This is also known as a *minimization* procedure or *dynamic allocation.*

in the design phase. These covariates can be adjusted for in a post-hoc manner through multiple regression modeling or post-stratification, or strict eligibility criteria can be implemented for more homogeneous populations [15, 19]. However, the feasibility of these methods is restricted to discerning — and thus having measures for — the appropriate covariates to use out of all information available. It is also difficult to perform RA randomization if there are too many strata due to potentially small sample sizes within strata.

The CARA design, a generalization of RA randomization, adjusts the allocation for a set of covariates to get characteristic-specific assignments. Examples of these factors include subject age, gender, race/ethnicity, baseline disease severity or risk score. Here, treatment assignments, patient outcomes and covariates are used to update the allocation ratio. We will focus on extending CARA randomization for binary outcomes [20, 21, 22]. However, the CARA design has also been created for survival outcomes [23] and continuous treatment responses, where the outcome can be normal or a general class of continuous responses [24, 25, 26, 27]. Note that this design differs from covariate-adaptive randomization, which aims to balance treatment assignments among a set of covariates. Accordingly, an adaptive design that allocates patients to the better performing treatment group based on both treatment responses and patient characteristics is considered [15, 20].

Treatment effect mappings is a method that was previously developed for continuous [28], survival [29] and binary outcomes [20] to map the current treatment effect (treatment A - treatment B) calculated from accrued patients to a probability $w_n \in [0, 1]$, which is then used to assign the next patient to treatment group A. Balanced allocation occurs when $w_n = 0.5$, indicating that the two treatment groups are performing comparably and thus have similar numbers of successes. Consider two treatment groups, A and B, with a binary outcome ($Y_i = 1$ success, $Y_i = 0$ failure) and a success probability, p_i, for the i^{th} patient, $i = 1, \ldots, n$. Rosenberger et al. (2001) specify a standard logistic regression model as follows:

$$\text{logit}(p_i) = \alpha + \beta T_i + z_i^{'}\gamma + T_i z_i^{'}\delta \tag{2.5}$$

where T_i is the treatment group ($T_i = 1$ if A, $T_i = 0$ if B), z_i is a $k \times 1$ vector of covariates, γ is a $k \times 1$ vector of coefficients and δ is a $k \times 1$ vector of treatment-covariate interaction effects. The logit function is appropriate for binary outcomes and is used for ease in interpretability, though

other link functions (e.g. probit) can also be implemented.

Prior to implementing the CARA design, a hard lead-in is used where equal allocation ($w_n = 0.5$) is used to randomize the first m patients between treatment groups until the regression parameters in equation (2.5) become estimable. After this lead-in, adaptive allocation is implemented by updating the allocation ratio using a function of the estimated treatment effect, adjusted by the covariates included in the model (2.5). Let $\mathscr{F}_i$ be the accrued patient information including treatment assignments, covariates and responses for patients $i = 1, \ldots, m, m+1, \ldots, n$, where $m+1 \leq n$. The estimated adjusted odds ratio (OR) for treatment A compared to treatment B is defined as

$$OR(\mathscr{F}_i) = \exp\{\hat{\alpha} + \hat{\beta} + z_{i+1}^{'}\hat{\gamma} + z_{i+1}^{'}\hat{\delta}\} / \exp\{\hat{\alpha} + z_{i+1}^{'}\hat{\gamma}\} = \exp\{\hat{\beta} + z_{i+1}^{'}\hat{\delta}\}. \tag{2.6}$$

Since the covariate-adjusted OR in equation (2.6) is always positive, $x \in [0, \infty)$, the function,

$$f(x) = x/(x+1) = (1 + x^{-1})^{-1}, \tag{2.7}$$

can be used to map the OR to a probability in $[0, 1]$. An odds ratio with the null value of 1, indicating equivalent odds of success between the treatment groups, is thus mapped to a probability of $1/2$ resulting in balanced allocation ($w_\ell = 0.5$). Using equation (2.7) to map the odds ratio in equation (2.6), this allocation rule assigns the next patient recruited to the study to treatment group A with probability

$$\begin{aligned} w_{i+1,A}(\mathscr{F}_i) = E(T_{i+1}|\mathscr{F}_i) &= \{1 + \exp\{-(\hat{\beta} + z_{i+1}^{'}\hat{\delta})\}\}^{-1} \\ &= \frac{\exp\{\hat{\beta} + z_{i+1}^{'}\hat{\delta}\}}{1 + \exp\{\hat{\beta} + z_{i+1}^{'}\hat{\delta}\}}, \end{aligned} \tag{2.8}$$

assuming the model converged, and with probability $w_{i+1} = 0.5$ otherwise. Alternatively, the last weight w_i that converged could be used until a sequential model successfully converges. Likewise,

the estimated probability of success in treatment group B is

$$w_{i+1,B}(\mathscr{F}_i) = E(T_{i+1}|\mathscr{F}_i) = \{1 + \exp\{-\left(-(\hat{\beta} + z'_{i+1}\hat{\delta})\right)\}\}^{-1}$$

$$= \frac{\exp\{-(\hat{\beta} + z'_{i+1}\hat{\delta})\}}{1 + \exp\{-(\hat{\beta} + z'_{i+1}\hat{\delta})\}} \tag{2.9}$$

$$= 1 - w_{i+1,A}. \tag{2.10}$$

Rosenberger et al. provide a proof to show the MLE estimates of $\hat{\beta}$ and $\hat{\delta}$ are consistent and asymptotically normal. For large treatment effects, the adaptive rule has similar power to equal allocation with a small reduction in the number of treatment failures [20].

2.2 Methods

Treatment responses and effectiveness can vary considerably across sites in a multi-center clinical trial. First, we use the traditional CARA design where we suggest using site as a fixed covariate to account for inter-center variability in treatment responses. Next, we propose the Site-Adjusted Response-Adaptive (SARA) approach as an extension of the CARA design, where site is included as a random effect instead of a fixed effect. The success probabilities, conditional on site, are used to update the assignment probabilities for allocating patients between treatment groups as subjects are accrued.

By including a random effect in a GLM with a fixed treatment effect, the resulting GLMM is classified as a *random intercept model*. The random site effect adjusts the overall intercept, which becomes a random intercept, and allows for site-specific success probabilities to be calculated for each treatment group. A random intercept model accounts for site-based response variability, where treatment responses are allowed to differ across centers. A random treatment-by-site interaction could also be included in the model to allow the fixed treatment effect to vary across centers. Both the site and treatment-by-site terms are treated as random effects. A GLMM of this form is called a *random coefficients model* and accounts for inter-center variability in treatment effectiveness [10].

Previously, Rosenberger and colleagues fit a GLM with a logit link and mapped the covariate-adjusted odds ratio to a treatment assignment probability (*treatment effect mappings* approach) [20]. Alternatively, a natural approach is to calculate the estimated success probabilities for each

treatment group from the model and use these predicted probabilities to update one of the allocation equations (e.g. Optimal, Neyman) depending on the objective desired. The allocation weights from Rosenberger's approach are slightly inflated compared to using the optimal allocation equations, which naturally map the linear predictor to success probabilities. Using the odds ratio approach results in more adaptation, where more subjects are being allocated to the superior treatment group, but at an unknown cost. By increasing the number of treatment successes, more imbalance between treatment groups could result in a trade-off in bias and power. While the effects on bias and power could be severe, the magnitude is unclear. Because of these unknown ramifications, we focused solely upon using success rates rather than odds ratios, as optimal allocation (Eq. 2.3) was specifically designed to minimize treatment failures without increasing variability (e.g. optimal allocation).

Due to the potential for inter-site variability in treatment responses, two separate models were fit, one with a fixed site effect and another with a random site effect included in the linear predictor, to get site-specific assignments.

2.2.1 CARA Model

First, a GLM is fit with both fixed site and treatment-by-site interaction effects, the traditional CARA design where center is being treated as the covariate of interest. The model is specified as:

$$logit(p_{ij}) = \beta_0 + \beta_1 T_i + \boldsymbol{Z_i'}\boldsymbol{\gamma} + T_i\boldsymbol{Z_i'}\boldsymbol{\delta} + \boldsymbol{X_i'}\boldsymbol{\theta}, \tag{2.11}$$

where β_0 is the overall intercept, β_1 is the fixed treatment effect and T_i is the treatment group indicator ($T_i = 1$ if A, $T_i = 0$ if B) for patients $i = 1, \ldots, n$. Here, $\boldsymbol{\gamma}$ represents a $(k-1) \times 1$ vector of fixed site effects and $\boldsymbol{\delta}$ is a $(k-1) \times 1$ vector of treatment-by-site interaction effects for the j^{th} site, $j = 1, \ldots, k-1$. In addition, $\boldsymbol{Z_i'}$ is a $1 \times (k-1)$ vector of site indicator variables for the j^{th} site, where k is the total number of sites. The $\boldsymbol{X_i'}\boldsymbol{\theta}$ term is an optional set of individual covariates that can be incorporated into the model to adjust the estimated success probabilities, with corresponding slopes/effect sizes denoted by $\boldsymbol{\theta}$.

The conditional probability of success was calculated for each subject to be randomized to a treatment group at a particular site by plugging the exponential of the linear predictor into the

inverse-logit (Eq. 2.7). The estimated success probabilities $\hat{p}_{ij}^{t}$ for the t^{th} treatment group, $t = 1, 2$, were defined as follows:

$$\begin{aligned}
\hat{p}_{ij}^{1} &= \text{logit}^{-1}\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\gamma}_j + \hat{\delta}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\} \\
&= \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\gamma}_j + \hat{\delta}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{\gamma}_j + \hat{\delta}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}, \\
\hat{p}_{ij}^{2} &= \text{logit}^{-1}\{\hat{\beta}_0 + \hat{\gamma}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\} \\
&= \frac{\exp\{\hat{\beta}_0 + \hat{\gamma}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\gamma}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}.
\end{aligned} \tag{2.12}$$

These predicted probabilities of success (Eq. 2.12) were plugged into one of the allocation equations (Eqs. 2.3 + 2.4) to update the allocation probabilities for the next subject by using information about their treatment center.

2.2.2 SARA Model 1

Next, the CARA design was extended by incorporating random site and treatment-by-site interaction effects into the model.

$$logit(p_{ijt}) = \beta_0 + \beta_1 T_i + g_j + d_{jt} + \boldsymbol{X_i'}\boldsymbol{\theta} \tag{2.13}$$

Here, the fixed site effect in equation (2.11) is replaced with g_j, a random site effect for the j^{th} site, where $g_j \sim N(0, \sigma_g^2)\ \forall j$, $j = 1, \ldots, k$ and k is the number of sites. The fixed treatment-by-site interaction is replaced with d_{jt}, a random interaction effect corresponding to the t^{th} treatment, where $d_{jt} \sim N(0, \sigma_d^2)\ \forall j, t = 1, 2$. In this model, the site for the next patient was used to calculate the estimated success probabilities $\hat{p}_{ij}^{t}$ as follows:

$$\begin{aligned}
\hat{p}_{ij}^{1} &= \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \hat{d}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \hat{d}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}, \\
\hat{p}_{ij}^{2} &= \frac{\exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}.
\end{aligned} \tag{2.14}$$

The allocation ratio is dependent on the previous subjects accrued, however, the current subject is then randomized using these updated allocation probabilities so independence holds. If there is

no true inter-site variability in the data, SARA Model 1 (Eq. 2.13) with no additional (optional) covariates should decompose to the RA design (Eq. 2.1) as $\hat{g}_j \simeq 0$ and $\hat{d}_{jt} \simeq 0,\ \forall j, t$, where adaptation is guided only by the treatment assignments and outcomes.

A treatment-by-center interaction was not included in the simulations due to (i) our initial focus on heterogeneity in patient responses and not treatment effectiveness, and also (ii) issues in estimating many parameters introduced by the additional variance components to estimate using a Frequentist approach. The latter can occur early on in the study when sample sizes within the centers are small or when all of the outcomes at a center are all successes or failures. ML estimates of the parameters may be severely biased for highly sparse data [14]. We focus on simulations with both random site and treatment-by-site interactions in Chapter 3.

2.3 Simulation Study

First, we need to show that existing randomization methods will be adversely affected when site is not accounted for, particularly when inter-center variability exists. A small simulation study will be performed to examine the performance of balanced allocation and response-adaptive randomization in terms of the empirical power. We will consider trials with a small number of centers ($k = 3$), a small treatment group 1 success probability ($p_1 = 0.3$), a small effect size ($\delta = 0.1$) and several inter-site standard deviations ($\sigma = 0, 0.025, 0.05, 0.075$). When there is no inter-center variability ($\sigma = 0$), meaning that there are no differences in treatment responses and/or effects across centers, the site-specific treatment success probabilities are equal to the true success probabilities ($p_1 = 0.3,\ p_2 = 0.2$). However, as we increase the amount of inter-center variability, the distributions of the success probabilities and the corresponding effect size ($p_1 - p_2$) all spread out considerably around the true success probabilities (Figure 2.1). To ensure the success probabilities used in the simulations are contained in (0, 1), a function was used to bound the values between (0.01, 0.99).

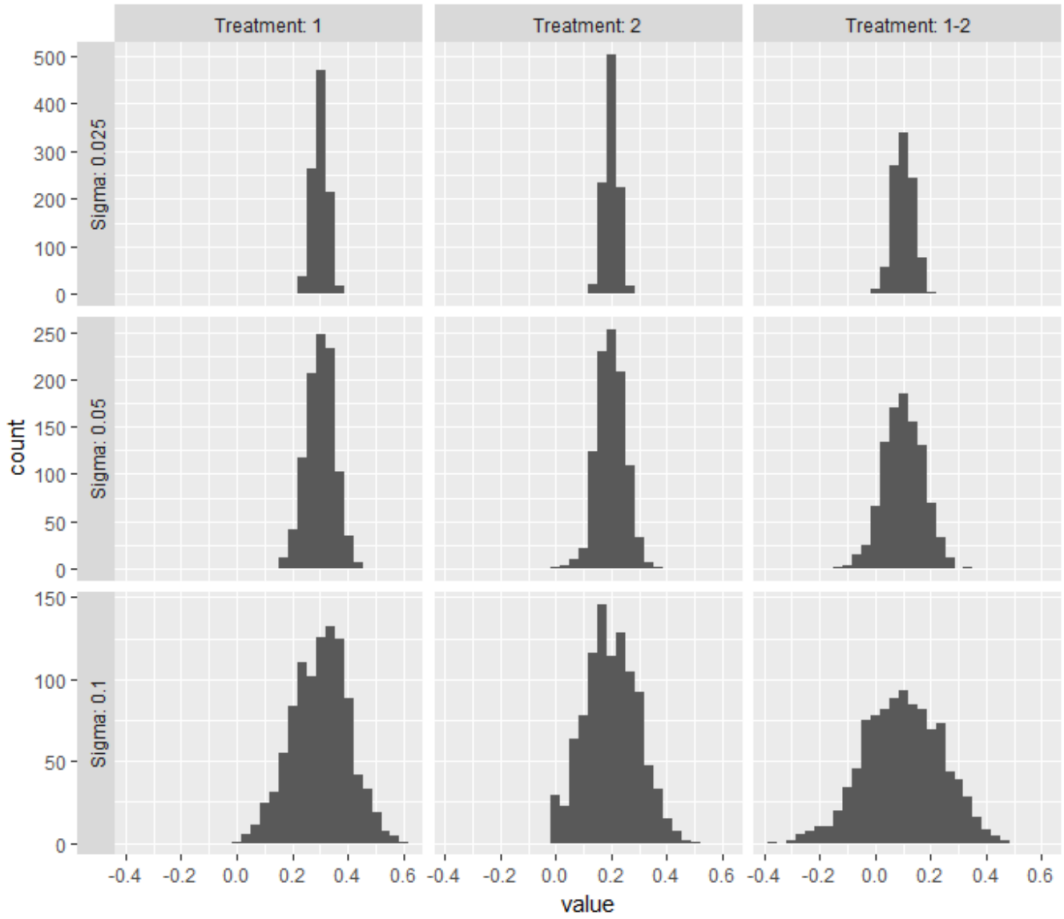


Figure 2.1: Simulated Success Probabilities for Center 1
Scenario: $N = 1000,\ p_1 = 0.3,\ p_2 = 0.2,\ \sigma = 0.025, 0.05, 0.1.$

Next, we will examine the performance of our proposed CARA and SARA models to existing allocation designs. The simulations were modeled as Phase 2 multi-center trials with a relatively small number of centers ($k = 3, 6, 9$). A phase 2 trial can recruit several hundred patients to examine efficacy and monitor side effects. The traditional sample size formula [6] for balanced two independent samples is as follows:

$$2N = \frac{2 \times \{Z_\alpha \sqrt{2 \times \bar{p}(1-\bar{p})} + Z_\beta \sqrt{p_2(1-p_2) + p_1(1-p_1)}\}^2}{(p_2 - p_1)^2}, \tag{2.15}$$

where $2N$ is the total sample size, $\bar{p} = (p_1 + p_2)/2$, Z_α is the critical value for the significance level α, Z_β is the critical value for the power $(1-\beta)$ and p_ℓ is defined as the proportion of successes in treatment group ℓ out of the total number of patients in that group, $\ell = 1, 2$. The second square root term represents the inherent variability. The total sample size $(2N)$ was rounded up to the

nearest whole number to assure enough patients were recruited to achieve the desired power.

Alternatively, the sample size for a dichotomous outcome can be determined for each trial by adding variability to account for clustering. Previous research has been conducted to power multi-center trials where the center sample size or number of centers is fixed in advance [2]. This approach ensures that balanced randomization achieves a nominal power level (e.g. 80%). However, we decided not to account for the inter-site variability so that we could see the power loss when site is not accounted for in the randomization, including in the balanced case.

Both the effect sizes and values of inter-site variability used in the simulation scenarios were determined based on the results of a brief simulation study to ensure the presence of a slight quantitative treatment-by-site interaction as defined by Gallo (2008). The number of sites ($k = 3, 6, 9$) and inter-site standard deviations in treatment response were varied with a 30% treatment group 1 success probability ($p_1 = 0.3$) and several effect sizes ($\delta = p_1 - p_2 = 0,\ 0.1,\ 0.15$). The percentages of flipped success probabilities, defined as P(*at least one* $\hat{p}_1 < \hat{p}_2$)×100, were averaged across 1000 iterations to indicate how frequent qualitative interactions occurred. The results of the simulation are displayed in Table 2.1. As expected, the results show that as the effect size increases, there are less qualitative treatment-by-site interactions. However, as the inter-site standard deviation increases, the number of qualitative interactions increases regardless of effect size. Accordingly, we select smaller values of inter-site SD to limit the amount of qualitative interactions to study the operating characteristics of the proposed methods.

Table 2.1: Percentage (%) of Qualitative Treatment-by-Site Interactions

			Inter-Site Standard Deviation (σ)					
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	**0**	**0.025**	**0.0375**	**0.05**	**0.075**	**0.1**
3	0.3	0	0	89.5	89.5	89.5	89.5	89.5
		0.1	0	0.6	8.1	21.1	43.9	56.0
		0.15	0	0	0.6	4.4	21.1	36.9
6		0	0	98.5	98.5	98.5	98.5	98.5
		0.1	0	0.9	16.7	39.3	68.2	80.1
		0.15	0	0	0.9	9.5	39.3	58.6
9		0	0	99.9	99.9	99.9	99.9	99.9
		0.1	0	1.9	24.0	50.9	80.2	89.5
		0.15	0	0	1.9	14.4	50.9	74.1

We considered small ($p_1 = 0.3$), medium ($p_1 = 0.5$) and large ($p_1 = 0.7$) treatment group

1 success probabilities. Null ($\delta = 0$), small ($\delta = 0.1$) and medium ($\delta = 0.15$) effect sizes were considered, where $\delta = p_1 - p_2$, to allow us to determine $p_2 = p_1 - \delta$. Null effect sizes were powered with $\delta = 0.1$ to allow a reasonable sample size to be recruited. The variability in success probabilities across centers was controlled by adjusting the amount of inter-site standard deviation ($\sigma = 0,\ 0.025,\ 0.05,\ 0.075$), which controls the magnitude and direction of treatment-by-site interactions.

The site-specific success probabilities $\hat{p}_{\ell j}$ were simulated from a normal distribution with a mean equal to the treatment group ℓ success probability p_ℓ and variance representing the inter-site variability σ^2, where $\ell = 1, 2$ and $j = 1, \ldots, k$ sites. A function was implemented to ensure success probabilities were bound between 0 and 1. A random variate v was generated from a uniform$(0, 1)$ distribution, denoted by $U(0, 1)$, for each subject and assigned to site k if $\frac{z-1}{k} < v \leq \frac{z}{k}$ $\forall z = 1, \ldots, k$. All patient sites were calculated prior to adaptation to ensure that the estimated success probabilities for the next patient could be calculated using their recruitment site. Next, as each patient was recruited in the simulation, another $U(0, 1)$ variate was drawn and compared to the previous treatment group 1 allocation weight to determine the assignment and outcome of the current patient being recruited. If the value was less than or equal to the previous weight (Eqns.2.3 and 2.4), the patient was assigned to treatment group 1, otherwise they were recruited to the second treatment group. The patient outcomes were then assigned from a Bernoulli distribution with success probability $\hat{p}_{\ell j}$ which depends on the j^{th} site they are being recruited from.

An artificial lead-in was used in the simulations to prevent estimator instability or convergence issues. This hard lead-in required that balanced allocation be performed for the first $m = 20$ subjects irrespective of the design. Additionally, for the CARA and SARA models, both outcomes must be observed in each group prior to adapting to ensure the GLMM can be fit using the *glmer* {lme4} function [30] in R, otherwise the function produces a constant response error. This is not a requirement for the RA/CARA models using the *glm* {stats} function [31], however, separation issues may arise so this lead-in criterion was implemented. For both CARA/SARA, an additional criterion required that all centers be observed to ensure that allocation weights could be calculated using the next patient's site. Alternatively, we could have used balanced allocation for patients recruited from unobserved sites to potentially adapt sooner. Due to the artificial lead-in, results may slightly under represent the adaptation effect of the allocation design, particularly with the

CARA/SARA designs which have additional criteria to be satisfied prior to adapting.

The GLM was fit in the *glm* function which obtains the ML estimates through iteratively reweighted least squares (IRLS) method. The GLMM model was fit in the *glmer* function with ML estimation using the adaptive Gauss-Hermite quadrature approximation (quadrature nodes = 10) of the likelihood for the mixed model, which is an integral over the random effect space. The variance-covariance matrices of the random effects were estimated using Powell's BOBYQA method, a nonlinaer optimizer [30]. A variance component was modeled for the random site effect.

If convergence warnings arose when fitting the SARA models iteratively for each trial, the previous allocation weights were used until the model successfully converged again [20]. This allows previous information about the trial to be used to update the allocation weights. Instead, an alternative approach we could have implemented is to use balanced allocation ($w_1 = 0.5$) until the model converges again, however, this ignores any previous patient data. The *try* {base} function [31] may be used to suppress any errors from terminating loops, such as non positive-definite variance/covariance matrices or for those that contain null values, when using the linux cluster. All simulations were coded using R [31] statistical software.

In our experience, fitting these iterative models for each parameter template (trial scenario) required significant computation time. Accordingly, R scripts were submitted on the departmental Beowulf Cluster using the most recent update of R available. To facilitate faster computation times, the trial scenarios were split up to be run on the cluster simultaneously and the results compiled.

To evaluate the proposed allocation designs, we use the conceptual framework established by Rosenberger et al. (2008). When comparing methods, the authors interpreted the operating characteristics in terms of balance (allocation proportion), efficiency (power, type I error) and ethics (total number of treatment successes) [19]. We are particularly interested in the mean total number of successes/failures ($\pm$ standard error), empirical power ($P\{\text{reject } H_0 | H_1 \text{ true}\}$) and type I error rate ($P\{\text{reject } H_0 | H_0 \text{ true}\}$). The type I error rate and empirical power were plotted to summarize patterns in the operating characteristics across sites among the methods. These figures were calculated using the naive BC/RA estimates, which do not account for site, while the CARA/SARA approaches use estimates from the GLMMs, which incorporate site effects, to assess whether they perform as intended. We compared the results of the balanced case, RA (Eq. 2.1), CARA (Eq. 2.11) and SARA 1 (Eq. 2.13) designs for all scenarios considered. The balanced and RA designs do

not account for site variability while both the CARA and SARA designs account for the clustering differently, through either a fixed or random site effect, respectively. We hypothesize that the CARA and SARA methods, which account for inter-center variability, will reduce the variability of treatment responses in multi-center trials, thus reducing biased or improper patient allocation.

There were 1000 trials simulated for each scenario and the results aggregated. Both Optimal and Neyman allocation were considered. For a single trial scenario, the following metrics were stored: sample size, number of patients assigned to each treatment, number of successes for each treatment, total number of successes and the number of convergence issues (indicates previous allocation weights were used until model converged). The allocation proportion calculated at the end of each trial was averaged across the simulated trials $\left(E\left[\frac{n_1}{n_1+n_2}\right]\right)$ and examined to compare both the amount and variability of adaptation, where n_i is the number of patients allocated to treatment group $\ell, \ell = 1, 2$.

After all patients were recruited, a two-sided chi-squared test for the treatment effect was performed. For the BC/RA designs, a naive Wald's chi-square test from the GLM was used to test the treatment effect ($H_0 : \hat{\beta}_1 = 0$). This approach ignores site when testing the treatment effect as this is convention when analyzing multi-center trials. For the CARA/SARA designs, in addition to a naive chi-squared test, a chi-squared test from a GLM or GLMM, which incorporates fixed or random site effects, respectively, was used. Specifically, for the naive approach used in the CARA/SARA methods, the proportions of successes were compared using a Pearson's chi-squared ($H_0 : p_1 = p_2$) test to assess the treatment effect. There was no need to worry about violating assumptions due to low expected cell counts for the CARA/SARA approaches. However, RA randomization may result in severely skewed treatment groups and not satisfy the assumptions for a chi-squared test. An indication of whether or not the null hypothesis (H_0 : no treatment effect) was rejected was retained in order to calculate the operating characteristics of the method. These metrics were averaged over all iterations.

2.3.1 Simulation Algorithm

The basic algorithm for our simulation study is as follows.

1. Hard lead-in for the first m patients, where individuals are allocated equally between the two treatment groups.

2. Let $\mathscr{F}_i$ be accrued information which includes treatment assignments, observed patients outcomes and center information. Use $\mathscr{F}_i$ for patients $i = 1, \ldots, m, m+1, \ldots, n$ to begin adaptation, starting with patient $m+1$.

3. Using the center for patient $m+1$, calculate the conditional probability of success $\hat{p}_{ij}^t$ for the t^{th} treatment using the CARA (Eqns. 2.11, 2.12) or SARA model (Eqns. 2.13, 2.14).

4. Using the estimated success probabilities $\hat{p}_i^1, \hat{p}_i^2$, update the allocation weights W_1, W_2 and assign the next patient to a treatment.

$$W_1(\mathscr{F}_i) = \sqrt{\hat{p}_i^1}/\sqrt{\hat{p}_i^1 + \hat{p}_i^2},\ W_2(\mathscr{F}_i) = 1 - W_1.$$

5. Repeat Steps 2-4 to randomize patients $m+2, \ldots, n$.

For a single scenario, this algorithm is repeated for $N = 1000$ trials.

2.4 Results

The results of our simulation study of existing methods shows that the allocation probabilities and treatment effectiveness for the balanced case and RA design are adversely affected when multiple sites are not accounted for during randomization, particularly when inter-site variability exists. This is demonstrated by the simulations presented in Figure 2.2. When there is no variability across centers, the two approaches perform similarly and achieve close to nominal levels of power. Balanced allocation results in slightly more power due to equal sample sizes in the two treatment groups. However, as the inter-site variability increases, the empirical power decreases when we do not adjusts for this clustering. Thus, adaptive methods should be flexible enough to address potential variability across centers.

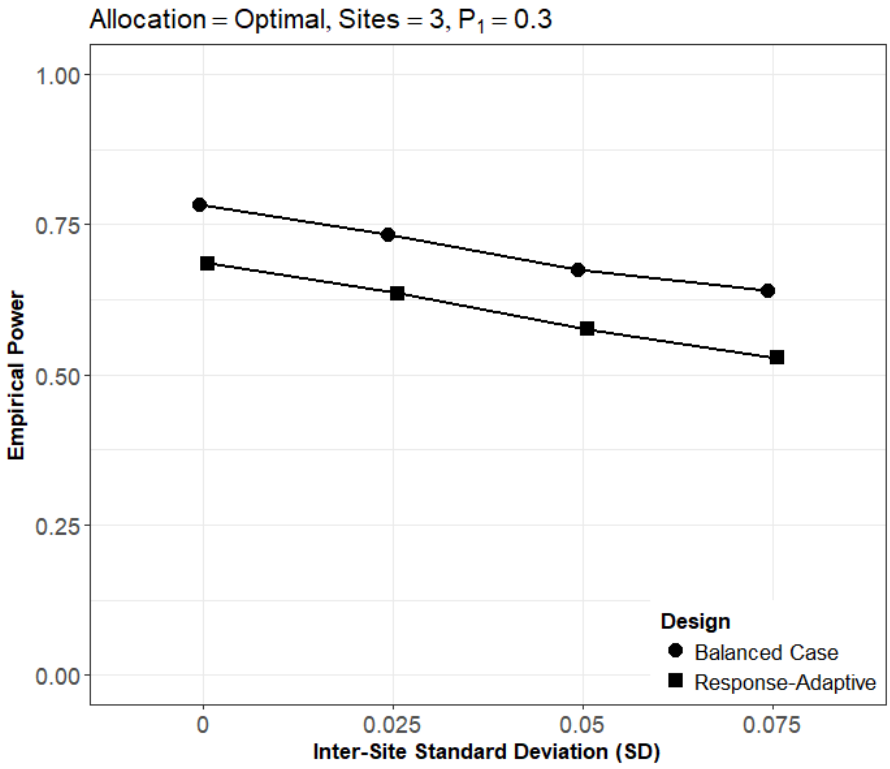


Figure 2.2: Not Accounting for Clustering

Next, we explore the the mean total number of treatment successes (SE), the expected allocation proportions to treatment group 1 (SE) and the empirical type I error and power for SARA 1 (Eqn. 2.13) compared to three existing allocation designs.

2.4.1 Scenario 1: Optimal Allocation, 3 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 1 are displayed in Table 2.2. When the effect size is null ($\delta = p_1 - p_2 = 0$), the four methods are comparable. There is a slight increase in the total number of successes for the adaptive designs as the amount of inter-site standard deviation (σ) increases, since the success probabilities are simulated from a right skewed binomial distribution (more mass above 0.3 since left truncated at 0) and are more likely to vary greater in one direction (e.g. values greater than P_1, P_2).

When the effect size is increased to 10% ($p_1 = 0.3, p_2 = 0.2$), the adaptive designs result in more treatment successes than the balanced case, however, the improvement is lower for SARA compared to the RA/CARA designs. SARA has lower variability than the other adaptive approaches and it

is similar to the variability from the balanced case. For an effect size of 15%, the pattern is similar to the medium effect size. The adaptive designs result in more treatment successes than the BC and there are fewer treatment successes for SARA compared to the other adaptive approaches. SARA maintains variability similar to the balanced case, both of which have less variability than the RA/CARA designs. In addition, as σ increases, the variability in the expected total number of treatment successes increases for all methods.

Table 2.2: Scenario 1 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)
		(n=587)	0.025	175.7 (12.6)	175.8 (12.7)	175.8 (12.6)	175.8 (12.6)
			0.05	175.5 (16.3)	176.1 (16.6)	176.0 (16.4)	176.0 (16.3)
			0.075	175.3 (21.0)	176.7 (21.5)	176.5 (21.4)	176.3 (21.1)
		0.1	0	146.6 (10.6)	151.1 (14.3)	150.7 (12.5)	149.5 (10.7)
		(n=587)	0.025	146.2 (12.3)	150.5 (15.7)	150.3 (13.6)	149.2 (12.3)
			0.05	146.0 (16.1)	151.0 (19.6)	150.7 (17.3)	149.5 (16.3)
			0.075	145.9 (20.8)	152.5 (24.0)	151.8 (22.0)	150.0 (20.8)
		0.15	0	54.2 (6.7)	59.1 (9.1)	58.4 (7.8)	57.2 (6.8)
		(n=241)	0.025	54.3 (6.8)	59.1 (9.6)	58.5 (8.5)	57.4 (7.1)
			0.05	54.4 (7.9)	59.6 (10.4)	58.9 (9.4)	57.7 (8.1)
			0.075	54.7 (9.5)	60.1 (11.9)	59.5 (10.7)	58.2 (9.6)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

The allocation proportions are displayed in Table 2.3. For a null effect size, the four methods have comparable allocation proportions. The SARA design has lower variability in allocating than the other adaptive designs, while the RA design has the highest variability. When $\delta = 0.10$, the adaptive designs have a larger allocation ratio relative to the balanced case, though the improvement is lower for SARA compared to the other adaptive approaches. This explains the increase in expected treatment successes in the adaptive methods with a slightly lower increase in the SARA approach. SARA is less variable than the RA/CARA designs in all cases. For $\delta = 0.15$, the

adaptive designs have larger allocation ratios than the BC, particularly RA randomization, but the improvement is lower for CARA and SARA. SARA has lower variability compared to the other adaptive designs. Overall, the amount of inter-site standard deviation does not affect the allocation proportions or the variability in allocation proportions.

Table 2.3: Scenario 1 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**
3	0.3	0	0	0.50 (0.02)	0.50 (0.13)	0.50 (0.09)	0.50 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.49 (0.12)	0.50 (0.08)	0.50 (0.03)
			0.05	0.50 (0.02)	0.49 (0.13)	0.50 (0.09)	0.50 (0.04)
			0.075	0.50 (0.02)	0.49 (0.13)	0.50 (0.08)	0.50 (0.04)
		0.1	0	0.50 (0.02)	0.58 (0.17)	0.57 (0.12)	0.55 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.57 (0.17)	0.57 (0.11)	0.55 (0.04)
			0.05	0.50 (0.02)	0.57 (0.18)	0.57 (0.12)	0.55 (0.05)
			0.075	0.50 (0.02)	0.58 (0.18)	0.57 (0.12)	0.55 (0.05)
		0.15	0	0.50 (0.03)	0.63 (0.19)	0.62 (0.13)	0.58 (0.05)
		($n = 241$)	0.025	0.50 (0.03)	0.63 (0.18)	0.61 (0.13)	0.58 (0.05)
			0.05	0.50 (0.03)	0.63 (0.19)	0.61 (0.14)	0.58 (0.06)
			0.075	0.50 (0.03)	0.63 (0.20)	0.61 (0.13)	0.58 (0.07)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

The empirical type I error and power results are displayed in Table 2.4 and Figure 2.3. For $\delta = 0$, all four methods perform similarly where there is an inflation in the type I error rate as σ increases. RA randomization appears to control the type I error rate slightly better than the other designs particularly when there is more inter-site variability. The pattern is similar for small and medium effect sizes, where the power decreases with increasing σ. Specifically, for $\delta = 0.10$, the CARA design is similar to balanced randomization and gains back some of the power lost with RA randomization. The SARA design outperforms the balanced case for smaller σ and is comparable when $\sigma = 0.075$. For $\delta = 0.15$, the SARA design is similar to balanced randomization and thus

gains back some of the power lost with the other adaptive methods. There is a notable loss of power for RA randomization relative to the other methods.

We now discuss the results from the GLM/GLMM approach, where either a fixed (CARA) or random (SARA) site effect is incorporated into the modeling to account for clustering. For a null effect size, the type I error results are comparable to the naive approach. When $\delta = 0.10$, the SARA design performs similarly, where there is a slight improvement over balanced randomization, but the CARA has more of a loss of power relative to the BC. For $\delta = 0.15$, the CARA design has an even more notable loss of power relative to the BC/SARA approaches using the GLM/GLMM results.

Table 2.4: Scenario 1 - Empirical Type I Error and Power

				Naive				**GLM/GLMM**	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
3	0.3	0	0	0.06	0.06	0.07	0.06	0.06	0.06
		($n = 587$)	0.025	0.09	0.08	0.09	0.09	0.08	0.09
			0.05	0.20	0.19	0.21	0.20	0.20	0.20
			0.075	0.31	0.29	0.32	0.32	0.30	0.31
		0.1	0	0.78	0.68	0.76	0.79	0.73	0.79
		($n = 587$)	0.025	0.73	0.64	0.72	0.75	0.69	0.74
			0.05	0.68	0.57	0.67	0.68	0.63	0.68
			0.075	0.64	0.53	0.62	0.63	0.58	0.63
		0.15	0	0.80	0.61	0.78	0.81	0.70	0.80
		($n = 241$)	0.025	0.76	0.58	0.73	0.76	0.67	0.75
			0.05	0.73	0.54	0.69	0.72	0.63	0.72
			0.075	0.67	0.50	0.66	0.69	0.59	0.68

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

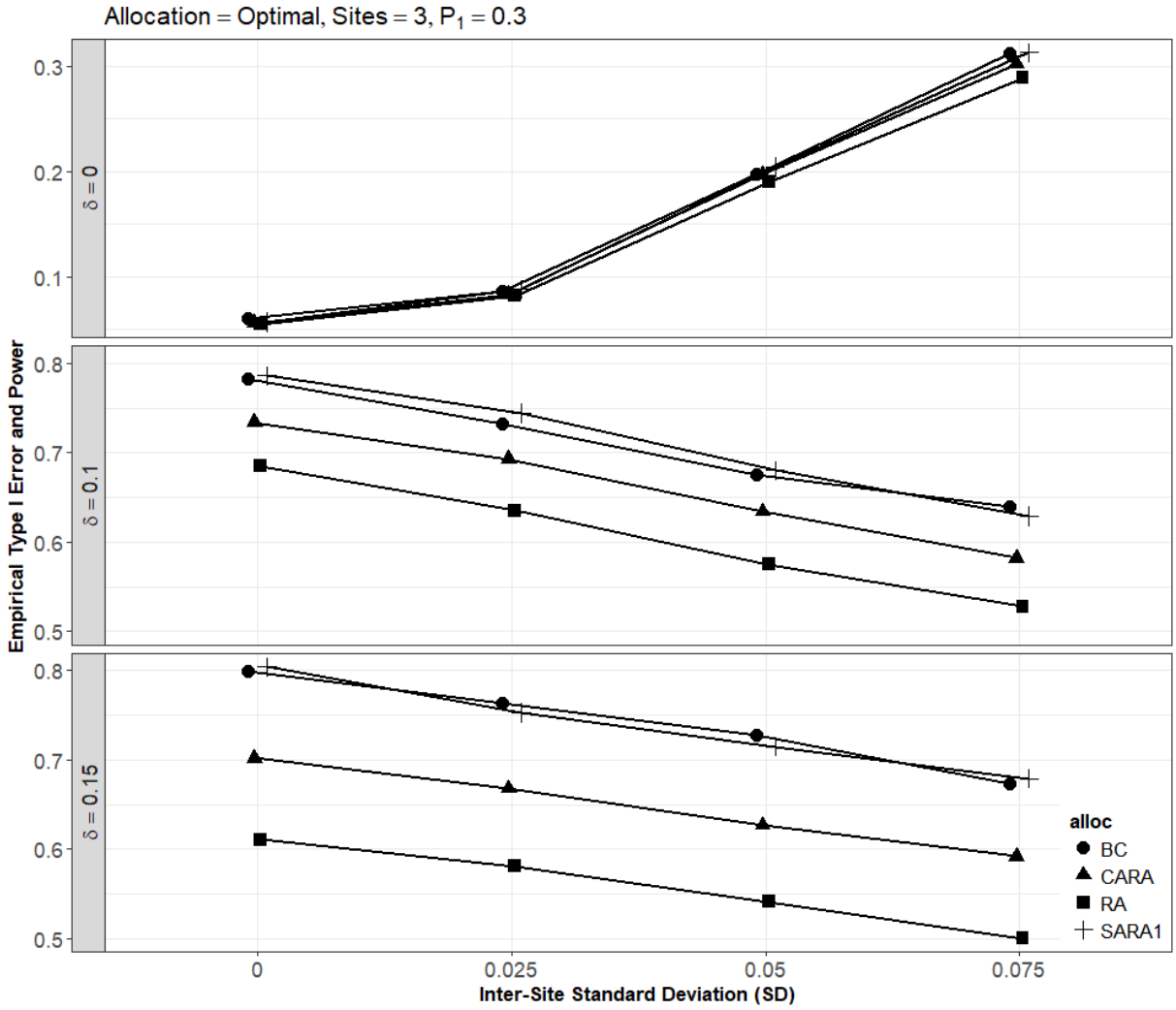


Figure 2.3: Scenario 1 - Empirical Type I Error and Power

2.4.2 Scenario 2: Optimal Allocation, 6 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 2 are displayed in Table 2.5. When the effect size is null, the four methods are comparable with respect to the expected number of treatment successes and variability. There is a slight increase in the expected successes as σ increases for the adaptive approaches. When the effect size is increased to 10%, the adaptive designs all produce greater expected treatment successes than the balanced case, however, the RA design results in slightly more treatment successes than CARA/SARA. The CARA/SARA designs have lower variability in treatment successes than RA randomization, with SARA having more improvement than CARA with variabilities closer to the BC. For an effect size of 15%, while all of the adaptive designs result in more treatment successes than the balanced case, the RA design performs better than both the CARA/SARA approaches with respect to treatment successes. The variability in treatment successes for the RA method is higher than that of the variability for the other methods which perform similarly while the variability for SARA is closest to the BC. Overall, as the inter-site standard deviation increases, the variability in the expected total number of treatment successes increases for all of the designs.

Table 2.5: Scenario 2 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
6	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)
		(n=587)	0.025	176.3 (11.8)	176.3 (11.8)	176.3 (11.8)	176.3 (11.8)
			0.05	176.3 (14.1)	176.6 (14.1)	176.5 (14.2)	176.5 (14.1)
			0.075	176.1 (17.2)	176.7 (17.3)	176.6 (17.2)	176.6 (17.2)
		0.1	0	146.6 (10.6)	151.1 (14.3)	149.8 (10.9)	149.5 (10.7)
		(n=587)	0.025	146.9 (11.3)	151.2 (14.8)	150.0 (11.8)	149.8 (11.4)
			0.05	146.9 (13.6)	151.6 (16.9)	150.4 (14.4)	150.0 (13.8)
			0.075	146.9 (16.6)	151.6 (19.8)	150.6 (17.4)	150.2 (16.9)
		0.15	0	54.2 (6.7)	59.1 (9.1)	57.5 (7.0)	57.2 (6.9)
		(n=241)	0.025	54.2 (6.6)	59.1 (9.4)	57.4 (7.2)	57.1 (7.0)
			0.05	54.1 (7.2)	59.1 (9.9)	57.4 (7.8)	57.1 (7.7)
			0.075	54.4 (8.2)	59.5 (10.9)	57.8 (8.9)	57.4 (8.7)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

The allocation proportions are displayed in Table 2.6. For $\delta = 0$, the four approaches have comparable allocation proportions. The CARA/SARA designs have lower variability in allocation proportions than the RA design, with SARA closer to balanced randomization. When $\delta = 0.10$, the adaptive designs have larger allocation ratios relative to the BC, with the improvement lower for the CARA/SARA designs than that of RA randomization. While the CARA/SARA designs are less variable than the RA design, SARA maintains the lowest variability, which is closer to the BC. For an effect size of 15%, while all of the adaptive designs have larger allocation ratios than the balanced case, the RA design performs better than both the CARA/SARA designs. The CARA/SARA approaches here are less than half the variability of the RA design, however, the SARA approach has more of an improvement with variability closer to balanced randomization. The amount of inter-site standard deviation does not affect the allocation proportions, though there is a slight increase in the variability of the allocation proportions.

Table 2.6: Scenario 2 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
6	0.3	0	0	0.50 (0.02)	0.50 (0.13)	0.50 (0.06)	0.50 (0.04)
		($n = 587$)	0.025	0.50 (0.02)	0.49 (0.12)	0.50 (0.04)	0.50 (0.03)
			0.05	0.50 (0.02)	0.49 (0.12)	0.50 (0.04)	0.50 (0.04)
			0.075	0.50 (0.02)	0.49 (0.13)	0.50 (0.05)	0.50 (0.04)
		0.1	0	0.50 (0.02)	0.58 (0.17)	0.55 (0.06)	0.55 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.57 (0.16)	0.55 (0.05)	0.55 (0.04)
			0.05	0.50 (0.02)	0.57 (0.17)	0.55 (0.06)	0.55 (0.04)
			0.075	0.50 (0.02)	0.57 (0.17)	0.55 (0.06)	0.55 (0.05)
		0.15	0	0.50 (0.03)	0.63 (0.19)	0.59 (0.07)	0.58 (0.06)
		($n = 241$)	0.025	0.50 (0.03)	0.63 (0.18)	0.59 (0.07)	0.58 (0.06)
			0.05	0.50 (0.03)	0.63 (0.18)	0.59 (0.07)	0.58 (0.06)
			0.075	0.50 (0.03)	0.63 (0.19)	0.59 (0.08)	0.58 (0.06)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

The empirical type I error and power results are displayed in Table 2.7 and Figure 2.4. For $\delta = 0$, the type I error rate increases as the amount of inter-site variability increases. For smaller σ, the four methods perform similarly. However, as σ increases, RA randomization controls the type I error rate slightly better than all of the other approaches which perform similarly. For $\delta = 0.10$, the RA design has a notable loss of power compared to the other methods. Before $\sigma = 0.05$, the CARA/SARA designs have marginally higher power than the balanced case. For larger σ, the three methods are comparable. As the amount of inter-site variability increases, the power decreases for all methods though it is maintained better by BC/SARA/CARA. For $\delta = 0.15$, the CARA/SARA methods gain back most of the power lost with the RA design, but the improvement is slightly lower for CARA for $\sigma = 0.075$. Overall, as σ increases, there is a slight loss of power for all four approaches.

For a null effect size under the GLM/GLMM approach, the CARA design has the highest inflation in type I error while RA randomization has the lowest, particularly as σ increases. The

SARA design performs similarly to the balanced case. For $\delta = 0.10$, the CARA/SARA designs have marginally higher power than the balanced case, with a slight improvement for the SARA design, before $\sigma = 0.05$. For larger σ, the methods are comparable to the naive approach, where the CARA/SARA designs have marginally higher power than the BC before $\sigma = 0.05$ but all perform similarly for larger σ. RA randomization has a notable loss of power compared to the other methods. For $\delta = 0.15$, the CARA/SARA methods gain back most of the power lost with the RA design, though they perform similarly to balanced allocation, but the improvement is slightly lower for CARA.

Table 2.7: Scenario 2 - Empirical Type I Error and Power

				Naive				**GLM/GLMM**	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
6	0.3	0	0	0.06	0.06	0.06	0.06	0.06	0.06
		($n = 587$)	0.025	0.08	0.08	0.08	0.08	0.08	0.08
			0.05	0.12	0.11	0.12	0.12	0.13	0.12
			0.075	0.18	0.17	0.18	0.18	0.19	0.18
		0.1	0	0.78	0.68	0.78	0.79	0.78	0.78
		($n = 587$)	0.025	0.76	0.68	0.78	0.78	0.77	0.78
			0.05	0.73	0.64	0.73	0.72	0.72	0.72
			0.075	0.69	0.61	0.69	0.69	0.69	0.68
		0.15	0	0.80	0.61	0.81	0.80	0.79	0.80
		($n = 241$)	0.025	0.79	0.61	0.78	0.79	0.78	0.78
			0.05	0.76	0.58	0.76	0.76	0.74	0.75
			0.075	0.71	0.54	0.71	0.72	0.70	0.71

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

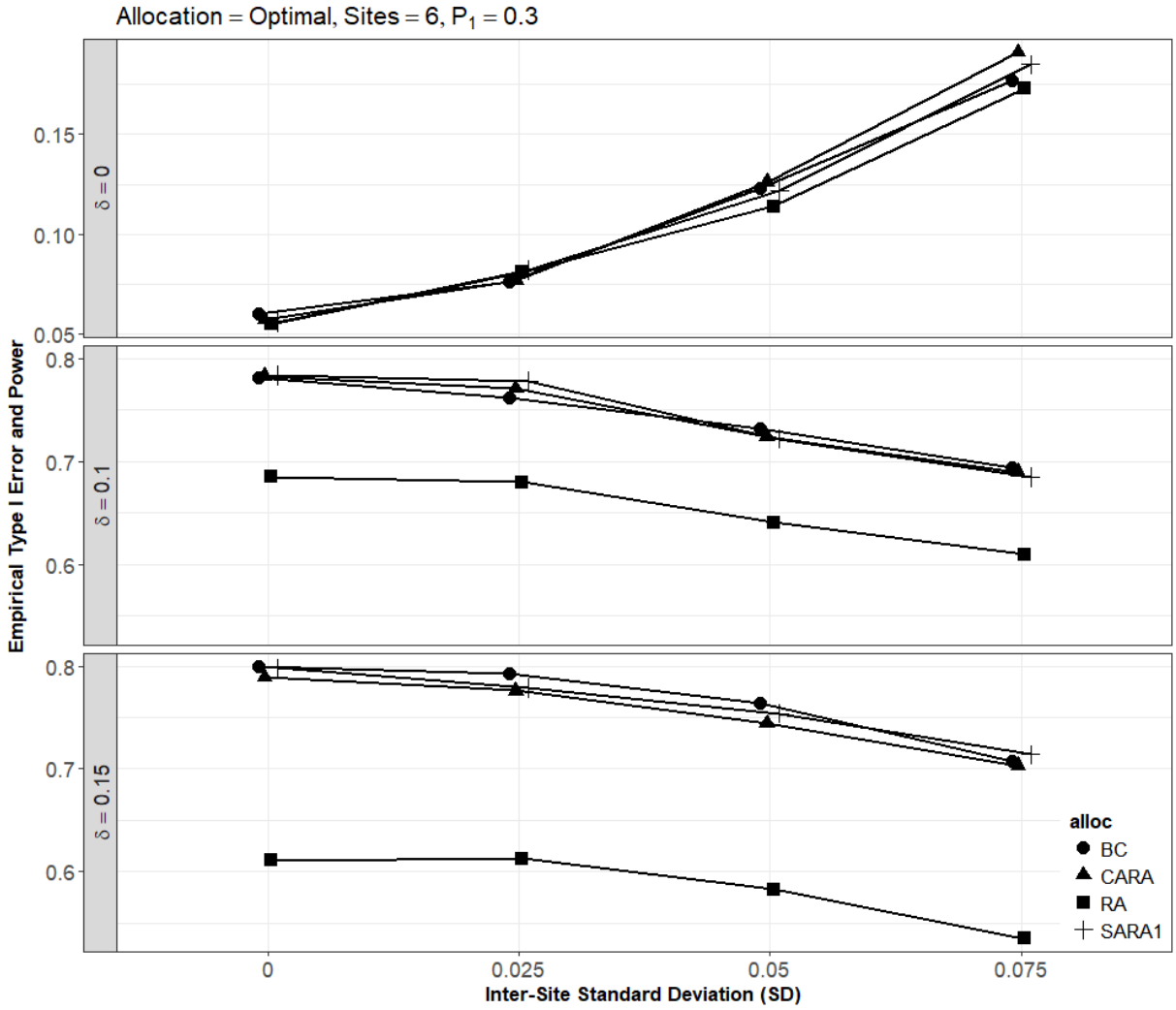


Figure 2.4: Scenario 2 - Empirical Type I Error and Power

2.4.3 Scenario 3: Optimal Allocation, 3 Sites, $p_1 = 0.5$

The mean total number of treatment successes for Scenario 3 are displayed in Table 2.8. When the effect size is null, the four methods perform similarly with respect to both the expected total number of treatment successes and the variability in successes. As the inter-site standard deviation increases, there is a slight increase in treatment successes accompanied by a greater increase in treatment success variability. When the effect size is increased to 10%, the adaptive methods increase the expected treatment successes relative to the BC and they maintain the same variability. For $\delta = 0.15$, the pattern is similar to a small effect size, where the adaptive approaches result in more treatment successes. SARA has a slightly smaller variability in expected successes similar to the balanced case compared to the other adaptive designs. As σ increases, there is a slight increase in expected treatment successes along with a greater increase in the variability in treatment successes.

Table 2.8: Scenario 3 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.5	0	0	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)
		(n=775)	0.025	388.5 (15.9)	388.6 (15.9)	388.5 (15.9)	388.6 (15.9)
			0.05	389.1 (20.9)	389.4 (20.9)	389.4 (20.9)	389.4 (20.9)
			0.075	389.3 (27.1)	390.1 (27.0)	390.1 (27.0)	390.1 (27.0)
		0.1	0	348.9 (14.1)	351.2 (14.1)	351.2 (14.1)	351.0 (14.0)
		(n=775)	0.025	349.0 (15.8)	351.5 (16.0)	351.5 (16.0)	351.4 (15.9)
			0.05	349.3 (21.1)	352.1 (21.2)	352.2 (21.2)	352.0 (21.1)
			0.075	350.0 (27.5)	353.4 (27.8)	353.4 (27.6)	353.2 (27.5)
		0.15	0	143.9 (9.1)	146.3 (9.4)	146.3 (9.2)	146.1 (9.1)
		(n=339)	0.025	143.7 (9.8)	146.1 (10.2)	146.1 (10.0)	145.8 (9.8)
			0.05	143.8 (11.6)	146.4 (11.9)	146.4 (11.7)	146.1 (11.5)
			0.075	143.7 (13.9)	146.6 (14.3)	146.5 (14.1)	146.3 (13.9)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

The allocation proportions are displayed in Table 2.9. For $\delta = 0$, the four approaches have

comparable allocation proportions. The SARA design has lower variability similar to the BC compared to the RA/CARA designs. As the amount of inter-site variability increases, the allocation proportions remain constant but there is a marginal increase in variability in allocation. When $\delta = 0.10$, the adaptive designs have a larger allocation ratio relative to the BC. SARA is less variable than the RA/CARA designs with variabilities comparable to the balanced case. For an effect size of 15%, the adaptive methods have higher allocation ratios, however, the improvement is lower for SARA compared to the RA/CARA designs. SARA has lower variability compared to the other adaptive designs but has marginally higher variability than the BC. The amount of inter-site standard deviation does not affect the allocation proportions and only has the slightest of effects in the variability in allocation for medium effect sizes.

Table 2.9: Scenario 3 - E$[n_1/(n_1 + n_2)]$ (Standard Error)

k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**BC**	**RA**	**CARA**	**SARA 1**
3	0.5	0	0	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.50 (0.02)
		$(n = 775)$	0.025	0.50 (0.02)	0.50 (0.04)	0.50 (0.04)	0.50 (0.02)
			0.05	0.50 (0.02)	0.50 (0.04)	0.50 (0.04)	0.50 (0.02)
			0.075	0.50 (0.02)	0.50 (0.04)	0.50 (0.04)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.53 (0.04)	0.53 (0.04)	0.53 (0.02)
		$(n = 775)$	0.025	0.50 (0.02)	0.53 (0.04)	0.53 (0.04)	0.53 (0.02)
			0.05	0.50 (0.02)	0.53 (0.05)	0.53 (0.04)	0.53 (0.03)
			0.075	0.50 (0.02)	0.53 (0.05)	0.53 (0.04)	0.53 (0.03)
		0.15	0	0.50 (0.03)	0.55 (0.07)	0.55 (0.06)	0.54 (0.04)
		$(n = 339)$	0.025	0.50 (0.03)	0.55 (0.08)	0.55 (0.06)	0.54 (0.04)
			0.05	0.50 (0.03)	0.55 (0.08)	0.55 (0.06)	0.54 (0.04)
			0.075	0.50 (0.03)	0.55 (0.08)	0.55 (0.06)	0.54 (0.04)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

The empirical type I error and power results are displayed in Table 2.10 and Figure 2.5. For $\delta = 0$, the four approaches perform similarly where the type I error rate increases as the amount

of inter-site variability increases. For small and medium effect sizes, there is a loss of power for all four methods as σ increases. For $\delta = 0.10$, RA randomization has slightly lower power than the other designs for smaller σ and this pattern is more pronounced when $\delta = 0.15$. For $\sigma = 0.075$, SARA begins to have marginally more power than the other methods. For an effect size of 15%, the BC/CARA/SARA designs perform comparably, whereas the RA design has a notable loss of power relative to the other designs.

For a null effect size, the type I error results from the GLM/GLMM are comparable to the naive approach. For $\delta = 0.10$, RA randomization has slightly lower power than the other designs for smaller σ as before, however, SARA outperforms the other methods for larger σ. For an effect size of 15%, the BC/SARA designs have higher power compared to the RA/CARA approaches. As σ increases, SARA maintains slightly more power than the balanced case along with the other methods, whereas the RA design has the greatest loss of power.

Table 2.10: Scenario 3 - Empirical Type I Error and Power

				Naive				**GLM/GLMM**	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
3	0.5	0	0	0.06	0.05	0.06	0.05	0.06	0.05
		($n = 775$)	0.025	0.09	0.09	0.09	0.09	0.08	0.09
			0.05	0.20	0.20	0.21	0.20	0.20	0.20
			0.075	0.33	0.34	0.34	0.34	0.34	0.34
		0.1	0	0.78	0.78	0.79	0.79	0.78	0.79
		($n = 775$)	0.025	0.77	0.76	0.77	0.77	0.77	0.77
			0.05	0.72	0.71	0.72	0.71	0.71	0.72
			0.075	0.66	0.66	0.66	0.67	0.66	0.67
		0.15	0	0.80	0.78	0.80	0.80	0.79	0.80
		($n = 339$)	0.025	0.76	0.74	0.77	0.77	0.75	0.76
			0.05	0.74	0.72	0.74	0.75	0.73	0.75
			0.075	0.71	0.69	0.71	0.71	0.70	0.71

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

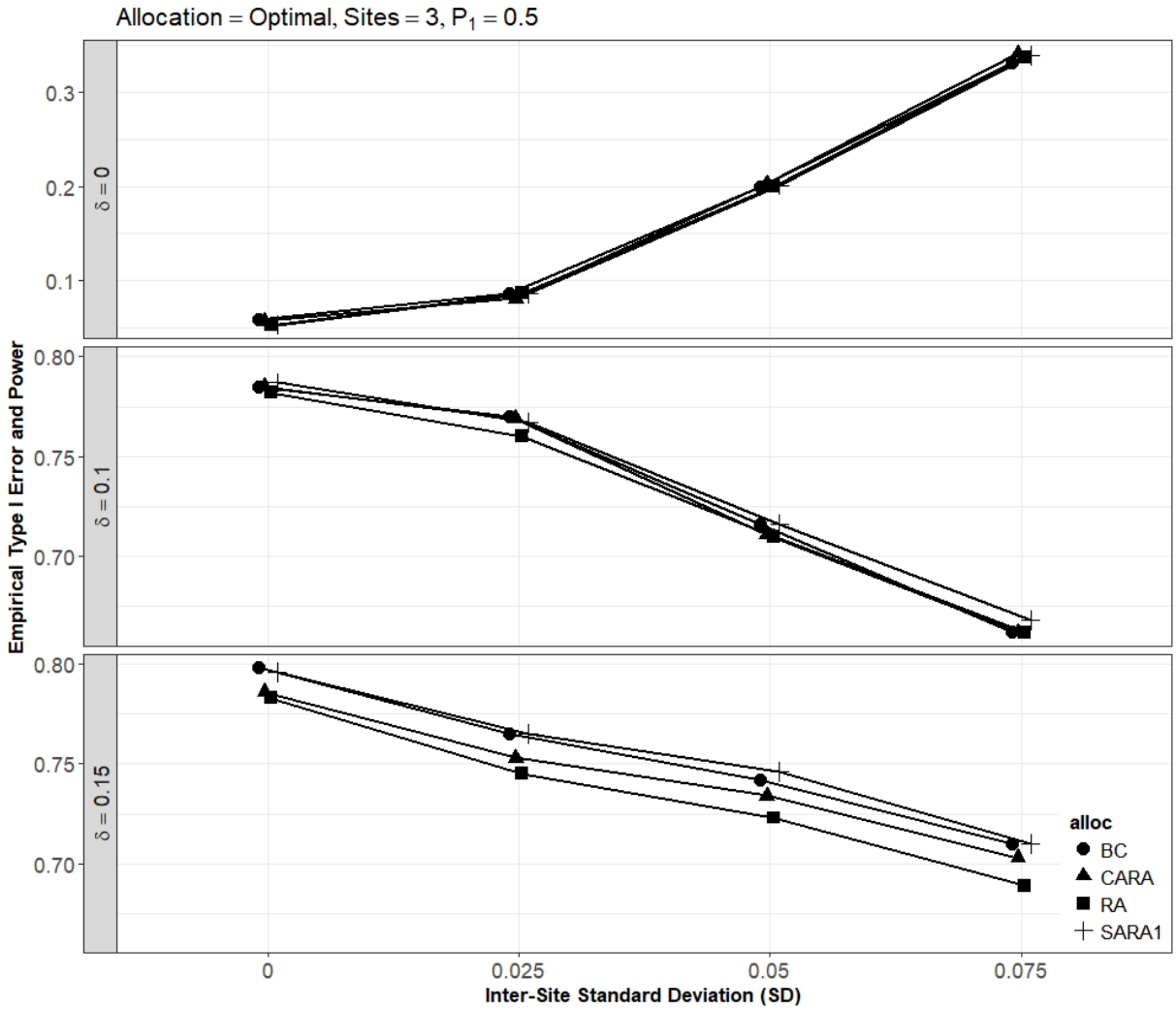


Figure 2.5: Scenario 3 - Empirical Type I Error and Power

2.4.4 Scenario 4: Neyman Allocation, 3 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 4 are displayed in Table 2.11. When the effect size is null, the four approaches are comparable, with a slight increase in the average total number of treatment successes for the adaptive designs as the amount of inter-site standard deviation increases. There is a slight increase in variability for RA/CARA over the BC but not for SARA. For $\delta = 0.10$, the adaptive designs have more expected successes than the balanced case, however, the improvement is lower for SARA compared to the RA/CARA designs. SARA has slightly less variability than the balanced case while RA randomization is the most variable. For an effect size of 15%, the adaptive methods result in more treatment successes than the BC but the

improvement is slightly lower for CARA and even less for SARA. The BC/SARA approaches have similarly low variability while the other designs have higher variability, particularly the RA design. Overall, as σ increases, the variability in expected treatment successes increases for all methods.

Table 2.11: Scenario 4 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)
		(n=587)	0.025	175.7 (12.6)	175.8 (12.7)	175.8 (12.7)	175.7 (12.6)
			0.05	175.5 (16.3)	176.0 (16.6)	175.8 (16.5)	175.8 (16.2)
			0.075	175.3 (21.0)	176.3 (21.4)	176.1 (21.5)	175.9 (20.9)
		0.1	0	146.6 (10.6)	150.3 (14.3)	149.8 (12.5)	148.6 (10.6)
		(n=587)	0.025	146.2 (12.3)	149.7 (15.6)	149.5 (13.7)	148.2 (12.2)
			0.05	146.0 (16.1)	150.1 (19.4)	149.8 (17.3)	148.4 (16.0)
			0.075	145.9 (20.8)	151.5 (23.8)	150.8 (22.1)	148.8 (20.5)
		0.15	0	54.2 (6.7)	58.5 (9.1)	57.8 (7.8)	56.4 (6.7)
		(n=241)	0.025	54.3 (6.8)	58.5 (9.6)	57.9 (8.5)	56.6 (6.8)
			0.05	54.4 (7.9)	59.0 (10.3)	58.3 (9.4)	56.9 (7.8)
			0.075	54.7 (9.5)	59.5 (11.9)	58.9 (10.7)	57.3 (9.3)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Neyman Allocation

The allocation proportions are displayed in Table 2.12. For $\delta = 0$, the four approaches have comparable allocation proportions. SARA has low variability, comparable to the balanced case, while the RA design has the largest variability. When $\delta = 0.10$, the adaptive designs have larger allocation ratios compared to the balanced case, however, the improvement is lower for SARA compared to the RA/CARA designs. While SARA is less variable than the other adaptive designs, similar to the balanced case, the RA method has the highest variability. When the effect size is 15%, the adaptive designs have even higher allocation ratios, especially RA randomization, but the improvement is lower for the CARA and SARA designs. Although SARA adapts less than the RA/CARA designs, it has the lowest variability similar to the balanced case while the RA

design has the largest variability. The amount of inter-site standard deviation does not affect the allocation proportions but slightly increases the variability in allocation.

Table 2.12: Scenario 4 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.3	0	0	0.50 (0.02)	0.51 (0.13)	0.50 (0.09)	0.50 (0.02)
		($n = 587$)	0.025	0.50 (0.02)	0.49 (0.12)	0.50 (0.08)	0.50 (0.03)
			0.05	0.50 (0.02)	0.49 (0.13)	0.50 (0.09)	0.50 (0.03)
			0.075	0.50 (0.02)	0.50 (0.13)	0.50 (0.08)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.56 (0.17)	0.55 (0.12)	0.53 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.56 (0.17)	0.55 (0.12)	0.53 (0.03)
			0.05	0.50 (0.02)	0.56 (0.18)	0.56 (0.12)	0.54 (0.04)
			0.075	0.50 (0.02)	0.57 (0.18)	0.56 (0.12)	0.54 (0.05)
		0.15	0	0.50 (0.03)	0.62 (0.19)	0.60 (0.13)	0.56 (0.05)
		($n = 241$)	0.025	0.50 (0.03)	0.61 (0.18)	0.60 (0.14)	0.56 (0.05)
			0.05	0.50 (0.03)	0.62 (0.19)	0.60 (0.14)	0.56 (0.05)
			0.075	0.50 (0.03)	0.62 (0.20)	0.60 (0.14)	0.56 (0.06)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Neyman Allocation

The empirical type I error and power results are displayed in Table 2.13 and Figure 2.6. For a null effect size, the four methods perform similarly where there is an inflation in type I error rate as σ increases. RA randomization has marginal improvement in the type I error as σ increases while SARA begins to show slight improvement at $\sigma = 0.075$. For both small and medium effect sizes, the power decreases as σ increases. For $\delta = 0.10$, the BC/SARA approaches have similar power, with a slight advantage for SARA, while CARA has less improvement. RA randomization maintains the lowest power. For $\delta = 0.15$, the BC/CARA/SARA designs have higher power than the RA method, but SARA slightly outperforms the balanced case. The CARA design has slightly less power than the BC/SARA designs whereas RA randomization has a notable loss of power.

Using the results from the GLM/GLMM approach, both CARA/SARA show slight improve-

ment in type I error rate at $\sigma = 0.075$ for a null effect size. For $\delta = 0.10$, the SARA design gains back even more power than the balanced case but CARA has a notable loss of power compared to the naive approach. For $\delta = 0.15$, the SARA design maintains similar power to balanced randomization, however, SARA only slightly outperforms the balanced case for large σ. The CARA design has a greater loss of power relative to the BC/SARA designs.

Table 2.13: Scenario 4 - Empirical Type I Error and Power

				Naive				GLM/GLMM	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
3	0.3	0	0	0.06	0.05	0.07	0.05	0.06	0.05
		($n = 587$)	0.025	0.09	0.09	0.10	0.09	0.08	0.09
			0.05	0.20	0.18	0.20	0.20	0.20	0.21
			0.075	0.31	0.28	0.31	0.31	0.30	0.31
		0.1	0	0.78	0.69	0.76	0.79	0.73	0.79
		($n = 587$)	0.025	0.73	0.63	0.73	0.74	0.68	0.73
			0.05	0.68	0.57	0.66	0.68	0.62	0.68
			0.075	0.64	0.53	0.62	0.64	0.58	0.64
		0.15	0	0.80	0.61	0.79	0.80	0.69	0.80
		($n = 241$)	0.025	0.76	0.60	0.74	0.78	0.66	0.76
			0.05	0.73	0.56	0.70	0.73	0.62	0.73
			0.075	0.67	0.50	0.66	0.68	0.59	0.68

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Neyman Allocation

2.4.5 Scenario 5: Neyman Allocation, 6 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 4 are displayed in Table 2.14. When the effect size is null, the four approaches are comparable with respect to the expected number of treatment successes and variability in treatment successes. For an effect size of 10%, the adaptive designs perform better than the balanced case with respect to expected treatment successes. The

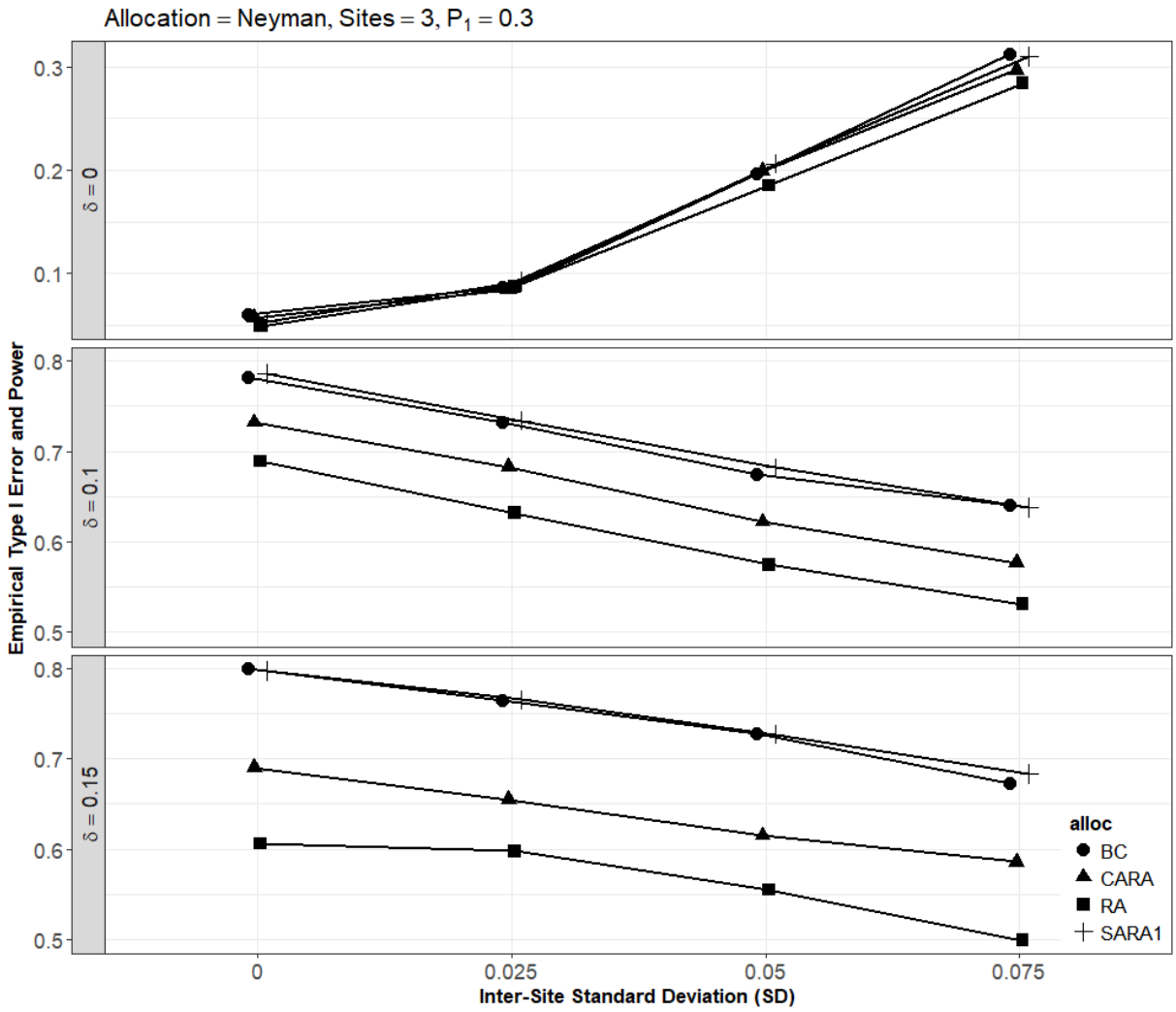


Figure 2.6: Scenario 4 - Empirical Type I Error and Power

RA design has slightly more expected treatment successes than the CARA/SARA designs, which perform similarly. The CARA/SARA designs have lower variability than RA randomization, with SARA having a slight improvement towards the balanced case, while the RA method has the highest variability. For $\delta = 0.15$, the adaptive designs result in more treatment successes than the balanced case, however, RA randomization has a greater improvement over the CARA/SARA designs, which both perform similarly. The CARA/SARA designs are less variable than the RA method. Overall, as the amount of inter-site standard deviation increases, the variability in the mean total number of treatment successes increases for all methods. However, SARA maintains less variability than the other methods and keeps the variance closest to the balanced case.

Table 2.14: Scenario 5 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**
6	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)
		(n=587)	0.025	176.3 (11.8)	176.3 (11.8)	176.3 (11.8)	176.3 (11.8)
			0.05	176.3 (14.1)	176.5 (14.1)	176.4 (14.1)	176.4 (14.1)
			0.075	176.1 (17.2)	176.6 (17.3)	176.3 (17.2)	176.4 (17.1)
		0.1	0	146.6 (10.6)	150.3 (14.3)	149.0 (10.9)	148.6 (10.6)
		(n=587)	0.025	146.9 (11.3)	150.4 (14.8)	149.2 (11.6)	148.9 (11.3)
			0.05	146.9 (13.6)	150.7 (16.8)	149.5 (14.2)	149.0 (13.6)
			0.075	146.9 (16.6)	150.7 (19.6)	149.6 (17.2)	149.1 (16.6)
		0.15	0	54.2 (6.7)	58.5 (9.1)	56.8 (6.8)	56.4 (6.7)
		(n=241)	0.025	54.2 (6.6)	58.5 (9.4)	56.7 (7.0)	56.3 (6.7)
			0.05	54.1 (7.2)	58.5 (9.8)	56.7 (7.6)	56.3 (7.4)
			0.075	54.4 (8.2)	58.9 (10.8)	57.1 (8.7)	56.6 (8.4)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Neyman Allocation

The allocation proportions are displayed in Table 2.15. For $\delta = 0$, RA randomization has a slightly higher allocation proportion when there is no inter-site standard deviation, however, as σ increases, the four approaches have comparable allocation proportions. The CARA/SARA designs

have variability similar to the balanced case, with better performance for SARA, while the RA design has the highest variability. When $\delta = 0.10$, the adaptive designs have larger allocation ratios compared to the balanced case, however, the improvement is lower for SARA compared to the RA/CARA designs. The SARA approach is less variable than the other adaptive designs and has variance similar to the balanced case, while the RA design has the highest variability. For $\delta = 0.15$, RA randomization has a higher allocation ratio than the CARA/SARA methods. Although SARA has a smaller allocation ratio compared to the other adaptive designs, it has similar variability to balanced randomization while the RA design has the highest variability. As the effect size increases, the allocation proportions increase for the adaptive designs while the variability in allocation increases for all four methods. Additionally, the amount of inter-site variability does not affect the allocation proportions or variability.

Table 2.15: Scenario 5 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**BC**	**RA**	**CARA**	**SARA 1**
6	0.3	0	0	0.50 (0.02)	0.51 (0.13)	0.50 (0.06)	0.50 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.49 (0.11)	0.50 (0.05)	0.50 (0.03)
			0.05	0.50 (0.02)	0.50 (0.12)	0.50 (0.05)	0.50 (0.03)
			0.075	0.50 (0.02)	0.49 (0.13)	0.50 (0.05)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.56 (0.17)	0.54 (0.06)	0.53 (0.03)
		($n = 587$)	0.025	0.50 (0.02)	0.56 (0.16)	0.54 (0.06)	0.53 (0.03)
			0.05	0.50 (0.02)	0.56 (0.17)	0.54 (0.06)	0.53 (0.04)
			0.075	0.50 (0.02)	0.55 (0.16)	0.54 (0.06)	0.53 (0.04)
		0.15	0	0.50 (0.03)	0.62 (0.19)	0.57 (0.07)	0.56 (0.05)
		($n = 241$)	0.025	0.50 (0.03)	0.62 (0.18)	0.57 (0.07)	0.56 (0.05)
			0.05	0.50 (0.03)	0.62 (0.18)	0.57 (0.07)	0.56 (0.06)
			0.075	0.50 (0.03)	0.62 (0.19)	0.57 (0.08)	0.56 (0.06)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Neyman Allocation

The empirical type I error and power results are displayed in Table 2.16 and Figure 2.7. For a null

effect size, there is an inflation of the type I error rate as the amount of inter-site standard deviation increases for all methods. When there is no inter-site variability, there is a slight improvement in the type I error rate for the RA/SARA designs. As σ increases, RA randomization maintains the lowest type I error rate. The BC/SARA designs perform similarly while the CARA approach has the highest inflation in type I error rate. For small and medium effect sizes, the power decreases as σ increases. When $\delta = 0.10$, RA randomization has a notable loss of power compared to the other approaches. For small σ, the CARA/SARA designs perform comparably while for larger values of σ, the BC/SARA perform similarly. When $\delta = 0.15$, the BC/CARA/SARA designs perform comparably while there is a notable loss of power for the RA method relative to the other designs. For larger σ, both the CARA/SARA designs outperform balanced randomization, with slight improvement for the SARA approach.

For a null effect size, the type I error results from the GLM/GLMM are comparable to the naive approach, where CARA maintains the highest inflation in type I error for increasing σ. When $\delta = 0.10$, the SARA design has a slight improvement in power compared to CARA. For $\delta = 0.15$, the BC/SARA approaches have comparable performance, with slightly less improvement for SARA when σ is small and marginal improvement for larger σ. There is slightly less improvement for the CARA design relative to the BC/SARA approaches.

Table 2.16: Scenario 5 - Empirical Type I Error and Power

				Naive				GLM/GLMM	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
6	0.3	0	0	0.06	0.05	0.06	0.05	0.06	0.05
		($n = 587$)	0.025	0.08	0.08	0.08	0.08	0.08	0.08
			0.05	0.12	0.12	0.13	0.13	0.14	0.13
			0.075	0.18	0.17	0.19	0.18	0.18	0.18
		0.1	0	0.78	0.69	0.78	0.79	0.78	0.79
		($n = 587$)	0.025	0.76	0.68	0.78	0.78	0.78	0.78
			0.05	0.73	0.64	0.72	0.73	0.72	0.73
			0.075	0.69	0.62	0.68	0.69	0.69	0.7
		0.15	0	0.8	0.61	0.8	0.8	0.78	0.8
		($n = 241$)	0.025	0.79	0.6	0.78	0.79	0.77	0.78
			0.05	0.76	0.6	0.76	0.77	0.75	0.76
			0.075	0.71	0.54	0.72	0.72	0.71	0.72

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Neyman Allocation

2.4.6 Scenario 6: Neyman Allocation, 3 Sites, $p_1 = 0.5$

The mean total number of treatment successes for Scenario 4 are displayed in Table 2.17. When the effect size is null, the four approaches perform similarly with respect to the expected total number of treatment successes and the variability in treatment successes. When the effect size is increased to 10%, the adaptive designs result in slightly more expected treatment successes than the balanced case. While all approaches have similar variability, SARA has a slight improvement over the others, including the balanced case. For $\delta = 0.15$, the adaptive designs result in more treatment successes than the balance case. SARA has the lowest variability while the RA design has the highest variability. As σ increases, there is an increase in the variability in treatment successes for all methods.

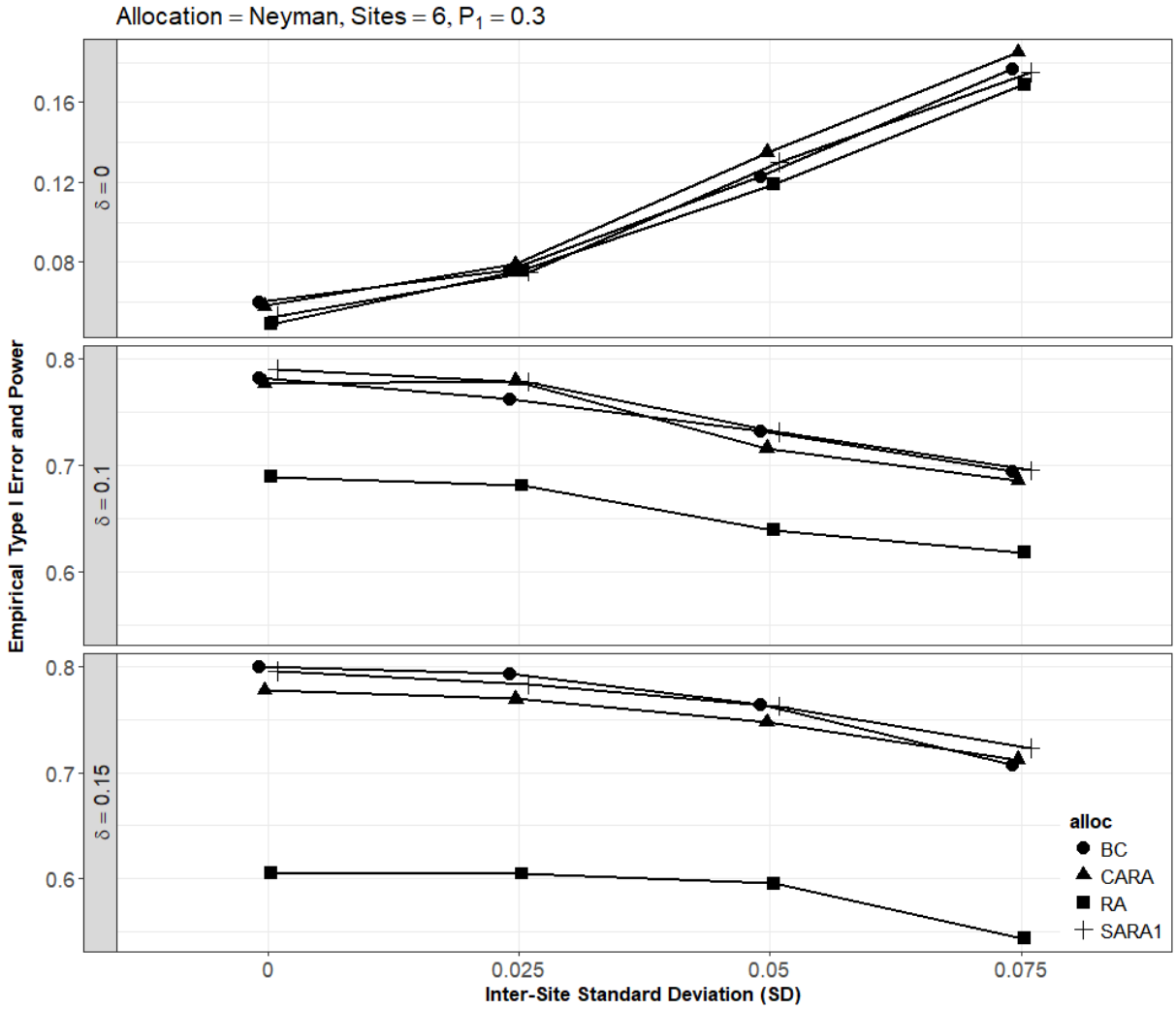


Figure 2.7: Scenario 5 - Empirical Type I Error and Power

Table 2.17: Scenario 6 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.5	0	0	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)
		(n=775)	0.025	388.5 (15.9)	388.5 (15.8)	388.6 (15.8)	388.5 (15.9)
			0.05	389.1 (20.9)	389.1 (20.8)	389.1 (20.9)	389.1 (20.8)
			0.075	389.3 (27.1)	389.4 (27.0)	389.2 (27.3)	389.3 (26.9)
		0.1	0	348.9 (14.1)	349.5 (14.1)	349.4 (14.1)	349.3 (13.9)
		(n=775)	0.025	349.0 (15.8)	349.5 (15.9)	349.5 (16.1)	349.5 (15.7)
			0.05	349.3 (21.1)	349.9 (21.0)	349.9 (21.4)	349.7 (20.9)
			0.075	350.0 (27.5)	350.8 (27.4)	350.5 (27.7)	350.6 (27.1)
		0.15	0	143.9 (9.1)	144.7 (9.2)	144.7 (9.1)	144.4 (8.9)
		(n=339)	0.025	143.7 (9.8)	144.6 (10.1)	144.5 (10.0)	144.3 (9.5)
			0.05	143.8 (11.6)	144.9 (11.8)	144.7 (11.6)	144.5 (11.3)
			0.075	143.7 (13.9)	144.8 (14.1)	144.7 (13.9)	144.5 (13.5)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Neyman Allocation

The allocation proportions are displayed in Table 2.18. For $\delta = 0$, the four methods have comparable allocation proportions. The BC/SARA approaches have similarly lower variability in allocation proportions while the RA/CARA designs have comparably higher variability. When the effect size is increased to 10%, the adaptive designs have larger allocation ratios compared to the balanced case. The BC/SARA approaches have similarly low variability while the RA/CARA designs have comparably high variability. For $\delta = 0.15$, the adaptive designs have larger allocation ratios compared to the BC, however, the improvement is marginally lower for SARA compared to the other designs. The BC/SARA approaches are comparable with respect to lower variability in allocation proportions compared to the other designs, particularly RA randomization. The amount of inter-site variability does not affect the allocation proportions or the variability in allocation.

Table 2.18: Scenario 6 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1
3	0.5	0	0	0.50 (0.02)	0.50 (0.04)	0.50 (0.04)	0.50 (0.02)
		($n = 775$)	0.025	0.50 (0.02)	0.50 (0.05)	0.50 (0.05)	0.50 (0.02)
			0.05	0.50 (0.02)	0.50 (0.04)	0.50 (0.05)	0.50 (0.02)
			0.075	0.50 (0.02)	0.50 (0.05)	0.50 (0.05)	0.50 (0.02)
		0.1	0	0.50 (0.02)	0.51 (0.04)	0.51 (0.04)	0.51 (0.02)
		($n = 775$)	0.025	0.50 (0.02)	0.51 (0.05)	0.51 (0.06)	0.51 (0.02)
			0.05	0.50 (0.02)	0.51 (0.05)	0.51 (0.06)	0.51 (0.02)
			0.075	0.50 (0.02)	0.51 (0.05)	0.51 (0.05)	0.51 (0.02)
		0.15	0	0.50 (0.03)	0.52 (0.07)	0.52 (0.06)	0.51 (0.03)
		($n = 339$)	0.025	0.50 (0.03)	0.52 (0.08)	0.52 (0.07)	0.51 (0.03)
			0.05	0.50 (0.03)	0.52 (0.08)	0.52 (0.07)	0.51 (0.03)
			0.075	0.50 (0.03)	0.52 (0.08)	0.52 (0.06)	0.51 (0.03)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Neyman Allocation

The empirical type I error and power results are displayed in Table 2.19 and Figure 2.8. For a null effect size, the four methods perform similarly where the type I error rate increases as the amount of inter-site standard deviation increases. As σ increases, CARA has a slight inflation in the type I error rate compared to the other approaches. Overall, for small and medium effect sizes, there is a loss of power for all methods as σ increases. Specifically, when the effect size is increased to 10%, the BC/CARA/SARA designs maintain slightly higher power relative to the RA randomization for smaller σ. As σ increases, the four approaches perform similarly. For $\delta = 0.15$, the BC/CARA/SARA approaches have the highest power. There is a notable loss of power for RA randomization.

For a null effect size, the type I error results from the GLM/GLMM are comparable to the naive approach, where there is an inflation in the type I error rate as σ increases. When the effect size is increased to 10%, the CARA design has notably lower power compared to the other approaches. The BC/RA/SARA methods perform similarly with marginal improvement in power for the SARA

design. For $\delta = 0.15$, the BC/SARA approaches have the highest power, with slight improvement for SARA with increasing σ. There is a notable loss of power for the RA/CARA designs, where CARA has worst performance when accounting for clustering.

Table 2.19: Scenario 6 - Empirical Type I Error and Power

				Naive				**GLM/GLMM**	
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**CARA**	**SARA 1**
3	0.5	0	0	0.06	0.06	0.06	0.06	0.06	0.06
		($n = 775$)	0.025	0.09	0.09	0.10	0.09	0.09	0.09
			0.05	0.20	0.20	0.21	0.20	0.20	0.20
			0.075	0.33	0.33	0.33	0.33	0.32	0.33
		0.1	0	0.78	0.78	0.79	0.79	0.77	0.79
		($n = 775$)	0.025	0.77	0.76	0.77	0.77	0.74	0.77
			0.05	0.72	0.72	0.72	0.72	0.69	0.72
			0.075	0.66	0.66	0.66	0.66	0.64	0.66
		0.15	0	0.80	0.78	0.80	0.80	0.77	0.79
		($n = 339$)	0.025	0.76	0.74	0.76	0.76	0.73	0.76
			0.05	0.74	0.72	0.74	0.74	0.71	0.74
			0.075	0.71	0.69	0.71	0.71	0.68	0.71

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Neyman Allocation

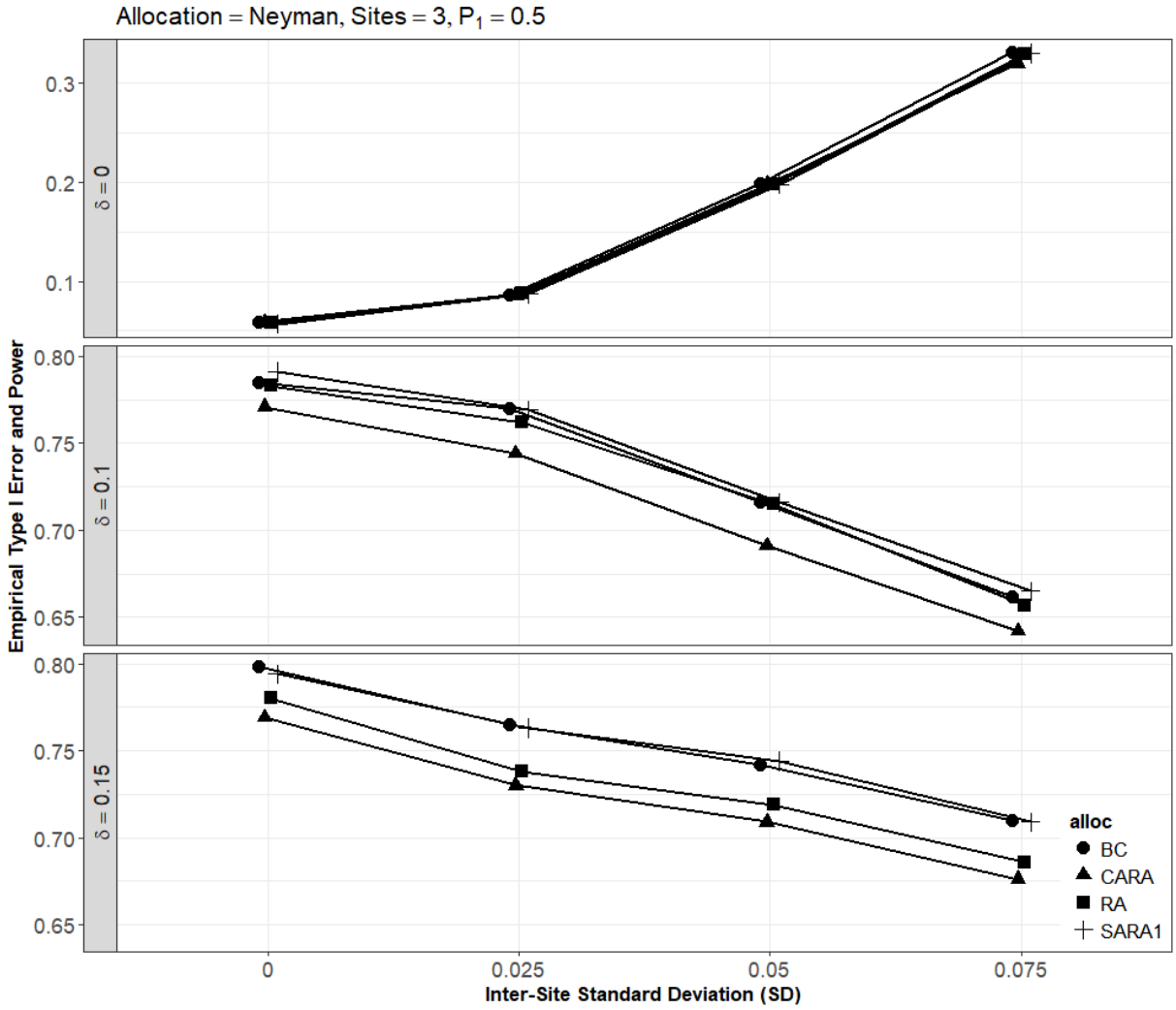


Figure 2.8: Scenario 6 - Empirical Type I Error and Power

2.4.7 Empirical Type I Error and Power by Number of Sites (k)

In Section 2.4.7, we explore the relationship between the number of sites and both empirical type I error and power. The GLM/GLMM approach, which incorporates either fixed or random site effects, is used to test the treatment effect ($H_0 : \hat{\beta}_1 = 0$) to account for inter-center variability in treatment responses. An indication of whether the null hypothesis was rejected was retained to calculate the operating characteristics of the methods.

Scenario: Allocation = Optimal, $P_1 = 0.3$

The empirical type I error results for $\delta = 0$ are displayed in Figure 2.9. For $k = 3$, there is an inflation in the type I error rate as σ increases for all four methods. RA randomization appears to control the type I error rate slightly better than the other designs particularly when there is more inter-site variability. When $k = 6$, the four methods are comparable for lower inter-site variability. For larger σ, the BC/SARA designs perform similarly whereas the CARA method has the highest type I error rate and RA randomization has the lowest. For $k = 9$, all methods are comparable regardless of the inter-site variability. As the number of sites increases, the type I error rate decreases, particularly in cases where there is more inter-site standard deviation. When σ is low, the methods perform close to the nominal level of 0.05 regardless of k.

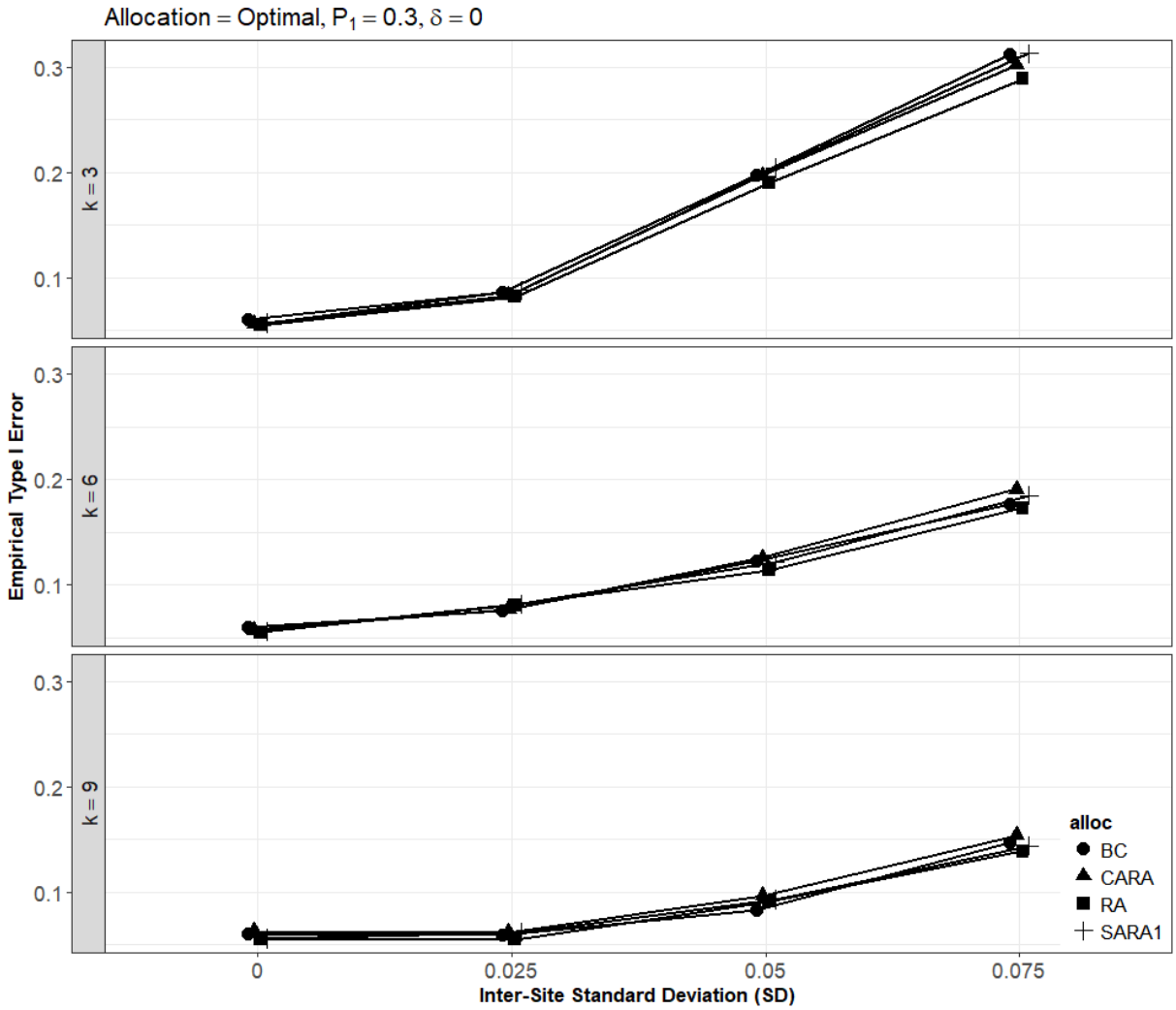


Figure 2.9: Empirical Type I Error by Site, $P_1 = 0.3, \delta = 0$

The empirical power results for $\delta = 0.10$ are displayed in Figure 2.10. For $k = 3$, the SARA design is similar to balanced case and thus gains back some of the power lost with the other adaptive methods, particularly RA randomization. For $k = 6$, the RA design has a notable loss of power relative to the other methods which perform comparably. Before $\sigma = 0.05$, the CARA/SARA designs have marginally higher power than the balanced case, where SARA has a slightly improvement. For larger σ, the three methods are comparable. When $k = 9$, the RA method maintains a notable loss of power relative to the other methods which perform comparably. As the number of sites increases, there is less of a decrease in power for all four methods though power is maintained better by BC/SARA/CARA, particularly as the inter-site variability increases.

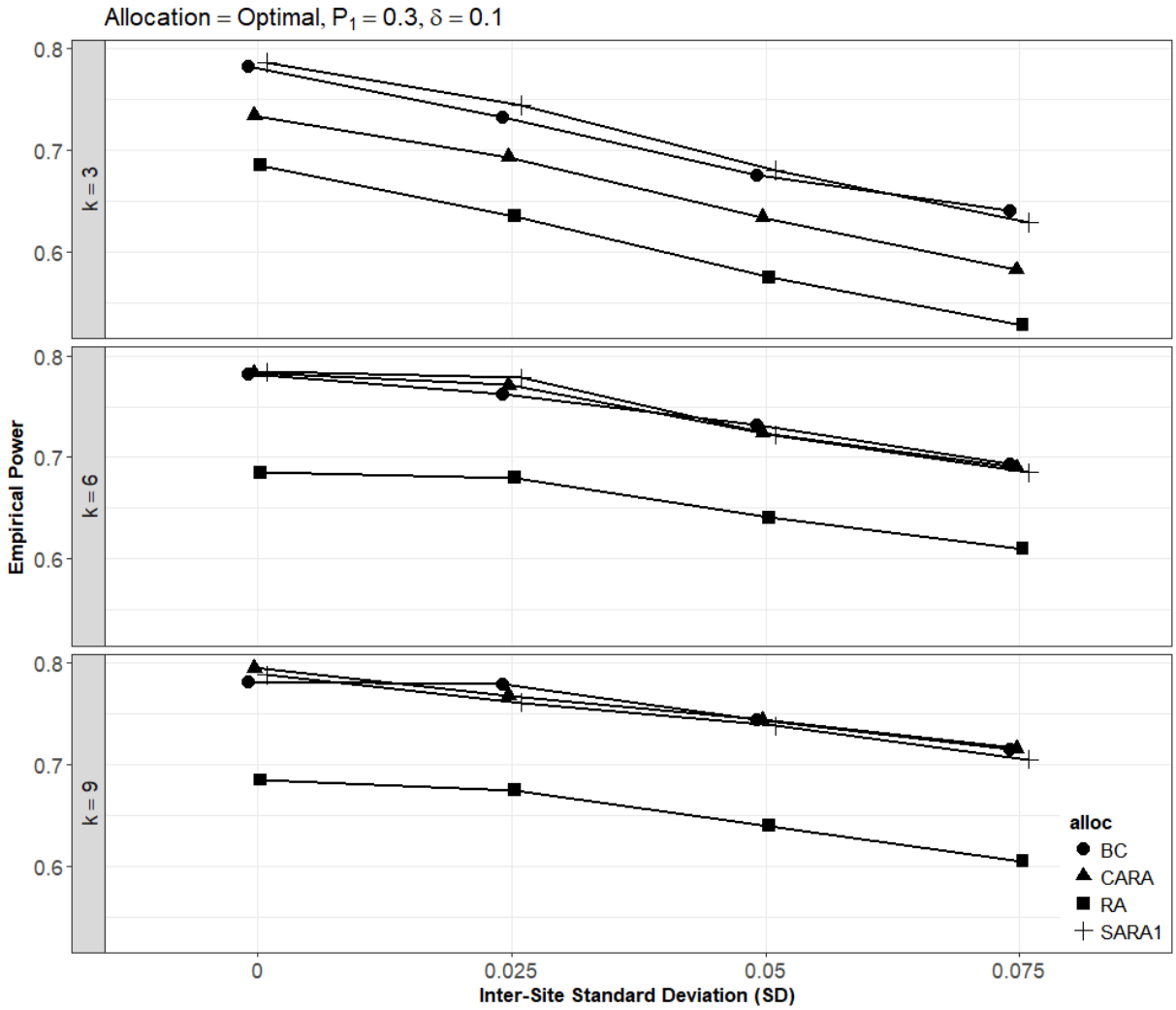


Figure 2.10: Empirical Power by Site, $P_1 = 0.3, \delta = 0.10$

The empirical power results for $\delta = 0.15$ are displayed in Figure 2.11. For $k = 3$, there is a notable loss of power for the CARA/RA designs while the SARA approach gains back most, if not all, of the power lost with the other adaptive methods. For $k = 6$, the CARA/SARA methods gain back most of the power lost with the RA design, but the improvement is slightly lower for CARA. When $k = 9$, RA randomization has the lowest power while the BC/CARA/SARA methods have comparable power when there is little inter-site variability. SARA has the highest power when there is no variability, however, there is slight loss of power for the SARA design compared to the BC/CARA methods as σ increases. As k increases, there is less of a power loss for all four methods, especially as σ increases.

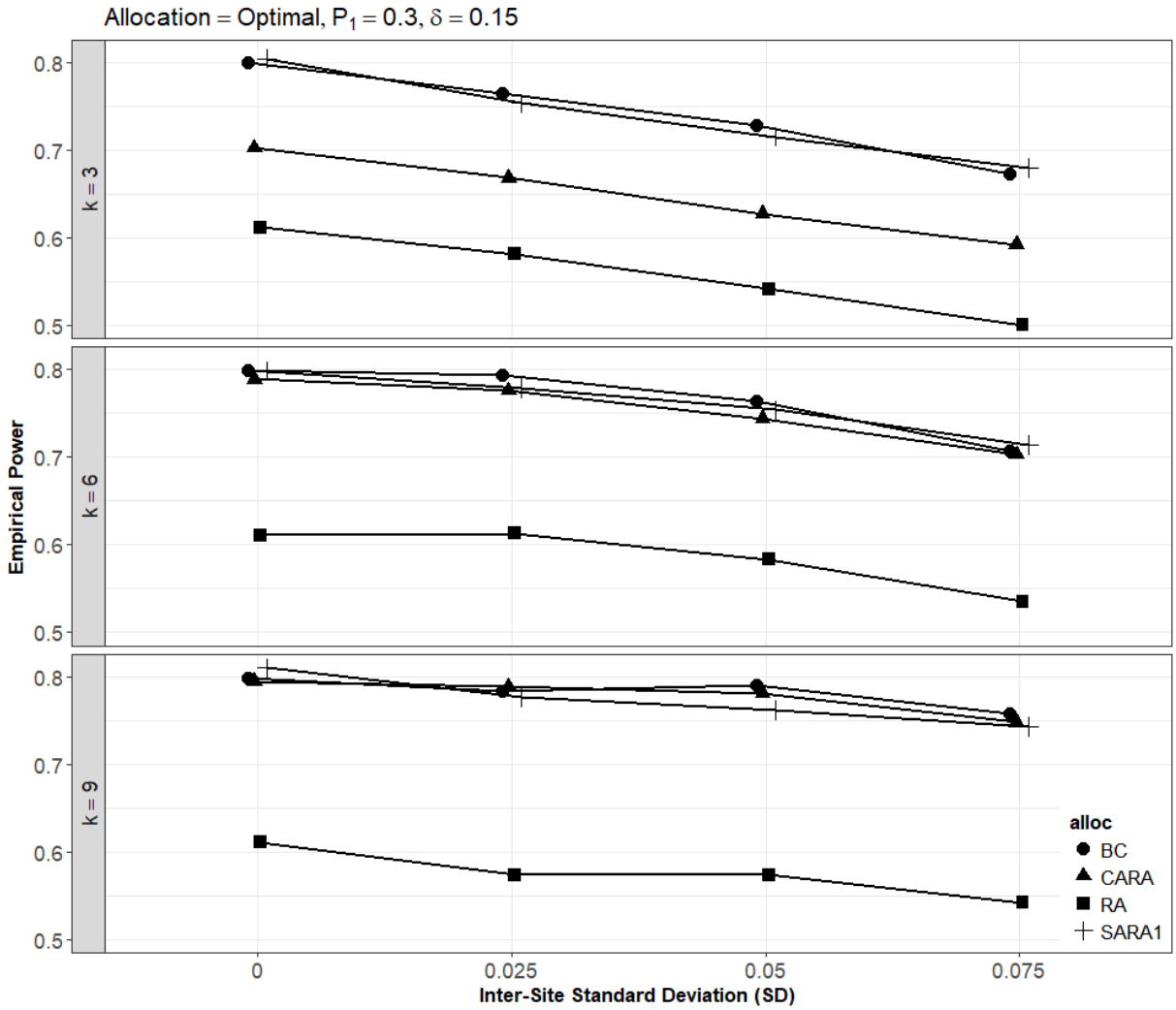


Figure 2.11: Empirical Power by Site, $P_1 = 0.3, \delta = 0.15$

Scenario: Allocation = Optimal, $P_1 = 0.5$

The empirical type I error results for $\delta = 0$ are displayed in Figure 2.12. For $k = 3$, the four approaches perform similarly where the type I error rate increases as the amount of inter-site variability increases. For $k = 6$, the four methods are comparable, however, there is a slight improvement for the BC/RA/SARA designs when $\sigma = 0.075$. When $k = 9$, there is a slight increase in type I error for the CARA design for larger values of σ compared to the other methods which perform similarly. As the number of sites increases, the type I error rate decreases, particularly with higher σ. When σ is low, the methods perform closer to the nominal level of 0.05 regardless of the number of sites.

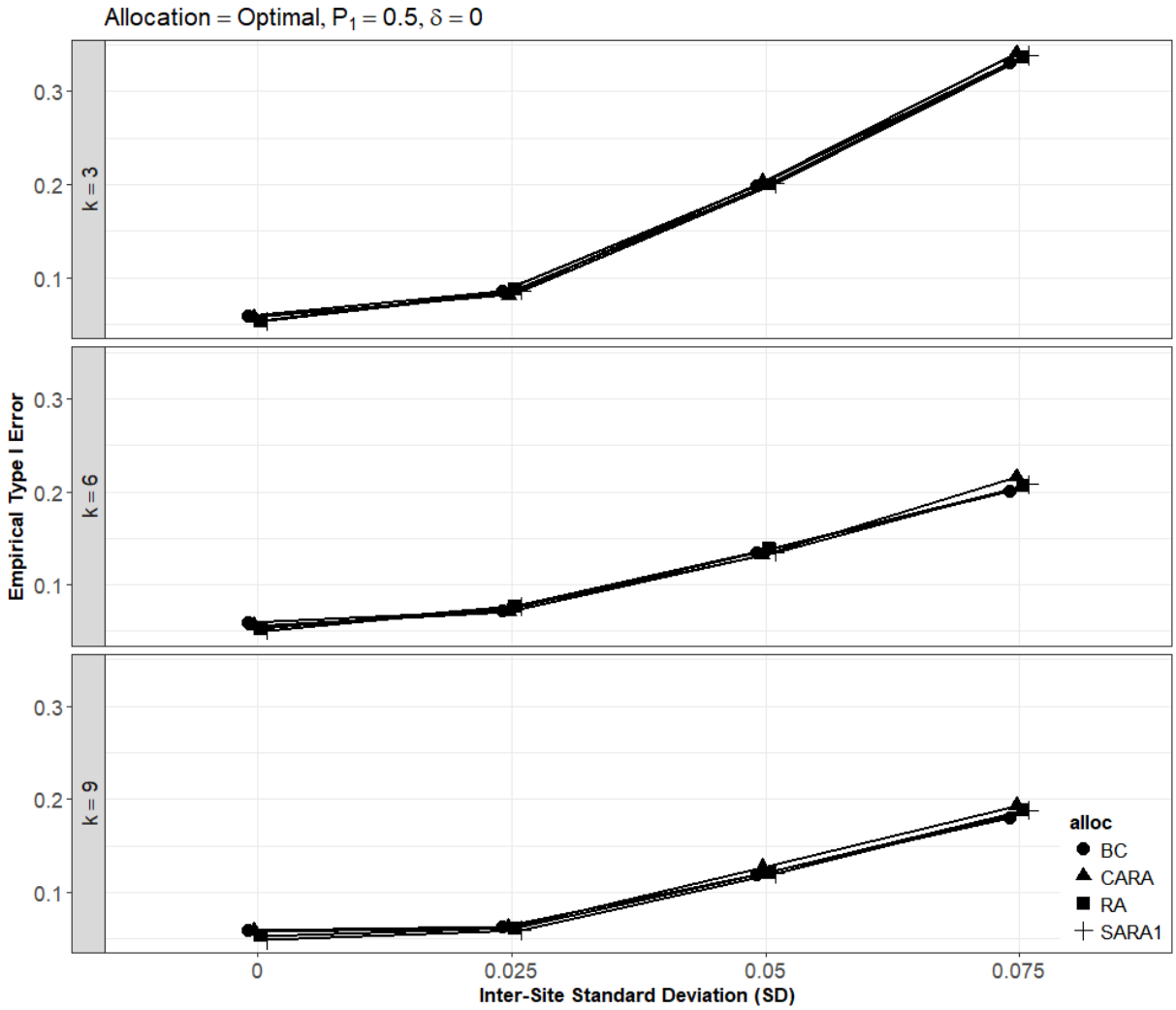


Figure 2.12: Empirical Type I Error by Site, $P_1 = 0.5, \delta = 0$

The empirical power results for $\delta = 0.10$ are displayed in Figure 2.13. For $k = 3$, RA randomization has slightly lower power than the other designs for smaller σ and this pattern is more pronounced with increasing k. For larger inter-site standard deviation, SARA has marginally more power than the other methods which perform similarly. For $k = 6$, the CARA/SARA designs gain back the power lost with RA randomization, particularly for smaller σ. For larger σ, the SARA design has slightly higher power than the other approaches, include the BC. When $k = 9$, the CARA/SARA designs gain back most, if not all, of the power lost with the RA design, compared to the balanced case. There is a slight improvement in power for the CARA method. As k increases, the power loss stabilizes for all four methods, especially for larger σ.

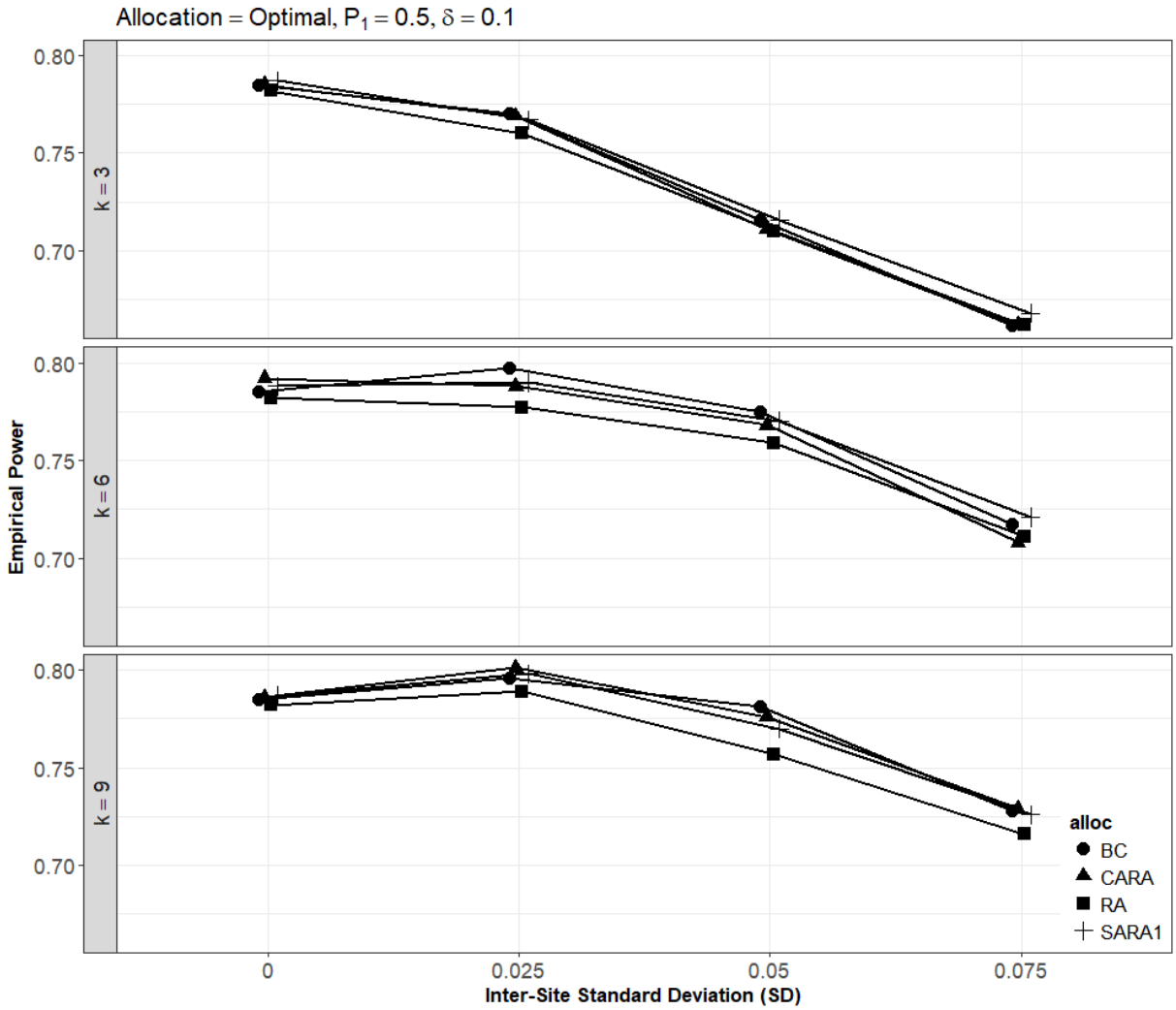


Figure 2.13: Empirical Power by Site, $P_1 = 0.5, \delta = 0.10$

The empirical power results for $\delta = 0.15$ are displayed in Figure 2.14. For $k = 3$, the BC/SARA designs have higher power compared to the RA/CARA approaches. As σ increases, SARA maintains slightly more power than balanced randomization, whereas the RA design has the lowest power. When $k = 6$, the BC/CARA/SARA approaches have higher power than RA randomization. For smaller inter-site standard deviation, there is less improvement for SARA. However, as σ increases, the SARA design gains back most of the power while there is a loss of power for the CARA method. For $k = 9$, the BC/CARA designs perform comparably when there is no inter-site variability. There is a loss of power for the RA/SARA approaches, with a slight improvement for the SARA design. As σ increases, the CARA/SARA designs outperform the balanced case with respect to power, however, the improvement is lower for SARA. There is a consistent loss of power

for the RA method compared to balanced allocation. Overall, as k increases, there is less of a power loss for all four approaches. For a larger number of sites and smaller σ, there are instances where there is an increase in power but, as σ increases, there is a slight loss of power for all methods.

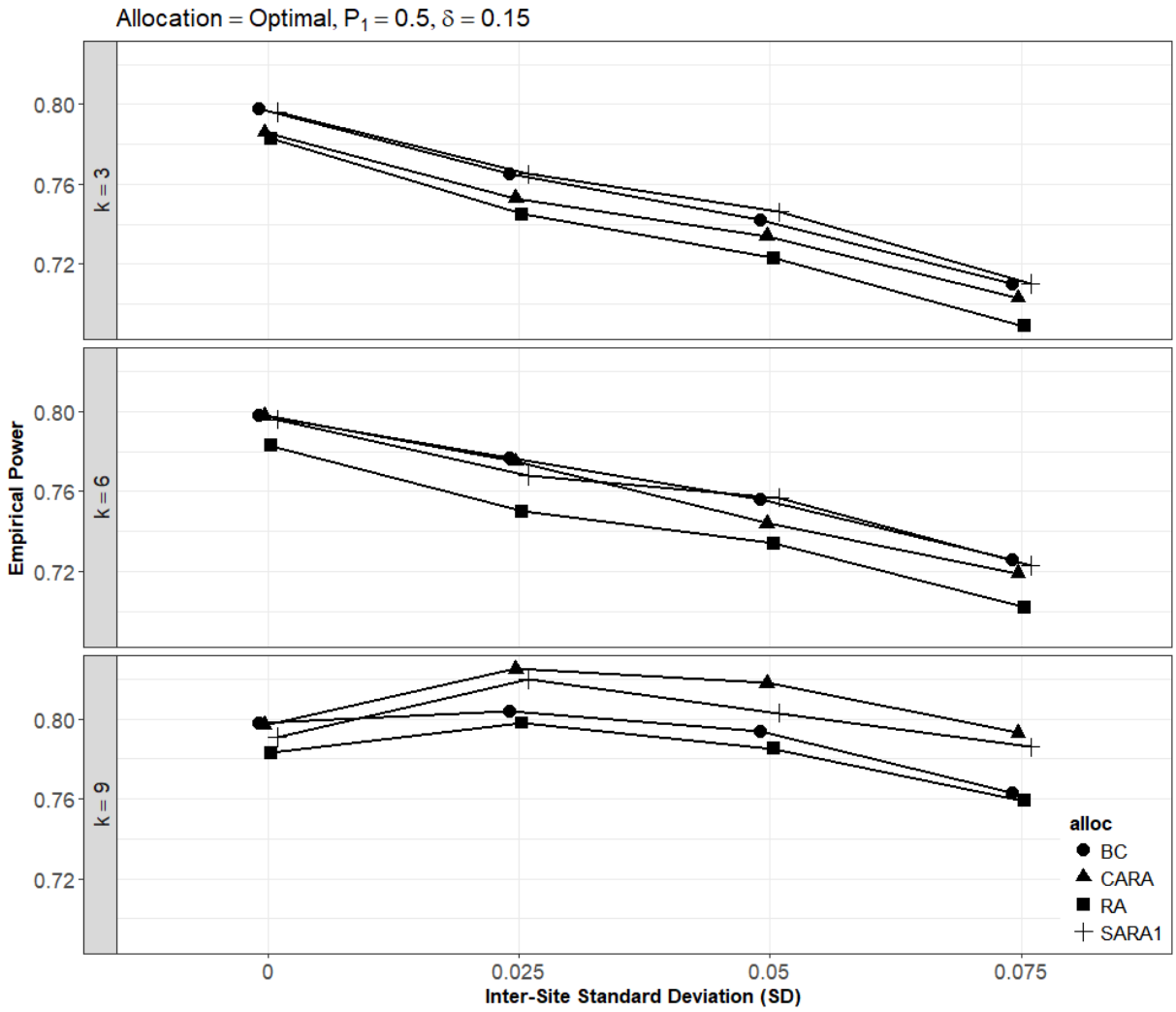


Figure 2.14: Empirical Power by Site, $P_1 = 0.5, \delta = 0.15$

Scenario: Allocation = Neyman, $P_1 = 0.3$

The empirical type I error results for $\delta = 0$ are displayed in Figure 2.15. For $k = 3$, the four methods perform similarly where there is an inflation in type I error rate as the amount of inter-site standard deviation increases. The adaptive designs have marginal improvement in the type I error rate as σ increases, although there is less improvement for the CARA/SARA approaches compared to RA randomization. For $k = 6$, the four methods perform comparably when σ is small. As σ increases, the BC/SARA approaches have similar type I error rates with a marginal improvement

over CARA. RA Randomization has a slightly lower type I error relative to the other methods with increasing inter-site variability. For $k = 9$, the CARA design performs comparable to balanced allocation with the highest type I error rates when there is little to no inter-site standard deviation. However, as σ increases, the BC/RA/SARA approaches perform similarly with slight improvement compared to CARA. As the number of sites increases, the type I error rate decreases, particularly in the presence of inter-site variability.

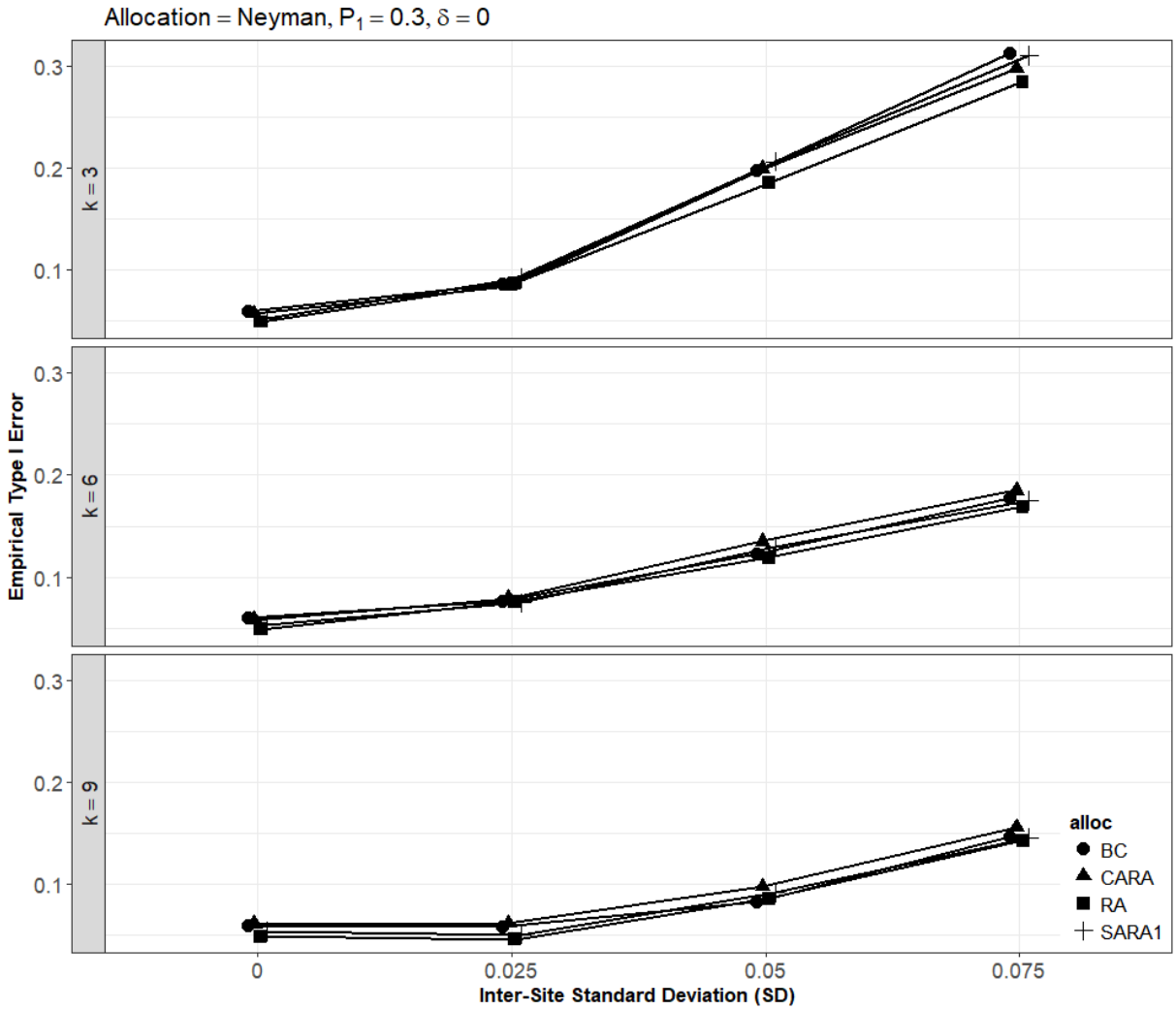


Figure 2.15: Empirical Type I Error by Site, $P_1 = 0.3, \delta = 0$

The empirical power results for $\delta = 0.10$ are displayed in Figure 2.16. For $k = 3$, the BC/SARA approaches have comparable power, with a slight advantage for SARA, while there is a notable loss of power for the other adaptive designs. RA randomization has the lowest power. For $k = 6$, the BC/CARA/SARA approaches perform similarly while the RA method maintains a notable

decrease in power. In cases with little to no inter-site variability, the CARA/SARA designs have a slight increase in power, however, as σ becomes larger, the BC/SARA approaches have a marginal improvement compared to CARA. When $k = 9$, the BC/CARA/SARA designs have similar power and outperform RA randomization. As k increases, the power loss stabilizes for all methods, particularly for larger inter-site variability.

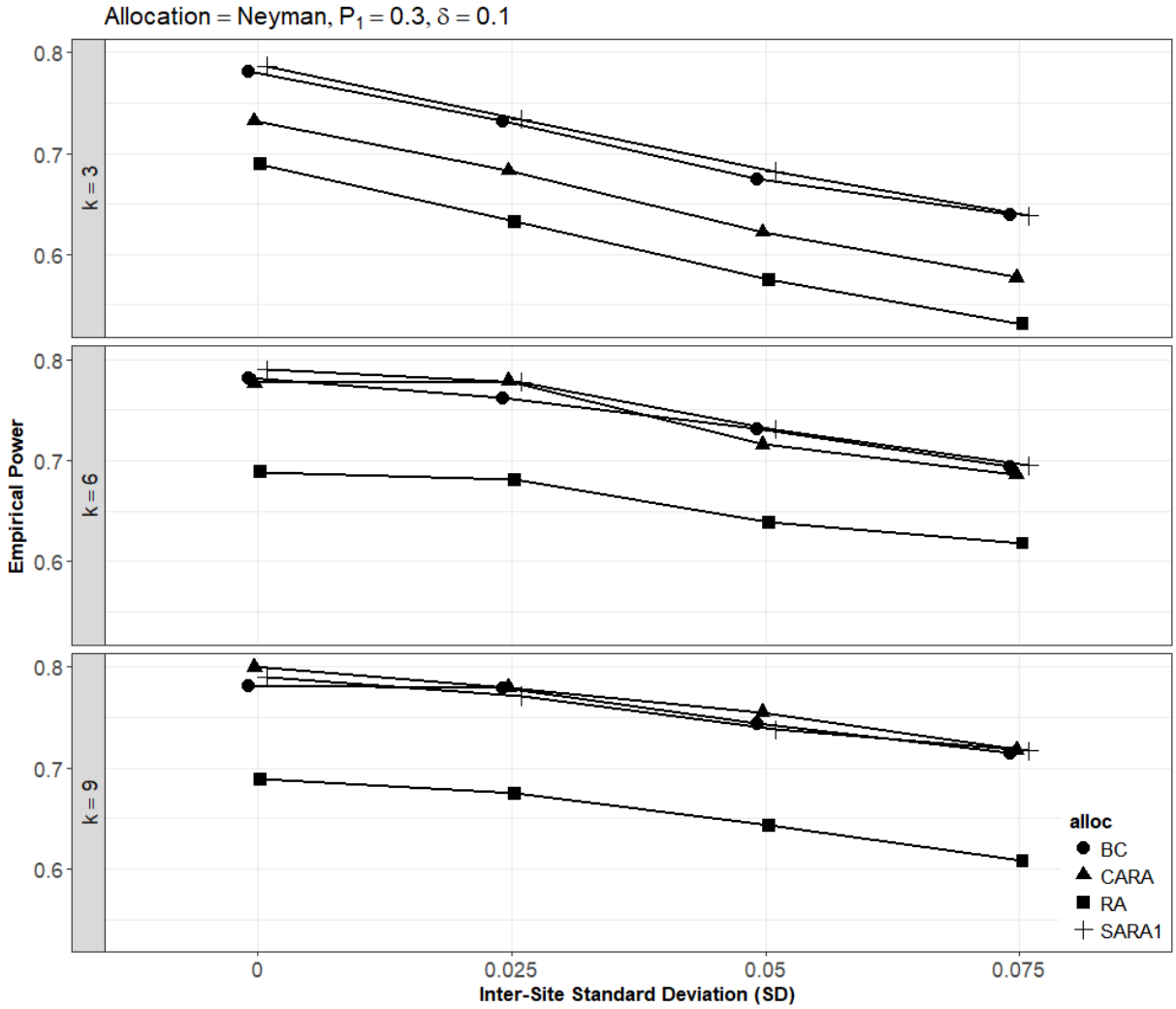


Figure 2.16: Empirical Power by Site, $P_1 = 0.3, \delta = 0.10$

The empirical power results for $\delta = 0.15$ are displayed in Figure 2.17. For $k = 3$, the BC/SARA designs have higher power than the other adaptive methods, but SARA slightly outperforms the balanced case as the inter-site standard deviation increases. The CARA design has notably less power than the BC/SARA designs but performs better than RA randomization. For $k = 6$, there is notable loss of power for the RA method relative to the other designs. For smaller σ, the BC/SARA

approaches have marginal improvement over the CARA design. However, as σ increases, the SARA approach has the highest power. When $k = 9$, BC/CARA/SARA perform comparably, particularly for lower σ, while RA randomization maintains a notable loss of power. As σ increases, balanced randomization and the CARA design slightly outperform SARA. As k increases, the power loss stabilizes for the BC/CARA/SARA approaches, especially as σ increases. For RA randomization, there is less of a stabilizing effect on power as the number of sites increases.

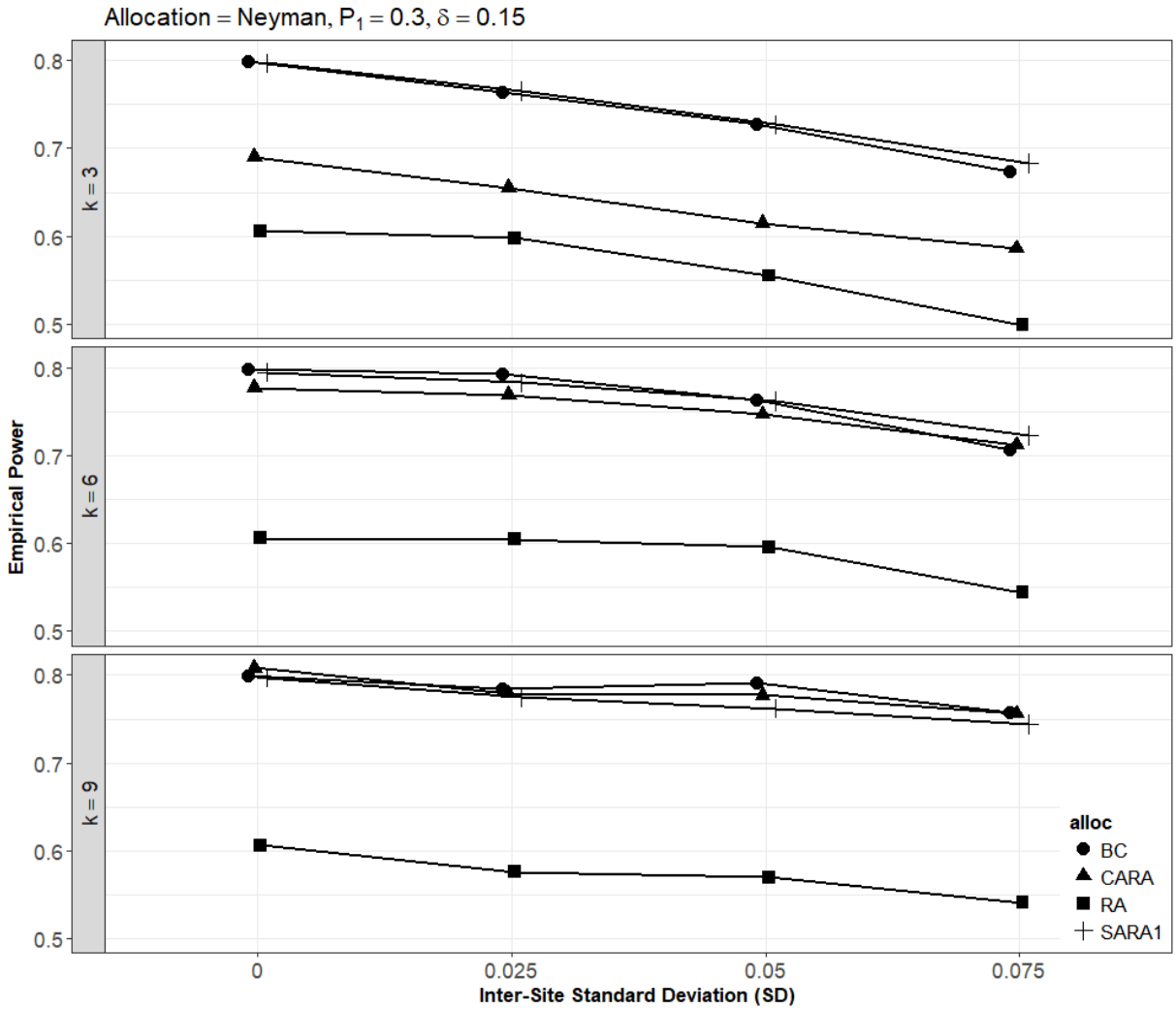


Figure 2.17: Empirical Power by Site, $P_1 = 0.3, \delta = 0.15$

Scenario: Allocation = Neyman, $P_1 = 0.5$

The empirical type I error results for $\delta = 0$ are displayed in Figure 2.18. For $k = 3$, all methods perform similarly where the type I error rate increases as the amount of inter-site standard deviation increases. For larger σ, the adaptive designs have a marginal improvement in type I error compared

to balanced randomization. For $k = 6$ and $k = 9$, the four approaches perform comparably. For $k = 9$, there is a slight improvement for the RA/SARA designs for larger σ. As the number of sites increases, the type I error rate decreases, especially when larger inter-site variability exists.

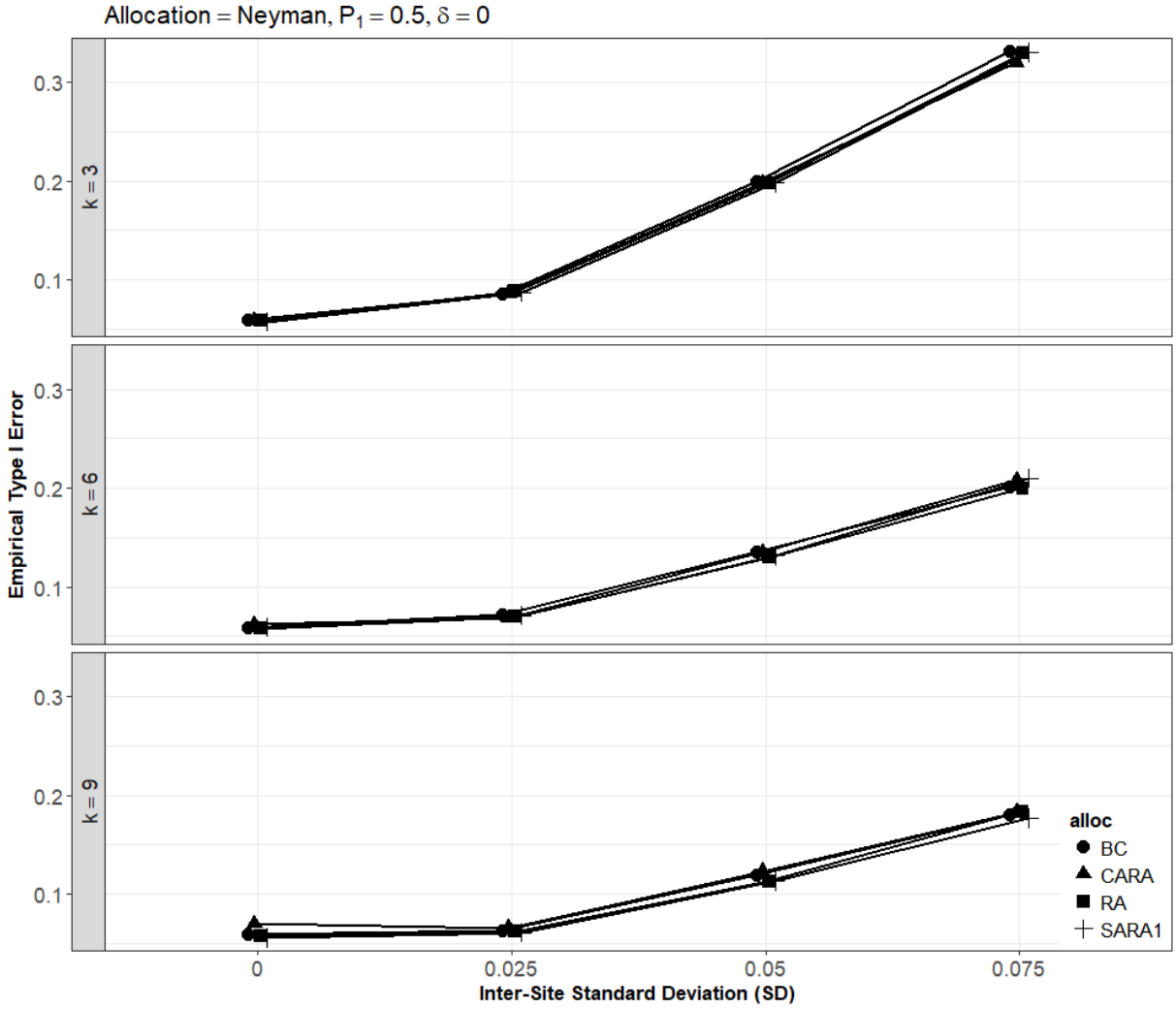


Figure 2.18: Empirical Type I Error by Site, $P_1 = 0.5, \delta = 0$

The empirical power results for $\delta = 0.1$ are displayed in Figure 2.19. When $k = 3$, the CARA design has lower power compared to the other approaches. The BC/RA/SARA methods perform similarly with marginal improvement in power for the SARA design. For $k = 6$, SARA gains back all of the power lost with the other adaptive designs and has comparable power to the balanced case. The RA/CARA approaches perform similarly, both with less power than the BC/SARA designs. When $k = 9$ and there is no inter-site variability, the CARA/SARA designs outperform the balanced case while there is a slight loss of power for the RA method. As σ increases, the BC/CARA/SARA

approaches perform comparably while there is a slight loss of power for RA randomization. As the number of sites increases, the power loss stabilizes, particularly as σ increases. In some cases, there appears to be a slight increase in power for a larger number of sites and little inter-site variability.

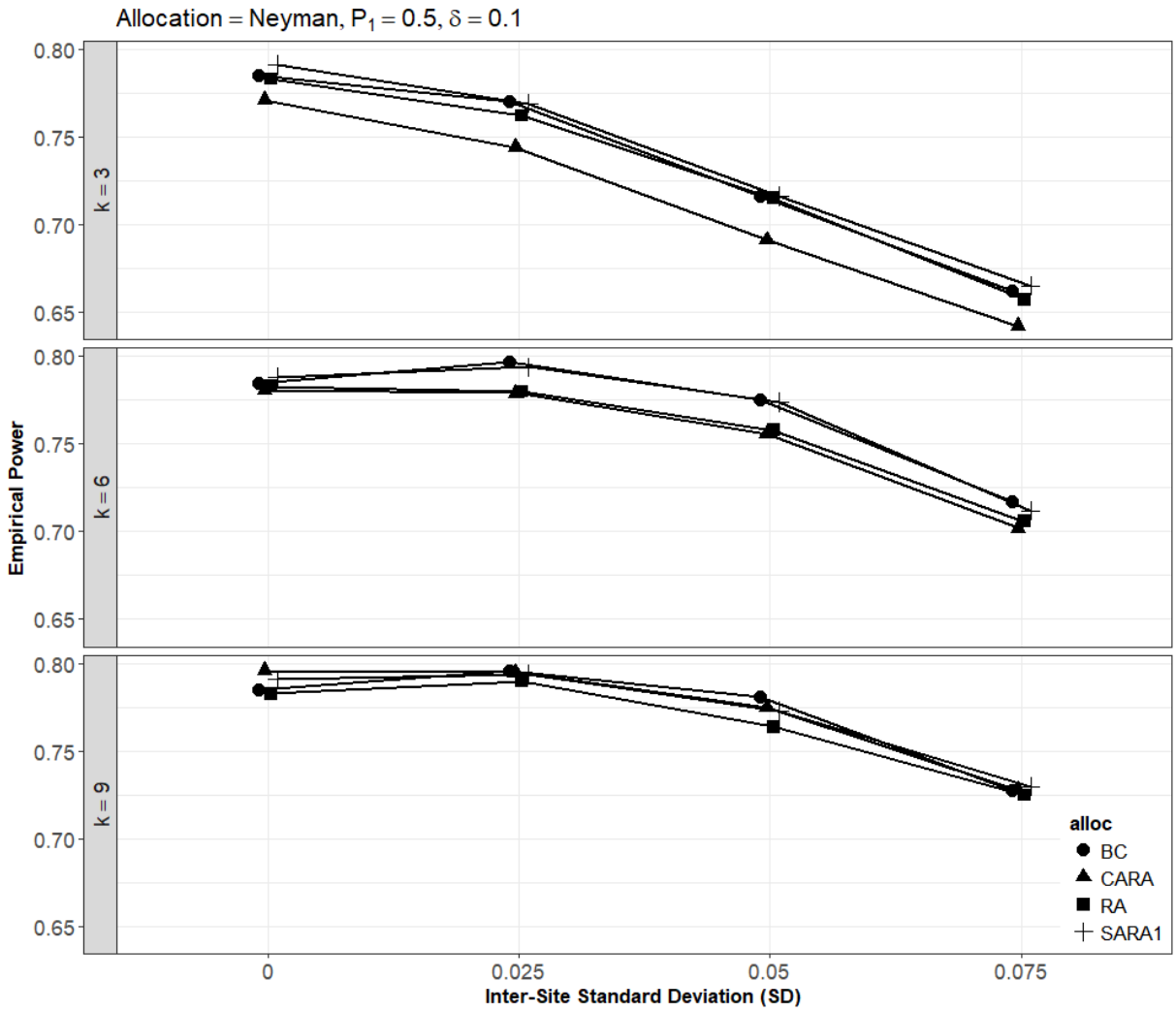


Figure 2.19: Empirical Power by Site, $P_1 = 0.5, \delta = 0.10$

The empirical power results for $\delta = 0.15$ are displayed in Figure 2.20. For $k = 3$, the BC/SARA approaches have the highest power, with SARA performing marginally better with increased σ. There is a notable loss of power for the RA/CARA designs, with slight improvement for RA randomization. For $k = 6$, the BC/SARA designs have comparably higher power than the RA/CARA methods. CARA outperforms RA randomization and is more similar to the other deigns when there is a no inter-site variability or when σ is high. For $k = 9$, the BC/CARA designs have comparable power when $\sigma = 0$ while there is a slight loss of power for SARA. As the amount

of inter-site variability increases, the CARA/SARA designs outperform the balanced case, with a slight improvement for CARA. RA randomization maintains a notable loss of power relative to the other designs for all σ. As k increases, there is a stabilizing effect on the power loss for all approaches. When $k = 9$, there is a slight improvement in power for all methods when inter-site variability is small, however, the power continues to decline as σ increases.

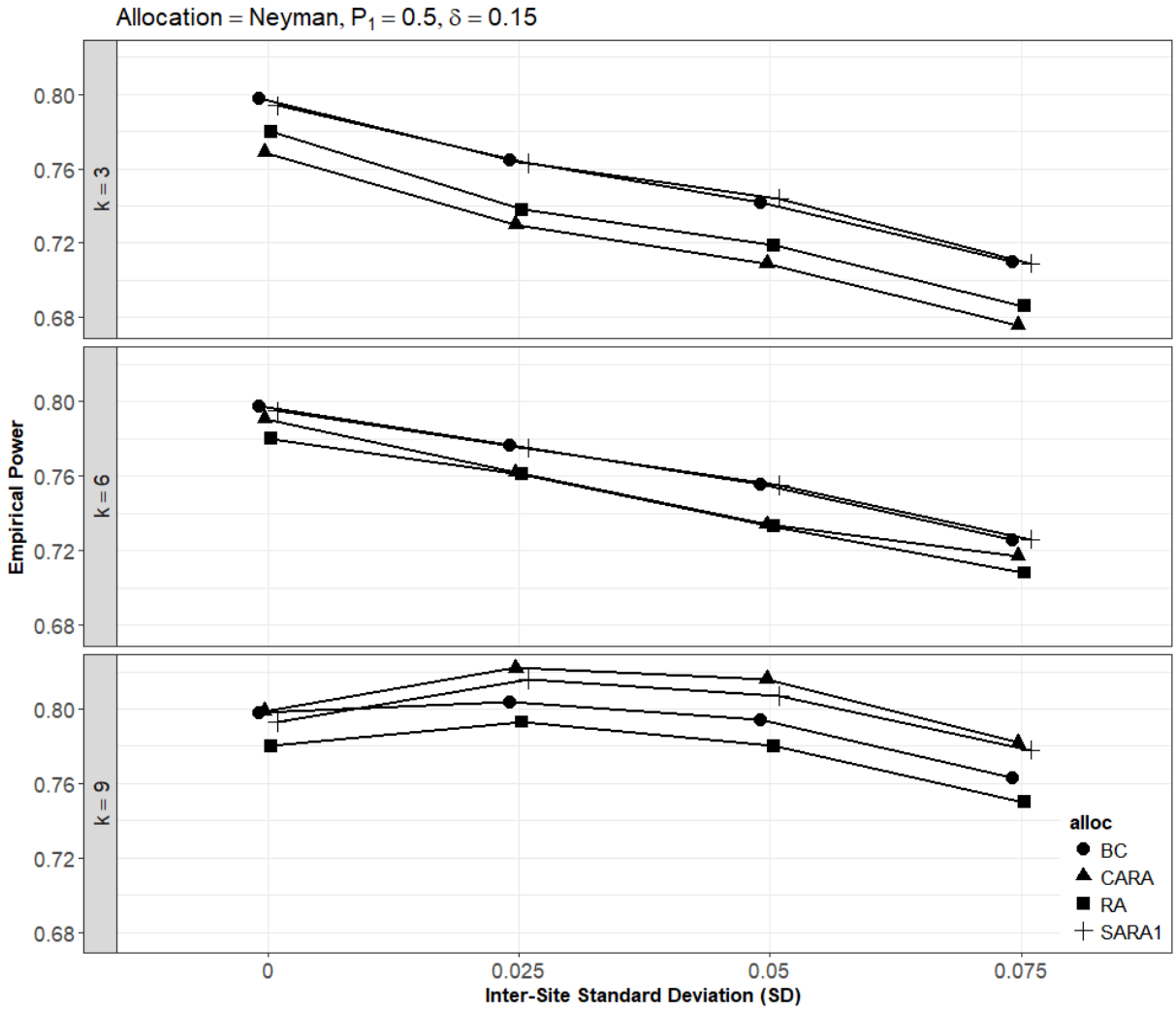


Figure 2.20: Empirical Power by Site, $P_1 = 0.5, \delta = 0.15$

2.5 Discussion

We proposed two methods to account for site-based response variability by calculating conditional success probabilities specific to each center that were incorporated in a novel approach to skew the treatment allocations. When there is a null effect size, the results show that the four allocation designs (equal allocation, RA, CARA, SARA) perform similarly in terms of mean total number of successes, allocation proportions, and type I error rates, regardless of the true success rates and amount of inter-site standard deviation. The SARA approach outperforms the other designs for small success probabilities with both small and medium effect sizes. In these cases, the other adaptive designs result in slightly more treatment successes and larger allocation proportions, however, the SARA design has more treatment success and adaptation than balanced allocation while also achieving lower variability compared to RA randomization and the CARA approach. SARA gains back most, if not all, of the power lost with the other adaptive designs. It is encouraging to note that the adaptive designs work best at lower success probabilities, when reducing treatment failures is more important.

For larger success probabilities with small and medium effect sizes, the adaptive designs outperform the balanced case in regards to treatment successes and adaptation, with slightly less improvement for SARA, with all methods achieving similar variability. SARA may result in less adaptation than the other adaptive methods due to accounting for site heterogeneity. As the amount inter-site variability increases, SARA maintains slightly more power than the other designs, including the balanced case. SARA may have gained back power relative to the other adaptive designs since here we are accounting for additional variability introduced from the centers. By incorporating the random site effects in the model to account for between-center variability, we better control the type I error and increase power.

For small and medium success probabilities, Neyman allocation performs similarly to Optimal Allocation because for $p_1, p_2 < 0.5$, the variance is greater in the more successful treatment group. Accordingly, Neyman allocation produces a larger allocation weight to the superior treatment. For large success probabilities, the balanced case outperforms the adaptive designs with respect to treatment successes for small and medium efficacy, however, SARA has a slight improvement in expected treatment successes and the lowest variability compared to the other methods, including

equal allocation, even in the presence of inter-site variability.

Overall, for increasing inter-site standard deviation, the variability in expected treatment successes increases, the type I error rate increases and the power decreases for all methods. However, as you increase the number of sites, the variability in average treatment successes decreases and there is less of an inflation in type I error and loss of power, particularly as σ increases. The variability in expected treatment successes may decrease with increasing sites because there is less variability in the site-specific success probabilities as they regress towards the true mean success probabilities.

Selvaratnam et. al (2018) established the consistency and asymptotic normality (multivariate Gaussian distribution) of the ML estimators of GLMs implemented in adaptive randomization designs. A simulation study was conducted to compare the performance of the completely randomized, RA and CARA approaches for a binary outcome and categorical covariates. The results showed that the RA method is more ethical (success rates) and has higher statistical efficiency (ML estimates) if a treatment-by-covariate interaction does not exist. However, the CARA design is more ethical in the presence of a interaction, particularly later on in a trial when a significant covariate interaction is detected [11].

In this chapter, the CARA and SARA designs incorporated fixed and random site effects, respectively, to account for site-based response variability when allocating patients in a multi-center trial. We expect that as the inter-site standard deviation increases, the magnitude of qualitative interactions increases regardless of effect size. The effect of not accounting for potential variability in effect sizes between centers is unknown. Accordingly, a random treatment-by-site interaction could be included in the model. We did not do so because another limitation of this approach is issues in estimating many random effects, particularly early on in the study when sample sizes within centers are small or when there is a constant response. Because the simulations were modeled as a Phase 2 efficacy trial, there are a small number of centers that require that enough patients be recruited to each. We expect there would be additional convergences issues when including a random treatment-by-site interaction in the frequentist setting. Finally, the simulations were designed with 80% power, assuming no inter-site variability, which requires a relatively large sample size which may be unrealistic when conducting a clinical trial in real life.

The trade-off between power and treatment failures is an expected consequence of adaptive

designs. By accounting for an extra source of variation in the randomization process, the variability in the allocation weights and thus assignments were reduced. Since adaptive methods skew the allocation ratio depending on the aim, the sample sizes for the groups are often unbalanced and may result in a loss of power compared to balanced randomization. However, the loss of power for the SARA approach was not as notable as the other methods, including RA randomization. Therefore, our methods reduced variability but only at the slight cost of power in some cases compared to the balanced case. Though SARA gains back much of the power lost when using either the RA or CARA approaches, it does not gain back all of the power lost due to inter-site variability. Thus, it is imperative to account for inter-site variability in the design stage by inflating the anticipated response variability, which will increase the required sample size.

CHAPTER 3

ACCOUNTING FOR VARIABILITY IN TREATMENT EFFECTIVENESS

In this chapter, we introduce a Bayesian approach to the SARA design to account for variability in treatment effectiveness in the presence of qualitative interactions. A Bayesian approach is considered to be flexible enough to overcome any computational challenges from modeling a large number of centers as well as the additional parameters introduced by the treatment-by-site interaction effects. In Section 3.1, we summarized the existing Bayesian adaptive methods and treatment-by-center interactions in clinical trials. In Section 3.2, we introduce two approaches, one that extends Chapter 2 under the Bayesian framework and another that models a random treatment-by-site interaction effect to account for inter-site variability in treatment effectiveness in addition to site-based response variability. A simulation study is conducted in Section 3.3 to compare the performance of our proposed methods to those in Chapter 1 as well as both balanced and RA randomization. The results are explored and a brief discussion of the implications follows.

3.1 Background and Existing Methods

3.1.1 Bayesian Covariate-Adjusted Response-Adaptive (CARA) Design

Thall and Wathen (2005) use a between-patient adaptive rule, namely outcome-adaptive randomization[1], a generalization of Thompson's (1933) method for two binomial distributions with beta priors specified [32, 33]. Thompson randomized the next patient to treatment group A with probability $w_{n+1,A}(\mathscr{F}_n) = \Pr(p_B < p_A|\mathscr{F}_n)$, where p_i = treatment group i success probability, $i = A, B$, given the previous subjects' data ($\mathscr{F}_n$). Similarly, the next patient is randomized to treatment group B with probability $w_{n+1,B}(\mathscr{F}_n) = 1 - w_{n+1,A}(\mathscr{F}_n)$. Thall and Wathen's generalization of this approach estimates the posterior probabilities used to calculate the allocation weights using a model that includes treatment-covariate interactions [33]. This ensures that covariates are not ignored and separate trials for each covariate subgroup are not necessary. In their follow-up paper reviewing Bayesian Adaptive Randomization, the allocation probabilities were stated to be

[1]This is a less commonly used name for response-adaptive randomization.

highly variable and may favor the inferior treatment [22]. Instead, they suggest to randomize the next patient to treatment group A with the following probability:

$$w_{n+1,A}(\mathscr{F}_n) = \frac{\{P_{A>B}(\mathscr{F}_n)\}^c}{\{P_{A>B}(\mathscr{F}_n)\}^c + \{P_{A<B}(\mathscr{F}_n)\}^c},$$

where c is a positive tuning parameter. Similarly, the next patient should be randomized to treatment B with probability $w_{n+1,B}(\mathscr{F}_n) = 1 - w_{n+1,A}(\mathscr{F}_n)$. They suggest to use $c = n/2N$ to preserve power and reduce variability, where n is the current sample size and N is the total planned sample size. The authors discuss the 'drift' phenomenon and express the importance of adjusting for covariates to reduce negative effects on the adaptive randomization [22].

Using the posterior probability that $p_1 > p_2$ is an intuitively appealing quantity to use, however, theoretical justification requires a standardized treatment difference $\phi\left[\frac{\hat{p}_1-\hat{p}_2}{SE(\hat{p}_1-\hat{p}_2)}\right]$ that optimizes the allocation, where ϕ is a link function (e.g. probit) [34]. Instead, while a different approach than efficacy comparisons ($p_1 > p_2$), we choose to plug the estimated probabilities of success into the allocation equations (Eqns. 2.3 and 2.4) since they were designed to minimize treatment failures or maximize power, for example.

3.1.2 Treatment-by-Center Interactions in Clinical Trials

Clinical trials are typically performed at several centers in to recruit a sufficient number of patients in a relatively short amount of time. This is often difficult to do with a single site. An added benefit of multi-center trials is that they reflect natural differences in clinical practice between sites. A wider population of patients receive treatments in a variety of settings that are often expected to differ slightly due to patient, provider and clinic characteristics. Multi-center trials often use stratified randomization to balance allocation within each center at the design phase but rarely adjust for center in the analysis. In fact, Pocock et al. (2002) go as far to anecdotally claim that adjusting for centers in multi-center trials rarely makes a difference [35]. The added heterogeneity introduced in these trials may lead to a significant treatment-by-center interaction effect where the treatment effects vary across locations. Consequently, any inference made is subject to problems with interpretability of the treatment effects and generalizability of

the results to a broader population. This requires further investigation into the identification of group-specific effects [1, 36].

In 1998, the Food and Drug Administration published guidance for the pharmaceutical industry on the "Statistical Principles for Clinical Trials", developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) [36]. Included in the ICH-E9 statements are guidance for how to investigate the presence and cause of treatment-by center interactions. First, the main treatment effect should be fit in a model that allows center differences but does not include a treatment-by-center interaction. Including the interaction would use up degrees of freedom and reduce the power of the test for the main treatment effect if it were truly homogeneous. In the presence of true treatment heterogeneity, graphical displays and/or significance testing should be used to help explore characteristics that vary across sites. However, these hypothesis tests may be underpowered if the interaction effects are not powered for in advance; some have suggested using a higher significance level ($p = 0.10$) than convention for significance testing [1]. A significant interaction can be caused by the measurement scale of the data or outliers caused by extreme values of the response variable, neither of which are clinically relevant. Patient characteristics not accounted for in the analysis and/or variable conditions among the clinical centers could also be the source [1].

Without a specific explanation for these site differences, further trials may be needed when a significant qualitative interaction is observed. Alternatively, weighting centers in the analysis to justify the robustness of the treatment effect estimates may be necessary for a significant quantitative interaction [36]. In a typical fixed-effects analysis of variance (ANOVA) model, this weighting is performed by the inclusion of a treatment-by-center interaction effect. By including the interaction in the model, equal weight is given to the within-center effect estimates to produce an overall treatment effect where patient responses are weighted according to cluster size. Without the interaction term, within-center estimates are weighted by their precision, a function of the within-center sample sizes, and thus patient responses are equally weighted. Gallo (2000) recommends omitting the interaction term for the primary analysis in trials with unbalanced centers for a more precise overall treatment effect as discussed in the ICH-E9 guidance document [1, 12].

In multi-center trials, both the RA and CARA algorithms use all available data irrespective of site to calculate the allocation assignments. However, these existing methods do not account for

inter-site variability in treatment effectiveness. However, the effect of not accounting for potential variability between centers is unknown.

3.2 Methods

We considered a slightly more complex modeling approach under the Bayesian framework to overcome the anticipated issues in estimating many parameters as in previous research (e.g. noted in Rosenberger et al., 2001 [20]). Gelman et al. recommend using Bayesian inference to overcome probable issues with non-identifiability and separation that are commonplace when fitting frequentist logistic regression models [37]. The proposed CARA model (Eqn. 2.11) which accounts for clustering with a fixed site effect was not considered because all regression parameters are considered to be random in the Bayesian setting. SARA Model 1 (Eqn.2.13) was extended to a Bayesian GLMM by incorporating priors on the model parameters. Two models were fit: one model without an interaction and one model that includes a treatment-by-site interaction. The performance of these models will be assessed under scenarios that promote 'high' magnitude qualitative interactions based on varying effect sizes and inter-site standard deviations.

First, we extended SARA Model 1 (Eqn.2.13) to account for site-based response variability under the Bayesian framework.

3.2.1 SARA Model 2 - No Treatment-by-Site Interaction

$$logit(p_{ij}) = \beta_0 + \beta_1 T_i + g_j + \boldsymbol{X_i'\theta} \tag{3.1}$$

Here, β_0 is the overall intercept, β_1 is the treatment effect, T_i is the treatment group indicator ($T_i = 1$ if A, $T_i = 0$ if B), for patients $i = 1, \ldots, n$. Here, g_j is a random site effect for the j^{th} site, where $g_j \sim N(0, \sigma_g^2)$ $\forall j$, $j = 1, \ldots, k$ and k is the number of sites. The $\boldsymbol{X_i'\theta}$ is an optional set of covariates that can be incorporated into the model to adjust the estimated success probabilities, with corresponding slopes/effect sizes denoted by $\boldsymbol{\theta}$.

The priors for the regression coefficients and random effects are specified as $N(\mu, \sigma^2)$, where μ

is the mean and σ^2 is the variance. The prior on the overall intercept and the treatment effect is an informative normal: $\beta_l \sim N(0, 5),\ l = 0, 1$. In addition, the prior for the site random effects is specified as an informative normal: $g_j \sim N(0, 1)\ \forall j$. For a normal distribution, $N(\mu, \sigma^2)$, this implies that 99.7% of the probability mass for the prior distribution would fall within 3σ ($\sigma = 1$) of the mean ($\mu = 0$). In our experience, more informative priors with smaller variances were necessary on the random site effects to avoid issues with convergence as the effect for each site was small relative to traditional non-informative Gaussian priors, as demonstrated in a previous study by Longini et al. [4].

Traditionally, a noninformative or flat prior distribution is specified for parameters in a generalized linear model, such as a uniform prior density, so that the posterior distribution depends mostly on the data and not prior knowledge. Accordingly, the posterior mode of this noninformative density is the ML estimate of the parameter. However, with logistic regression, Gelman et al. suggest selecting priors under the assumption that effects will typically be relatively small and that coefficients are independent, thus more informative. An example is provided where separation occurs, particularly with binary predictors, resulting in unstable ML estimates. Using Bayesian inference, specifying a uniform prior distribution is shown to also fail in certain scenarios. Instead, the t family, such as the normal distribution (a special case when the degrees of freedom equals infinity), is recommended for regression coefficients because flat-tailed distributions allow for more robust inference. An approximate EM algorithm using a t prior distribution is also provided by the authors but not used for our purposes [37].

Spiegelhalter (2004) refers to these informative priors as 'skeptical' since there are doubts that large effects exist [38]. As such, the prior distribution is centered as zero with the scale parameter controlling the degree of skepticism to incorporate. Gelman et al. describes these priors as 'weakly informative' priors, where enough previous information is provided through the prior to keep the posterior distribution within reasonable bounds ('regularized') without dominating it with the prior specification. As the sample size approaches infinity, the posterior distribution will depend less on the prior specified. As an example, the authors consider a logistic regression model, where the predictors are either binary or standardized, and assume that an effect size will be no larger than 10 for most scenarios. This difference corresponds to a shift on the logit scale in a probability from $\text{logit}^{-1}(-5) = 0.01$ to $\text{logit}^{-1}(5) = 0.99$. It should be noted that a weakly informative prior may

affect inference when there is not enough data (i.e. small sample size) and skew it in a particular direction [37].

After the chains converge to the stationary distribution, the posterior medians of the model parameters were used to calculate the conditional probability of success for each subject to be randomized to a treatment group at a particular site. These estimated success probabilities $\hat{p}_{ij}^{t}$ for the t^{th} treatment, $t = 1, 2$, conditional on the site for the next patient were defined as follows:

$$\begin{aligned} \hat{p}_{ij}^{1} &= \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}, \\ \hat{p}_{ij}^{2} &= \frac{\exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}. \end{aligned} \tag{3.2}$$

Two methods were considered in the instance that a patient was being recruited from a particular site with no previous subjects. First, a hard lead-in criterion could require that all sites be observed prior to adapting, however, this may limit the benefits of adaptation. Alternatively, balanced allocation could be used to calculate the predicted probabilities for subjects coming from these 'new' sites. We decided to use a lead-in that required all sites be observed prior to beginning adaptation. While this is not a requirement for the Bayesian models, this ensures the results will be consistent with Chapter 2 and that a sufficient number of patients will be randomized prior to adapting.

Naturally, additional parameters can include treatment-by-covariate interactions [20, 25]. In addition to a random site effect, a random quantitative or qualitative treatment-by-site interaction, depending on the parameter template, was included to extend SARA Model 2 to allow the treatment effects to vary across sites.

3.2.2 SARA Model 3 - Treatment-by-Site Random Interaction

$$logit(p_{ijt}) = \beta_0 + \beta_1 T_i + g_j + d_{jt} + \boldsymbol{X_i'}\boldsymbol{\theta} \tag{3.3}$$

Here, we incorporated a random treatment-by-site interaction δ_{jt} corresponding to the j^{th} site and the t^{th} treatment, $t = 1, 2$, into equation (3.1). Priors were specified for the $2+2^k$ model parameters,

excluding the optional set of covariates. In addition to the previous priors for SARA Model 2, an additional prior was specified for the interaction term:

$$d_{jt} \sim N(0, \sigma_d^2) \ \forall j, t.$$

We expect the treatment-by-site effects to be small similar to the site effects. Accordingly, a 'weakly informative' normal prior is used for the treatment-by-site interaction effects: $d_{jt} \sim N(0,1) \ \forall j, t$. Sensitivity analyses will be performed with other specifications and larger variances to see how convergence and robustness are affected. In this model, the estimated success probabilities $\hat{p}_{ij}^t$ were calculated as follows:

$$\begin{aligned} \hat{p}_{ij}^1 &= \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \hat{d}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 + \hat{g}_j + \hat{d}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}, \\ \hat{p}_{ij}^2 &= \frac{\exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}{1 + \exp\{\hat{\beta}_0 + \hat{g}_j + \boldsymbol{X_i'}\hat{\boldsymbol{\theta}}\}}. \end{aligned} \tag{3.4}$$

These conditional probabilities of success will be substituted into the Optimal allocation equations to update the site-specific randomization probabilities.

3.3 Simulation Study

The methods of Chapter 2 were extended to incorporate a treatment-by-site interaction under the Bayesian framework. The simulation template is similar, however, 'low' and 'high' magnitudes of both qualitative and quantitative treatment-by-site interactions were considered by adjusting the amount of inter-site standard deviation (σ) appropriately. We conducted a brief simulation study to explore scenarios that promote more qualitative treatment-by-site interactions as defined by Gallo to select our parameter template. The number of sites ($k = 3, 6, 9$) and inter-site standard deviations in treatment response ($\sigma = 0, 0.025, 0.0375, 0.05, 0.075, 0.1, 0.125, 0.15$) were varied with a 30% treatment group 1 success probability ($p_1 = 0.3$) and several effect sizes ($\delta = p_1 - p_2 = 0, 0.1, 0.15$). The percentages of flipped success probabilities, defined as P(*at least one* $\hat{p}_1 < \hat{p}_2$)×100, were averaged across 1000 iterations to indicate how frequent qualitative interactions occurred.

The results of the interaction simulation are displayed in Table 3.1. As expected, the results show that as the effect size increases, there are less qualitative treatment-by-site interactions.

However, as the inter-site standard deviation increases, the number of qualitative interactions increases regardless of effect size. Accordingly, we selected smaller values of inter-site SD for Chapter 2, to limit the amount of qualitative interactions, and use larger values for Aim 2, to promote more qualitative interactions, to study the operating characteristics of the proposed methods.

Table 3.1: Percentage (%) of Qualitative Treatment-by-Site Interactions

			Inter-Site Standard Deviation (σ)							
k	p_1	δ	**0**	**0.025**	**0.0375**	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
3	0.3	0	0	89.5	89.5	89.5	89.5	89.5	89.5	89.5
		0.1	0	0.6	8.1	21.1	43.9	56.0	65.0	69.5
		0.15	0	0	0.6	4.4	21.1	36.9	49.4	55.9
6		0	0	98.5	98.5	98.5	98.5	98.5	98.5	98.5
		0.1	0	0.9	16.7	39.3	68.2	80.1	86.5	89.7
		0.15	0	0	0.9	9.5	39.3	58.6	73.3	79.8
9		0	0	99.9	99.9	99.9	99.9	99.9	99.9	99.9
		0.1	0	1.9	24.0	50.9	80.2	89.5	94.2	95.8
		0.15	0	0	1.9	14.4	50.9	74.1	85.2	89.5

Selecting large enough values for the inter-site standard deviation ($\sigma = 0, 0.05, 0.1, 0.15$) will result in differences in both the magnitude and direction of the treatment effectiveness across centers, indicating a qualitative interaction. Accordingly, we considered small ($p_1 = 0.3$), medium ($p_1 = 0.5$) and large ($p_1 = 0.7$) treatment group 1 success probabilities along with null ($\delta = 0$), small ($\delta = 0.1$) and medium ($\delta = 0.15$) effect sizes. The simulated Phase 2 multi-center trials were limited to a relatively small number of centers ($k = 3, 6$).

By specifying priors on the model parameters, the joint posterior distribution of the regression coefficients and random effects were estimated using the MCMC techniques coded in OpenBUGS through the BRugs interface in R. Based on our experiences, we specified 'weakly informative' prior distributions on the random site and treatment-by-site interaction effects. Accordingly, a sensitivity analysis was performed to explore different prior specifications, including slightly larger variances (less informative), for both SARA 2/SARA 3 models. The priors considered are displayed in Table 3.2. Prior C was the specification used for both the random site and interaction effects in our simulation study. In the sensitivity analysis, Bayesian GLMM models were fitted as subjects were randomized for trials with a small number of centers ($k = 3$), a small treatment group 1 success probability ($p_1 = 0.3$), a medium effect size ($\delta = 0.15$) and larger values for the inter-site standard

deviation ($\sigma = 0.1, 0.15$). Due to the anticipated significant computation time, we simulated only 10 trials and for each, estimated the model parameters using 2000 iterations for burn-in and 4000 iterations from two parallel chains with no thinning from the joint posterior distribution. Gelman and Rubin's convergence diagnostic was used as the criteria for evaluating convergence of the two Markov chains generated [39, 40]. We examined the robustness of the mean total number of treatment successes (SE), expected allocation proportions for treatment 1, empirical power and the average number of non-convergence warnings per trial to compare the priors being considered.

Table 3.2: Sensitivity Analysis: Priors Considered

Prior	Specification
A	$g_j \sim N(0, \sigma^2_{g_j}), \sigma^2_{g_j} \sim G(1,1)\ \forall j$
B	$g_j \sim U(-2, 2)$
C	$g_j \sim N(0, \sigma^2_{g_j}), \sigma^2_{g_j} = 1\ \forall j$
D	$g_j \sim N(0, \sigma^2_{g_j}), \sigma^2_{g_j} = 5\ \forall j$
E	$g_j \sim N(0, \sigma^2_{g_j}), \sigma^2_{g_j} \sim U(0,2)\ \forall j$

[1] The normal priors are specified as $N(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance.

The models were implemented into the adaptive design using the *BRugsFit* {BRugs} function [41]. The initial values must be specified for the random site and treatment-by-site interaction effects to ensure they are reasonable and within the bounds of the more informative priors that were used. Instead of using fixed starting values throughout all of the simulations, ML estimates were used to begin and then the posterior estimates were used for subsequent individuals randomized in an attempt to reduce computation time by allowing the chains to converge faster to the stationary distribution. For each scenario, a frequentist GLM was fit, which ignores the site effect, after the first subject is randomized following the lead-in (e.g. patient 21) to get ML estimates to initialize the starting values for the Markov chains. The estimates of the intercept and treatment effect were used as the initial values for the first chain, while 0.5 was added to the regression coefficients for the starting values for the second chain. For both chains, a fixed starting value of zero was used for the random site effects. For subsequent patients, the previous individual's posterior estimates

were used to initialize chain 1. For the second chain, 0.5 was added to the posterior estimates for both the intercept and treatment effect while 0.05 was added for the site effects. However, if the model did not converge, the last subject's estimates to converge were used for the starting values.

We estimated the posterior median of the model parameters using 2000 iterations for burn-in and 4000 iterations from two parallel Markov chains from the joint posterior distribution. The posterior median, which minimizes the expected posterior absolute loss function, was used for point estimates because it is equivalent to the mean for symmetric posterior densities. For asymmetric posteriors, the median lies between the mean and the mode (which considers only the maximum value) and is not skewed by extreme values [42]. Thinning was not performed because it has been shown to be unnecessary and inefficient, particularly due to the cost associated with increased computation time in simulations [43].

An artificial lead-in was used in the simulations to prevent estimator instability or convergence issues. This hard lead-in required that balanced allocation be performed for the first $m = 20$ subjects irrespective of the design. An advantage of the Bayesian framework is that it no longer requires lead-in criteria, specifically that both responses or that all sites be observed prior to adapting. Gelman and Rubin's convergence diagnostic $\hat{R}$, or the potential scale reduction factor (PSRF), was used as the criteria for evaluating convergence of the two Markov chains generated. This was implemented in R using the *gelman.diag* {coda} function [44]. Adaptation continued with the more strict cutoff of $\hat{R} < 1.1$ for all parameters [39, 40], otherwise, the previous patient's allocation weights were used until convergence was achieved [20]. In practice, visual assessment of trace plots should be examined for convergence in addition to looking at convergence diagnostic test statistics, however, this was not feasible in our simulation study.

To evaluate the proposed allocation designs, we use the operating characteristics established by Rosenberger et al. (2008), namely balance (allocation proportion), efficiency (power, type I error) and ethics (total number of treatment successes) [19]. We are particularly interested in the mean total number of successes/failures ($\pm$ standard error), empirical power ($P\{\text{reject } H_0 | H_1 \text{ true}\}$) and type I error rate ($P\{\text{reject } H_0 | H_0 \text{ true}\}$). In addition, we are interested in comparing the computation times for the proposed competing approaches. Based on our previous simulations in Chapter 2, each scenario for SARA Model 1 takes a significant amount of time to run. We expect that the Bayesian extensions will require even more computation time due to the added complexity of the

two models. Therefore, for all SARA models, we calculate the average simulation time (seconds) per subject to compare the effect of the increasing complexity of the models on computation times. Balanced allocation, RA and CARA randomization are not considered because these models are fit almost instantaneously for each patient. The simulation times represent the amount of time (seconds) to fit each model and update the allocation probabilities to randomize the subsequent patient in a trial. The times only account for the $n - 20$ patients adaptively randomized after the initial lead-in.

As a natural comparator, the performance of SARA Model 2 (Eqn. 3.1) was compared to SARA Model 1 (Eqn. 2.13), its equivalent under the frequentist framework in Chapter 2. In addition, SARA Model 3 (Eqn. 3.3) was compared to SARA Model 2. These methods were also compared to the CARA model (Eqn. 2.11), where we treat site as a fixed effect instead and do not include an interaction, as well as the existing designs, balanced and RA randomization (Eqn. 2.1). Due to our focus on accounting for treatment-by-site interactions in this chapter, we selected larger values of inter-site SD to promote the amount of qualitative interactions and thus had to extend the simulations for BC, RA, CARA and SARA Model 1. The balanced and RA designs do not account for site variability while both the CARA and SARA designs account for the clustering differently. We hypothesize that SARA Model 3, which accounts for inter-center variability in treatment responses and effectiveness, will reduce the variability in multi-center trials, especially in the presence of qualitative interactions, thus reducing improper patient allocation.

There were 1000 trials simulated for each scenario and the results aggregated. Due to our focus on ethics, we only considered Optimal allocation in Aim 2. For a single trial scenario, the following metrics were stored: sample size, number of patients assigned to each treatment, number of successes for each treatment, total number of successes and the number of convergence issues (indicates previous allocation weights were used until the model converged with additional data). After all patients were recruited in a trial, several methods were considered to test the treatment group, site and treatment-by-site interaction effects.

To be consistent with Chapter 2, the proportions of successes were compared using a Pearson's chi-squared test to assess the treatment effect and an indication of whether the null hypothesis ($H_0 : p_1 = p_2$) was rejected was retained to calculate the operating characteristics of the method. The expected cell counts were checked (> 5) to ensure the assumptions of a chi-squared test were

satisfied, otherwise, a Fisher's Exact test was performed. This approach ignores site when testing the treatment effect as this is convention when analyzing multi-center trials. In addition to a naive chi-squared test, a chi-squared test from the GLMM, which incorporates random site effects to account for inter-center variability in treatment responses, was performed. A random treatment-by-site interaction effect was not included in the GLMM for SARA 3 model due to concerns of non-convergence.

The frequentist GLMMs were fit in the *glmer* function using ML estimation with an Laplace approximation (quadrature nodes = 1) of the likelihood for the mixed model, which is an integral over the random effect space. The variance-covariance matrices of the random effects were estimated using Powell's BOBYQA method, a nonlinear optimizer [30]. A separate variance component was modeled for each random effect, the site and treatment-by-site interaction effects, separately. A likelihood-ratio test (LRT) was used to test the random site and treatment-by-site interaction effects for models that successfully converged, otherwise, the LRT was not performed. To test the random site effect, a GLM was fit using the GLM function for the null model. The frequency of statistically significant site and treatment-by-site interaction effects were used to confirm the presence of quantitative and qualitative interactions for each scenario and will not be reported here.

The LRT statistic [10] used to compare two models is shown to be the difference in the deviances as follows:

$$\begin{aligned} D &= -2\ln\left(\frac{\text{likelihood of } M_0}{\text{likelihood of } M_1}\right) \\ &= -2\left[\mathcal{L}_0 - \mathcal{L}_1\right] \\ &= -2\left[\mathcal{L}_0 - \mathcal{L}_s\right] - \left\{-2\left[\mathcal{L}_1 - \mathcal{L}_s\right]\right\} \\ &= D_0 - D_1 \sim \chi^s_{(df_1 - df_0)} \end{aligned} \tag{3.5}$$

where $\mathcal{L}_0$ is the maximized log likelihood for the null model, $\mathcal{L}_1$ is the maximized log likelihood for the alternative model, $\mathcal{L}_s$ is the maximized log likelihood for the saturated model and M_0 is a simpler model nested within M_1. The difference in deviances follows approximately a chi-squared null distribution with degrees of freedom equal to the difference in the number of parameters between the two models. A convergence indicator was used to determine the proportion of trials

where the LRT was performed.

A Bayesian approach was also considered to test the treatment, site and treatment-by-site interaction effects but not reported here as frequentist inferences are typically used in practice for adaptive designs. A z-test statistic was calculated using the posterior mean and variance of the treatment group to test the treatment effect $\hat{\beta}_1 = 0$. The posterior mean and variance estimates were calculated as follows:

$$\begin{aligned} \hat{E}(\theta_i) = \hat{\theta}_i &= \sum_{j=1}^{m} \frac{\theta_{ij}^*}{m}, \\ \hat{Var}(\theta_i) &= \frac{1}{m-1} \sum_{j=1}^{m} (\theta_{ij}^* - \hat{\theta}_i)^2, \end{aligned} \tag{3.6}$$

where θ_{ij}^* is the observed estimate for the i^{th} parameter and the j^{th} iteration for $j = 1, \ldots, m$ draws from the posterior distribution. An LRT was used to perform hypothesis testing on the random site and treatment-by-site interaction effects using the difference in the posterior mean of the deviance from the models, regardless of whether the posterior estimates converged after all subjects had been randomized. To account for the difference in the additional parameters estimated (regressions coefficients, variances for priors) between the nested models, the effective degrees of freedom (pD) from each model was used to calculate the estimated difference in parameters ($df_1 - df_0$). The effective degrees of freedom, $pD = \bar{D} - \hat{D}$, is the difference in the posterior mean of the deviance ($\bar{D}$) and the point estimate of the deviance ($\hat{D}$) when plugging in the posterior means of the parameters [45].

3.3.1 Simulation Algorithm

The basic algorithm for our simulation study is as follows.

1. Hard lead-in for the first m patients, where individuals are allocated equally between the two treatment groups.

2. Let $\mathscr{F}_i$ be accrued information which includes treatment assignments, observed patients outcomes and center information. Use $\mathscr{F}_i$ for patients $i = 1, \ldots, m, m+1, \ldots, n$ to begin adaptation, starting with patient $m + 1$.

3. Using the center for patient $m+1$, calculate the conditional probability of success $\hat{p}_{ij}^t$ for the t^{th} treatment using the SARA 2 (Eqns. 3.1, 3.2) or SARA 3 model (Eqns. 3.3, 3.4).

4. Using the estimated success probabilities $\hat{p}_i^1, \hat{p}_i^2$, update the allocation weights W_1, W_2 and assign the next patient to a treatment.

$$W_1(\mathscr{F}_i) = \sqrt{\hat{p}_i^1}/\sqrt{\hat{p}_i^1 + \hat{p}_i^2},\ W_2(\mathscr{F}_i) = 1 - W_1.$$

5. Repeat Steps 2-4 to randomize patients $m+2, \ldots, n$.

For a single scenario, this algorithm is repeated for $N = 1000$ trials.

3.4 Results

In Sections 3.4.1 - 3.4.6, we compare our two proposed approaches from Chapter 2, which account for clustering through either a fixed or random site effect, and an extension under the Bayesian framework with the two existing approaches, balanced randomization and response-adaptive randomization. We investigate these methods in the presence of qualitative interactions. In Section 3.4.7, we then investigate the performance of our SARA model which incorporates a random treatment-by-site interaction to account for variability in treatment effectiveness.

3.4.1 Scenario 1: No Interaction, 3 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 1 are displayed in Table 3.3. When the effect size is null $\delta = p_1 - p_2 = 0$ and there is no inter-site standard deviation ($\sigma = 0$), the five methods are comparable. However, as σ increases, there is an increase in the total number of successes for the adaptive designs, with the improvement slightly lower for both SARA designs. The SARA 1 design has lower variability than the other adaptive approaches and matches the variability from the balanced case. When the effect size is increased to 10% ($p_1 = 0.3, p_2 = 0.2$), the adaptive designs result in more treatment successes than the balanced case, particularly as σ increases. However, the improvement is lower for the SARA approaches compared to the RA/CARA designs. Both SARA designs are similar in that they have lower variability than the other adaptive

approaches and they are similar to the variability from the balanced randomization. For an effect size of 15%, the pattern is similar to the medium effect size. The adaptive designs result in more treatment successes than the BC but the improvement is lower for the SARA approaches. The SARA designs perform similarly in terms of treatment successes and variability, both of which have lower variability than the RA/CARA designs. Overall, as σ increases, there is an increase in expected treatment successes along with a greater increase in the variability in treatment successes for all methods.

Table 3.3: Scenario 1 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	176.1 (10.7)
		(n=587)	0.05	175.5 (16.3)	176.1 (16.6)	176.0 (16.4)	176.0 (16.3)	176.4 (16.6)
			0.1	175.2 (26.3)	177.5 (27.1)	177.2 (26.6)	177.0 (26.2)	177.3 (26.9)
			0.15	176.0 (36.5)	181.6 (37.8)	180.9 (37.1)	179.9 (36.5)	179.7 (37.6)
		0.1	0	146.6 (10.6)	151.1 (14.3)	150.7 (12.5)	149.5 (10.7)	150.0 (10.2)
		(n=587)	0.05	146.0 (16.1)	151.0 (19.6)	150.7 (17.3)	149.5 (16.3)	150.2 (16.3)
			0.1	146.0 (25.7)	153.6 (28.8)	152.6 (27.1)	150.8 (25.9)	151.7 (26.5)
			0.15	148.0 (35.6)	158.7 (39.9)	157.2 (38.3)	155.0 (36.3)	155.7 (37.0)
		0.15	0	54.2 (6.7)	59.1 (9.1)	58.4 (7.8)	57.2 (6.8)	57.5 (6.7)
		(n=241)	0.05	54.4 (7.9)	59.6 (10.4)	58.9 (9.4)	57.7 (8.1)	57.6 (8.2)
			0.1	55.1 (11.4)	61.0 (13.4)	60.2 (12.4)	58.8 (11.5)	58.6 (11.6)
			0.15	56.7 (15.0)	63.3 (16.9)	62.2 (16.3)	60.7 (15.5)	60.5 (15.4)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

The allocation proportions are displayed in Table 3.4. For a null effect size, the five methods have comparable allocation proportions. The SARA designs perform similarly in that they have lower variability in allocating than the other adaptive designs, while RA randomization has the highest variability. When $\delta = 0.10$, the adaptive designs have a larger allocation ratio relative to the balanced case, though the improvement is lower for the SARA designs. This explains the increase in expected treatment successes in the adaptive methods with a slightly lower increase for both

SARA approaches. The SARA designs perform similarly with respect to variability in allocating and are less variable than the RA/CARA designs in all cases. For $\delta = 0.15$, the adaptive designs have notably larger allocation ratios than balanced randomization, particularly RA randomization followed by the CARA design. The improvement is lower for the SARA approaches, though SARA 2 results in slightly higher allocation ratios than SARA 1. The SARA designs have lower variability compared to the other adaptive designs. The RA method has the highest variability in allocation ratios across all effect sizes. Overall, there appears to be a marginal increase in the variability in allocation proportion with as the amount of inter-site standard deviation increases.

Table 3.4: Scenario 1 - E$[n_1/(n_1 + n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.3	0	0	0.50 (0.02)	0.50 (0.13)	0.50 (0.09)	0.50 (0.03)	0.50 (0.03)
		(n=587)	0.05	0.50 (0.02)	0.49 (0.13)	0.50 (0.09)	0.50 (0.04)	0.50 (0.03)
			0.1	0.50 (0.02)	0.49 (0.14)	0.49 (0.09)	0.50 (0.05)	0.50 (0.05)
			0.15	0.50 (0.02)	0.49 (0.16)	0.49 (0.11)	0.50 (0.07)	0.50 (0.06)
		0.1	0	0.50 (0.02)	0.58 (0.17)	0.57 (0.12)	0.55 (0.03)	0.56 (0.03)
		(n=587)	0.05	0.50 (0.02)	0.57 (0.18)	0.57 (0.12)	0.55 (0.05)	0.55 (0.04)
			0.1	0.50 (0.02)	0.58 (0.19)	0.57 (0.13)	0.55 (0.06)	0.56 (0.06)
			0.15	0.50 (0.02)	0.57 (0.21)	0.57 (0.14)	0.55 (0.08)	0.55 (0.08)
		0.15	0	0.50 (0.03)	0.63 (0.19)	0.62 (0.13)	0.58 (0.05)	0.59 (0.05)
		(n=241)	0.05	0.50 (0.03)	0.63 (0.19)	0.61 (0.14)	0.58 (0.06)	0.59 (0.06)
			0.1	0.50 (0.03)	0.63 (0.20)	0.61 (0.14)	0.58 (0.07)	0.59 (0.07)
			0.15	0.50 (0.03)	0.63 (0.21)	0.61 (0.14)	0.58 (0.09)	0.59 (0.09)

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.5. For $\delta = 0$, all five methods perform similarly where there is an inflation in the type I error rate as σ increases. When there is no inter-site variability, SARA 2 appears to control the type I error rate slightly better than the other designs. However, as σ increases, both RA/SARA 2 approaches control the type I

error better than the other designs, with more improvement for RA randomization. For a small and medium effect size, there is a notable loss of power for RA allocation relative to the other adaptive designs. The CARA/SARA 1/SARA 2 designs gain back most of the power lost with the RA randomization. Specifically, the SARA designs outperform the balanced case for most, if not all, values of σ. Specifically, SARA 2 has the highest power compared to all other methods in all cases.

We now discuss the results from the GLM/GLMM approach, where either a fixed (CARA) or random (SARA 1/SARA 2) site effect is incorporated into the modeling to account for clustering. For a null effect size, the type I error results are comparable to the Naive approach, where SARA 2 has the best control of the type I error rate. For small and medium effect sizes, the results are similar to the naive approach, where RA randomization has the greatest power loss relative to the BC. The other adaptive designs gain back most of the power lost, particularly the SARA approaches. SARA 2 has a notable increase in power relative to all other methods, including the balanced case.

Table 3.5: Scenario 1 - Empirical Type I Error and Power

				Naive					GLM/GLMM		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
3	0.3	0	0	0.06	0.06	0.07	0.06	0.05	0.06	0.06	0.05
		(n=587)	0.05	0.20	0.19	0.21	0.20	0.17	0.20	0.20	0.17
			0.1	0.41	0.39	0.43	0.43	0.39	0.42	0.43	0.40
			0.15	0.57	0.50	0.56	0.57	0.53	0.56	0.59	0.53
		0.1	0	0.78	0.68	0.76	0.79	0.84	0.73	0.79	0.84
		(n=587)	0.05	0.68	0.57	0.67	0.68	0.74	0.63	0.68	0.74
			0.1	0.62	0.51	0.61	0.62	0.68	0.58	0.62	0.68
			0.15	0.63	0.51	0.62	0.64	0.68	0.59	0.65	0.69
		0.15	0	0.80	0.61	0.78	0.81	0.81	0.70	0.80	0.80
		(n=241)	0.05	0.73	0.54	0.69	0.72	0.75	0.63	0.72	0.75
			0.1	0.65	0.46	0.62	0.65	0.70	0.55	0.64	0.70
			0.15	0.62	0.43	0.60	0.62	0.65	0.53	0.62	0.65

[1] Trial Scenario: 3 Sites, $p_1 = 0.3$, Optimal Allocation

3.4.2 Scenario 2: No Interaction, 6 Sites, $p_1 = 0.3$

The mean total number of treatment successes for Scenario 2 are displayed in Table 3.6. When the effect size is null and smaller values of inter-site standard deviation ($\sigma = 0, 0.05$), the five methods are comparable with respect to the expected number of treatment successes and variability, with slightly lower variability for SARA 2. For larger values of σ, the adaptive designs have a modest improvement in treatment successes compared to the balanced case. SARA 2 has lower variability than all other methods while RA randomization has the highest variability. When the effect size is increased to 10%, the adaptive designs all produce greater expected treatment successes than the balanced case, however, RA randomization has slightly more treatment successes than the other methods. The CARA/SARA designs have lower variability in treatment successes than RA randomization, with more improvement for the SARA approaches. Specifically, SARA 2 has the lowest variability compared to all other designs, including the BC, particularly as σ increases. For

an effect size of 15%, the adaptive designs result in more treatment successes than the balanced case, with more improvement for RA randomization. The variability in treatment successes for the RA method is higher than all other approaches. When there is no inter-site standard deviation, the CARA/SARA designs perform similarly in terms of treatment success and variability. However, as σ increases, SARA 2 results in slightly more treatment successes yet maintains similar variability compared to CARA/SARA 1, while SARA 1 has slightly lower improvement in expected successes. Overall, as the inter-site standard deviation increases, there is an increase in expected treatment successes along with a greater increase in the variability in treatment successes for all methods, particularly for null and small effect sizes.

Table 3.6: Scenario 2 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
6	0.3	0	0	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	175.9 (11.2)	176.1 (10.7)
		(n=587)	0.05	176.3 (14.1)	176.6 (14.1)	176.5 (14.2)	176.5 (14.1)	176.0 (13.5)
			0.1	176.3 (20.8)	177.4 (21.1)	177.2 (20.8)	177.2 (20.7)	177.1 (19.8)
			0.15	176.9 (27.7)	179.1 (28.6)	178.7 (27.8)	178.6 (27.7)	178.9 (26.7)
		0.1	0	146.6 (10.6)	151.1 (14.3)	149.8 (10.9)	149.5 (10.7)	150.1 (10.2)
		(n=587)	0.05	146.9 (13.6)	151.6 (16.9)	150.4 (14.4)	150.0 (13.8)	149.8 (13.2)
			0.1	147.3 (20.2)	152.6 (23.2)	151.4 (21.1)	151.0 (20.5)	151.6 (19.5)
			0.15	149.4 (27.0)	155.5 (30.1)	154.1 (27.9)	153.8 (27.7)	154.3 (25.9)
		0.15	0	54.2 (6.7)	59.1 (9.1)	57.5 (7.0)	57.2 (6.9)	57.6 (6.8)
		(n=241)	0.05	54.1 (7.2)	59.1 (9.9)	57.4 (7.8)	57.1 (7.7)	58.0 (7.7)
			0.1	54.6 (9.2)	59.9 (11.5)	58.2 (9.8)	57.9 (9.8)	58.5 (9.7)
			0.15	56.0 (11.4)	61.0 (13.5)	59.3 (12.0)	59.1 (12.0)	59.9 (12.0)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

The allocation proportions are displayed in Table 3.7. For a null effect size, the five approaches have comparable allocation proportions. The CARA/SARA approaches have lower variability in allocation ratios than RA randomization, with SARA 2 closer to the BC for smaller σ. When

$\delta = 0.10$, the adaptive designs have larger allocation ratios relative to the BC, with the improvement lower for the CARA/SARA 1/SARA 2 designs than that of the RA method. However, a trend emerges where SARA 2 results in slightly more adaptation than SARA 1. The CARA/SARA 1/SARA 2 designs are less variable than the RA approach, though there is less improvement for CARA. For an effect size of 15%, the pattern is similar to a small effect size, where the adaptive designs result in slightly more adaptation, with less improvement for the CARA/SARA 1/SARA 2 designs. The CARA/SARA 2 approaches have slightly higher allocation ratios but the SARA 1/SARA 2 designs maintain lower variability. The RA method has more than double the variability in allocation ratios across all effect sizes. The amount of inter-site standard deviation does not affect the allocation ratios, though there is a marginal increase in the variability in allocation proportions as σ increases.

Table 3.7: Scenario 2 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
6	0.3	0	0	0.50 (0.02)	0.50 (0.13)	0.50 (0.06)	0.50 (0.04)	0.50 (0.03)
		(n=587)	0.05	0.50 (0.02)	0.49 (0.12)	0.50 (0.04)	0.50 (0.04)	0.50 (0.03)
			0.1	0.50 (0.02)	0.50 (0.14)	0.50 (0.05)	0.50 (0.04)	0.50 (0.04)
			0.15	0.50 (0.02)	0.49 (0.15)	0.50 (0.06)	0.50 (0.05)	0.50 (0.05)
		0.1	0	0.50 (0.02)	0.58 (0.17)	0.55 (0.06)	0.55 (0.03)	0.56 (0.03)
		(n=587)	0.05	0.50 (0.02)	0.57 (0.17)	0.55 (0.06)	0.55 (0.04)	0.55 (0.04)
			0.1	0.50 (0.02)	0.57 (0.17)	0.55 (0.07)	0.55 (0.05)	0.56 (0.05)
			0.15	0.50 (0.02)	0.56 (0.18)	0.55 (0.07)	0.55 (0.06)	0.55 (0.06)
		0.15	0	0.50 (0.03)	0.63 (0.19)	0.59 (0.07)	0.58 (0.06)	0.59 (0.05)
		(n=241)	0.05	0.50 (0.03)	0.63 (0.18)	0.59 (0.07)	0.58 (0.06)	0.59 (0.06)
			0.1	0.50 (0.03)	0.63 (0.19)	0.59 (0.08)	0.58 (0.07)	0.59 (0.07)
			0.15	0.50 (0.03)	0.62 (0.19)	0.58 (0.08)	0.57 (0.07)	0.58 (0.07)

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.8. For $\delta = 0$, the the type I error rate increases as the amount of inter-site variability increases. For smaller σ, the five

methods perform similarly, with slight improvement for SARA 2. However, as σ increases, the RA design controls the type I error rate better than all of the other approaches which perform similarly. For $\delta = 0.10$, the RA design has a notable loss of power compared to all other methods. SARA 2 has higher power than balanced allocation in all cases while the CARA/SARA 1 designs perform similarly and gain back most of the power lost. For $\delta = 0.15$, the CARA approach outperforms all other methods when there is no inter-site variability. However, as σ increases, the SARA designs perform similarly to balanced randomization, with slight improvement for SARA 2. In all cases, RA randomization has the lowest power compared to the other allocation designs. Overall, as σ increases, there is a notable loss of power for all five approaches.

For a null effect size under the GLM/GLMM approach, SARA 2 has the smallest type I error rate that is slightly below 0.05 when there is no inter-site variability. However, the other adaptive designs perform marginally better with increasing σ, particularly RA randomization. For $\delta = 0.10$, the results are comparable to the naive approach where SARA 2 has the highest power and the CARA/SARA 1 designs gain back most of the power lost with the RA method. For $\delta = 0.15$, the CARA/SARA methods perform similarly to balanced allocation, but improvement in is more notable for SARA 2.

Table 3.8: Scenario 2 - Empirical Type I Error and Power

				Naive					GLM/GLMM		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
6	0.3	0	0	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.04
		(n=587)	0.05	0.12	0.11	0.12	0.12	0.11	0.13	0.12	0.11
			0.1	0.28	0.24	0.26	0.27	0.27	0.27	0.27	0.29
			0.15	0.43	0.40	0.42	0.43	0.42	0.44	0.43	0.44
		0.1	0	0.78	0.68	0.78	0.79	0.84	0.78	0.78	0.84
		(n=587)	0.05	0.73	0.64	0.73	0.72	0.75	0.72	0.72	0.75
			0.1	0.65	0.58	0.65	0.65	0.69	0.65	0.66	0.70
			0.15	0.64	0.54	0.63	0.63	0.67	0.64	0.64	0.67
		0.15	0	0.80	0.61	0.81	0.80	0.80	0.79	0.80	0.80
		(n=241)	0.05	0.76	0.58	0.76	0.76	0.76	0.74	0.75	0.76
			0.1	0.68	0.50	0.68	0.68	0.70	0.68	0.68	0.71
			0.15	0.62	0.47	0.60	0.62	0.63	0.62	0.62	0.65

[1] Trial Scenario: 6 Sites, $p_1 = 0.3$, Optimal Allocation

3.4.3 Scenario 3: No Interaction, 3 Sites, $p_1 = 0.5$

The mean total number of treatment successes for Scenario 3 are displayed in Table 3.9. When the effect size is null and there is no inter-site variability, the five methods perform similarly with respect to both the expected total number of treatment successes and the variability in successes. However, for larger values of σ, there is an increase in treatment successes for the RA/CARA/SARA 1 designs and a slight decrease in successes for SARA 2, relative to the BC. All methods have similar variability in treatment successes. When the effect size is increased to 10%, the adaptive methods increase the expected treatment successes relative to the BC and they maintain similar variability. As the amount of inter-site standard deviation increases, there is slightly less improvement in treatment successes for SARA 2. For $\delta = 0.15$, the pattern is similar to a small effect size, where the adaptive approaches result in more treatment successes relative to balanced randomization. When $\sigma = 0$, SARA 2 results in slightly more treatment successes than the other adaptive designs

but maintains similar variability in expected successes. For $\sigma = 0.05, 0.1$, the adaptive designs perform similarly in terms of treatment successes, but SARA 2 has slightly lower variability. When $\sigma = 0.15$, the SARA designs result in slightly lower expected treatment success relative to the RA/CARA approaches, but SARA 3 has lower variability than all other methods, including BC. Overall, there is an increase in the expected streamline successes along with a greater increase in the variability in treatment successes as the amount of inter-site variability increases.

Table 3.9: Scenario 3 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.5	0	0	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.8 (13.9)
		(n=775)	0.05	389.1 (20.9)	389.4 (20.9)	389.4 (20.9)	389.4 (20.9)	387.3 (21.0)
			0.1	389.7 (34.1)	391.0 (34.2)	391.0 (33.9)	391.0 (33.9)	388.1 (33.9)
			0.15	390.3 (48.6)	393.2 (48.7)	393.3 (48.3)	393.2 (48.2)	389.8 (48.4)
		0.1	0	348.9 (14.1)	351.2 (14.1)	351.2 (14.1)	351.0 (14.0)	351.4 (14.0)
		(n=775)	0.05	349.3 (21.1)	352.1 (21.2)	352.2 (21.2)	352.0 (21.1)	350.9 (21.4)
			0.1	350.6 (34.4)	354.5 (34.8)	354.5 (34.4)	354.3 (34.4)	351.3 (34.4)
			0.15	351.7 (49.0)	357.8 (49.7)	357.8 (49.2)	357.3 (48.8)	353.3 (48.8)
		0.15	0	143.9 (9.1)	146.3 (9.4)	146.3 (9.2)	146.1 (9.1)	147.0 (9.0)
		(n=339)	0.05	143.8 (11.6)	146.4 (11.9)	146.4 (11.7)	146.1 (11.5)	146.3 (10.9)
			0.1	143.7 (16.9)	146.9 (17.3)	146.9 (17.0)	146.6 (16.9)	146.5 (16.0)
			0.15	144.0 (22.9)	148.5 (23.3)	148.4 (23.2)	147.8 (23.0)	147.5 (22.3)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

The allocation proportions are displayed in Table 3.10. For a null effect size, all five approaches have the same allocation proportion. The two SARA designs have lower variability similar to the BC compared to the other adaptive approaches. When $\delta = 0.10$, the adaptive methods all share a larger allocation ratio relative to the BC but the SARA approaches are maintaining less variability than RA/CARA. For an effect size of 15%, the adaptive methods have higher allocation ratios, though the improvement is lower for SARA 1/SARA 2. The SARA designs have lower variability

compared to the other adaptive approaches, with slightly more improvement for SARA 2 in some cases. The amount of inter-site standard deviation does not affect the allocation proportions and only has the slightest of effects in the variability in allocation.

Table 3.10: Scenario 3 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.5	0	0	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
		(n=775)	0.05	0.50 (0.02)	0.50 (0.04)	0.50 (0.04)	0.50 (0.02)	0.50 (0.02)
			0.1	0.50 (0.02)	0.50 (0.05)	0.50 (0.04)	0.50 (0.03)	0.50 (0.03)
			0.15	0.50 (0.02)	0.50 (0.06)	0.50 (0.05)	0.50 (0.04)	0.50 (0.04)
		0.1	0	0.50 (0.02)	0.53 (0.04)	0.53 (0.04)	0.53 (0.02)	0.53 (0.02)
		(n=775)	0.05	0.50 (0.02)	0.53 (0.05)	0.53 (0.04)	0.53 (0.03)	0.53 (0.03)
			0.1	0.50 (0.02)	0.53 (0.05)	0.53 (0.04)	0.53 (0.03)	0.53 (0.03)
			0.15	0.50 (0.02)	0.53 (0.07)	0.53 (0.06)	0.53 (0.04)	0.53 (0.04)
		0.15	0	0.50 (0.03)	0.55 (0.07)	0.55 (0.06)	0.54 (0.04)	0.54 (0.03)
		(n=339)	0.05	0.50 (0.03)	0.55 (0.08)	0.55 (0.06)	0.54 (0.04)	0.54 (0.04)
			0.1	0.50 (0.03)	0.55 (0.08)	0.55 (0.06)	0.54 (0.04)	0.55 (0.04)
			0.15	0.50 (0.03)	0.55 (0.10)	0.55 (0.08)	0.55 (0.06)	0.55 (0.05)

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.11. For $\delta = 0$, the five approaches perform similarly where the type I error rate increases as the amount of inter-site variability increases. SARA 2 appears to control the type I error rate slightly better than the other methods. For small and medium effect sizes, there is a loss of power for all five methods with the presence of inter-site variability. For $\delta = 0.10$, five approaches are comparable, with slightly lower power for the the BC/RA approaches when there is no inter-site variability and slightly higher power for SARA 2 as σ increases. For $\delta = 0.15$, the pattern is similar to a small effect size but more pronounced. When there is no inter-site variability, the BC/CARA/SARA approaches perform comparably, whereas the RA design has a loss of power relative to the other designs. As

σ increases, the CARA/SARA approaches gain back the power lost with RA randomization, with SARA 2 outperforming all other methods.

For a null effect size, the type I error results from the GLM/GLMM are comparable to the naive approach. For $\delta = 0.10$, the SARA approaches have slightly higher power when $\sigma = 0$ while RA randomization has the lowest power as σ increases. For an effect size of 15%, the results for empirical power are comparable to the naive approach, where SARA 2 outperforms all other methods with increasing amounts of inter-site standard deviation.

Table 3.11: Scenario 3 - Empirical Type I Error and Power

				Naive					**GLM/GLMM**		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
3	0.5	0	0	0.06	0.05	0.06	0.05	0.04	0.06	0.05	0.04
		(n=775)	0.05	0.20	0.20	0.21	0.20	0.20	0.20	0.20	0.19
			0.1	0.45	0.45	0.45	0.45	0.42	0.45	0.46	0.42
			0.15	0.59	0.58	0.58	0.58	0.57	0.58	0.58	0.59
		0.1	0	0.78	0.78	0.79	0.79	0.79	0.78	0.79	0.79
		(n=775)	0.05	0.72	0.71	0.72	0.71	0.70	0.71	0.72	0.70
			0.1	0.66	0.65	0.65	0.65	0.66	0.66	0.66	0.66
			0.15	0.69	0.68	0.68	0.68	0.69	0.69	0.69	0.70
		0.15	0	0.80	0.78	0.80	0.80	0.80	0.79	0.80	0.80
		(n=339)	0.05	0.74	0.72	0.74	0.75	0.77	0.73	0.75	0.77
			0.1	0.69	0.66	0.68	0.68	0.71	0.66	0.68	0.71
			0.15	0.65	0.62	0.65	0.65	0.68	0.65	0.66	0.69

[1] Trial Scenario: 3 Sites, $p_1 = 0.5$, Optimal Allocation

3.4.4 Scenario 4: No Interaction, 6 Sites, $p_1 = 0.5$

The mean total number of treatment successes for Scenario 4 are displayed in Table 3.12. When the effect size is null, the five approaches are comparable, with a slight increase in the average total number of treatment successes for the adaptive designs as the amount of inter-site standard deviation increases. There is a slight decrease in variability for the RA/CARA/SARA

1 designs relative to the balanced case. For $\delta = 0.10$, the adaptive designs have more expected successes than the balanced case, however, the improvement is slightly lower for SARA 2 compared to the RA/CARA/SARA 1 designs, particularly with increasing inter-site standard deviation. The variability in treatment successes is comparable for all of the methods. For an effect size of 15%, the adaptive designs result in more treatment successes than the BC, with slightly more improvement for SARA 2 when there is no inter-site variability. The variability in expected treatment successes is similar for all five methods. Overall, the variability in treatment successes increases as σ increases.

Table 3.12: Scenario 4 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
6	0.5	0	0	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.5 (13.9)	387.8 (13.9)
		(n=775)	0.05	387.0 (17.6)	387.2 (17.4)	387.2 (17.5)	387.2 (17.5)	387.5 (17.9)
			0.1	386.8 (25.9)	387.6 (25.7)	387.6 (25.7)	387.6 (25.7)	387.4 (26.1)
			0.15	386.8 (35.5)	388.4 (35.3)	388.4 (35.3)	388.3 (35.3)	387.7 (35.9)
		0.1	0	348.9 (14.1)	351.1 (14.0)	351.2 (14.1)	351.0 (14.0)	351.4 (14.0)
		(n=775)	0.05	347.9 (18.1)	350.6 (18.1)	350.5 (18.4)	350.5 (18.0)	350.7 (17.3)
			0.1	348.1 (26.4)	351.9 (26.9)	351.7 (26.2)	351.5 (26.2)	351.1 (25.9)
			0.15	348.3 (35.8)	353.2 (36.3)	352.8 (35.5)	352.6 (35.5)	351.7 (36.1)
		0.15	0	143.9 (9.1)	146.2 (9.1)	146.3 (9.4)	146.1 (9.1)	147.0 (9.0)
		(n=339)	0.05	143.8 (10.2)	146.2 (10.3)	146.6 (10.7)	146.1 (10.3)	146.3 (10.6)
			0.1	144.1 (13.3)	147.1 (13.4)	146.8 (13.3)	146.7 (13.3)	146.7 (13.4)
			0.15	144.2 (17.1)	147.6 (17.3)	147.4 (17.3)	147.2 (17.2)	147.3 (17.3)

[1] Trial Scenario: 6 Sites, $p_1 = 0.5$, Optimal Allocation

The allocation proportions are displayed in Table 3.13. For $\delta = 0$, the five methods have comparable allocation proportions. The SARA 1/SARA 2 approaches have similarly lower variability in allocation while the RA/CARA designs have comparably higher variability. When the effect size is increased to 10%, the adaptive designs result in larger allocation ratios compared to the balanced case. The SARA designs have lower variability than the RA/CARA methods, with marginal

improvement for SARA 1 in some cases. For $\delta = 0.15$, the pattern is similar to a small effect size, however, the improvement in allocation for the adaptive approaches is marginally lower for SARA 1/SARA 2 compared to the RA/CARA. The SARA 1/SARA 2 approaches have similarly lower variability in allocation proportions compared to the other adaptive designs. The amount of inter-site standard deviation does not affect the allocation proportions or the variability in allocation.

Table 3.13: Scenario 4 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
6	0.5	0	0	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
		(n=775)	0.05	0.50 (0.02)	0.50 (0.03)	0.50 (0.05)	0.50 (0.02)	0.50 (0.02)
			0.1	0.50 (0.02)	0.50 (0.05)	0.50 (0.03)	0.50 (0.03)	0.50 (0.03)
			0.15	0.50 (0.02)	0.50 (0.05)	0.50 (0.03)	0.50 (0.03)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.53 (0.02)	0.53 (0.04)	0.53 (0.02)	0.53 (0.02)
		(n=775)	0.05	0.50 (0.02)	0.53 (0.03)	0.53 (0.06)	0.53 (0.02)	0.53 (0.03)
			0.1	0.50 (0.02)	0.53 (0.07)	0.53 (0.04)	0.53 (0.03)	0.53 (0.03)
			0.15	0.50 (0.02)	0.53 (0.07)	0.53 (0.04)	0.53 (0.03)	0.53 (0.04)
		0.15	0	0.50 (0.03)	0.55 (0.04)	0.55 (0.07)	0.54 (0.04)	0.54 (0.04)
		(n=339)	0.05	0.50 (0.03)	0.55 (0.04)	0.55 (0.08)	0.54 (0.04)	0.54 (0.04)
			0.1	0.50 (0.03)	0.55 (0.07)	0.55 (0.04)	0.54 (0.04)	0.54 (0.04)
			0.15	0.50 (0.03)	0.55 (0.08)	0.55 (0.05)	0.54 (0.05)	0.54 (0.05)

[1] Trial Scenario: 6 Sites, $p_1 = 0.5$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.14. For a null effect size, the five methods perform similarly where the type I error rate increases as the amount of inter-site standard deviation increases. When $\sigma = 0$, the adaptive designs control the type I error rate better, with marginal improvement for SARA 2. When the effect size is increased to 10%, SARA 1 gains back most of the power lost relative to the other adaptive designs. For $\delta = 0.15$, the RA/CARA designs have slightly lower power compared to the other designs when σ is small. As σ increases, the five methods perform more similarly. Overall, there is a notable loss of power for all approaches

as σ increases.

For a null effect size, the type I error results from the GLM/GLMM approach are comparable to the naive approach, where there is an inflation in the type I error rate as σ increases. When $\sigma = 0$, the RA/SARA designs control the type I error rate better, with minimal improvement for SARA 2. For $\delta = 0.10$, the five methods perform similarly, with a loss of power for SARA 2 relative to the other designs with increasing inter-site variability. For $\delta = 0.15$, there is a loss of power for RA randomization compared to the other designs with increasing σ. This coincides with marginal improvement in power for the CARA/SARA 1 methods.

Table 3.14: Scenario 4 - Empirical Type I Error and Power

				Naive					**GLM/GLMM**		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
6	0.5	0	0	0.06	0.05	0.05	0.05	0.04	0.06	0.05	0.04
		(n=775)	0.05	0.14	0.13	0.14	0.14	0.13	0.13	0.14	0.14
			0.1	0.31	0.31	0.31	0.31	0.31	0.32	0.32	0.31
			0.15	0.45	0.45	0.44	0.45	0.46	0.46	0.46	0.48
		0.1	0	0.78	0.79	0.78	0.79	0.79	0.79	0.79	0.79
		(n=775)	0.05	0.78	0.76	0.76	0.77	0.76	0.77	0.77	0.77
			0.1	0.69	0.69	0.69	0.70	0.67	0.70	0.70	0.68
			0.15	0.68	0.67	0.67	0.68	0.65	0.68	0.67	0.65
		0.15	0	0.80	0.79	0.78	0.80	0.80	0.80	0.80	0.80
		(n=339)	0.05	0.76	0.75	0.73	0.76	0.76	0.74	0.76	0.76
			0.1	0.71	0.69	0.71	0.70	0.70	0.72	0.71	0.70
			0.15	0.67	0.66	0.67	0.68	0.66	0.69	0.69	0.67

[1] Trial Scenario: 6 Sites, $p_1 = 0.5$, Optimal Allocation

3.4.5 Scenario 5: No Interaction, 3 Sites, $p_1 = 0.7$

The mean total number of treatment successes for Scenario 5 are displayed in Table 3.15. When the effect size is null and there is little to no inter-site standard deviation, the five approaches perform similarly with respect to he expected number of treatment successes and the variability

in treatment successes. As σ increases, the adaptive designs result in more treatment successes relative to the balanced case, however, the variability remains comparable for all methods. For an effect size of 10%, the adaptive designs result in more treatment successes than the balanced case. The variability in treatment successes is similar for all methods expect for SARA 2, where there is slightly more variability in successes as σ increases relative to the other approaches. For $\delta = 0.15$, the adaptive designs increase the expected treatment successes relative to the BC, however the improvement is slightly lower for SARA 2. The variability in treatment successes is comparable for all five methods. Overall, as σ increases, there is a substantial increase in the variable in treatment successes. For balanced randomization, there is a slight decrease in the average total number of successes as the amount of inter-site standard deviation increases due to the nature of the distribution of high success probabilities being left skewed (more mass below 0.7 since right truncated at 1).

Table 3.15: Scenario 5 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.7	0	0	498.3 (12.5)	498.3 (12.5)	498.3 (12.5)	498.3 (12.5)	498.9 (12.0)
		(n=712)	0.05	498.4 (19.2)	498.6 (19.2)	498.6 (19.2)	498.6 (19.2)	498.5 (19.1)
			0.1	498.3 (31.1)	499.2 (31.1)	499.2 (31.1)	499.2 (31.1)	499.2 (31.0)
			0.15	497.2 (43.4)	499.1 (43.3)	499.2 (43.5)	499.1 (43.4)	499.1 (44.0)
		0.1	0	462.6 (13.1)	464.0 (13.1)	464.0 (13.2)	464.0 (13.1)	464.5 (12.3)
		(n=712)	0.05	463.0 (19.2)	464.6 (19.2)	464.7 (19.2)	464.6 (19.2)	464.3 (19.5)
			0.1	462.8 (31.0)	465.1 (30.9)	465.2 (30.9)	465.1 (30.9)	464.4 (31.4)
			0.15	461.9 (44.0)	465.5 (43.7)	465.7 (43.8)	465.5 (43.8)	465.3 (44.3)
		0.15	0	203.8 (8.3)	205.3 (8.3)	205.3 (8.3)	205.3 (8.3)	204.0 (8.6)
		(n=325)	0.05	203.2 (10.9)	204.7 (10.9)	204.7 (10.9)	204.7 (10.9)	204.1 (10.7)
			0.1	202.9 (16.1)	204.8 (15.7)	204.8 (15.8)	204.7 (15.8)	204.1 (16.0)
			0.15	202.7 (21.6)	205.2 (21.1)	205.4 (21.2)	205.2 (21.2)	204.7 (21.5)

[1] Trial Scenario: 3 Sites, $p_1 = 0.7$, Optimal Allocation

The allocation proportions are displayed in Table 3.16. For $\delta = 0$, the five approaches have comparable allocation proportions and variability, particularly as the amount of inter-site standard deviation increases. The CARA has slightly more variability than the other designs for smaller σ. When $\delta = 0.10$, the adaptive designs have a larger allocation ratio relative to balanced randomization. For smaller σ, CARA has slightly more variability compared to all other approaches. However, as σ increases the SARA designs have slightly lower variability than the other adaptive approaches, closer to the balanced case. For an effect size of 15%, the pattern is similar to a small effect size where adaptive methods share a higher allocation ratio relative to balanced randomization. All five methods have comparable variability in allocation when the amount of inter-site standard deviation is small. However, the adaptive designs are more variable than the BC for larger σ, though CARA has slightly more variability than the other adaptive approaches.

Table 3.16: Scenario 5 - E$[n_1/(n_1 + n_2)]$ (Standard Error)

k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**
3	0.7	0	0	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
		(n=712)	0.05	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
			0.1	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
			0.15	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.50 (0.03)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.52 (0.02)	0.52 (0.03)	0.52 (0.02)	0.52 (0.02)
		(n=712)	0.05	0.50 (0.02)	0.52 (0.02)	0.52 (0.03)	0.52 (0.02)	0.52 (0.02)
			0.1	0.50 (0.02)	0.52 (0.03)	0.52 (0.03)	0.52 (0.03)	0.52 (0.03)
			0.15	0.50 (0.02)	0.52 (0.04)	0.52 (0.04)	0.52 (0.03)	0.52 (0.03)
		0.15	0	0.50 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)
		(n=325)	0.05	0.50 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)
			0.1	0.50 (0.03)	0.53 (0.04)	0.53 (0.04)	0.53 (0.04)	0.53 (0.04)
			0.15	0.50 (0.03)	0.53 (0.04)	0.53 (0.05)	0.53 (0.04)	0.53 (0.04)

[1] Trial Scenario: 3 Sites, $p_1 = 0.7$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.17. For a null effect size,

the five methods perform similarly where the type I error rate increases with increasing inter-site standard deviation. As σ increases, there is a slight improvement in the type I error rate for the adaptive designs, particularly SARA 2, compared to the balanced case. When the effect size is increased to 10%, there is a negligible loss of power for the adaptive designs relative to balanced allocation, with less improvement for SARA 2. As σ increases, the adaptive designs maintain similar, and in some cases more, power relative to the balanced case. There is a modest increase in performance for SARA 2. For $\delta = 0.15$, the adaptive designs outperform the balanced case for smaller values of inter-site variability. As σ increases, the adaptive designs lose their advantage and perform similarly to the BC. Overall, for small and medium effect sizes, there is loss of power for all methods as the amount of inter-site variability increases.

For a null effect size, the results from the GLM/GLMM are similar to the naive approach. When the amount of inter-site variability is large, the RA/SARA 2 designs have lower type I error rates but this improvement is more notable for SARA 2. For $\delta = 0.10$, the pattern is more pronounced relative to the naive results, where the adaptive designs have less power when $\sigma = 0$, but outperform the balanced case as σ increases. SARA 2 has the highest power in the presence of the most inter-site variability. For an effect size of 15%, the pattern is similar to the naive results where the adaptive methods outperform the balanced allocation for smaller σ, particularly the RA/SARA 2 designs. For larger σ, the five methods perform more comparably to the balanced case.

Table 3.17: Scenario 5 - Empirical Type I Error and Power

				Naive					GLM/GLMM		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
3	0.7	0	0	0.06	0.06	0.06	0.06	0.05	0.06	0.05	0.05
		(n=712)	0.05	0.23	0.21	0.22	0.22	0.22	0.22	0.22	0.22
			0.1	0.45	0.45	0.45	0.45	0.45	0.46	0.46	0.45
			0.15	0.62	0.61	0.61	0.61	0.58	0.62	0.62	0.59
		0.1	0	0.81	0.80	0.80	0.80	0.79	0.80	0.80	0.79
		(n=712)	0.05	0.71	0.71	0.71	0.71	0.72	0.72	0.72	0.72
			0.1	0.67	0.67	0.67	0.67	0.67	0.68	0.68	0.68
			0.15	0.67	0.68	0.68	0.68	0.70	0.69	0.68	0.71
		0.15	0	0.80	0.82	0.82	0.81	0.82	0.81	0.81	0.82
		(n=325)	0.05	0.72	0.74	0.73	0.73	0.77	0.73	0.73	0.76
			0.1	0.68	0.68	0.68	0.68	0.68	0.69	0.69	0.68
			0.15	0.67	0.67	0.66	0.66	0.66	0.67	0.67	0.67

[1] Trial Scenario: 3 Sites, $p_1 = 0.7$, Optimal Allocation

3.4.6 Scenario 6: No Interaction, 6 Sites, $p_1 = 0.7$

The mean total number of treatment successes for Scenario 6 are displayed in Table 3.18. When the effect size is null, the BC/RA/CARA/SARA 1 methods perform similarly with respect to both the expected total number of treatment successes and the variability in successes. When there is no inter-site standard deviation, SARA 2 results in slightly more treatment successes while maintaining lower variability compared to all other approaches. However, as σ increases, the adaptive designs result in slightly more treatment successes while maintaining similar variability relative to the balanced case. There is slightly less improvement in both treatment successes and variability for SARA 2. When the effect size is increased to 10%, the RA/CARA/SARA 1 methods increase the expected treatment successes relative to the BC and they maintain the same variability. There is negligible improvement in treatment successes for SARA 2 compared to balanced randomization. For $\delta = 0.15$, the pattern is similar to a small effect size, where the RA/CARA/SARA 1 approaches

result in more treatment successes and comparable variability relative to balanced allocation. There is a marginal improvement in expected successes for SARA 2 compared to the BC, particularly as the amount of inter-site standard deviation increases. In addition, the variability in treatment successes for SARA 2 is slightly higher than the variability for all other methods.

Table 3.18: Scenario 6 - Mean Total Number of Treatment Successes (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1	SARA 2
6	0.7	0	0	498.3 (12.5)	498.3 (12.5)	498.3 (12.5)	498.3 (12.5)	498.9 (12.0)
		(n=712)	0.05	499.0 (16.0)	499.1 (16.0)	499.1 (16.0)	499.1 (16.0)	498.0 (15.8)
			0.1	498.8 (23.8)	499.3 (23.7)	499.3 (23.7)	499.3 (23.7)	498.1 (23.8)
			0.15	497.5 (32.5)	498.4 (32.4)	498.4 (32.4)	498.4 (32.4)	497.8 (32.5)
		0.1	0	462.6 (13.1)	464.0 (13.1)	464.0 (13.2)	464.0 (13.1)	464.5 (12.3)
		(n=712)	0.05	463.5 (16.5)	465.0 (16.4)	465.0 (16.4)	465.0 (16.4)	463.5 (16.5)
			0.1	463.3 (24.4)	464.9 (24.3)	465.1 (24.3)	465.0 (24.3)	463.6 (24.3)
			0.15	463.0 (32.6)	465.1 (32.6)	465.4 (32.6)	465.3 (32.6)	463.6 (33.2)
		0.15	0	203.8 (8.3)	205.3 (8.3)	205.3 (8.3)	205.3 (8.3)	204.0 (8.6)
		(n=325)	0.05	203.3 (9.7)	204.9 (9.6)	204.8 (9.7)	204.7 (9.7)	204.1 (9.9)
			0.1	203.0 (12.6)	204.6 (12.4)	204.7 (12.3)	204.6 (12.4)	204.4 (12.7)
			0.15	202.6 (15.9)	204.3 (15.7)	204.6 (15.7)	204.4 (15.7)	204.4 (16.5)

[1] Trial Scenario: 6 Sites, $p_1 = 0.7$, Optimal Allocation

The allocation proportions are displayed in Table 3.19. For $\delta = 0$, the five approaches have comparable performance. In some cases, the CARA/SARA 1/SARA 2 designs have a slight increase in the amount of variability in allocation relative to the other methods. When $\delta = 0.10$, the adaptive designs have larger allocation ratios compared to the balanced case. When the amount of inter-site standard deviation is low, CARA has more variability than all other approaches. However, as we increase σ, there is an increase in the variability for all adaptive designs relative to the BC. For $\delta = 0.15$, the pattern is similar where the adaptive approaches have larger allocation ratios compared to balanced randomization. As the amount of inter-site standard deviation increases,

there is slightly more variability in allocation for RA randomization and, at some point, the other adaptive designs.

Table 3.19: Scenario 6 - E$[n_1/(n_1+n_2)]$ (Standard Error)

k	p_1	δ	σ	BC	RA	CARA	SARA 1	SARA 2
6	0.7	0	0	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)	0.50 (0.02)
		(n=712)	0.05	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)
			0.1	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)
			0.15	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.03)	0.50 (0.03)
		0.1	0	0.50 (0.02)	0.52 (0.02)	0.52 (0.03)	0.52 (0.02)	0.52 (0.02)
		(n=712)	0.05	0.50 (0.02)	0.52 (0.02)	0.52 (0.03)	0.52 (0.02)	0.52 (0.02)
			0.1	0.50 (0.02)	0.52 (0.02)	0.52 (0.02)	0.52 (0.02)	0.52 (0.02)
			0.15	0.50 (0.02)	0.52 (0.03)	0.52 (0.03)	0.52 (0.03)	0.52 (0.03)
		0.15	0	0.50 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)
		(n=325)	0.05	0.50 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)
			0.1	0.50 (0.03)	0.53 (0.04)	0.53 (0.03)	0.53 (0.03)	0.53 (0.03)
			0.15	0.50 (0.03)	0.53 (0.04)	0.53 (0.04)	0.53 (0.04)	0.53 (0.04)

[1] Trial Scenario: 6 Sites, $p_1 = 0.7$, Optimal Allocation

The empirical type I error and power results are displayed in Table 3.20. For $\delta = 0$, there is a notable inflation in the type I error rates as the inter-site standard deviation increases. When there is no inter-site variability, there is an inflation in the type I error rate for all methods except for SARA 2. However, as σ increases, there is an inflation in the type I error rate for SARA 2 while the other methods perform comparably to balanced allocation. When $\delta = 0.10$, SARA 2 has lower power relative to the other approaches for smaller σ. For larger σ, the five methods have similar power. When $\delta = 0.15$, the five designs perform comparably, however, RA randomization has the highest power for larger σ. Overall, there is a loss of power as the amount of inter-site variability increases for small and medium effect sizes.

For a null effect size, SARA 1/SARA 2 are the only approaches that achieve nominal levels of

type I error when there is no inter-site variability. As σ increases, the adaptive designs perform similarly where there is in an inflation in the type I error rates relative to the BC, particularly SARA 2. When $\delta = 0.10$, the pattern is similar to the naive approach where SARA 2 has lower power for smaller σ but the methods perform comparable as the inter-site variability increases. For $\delta = 0.15$, the adaptive designs have higher power than the balanced case for small and large σ, otherwise all five methods perform comparably. SARA 2 consistently maintains the same, and in some cases more, power than balanced allocation.

Table 3.20: Scenario 6 - Empirical Type I Error and Power

				Naive					**GLM/GLMM**		
k	p_1	δ	σ	**BC**	**RA**	**CARA**	**SARA 1**	**SARA 2**	**CARA**	**SARA 1**	**SARA 2**
6	0.7	0	0	0.06	0.06	0.06	0.06	0.05	0.06	0.05	0.05
		(n=712)	0.05	0.15	0.16	0.15	0.15	0.16	0.16	0.15	0.16
			0.1	0.31	0.32	0.31	0.32	0.35	0.32	0.32	0.35
			0.15	0.46	0.46	0.45	0.45	0.49	0.47	0.47	0.50
		0.1	0	0.81	0.80	0.80	0.80	0.78	0.80	0.80	0.78
		(n=712)	0.05	0.73	0.73	0.73	0.73	0.71	0.74	0.74	0.71
			0.1	0.65	0.66	0.65	0.65	0.65	0.66	0.66	0.65
			0.15	0.65	0.64	0.63	0.64	0.65	0.65	0.65	0.65
		0.15	0	0.80	0.82	0.82	0.82	0.81	0.81	0.81	0.81
		(n=325)	0.05	0.77	0.76	0.76	0.76	0.77	0.76	0.76	0.77
			0.1	0.72	0.72	0.71	0.71	0.71	0.72	0.72	0.72
			0.15	0.66	0.67	0.66	0.66	0.66	0.68	0.67	0.67

[1] Trial Scenario: 6 Sites, $p_1 = 0.7$, Optimal Allocation

3.4.7 Incorporating a Treatment-by-Site Interaction

In Section 3.4.7, we compare the mean total number of treatment successes (SE), the expected allocation proportions for treatment group 1 (SE) and the empirical type I error and power for SARA models 2 and 3 (Eqns. 3.1 & 3.3, respectively). We only report the operating characteristics calculated using a frequentist GLMM (not a naive approach), which accounts for site-based response

variability using random site effects, to assess whether the approaches perform as intended. We present several scenarios to compare the results when incorporating a random treatment-by-site interaction effect into the adaptive design. The results are similar for additional scenarios found in Appendix B.

Scenario: 3 sites, $p_1 = 0.3, \delta = 0, 0.15$

The mean total number of treatment success, allocation proportions and empirical type I error/power for a 30% treatment group 1 success probability are presented in Table 3.21. When the effect size is null, the two SARA designs perform comparably when there is no inter-site variability. However, as σ increases, there is a notable increase in expected treatment successes for SARA 3 without an increase in the variability in successes. For $\delta = 0.15$, SARA 3 maintains more expected successes and similar variability relative to SARA 2, particularly as σ increases.

Comparison: For a null effect size, SARA 3 outperforms all other allocation approaches with respect to treatment successes and variability as σ increases (Table 3.3). When the effect size increases to 15%, SARA 3 gains back most of the treatment successes lost with SARA 1/SARA 2 compared to RA/CARA, with less improvement for CARA. SARA 3 performs similarly to the RA/CARA designs, but maintains a lower variability in successes that is comparable to BC/SARA 1/SARA 2.

For $\delta = 0$, the two SARA designs perform comparably with respect to expected allocation proportions and variability in allocation. When the effect size is increased to 15%, SARA 3 has a slightly larger allocation ratio relative to SARA 2 but maintains similar variability in allocation.

Comparison: For a null effect size, the allocations ratios for SARA 3 are comparable to the other approaches (Table 3.4). SARA 3 performs similarly to SARA 1/SARA 2 with respect to the variability in allocation. For $\delta = 0.15$, SARA 3 has slightly larger allocation ratios than SARA 1/SARA 2 that are closer to the RA/CARA designs. However, all three SARA designs have comparable variability in allocation ratios that is lower than the RA/CARA approaches.

For a null effect size, the two SARA designs perform similarly where there is an inflation in the type I error rate as the amount of inter-site variability increases. When σ is small the two methods are comparable but as σ increases, SARA 3 has better control of the type I error rate. For $\delta = 0.15$, the two SARA approaches have the same power when there is no inter-site variability.

As σ increases, there is a loss of power for SARA 3 compared to SARA 2.

Comparison: For $\delta = 0$, SARA 2/SARA 3 perform comparably and have the lowest type I error rates compared to the other approaches (Table 3.5) when the amount of inter-site variability is small. However, as σ increases, the RA/SARA 2/SARA 3 approaches perform similarly and have better control of the type I error rate, with less improvement with SARA 2. For $\delta = 0.15$, the SARA approaches have higher power than all other methods. SARA 2/SARA 3 have more improvement when σ is small while SARA 2 outperforms the other designs with increasing inter-site variability.

Table 3.21: Scenario - 3 sites, $p_1 = 0.3$, $\delta = 0, 0.15$

				SARA 2			**SARA 3**		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
3	0.3	0	0	176.1 (10.7)	0.50 (0.03)	0.05	176.1 (10.7)	0.50 (0.03)	0.05
		(n=587)	0.05	176.4 (16.6)	0.50 (0.03)	0.17	177.3 (16.6)	0.50 (0.04)	0.17
			0.1	177.3 (26.9)	0.50 (0.05)	0.40	180.7 (26.8)	0.50 (0.05)	0.38
			0.15	179.7 (37.6)	0.50 (0.06)	0.53	187.5 (37.6)	0.50 (0.07)	0.51
		0.15	0	57.5 (6.7)	0.59 (0.05)	0.80	57.8 (6.8)	0.60 (0.06)	0.80
		(n=241)	0.05	57.6 (8.2)	0.59 (0.06)	0.75	58.4 (8.3)	0.60 (0.06)	0.74
			0.1	58.6 (11.6)	0.59 (0.07)	0.70	60.4 (11.7)	0.60 (0.08)	0.67
			0.15	60.5 (15.4)	0.59 (0.09)	0.65	63.3 (15.6)	0.60 (0.09)	0.60

Scenario: 3 sites, $p_1 = 0.5, 0.7, \delta = 0.15$

The mean total number of treatment success, allocation proportions and empirical type I error/power for a large effect size ($\delta = 0.15$) are presented in Table 3.22. For a 50% treatment group 1 success probability ($p_1 = 0.5$), SARA 3 results in more expected treatment successes without an increase in the variability in successes, particularly as σ increases. When the treatment group 1 success probability is increased to 70%, the pattern is comparable but the improvement becomes notable for even larger σ.

Comparison: For $p_1 = 0.5$, SARA 3 outperforms all other allocation approaches with respect to expected treatment success, with more notable improvement as σ increases (Table 3.9). The SARA

2/SARA 3 designs have less variability in treatment successes than all other methods, including balanced allocation. For $p_1 = 0.7$, SARA 3 has slightly lower treatment success but comparable variability relative to the other adaptive designs (Table 3.15) when σ is small. However, as σ increases, SARA 3 gains back the loss in treatment successes, outperforming all other approaches and maintaining similar variability.

For $p_1 = 0.5$, SARA 3 has slightly larger allocation ratios when the inter-site variability is small, however, the two SARA designs perform similarly for larger σ. The variability in allocation is comparable, with marginal improvement for SARA 2. When the success probability increases to 70%, the two approaches perform similarly with respect to allocation proportions and variability.

Comparison: For $p_1 = 0.5$, SARA 3 gains back the slight loss in adaptation for the SARA 1/SARA 2 approaches relative to the RA/CARA designs (Table 3.10) for small σ. The SARA methods all have less variability in allocation relative to the RA/CARA designs. For larger inter-site variability, SARA 3 has similar allocation ratios to the RA/CARA approaches but maintains lower variability similar to BC/SARA 1/SARA 2. When the treatment group 1 success probability is increased to 70%, SARA 3 performs comparably to the other five approaches (Table 3.16) with respect to allocation ratios and variability.

For a 50% treatment group 1 success probability, SARA 3 has slightly higher power than SARA 2 when there is no inter-site variability. However, as σ increases, there is a negligible loss of power for SARA 3. For $p1 = 0.7$, there is a slight loss of power for SARA 3 which diminishes with increasing σ.

Comparison: For $p_1 = 0.5$, SARA 3 has the highest power relative to all other approaches (Table 3.11) when there is no inter-site variability. As σ increases, the three SARA designs have higher power compared to the other methods, with more improvement for SARA 2 and closely followed by SARA 3. For $p_1 = 0.7$, the RA/SARA 2 designs (Table 3.17) have the highest power while balanced allocation has the lowest power for smaller values of σ. However, the six approaches have similar power as σ increases.

Table 3.22: Scenario - 3 sites, $p_1 = 0.5, 0.7$, $\delta = 0.15$

				SARA 2			SARA 3		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
3	0.5	0.15	0	147.0 (9.0)	0.54 (0.03)	0.80	147.2 (8.9)	0.55 (0.04)	0.81
		(n=339)	0.05	146.3 (10.9)	0.54 (0.04)	0.77	146.7 (10.9)	0.55 (0.04)	0.76
			0.1	146.5 (16.0)	0.55 (0.04)	0.71	147.9 (15.9)	0.55 (0.05)	0.70
			0.15	147.5 (22.3)	0.55 (0.05)	0.69	150.5 (22.0)	0.55 (0.06)	0.67
	0.7	0.15	0	204.0 (8.6)	0.53 (0.03)	0.82	204.1 (8.5)	0.53 (0.03)	0.81
		(n=325)	0.05	204.1 (10.7)	0.53 (0.03)	0.76	204.3 (10.8)	0.53 (0.03)	0.75
			0.1	204.1 (16.0)	0.53 (0.04)	0.68	204.9 (16.0)	0.53 (0.04)	0.67
			0.15	204.7 (21.5)	0.53 (0.04)	0.67	206.4 (21.3)	0.53 (0.04)	0.67

Scenario: 6 sites, $p_1 = 0.3, \delta = 0, 0.15$

The mean total number of treatment success, allocation proportions and empirical type I error/power for a larger number of sites ($k = 6$) are presented in Table 3.23. When the effect size is null, the two SARA designs perform comparably when there is no inter-site variability. As σ increases, there is a notable increase in the mean total number of treatment successes for SARA 3 without an increase in the variability in expected successes. For $\delta = 0.15$, the pattern is similar to a null effect size where SARA 3 maintains more expected treatment successes without an increase in the variability.

Comparison: For a null effect size, SARA 2/SARA 3 have the highest mean total number of treatment successes and lowest variability when there is no inter-site variability. As σ increases, SARA 3 outperforms all other approaches with respect to the expected treatment successes (Table 3.6), achieving an additional eight successes more than next best method (RA/SARA 2) on average. SARA 2/SARA 3 achieve similar variability that is lower than all other designs. For $\delta = 0.15$, SARA 3 results in slightly more treatment successes than the CARA/SARA 1/SARA 2 designs and is closer to RA randomization when $\sigma = 0$. However, SARA 3 has more expected successes relative to all other approaches but maintains similar variability to the CARA/SARA 1/SARA 2 designs. The RA method has the highest variability in all cases.

For $\delta = 0$, the two SARA designs perform comparably with respect to the expected allocation proportions and variability in allocation. When the effect size is increased to 15%, SARA 3 has a larger allocation ratio relative to SARA 2 but stills maintains similarity variability in allocation.

Comparison: For a null effect size, the allocation ratios for SARA 3 are comparable to the other approaches (Table 3.7), with similar variability in allocation to the CARA/SARA approaches. RA randomization has slightly less adaptation relative to the other five approaches as σ increases. For $\delta = 0.15$, SARA 3 results in larger allocation ratios than the BC/CARA/SARA 1/SARA 2 approaches that are closer to RA allocation. However, the CARA/SARA designs maintains less variability than the RA design.

For a null effect size, SARA 3 has slightly inflated type I error rates relative to SARA 2 when the amount of inter-site variability is small. However, as σ increases, there is better control of the type I error rate for SARA 3. For $\delta = 0.15$, the two SARA approaches achieve similar power for smaller values of inter-site variability, with marginal improvement for SARA 3. As σ increases, there is a loss of power for SARA 3 relative to SARA 2.

Comparison: For $\delta = 0$, SARA 2/SARA 3 perform comparably and have lower type I error rates compared to the other approaches (Table 3.8) that are slightly inflated when there is no inter-site variability. However, as σ increases, the RA/SARA 3 designs perform similarly and have better control of the type I error rate than the other methods, with more improvement for RA randomization. When the effect size is increased to 15%, the CARA/SARA approaches have higher power than RA randomization. SARA 2/SARA 3 have more improvement when σ is small while SARA 2 outperforms all other designs with increasing values of inter-site variability.

Table 3.23: Scenario - 6 sites, $p_1 = 0.3$, $\delta = 0, 0.15$

				SARA 2			SARA 3		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
6	0.3	0	0	176.1 (10.7)	0.50 (0.03)	0.04	176.1 (10.7)	0.50 (0.03)	0.05
		(n=587)	0.05	176.0 (13.5)	0.50 (0.03)	0.11	177.0 (13.5)	0.50 (0.04)	0.12
			0.1	177.1 (19.8)	0.50 (0.04)	0.29	180.8 (19.7)	0.50 (0.05)	0.27
			0.15	178.9 (26.7)	0.50 (0.05)	0.44	186.6 (26.6)	0.50 (0.06)	0.41
		0.15	0	57.6 (6.8)	0.59 (0.05)	0.80	58.0 (6.8)	0.60 (0.06)	0.80
		(n=241)	0.05	58.0 (7.7)	0.59 (0.06)	0.76	58.7 (7.8)	0.60 (0.06)	0.77
			0.1	58.5 (9.7)	0.59 (0.07)	0.71	60.1 (9.8)	0.60 (0.07)	0.68
			0.15	59.9 (12.0)	0.58 (0.07)	0.65	62.5 (12.2)	0.59 (0.07)	0.62

3.4.8 Simulation Time per Patient

The simulation times per patient are compared to explore the trade-off of fitting increasingly complex SARA models and computation time when performing a simulation study.

Scenario: 3 sites, $p_1 = 0.3, \delta = 0, 0.15$

The results comparing null and medium effect sizes for three sites are displayed in Table 3.24. For $\delta = 0$, SARA 3 has the highest simulation times for all σ. The pattern is similar for an effect size of 15%, however, the simulation times are generally lower. Although the times are reported per patient, thus removing the effect of different sample sizes, scenarios with larger sample sizes ($\delta = 0$) have increased simulation times.

In a real life application, there would not be a notable difference in computation times across the three adaptive designs as only one patient is adaptively randomized at a time for a single trial. However, the computation times represent a significant burden when performing a simulation study with n patients, 1000 trials and a variety of scenarios. For example, a simulated multi-center trial with three sites, a 30% treatment group 1 success probability, no effect size or inter-site variability, and 1000 trials takes approximately 53 hours for SARA 1 and 2518 hours to run using the SARA 3 approach. This equates to 0.34 and 15.98 seconds for SARA 1 and SARA 3, respectively, per

patient adaptively randomized.

Table 3.24: Simulation Time (seconds) per Patient

k	p_1	δ	σ	SARA 1	SARA 2	SARA 3
3	0.3	0	0	0.34	10.28	15.98
		(n=587)	0.05	0.25	8.94	17.51
			0.1	0.24	9.07	16.31
			0.15	0.18	10.37	16.30
		0.15	0	0.27	4.88	9.71
		(n=241)	0.05	0.25	4.89	8.79
			0.1	0.15	5.56	8.77
			0.15	0.15	7.65	9.84

[1] Trial Scenario - 3 sites, $P_1 = 0.3$, $\delta = 0, 0.15$

Scenario: 3 sites, $p_1 = 0.5, 0.7, \delta = 0.15$

The results comparing medium ($p_1 = 0.5$) and large ($p_1 = 0.5$) treatment group 1 success probabilities for a large effect size are displayed in Table 3.25. The simulation times are comparable for both success probabilities where SARA 3 has the largest computation time while SARA 1 has the smallest. This may be explained by the fact that both scenarios have similar sample sizes.

Table 3.25: Simulation Time (seconds) per Patient

k	p_1	δ	σ	SARA 1	SARA 2	SARA 3
3	0.5	0.15	0	0.28	6.18	11.84
		(n=339)	0.05	0.16	6.16	10.42
			0.1	0.17	6.92	10.31
			0.15	0.17	7.91	11.71
	0.7	0.15	0	0.29	5.90	11.33
		(n=325)	0.05	0.15	5.93	9.99
			0.1	0.15	6.69	9.98
			0.15	0.16	7.82	11.35

[1] Trial Scenario - 3 sites, $P_1 = 0.5, 0.7$, $\delta = 0.15$

Scenario: 6 sites, $p_1 = 0.3, \delta = 0, 0.15$

The results comparing null and medium effect sizes for six sites are displayed in Table 3.26. The results are similar to three sites, where SARA 3 has the longest simulation times and a large effect size results in shorter computation times relative to a null effect size. Overall, the simulation times are longer for an increasing number of sites for the SARA 2/SARA 3 approaches but not for SARA 1.

Table 3.26: Simulation Time (seconds) per Patient

k	p_1	δ	σ	SARA 1	SARA 2	SARA 3
6	0.3	0	0	0.33	15.12	22.20
		(n=587)	0.05	0.25	13.03	24.44
			0.1	0.25	13.07	22.09
			0.15	0.19	14.95	22.15
		0.15	0	0.29	6.93	15.21
		(n=241)	0.05	0.27	6.94	13.16
			0.1	0.16	8.10	13.20
			0.15	0.15	9.57	15.45

[1] Trial Scenario - 6 sites, $P_1 = 0.3$, $\delta = 0, 0.15$

3.4.9 Sensitivity Analysis for Prior Specification of Random Effects

Several prior specifications were considered for the sensitivity analysis (Table 3.2). Prior C was the specification used for both the random site and treatment-by-site interaction effects in our simulation study. The results of the sensitivity analysis for SARA 2 and SARA 3 are displayed in Table 3.27.

For SARA 2, the results show that all five priors are comparable with respect to the mean total number of treatment successes (SE), expected allocation proportion for treatment group 1 (SE), empirical power. Prior B has slightly more average non-convergence warnings relative to the other distributions. For SARA 3, priors A and D result in less treatment successes, smaller allocation ratios and more expected non-convergence relative to the other priors, particularly as σ increases. Prior D has notably higher power than the other distributions that have comparable power. Overall, the prior used in our simulation study, Prior C, has similar inferences to the other specifications considered.

Table 3.27: Sensitivity Analysis for Prior Specification of Random Effects

Model	σ	Prior	Successes (SE)	Proportion (SE)	Power	Warnings
SARA 2	0.1	A	55.4 (15.3)	0.64 (0.08)	0.90	0.2
		B	55.4 (15.5)	0.64 (0.08)	0.90	4.1
		C	55.5 (15.3)	0.64 (0.08)	0.90	0.0
		D	55.5 (15.3)	0.64 (0.08)	0.90	0.0
		E	55.4 (15.5)	0.64 (0.08)	0.90	0.0
	0.15	A	55.8 (20.8)	0.65 (0.10)	0.78	0.4
		B	55.9 (20.8)	0.65 (0.10)	0.78	2.1
		C	55.8 (20.7)	0.65 (0.10)	0.78	0.0
		D	55.8 (20.7)	0.65 (0.10)	0.78	0.0
		E	55.7 (20.6)	0.65 (0.10)	0.78	0.0
SARA 3	0.1	A	56.8 (15.9)	0.65 (0.08)	0.80	1.0
		B	58.1 (17.2)	0.66 (0.09)	0.80	2.0
		C	57.5 (16.7)	0.66 (0.08)	0.80	0.3
		D	59.0 (17.2)	0.68 (0.09)	0.80	26.4
		E	56.9 (15.8)	0.65 (0.08)	0.80	1.0
	0.15	A	58.1 (21.1)	0.64 (0.10)	0.78	1.0
		B	60.4 (22.3)	0.66 (0.11)	0.78	2.7
		C	59.4 (21.8)	0.66 (0.10)	0.78	0.5
		D	61.2 (22.9)	0.67 (0.11)	0.89	28.8
		E	58.8 (21.1)	0.65 (0.10)	0.78	0.5

3.5 Discussion

We extended a method that accounts for site-based response variability under the Bayesian framework to avoid issues in estimating many parameters. We then proposed a method that also accounts for variability in treatment effectiveness by incorporating a random treatment-by-site interaction into the model. The simulations used larger values of inter-site standard deviation to

promote scenarios with qualitative interactions. Accordingly, the conditional probabilities of success specific to each treatment and center were estimated in a novel approach to skew the treatment allocations, thus allowing the treatment success probabilities to vary across centers.

For a small success rate and null effect size, SARA 3 outperforms all other allocation approaches with more expected treatment successes but maintains similar variability, particularly as the amount of inter-site variability increases. The allocations ratios are comparable across all six approaches, but the SARA approaches share lower variability in adaptation than the RA/CARA designs. The RA/SARA 2/SARA 3 approaches have similar control of the type I error rate.

For a large effect size, SARA 3 gains back the expected successes lost with SARA 1/SARA 2 relative to the RA/CARA but maintains lower variability. SARA 3 results in larger allocation ratios than the other SARA designs, closer to the RA/CARA approaches, but the SARA designs achieve lower variability in adaptation. As the amount of inter-site variability increases, SARA 2/SARA 3 maintain higher power than the other designs.

As the number of sites increases, SARA 3 notably outperforms all other approaches in terms of treatment successes while the CARA/SARA designs have the lowest variability in successes. SARA 3 results in lager allocation ratios than the BC/CARA/SARA approaches that are closer to RA allocation but with similar variability to the CARA/SARA designs. The RA/SARA 3 designs perform similarly and have better control of the type I error rates. The CARA/SARA approaches have higher power than RA randomization, with more improvement for SARA 2 with increasing inter-site variability.

When the success rate increases, the six approaches perform more comparably with respect to allocation ratios, variability in allocation and empirical power for a large effect size. However, SARA 3 maintains slightly more expected treatment successes than the other methods, but with similar variability, for increasing values of inter-site variability. It is encouraging to note that the adaptive designs generally work better at lower success probabilities when reducing treatment failures is most important.

By incorporating the random site and treatment-by-site effects in the SARA models, we account for additional sources of variability introduced from the centers. This leads to an increase in expected treatment successes, more adaptation, better control of the type I error rate and increased statistical power. SARA 3 outperforms SARA 1/SARA 2 with respect to treatment successes and

adaptation but at the cost of marginally less power. When accounting for site in the adaptive allocation design, we recommend assessing the operating characteristics, and thus make inferences, using the results from an approach that accounts for the clustering (e.g. GLMM).

In our simulation study, convergence was assessed after each model fit. In cases when the parameter estimates did not converge, we followed Rosenberger et al. (2001) and used the previous patient's allocation probabilities until convergence was achieved again [20]. This approach ignores the subsequent patient's site when 'updating' the allocation probabilities. Instead, we suggest using the last individual's allocation weights who was randomized at the same site as the next patient. Alternatively, the previous subject's parameter estimates can be used to calculate site-specific allocation weights, but instead using the subsequent patient's site. This would allow us to utilize the most recent data to update the randomization probabilities while also incorporating the appropriate site information. This would be expected to have a negligible effect on the results of our simulation study as the frequency of non-convergence is relatively low. However, this is a more ethical approach to handling non-convergence during the adaptation process since all available data, including site information, is being used to adaptively randomize.

The results of the simulation times demonstrate that SARA 3 has the longest computation time while SARA 1 has the shortest. Although the simulation times are reported per patient, thus removing the effect of varying sample sizes, scenarios with larger sample sizes have increased simulation times. This may be explained in part by the additional computation and storing of results for scenarios with larger sample sizes. In addition, SARA 3 takes longer due to estimating additional parameters from the random treatment-by-site interaction effect, particularly as the number of patients and centers increase. However, these simulation time results only pertain to empirical research. In a real life application, the increased computation times for the SARA designs relative to the other adaptive approaches and balanced allocation are trivial as only one patient is adaptively randomized at a time for a single trial.

There are several limitations associated with the simulation times that should be noted when interpreting the results. First, all simulations were run on the cluster, where scenarios were allocated to different nodes. Each node may vary in terms of performance and speed, especially depending on the current job queue. Second, each scenario was run as a single job containing 1000 simulations. This increases the elapsed time due to the additional computation and storing of the iterations.

In this chapter, Bayesian GLMMs were fit for the SARA 2/SARA 3 models and the conditional success probabilities calculated with the model parameters estimated using MCMC techniques. Weakly informative Gaussian priors were specified for the random site and treatment-by-site interaction effects as we anticipated small effects relative to a traditional uninformative prior. A sensitivity analysis demonstrated that the operating characteristics were robust to the selection of the prior distribution. However, due to long simulation times, this investigation was not exhaustive. Further work includes more extensive sensitivity analyses be performed with larger variances (smaller precisions) and with alternative prior specifications (e.g. hyperpriors) to test the robustness of the inferences.

CHAPTER 4

ACCOUNTING FOR LOCATION-BASED VARIABILITY

In addition to inter-center variability in treatment responses and effectiveness, treatment differences could also be attributed to social determinants of health or other lifestyle and landscape factors that that drive patient heterogeneity. When unmeasured, these treatment differences can manifest as spatial variability and are likely to complicate patient allocation. In Section 4.1, we discuss spatial correlation and summarize some existing methods to account for spatial correlation in regression modeling, particularly with incomplete covariate data. In Section 4.2, we introduce a novel approach to adaptive allocation in clinical trials to account for more coarse location-based variability in treatment effectiveness. Results from a simulation study are reported in Section 4.3 comparing the performance of our location-adjusted approach to both balanced and RA randomization. The results are explored and a brief discussion of the implications follows.

4.1 Background and Existing Methods

Traditionally, clinical trial researchers ignore geographic location – such as patient residential address – as a factor that may affect the outcome because these experiments are often randomized to reduce confounding. However, Lawson (2013) notes two cases in which location should be considered to incorporate any spatial effects into the design and/or analysis. During the design phase, there may be a geographical bias in the risk of disease resulting in selection bias. In the analysis phase, location may be a confounder in the relationship between the groups and the outcome. In addition, there may be unknown or known confounders that were not considered in the analysis. In these cases, accounting for spatial correlation in the analysis may be important [46].

Clayton et al. (1993) express the importance of including a spatial clustering term, in addition to an unstructured term, to account for residual geographical variation in risk when mapping disease risk. This spatially structured random effect acts as a surrogate for these unknown or known confounders that vary smoothly throughout space and were not accounted for otherwise in the model. The authors motivate this work with studies that explore the associations between environmental

and socioeconomic factors with cardiovascular mortality [47, 48] as well as the association between cancer and dietary intakes [49]. As an example, the authors fit a log-linear Poisson model of lip cancer incidence in Scotland that includes the percentage of the population employed in agriculture, fishing and forestry as a covariate along with both 'heterogeneity' and spatial 'clustering' terms. The results show that location acts as a confounder in the relationship between occupation and risk of lip cancer, where individuals with these outdoor occupations have more sunlight exposure and thus an increased risk of cancer [46, 50].

4.1.1 Spatial Correlation

Spatial autocorrelation (SA) is the correlation of a variable across space. Positive SA is characterized by similar outcome values for spatial units close together while negative spatial correlation occurs when outcome values differ when they are close together in space. The assumption of independence in traditional statistical methods is not satisfied when SA exists and thus spatial methods are necessary to account for this dependence [46].

Often, traditional regression models do not capture all of the variation present in the data by simply adding selected covariates. In order to account for over-dispersion or spatial correlation from unmeasured confounders, additional terms can be considered. One way of accounting for these sources of variability is to incorporate additive unstructured and/or spatially structured random effects to the linear, or even non-linear, predictor. An unstructured random effect can be used to model uncorrelated heterogeneity (UH) and is often given an exchangeable prior, such as an uninformative, zero-mean Gaussian distribution. Alternatively, correlated heterogeneity (CH) can be accounted for by adding a spatially structured random effect, using either intrinsic Gaussian conditionally autoregressive (CAR) or fully specified multivariate Gaussian prior distributions [46].

The intrinsic CAR, an improper prior distribution, was introduced into the Bayesian setting by Besag et al. (1991). A vector of random effects $\boldsymbol{u}$ with a CAR prior have a conditional distribution,

$$[u_i|u_{j\neq i}] \sim N\left(\bar{u}_i, r/n_{\xi_i}\right), i = 1, \ldots, n, \tag{4.1}$$

where u_i is conditional[1] on all other values of the vector $\boldsymbol{u}$ where $j \neq i$. The first and second

[1]This can also be denoted as $[u_i|u_{-i}]$ or $[u_i|\ldots]$.

conditional moments are functions of the number of neighbors (n_{ξ_i}) in the neighborhood (ξ_i) and their values. Specifically, the conditional variance (r) of the random effect is inversely proportional to the number of neighbors. The random effects have zero mean due to a sum-to-zero constraint ($\sum_{i=1}^{N} u_i = 0$), where the mean $\bar{u}_i$ is defined as follows:

$$\bar{u}_i = \sum_{j \in \xi_i} u_j / n_{\xi_i}, \tag{4.2}$$

which is the the neighborhood (ξ_i) average in the i^{th} region.

The univariate intrinsic CAR was previously extended to the multivariate setting [51]. A multivariate CAR (MCAR) can be used to model several spatially structured random effects or regression coefficients under the assumption that they are spatially dependent. For example, an MCAR prior can be used to specify multiple, dependent regression coefficients in a Bayesian hierarchical spatial model. This would allow the coefficients to vary across space. Random effects $S_i = (S_{i1}, \ldots, S_{ip})'$ with an MCAR prior specified have a multivariate conditional distribution,

$$\left[\boldsymbol{S}_i | \left(S_{(-i)1}, \ldots, S_{(-i)p}\right)\right] \sim N_p\left(\bar{\boldsymbol{S}}_i, \boldsymbol{\Sigma}/n_{\xi_i}\right), \tag{4.3}$$

where $\bar{\boldsymbol{S}}_i = \left(\bar{S}_{i1}, \ldots, \bar{S}_{ip}\right)'$ and $\bar{S}_i = \sum_{j \in \xi_i} S_j / n_{\xi_i}$ for the p dependent random effects in neighborhood ξ_i (n_{ξ_i} neighbors) for the i^{th} area. Here, the random effects are conditional on all other values of the vector $\boldsymbol{S_j}$, where $j \neq i$. Using a separable cross-covariance matrix $\boldsymbol{\Sigma}$ assumes the spatial dependence is the same across all coefficients while a non-separable form would allow this dependence to vary [52, 53, 54].

Since patient locations can be viewed as randomly located throughout a study surface and each patient experiences either a treatment success or failure, a spatial model designed for marked point pattern data will be used. Two types of Spatial Bayesian models will be covered that are hierarchical in their parameters, Geographically Adaptive Regression (GAR) and generalized linear spatial (GLS) models.

Geographically Adaptive Regression Model

The GAR Model is a type of geographically dependent regression, extended from geographically weighted regression (GWR), that allows the regression relationship to vary over space (parametric

spatial non-stationarity). Instead of global (fixed) parameter estimates in an ordinary least squares regression model, the model is fitted at each location to allow regression coefficients to be estimated at each location [55, 46, 56, 57]. Fotheringham et al. (2002) extended the GWR model for the Gaussian setting as follows:

$$y_i(s) = \beta_0(s) + \beta_1(s)x_{1i} + \beta_2(s)x_{2i} + \cdots + \epsilon(s), \tag{4.4}$$

where s is a particular spatial location and x_1, x_2 are assumed to have spatial dependence [56]. Under the Bayesian context, spatially-structured priors on the coefficients are necessary to allow for spatially-varying parameters. Equation (4.4) is a special case of a random-intercept and random-slope model. Using a single predictor, Lawson (2013) extended GWR to categorical outcomes under the Bayesian GLMM framework as follows:

$$\begin{aligned} E(y_i) &= \mu_i \\ \mu_i &= g(\eta_i) \\ \eta_i &= \beta_0 + \beta_1^* x_{1i} + R_i \\ R_i &= \beta_{01i} + \mu_i + v_i, \end{aligned} \tag{4.5}$$

where

$$\begin{aligned} \beta_0 &\sim N(0, \tau_{\beta_0}) \\ \beta_{01i} &\sim N(0, \tau_{\beta_{01}}) \\ \beta_1^* &= (\beta_1 + \beta_{11i}) \\ \beta_1 &\sim N(0, \tau_{\beta_1}) \\ \beta_{11i}|\cdots &\sim N(\mu_{\beta_{11}}, \tau_{\beta_{11}}/n_{\xi_i}). \end{aligned}$$

The intercept (β_0) includes an uncorrelated random component (β_{01i}) while the slope has a random main effect (β_1) and a correlated random component (β_{11i}), where $\mu_{\beta_{11}} = \bar{\beta}_{11\xi_i}$ with precision $\tau_{\beta_{11}}$. While the random slope can be unstructured or spatially structured, putting an improper CAR prior on this coefficient allows for a spatially-varying slope, as in the above example.

This is what characterizes a GAR model. Specifying a spatially structured prior on the random slope leads to a parsimonious model and allows for a more smooth (gradual changes in variation) slope across space. The two other (optional) random components (Eqn. 4.5), forming a convolution model, capture any residual error, or unmeasured confounding, in the data and are unstructured (μ_i) and spatially-structured (v_i).

The convolution model is less parsimonious and may have identifiability issues from the extra random effects compared to the GAR model; however, they may have similar deviance information criterion (DIC) values. Consequently, Lawson advises that $\mu_i + v_i$ may need to be removed when assuming a GAR model. The GLMM is flexible in that the outcome y_i is assumed to come from any one of the exponential family distributions. The linear predictor η_i can consist of several uncorrelated and spatially structured random components and is connected to the mean of the response via one of several link functions $g(\cdot)$ [46].

One approach to using an intrinsic CAR prior when working with point pattern data is tiling, such as Voronoi/Dirichlet Tessellation or Delauney Triangulation, to convert to areal data. The patient locations are used to construct the neighborhood structure and adjacency information needed to specify a CAR prior. In tessellation, the boundaries are calculated so that each tile around a point is closest to that point and no other. This creates natural neighborhoods where adjoining edges of the tiles consist of natural neighbors. Tessellation is preferred over Delauney triangulation, which uses bisectors of the tile edges to form a convex hull in the points, because triangulation may connect more distant points that may not be in the vicinity of the patient (not first order neighbors) [46].

4.1.2 Existing Interpolation Methods

We need to incorporate a method to predict the success probabilities at a new point in space. We discuss several existing approaches that handle incomplete covariate information, both deterministic and geostatistical interpolation.

Inverse Distance Weighted Interpolation

One simple approach to predict outcomes at a point, a set of points or a grid surface when the attribute values are only available at observed locations is to use Inverse-distance weighted (IDW)

interpolation, as mentioned in Lu and Wong (2008). IDW assumes that values closer together are more similar than values further apart. Specifically, IDW calculates a weighted average of the values at observed locations S_i to make a prediction at a new interpolation location S_0 as follows:

$$\hat{y}(S_0) = \sum_{i=1}^{n} \lambda_i \; y(S_i).$$

Here, the interpolated value $\hat{y}(S_0)$ is a linear combination of observed values and the weights, λ_i, where

$$\lambda_i = \frac{d_{0i}^{-p}}{\sum_i^n d_{0i}^{-p}},$$

and d_{oi}^{-p} is the inverse of the euclidean distance between the observed and new location with an IDW power of p, or $||s_i - s_0||^{-p}$. The weights for the observed locations are each calculated based on their distance to the new location. The value of the inverse distance weighting power is set by the user. For $p \to 0$, the weights are similar and thus the predicted value is an average of the neighboring observed values. However, when $p > 1$, weights diminish with increasing distance, with larger p giving more weight to closer observed values.

This method is easy to implement in R (see *gstat* {gstat} function) and is not computationally intensive. The default inverse distance weighting power is set to 2 to perform inverse distance squared weighted interpolation, though this is a arbitrary selection. A drawback of IDW interpolation is that it does not provide a confidence level of the interpolation (e.g. variance) nor are the distance weights calculated using empirical data. IDW assumes that the distance-decay is constant throughout the study area, and thus ignores the spatial variability of the distance-decay relationship [58, 59].

Fully Specified Covariance Model

The Generalized Linear Spatial Model (GLSM), also known as Bayesian Gaussian Kriging, is a type of correlated heterogeneity model that specifies only one random effect for both unstructured and spatially structured heterogeneity through two parameters in the prior distribution. A multiplicative relationship is specified for the uncorrelated and (spatial) correlated heterogeneity. Diggle et al. (1998) propose a kriging-type model, similar to universal kriging, and use a multivari-

ate normal (MVN) prior distribution to parametrically model the spatial variation in risk. In one application, they model the risk of *Campylobacter* infections at postcode locations using a spatially varying 'success' probability

$$\log\left[\frac{P(\boldsymbol{x})}{1-P(\boldsymbol{x})}\right] = \beta + S(\boldsymbol{x}), \tag{4.6}$$

where $S(\boldsymbol{x})$ is a zero-mean stationary Gaussian process with a specified correlation function [60]. The posterior estimates were used to predict spatially-varying log odds to a grid of interest. The authors suggest an approach to account for non-spatial extrabinomial variation by adding a non-spatial component Z to Eqn. 4.6, where $Z \sim N(0, \tau^2)$ and is described as using a 'convolution Gaussian prior'.

The GLSM incorporates a hierarchically-centered (zero-mean) random effect with a MVN prior that models a parametric function of the distances and spatial correlation parameter (ϕ) in the variance-covariance matrix $\mathbf{\Gamma}$. A powered exponential correlation function,

$$\rho(d_{ij}) = exp(-(d_{ij}/\phi)^{\kappa}),$$

where d_{ij} is the distance between the i^{th} and j^{th} locations, can be used. The simplifying assumption of a fixed smoothing parameter ($\kappa = 1$) can be considered for the spatial covariance matrix. Additional covariance functions can be included, including the Matérn correlation function and the disc function.

An advantage of the GLSM model is that the spatial bandwidth can be changed through the correlation parameter; smaller ϕ allows for a more variable surface for the effect [45, 46]. However, estimation in these models is often time-consuming, especially when trying to estimate some of the spatial parameters (e.g. ϕ). In addition, an unadjusted model with n patients requires the inversion and storage of an $n \times n$ matrix, which can be computationally demanding in certain settings [61, 62].

Spline-based Model

A spline-based model is a semi-parametric approach that uses a smoothing operator to incorporate location data [46]. Nychka et al. (1998) proposed the reduced knot model, or low-rank

kriging, which is an alternative to the kriging-type model by Diggle et al. [63]. Low-rank kriging is a geostatistical approach that combines interpolation, or kriging, with modeling of spatial variation in disease risk, for example. The spatially-correlated joint random effects can be used to capture the geographical variation in the disease risk and predict the spatial effects at new locations.

A Bayesian generalized additive model (GAM), or geoadditive model, approach can be used to achieve this. This method was first proposed by Kammann and Wand (2003) and combined geoadditive models, specifically the low-rank kriging formulation, with additive models [64]. French and Wand (2004) use a logistic GAM, with a mixed effects model representation, to model the spatial variation in cancer risk in Cape Cod, Massachusetts using splines. This approach accounts for incomplete covariate data when fitting a GAM [62]. Lawson (2013) define a Bayesian GAM for count data to model the risk of respiratory cancer mortality in Ohio. A low-rank kriging model is fitted with a planar trend in centroids (covariate) along with an additive spline term [46].

Using simulated case-control data to model spatial variation in disease risk, Wheeler and Calder (2016) define the low-rank kriging model as follows:

$$\begin{aligned} \log\left[P(y_i = 1)/P(y_i = 0)\right] &= \beta_0 + \sum_{j=i}^{p} \beta_j x_{ij} + \sum_{k=1}^{K} \psi_k C\{||s_i - \boldsymbol{\kappa}_k||\} \\ &= x_i'\boldsymbol{\beta} + z_i'\boldsymbol{\psi}, \end{aligned} \tag{4.7}$$

where x_{ij} are covariates, $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K$ are the K selected knot locations to represent the observed locations $(s_1, \ldots, s_n)$ and $\boldsymbol{z} = [C\{||s_i - \boldsymbol{\kappa}_k||/\rho\}]_{1\leq i\leq n, 1\leq k\leq K}$ [61]. The covariance function is defined as $C\{d\} = (1 + |d|)e^{-|d|}$ and comes from the Matérn family of covariance functions. The Gaussian random effect ψ_k is specified as follows:

$$\begin{aligned} \boldsymbol{\psi} &\sim N(\mathbf{0}, \tau\boldsymbol{\omega}^{-1}), \\ \boldsymbol{\omega} &= [C\{||\boldsymbol{\kappa}_i - \boldsymbol{\kappa}_k||/\rho\}]_{1\leq i,k\leq K}\,, \end{aligned}$$

where $\boldsymbol{\omega}$ is the covariance matrix, τ is the precision and ρ is the spatial range parameter. The model was used to predict spatially-varying odds ratios onto a grid, where the spatial odds at grid cell j was defined as $\theta_j = \exp(z_j'\boldsymbol{\psi})$.

French and Wand (2004) showed that the GAM has a GLMM representation using a reparam-

eterization of the following parameters:

$$\boldsymbol{z}_* = \boldsymbol{z}\boldsymbol{\omega}^{-1/2},$$
$$\boldsymbol{\psi}_* = \boldsymbol{\omega}^{1/2}\boldsymbol{\psi},$$

where $\text{Cov}(\boldsymbol{\psi}_*) = \tau\boldsymbol{I}$ with precision τ. Lawson (2013) and French and Wand (2004) fix the spatial range parameter ρ as the maximum observed inter-point distance,

$$\hat{\rho} = \max_{1 \leq i,j \leq n} \{||s_i - s_j||\},$$

for simplicity in fitting the GLM and to reduce the computational burden. Specifically, by fixing ρ and using the GLMM reparameterization, the covariance matrix can be precomputed and standard statistical software can be used, respectively.

French and Wand (2004) comment that the smoothness of their estimate was dependent on the choice of ρ but that the scale of their final map was insensitive. In addition, Lawson (2013) showed that the maximum interpoint distance produced very smooth surface effects and believes that an improved model fit could be achieved by estimating ρ [46, 62]. Accordingly, Wheeler and Calder (2016) considered a range of candidate ρ values from 0.1 to the maximum interpoint distance (2.76), in increments of 0.1, and selected the best model using the DIC. The authors discuss the importance of estimating the spatial range parameter since it is used to control the degree of smoothing in the spatial component. Estimating ρ in the model using MCMC would increase both the complexity of the model and computation time. A spline-based model approach is faster, especially when ρ is fixed, since a smaller covariance matrix ($K \times K$) corresponding to the number of knots is stored and inverted compared to a GLSM [61, 62, 64].

4.2 Methods

In Chapters 2 and 3, we incorporated site into the response-adaptive randomization framework to account for site-based response variability and inter-center variability in treatment effectiveness. In addition to these sources of variability across sites, treatment differences can also be attributed to social determinants of health. Here, we consider situations where treatment effectiveness may

vary throughout space in a clinical trial. However, SDH are often not measured or there is not enough information or available data to capture these phenomena. Accordingly, geographic location, specifically patient residential location, will be used as a proxy for SDH since they are often associated. For example, Wheeler et al. (2017) found an association between colonoscopy screening adherence and an area-level socioeconomic status index, where the SES index is known to have spatial components. Variables at both the census block group- and tract-levels were included in the index and it was shown that home ownership, race and income were among the most important [65]. Therefore, it is necessary to account for spatially-varying treatment effectiveness in adaptive allocation designs.

We update the allocation weights with estimated success probabilities using the subsequent patient's residential location to account for more coarse location-based variability when adapting. This allows us to randomize patients to the treatment group likely to perform best at their residence. One simple approach to calculate the predicted probabilities of success at a particular location is to use IDW interpolation. After estimating spatially-varying success probabilities for the current patients using a GAR model (Eqn. 4.5), IDW is used to predict the conditional probabilities to a new location. One of the drawbacks of IDW is that it does not account for spatial variability of the distance-decay relationship. Instead, a GLSM or low-rank kriging model (Eqn. 4.7) can be used to model the geographical variation in treatment successes, thus accounting for these unobserved or unknown SDH, and to interpolate the spatially-varying treatment effects. The LRK model was selected due to the significant computation time required to implement the GLSM model into adaptive allocation simulations.

One use of the low-rank kriging model in Spatial epidemiology is to model the spatial variation in disease risk. The splines are employed to estimate a spatially structured random effect to capture any unmeasured confounding while adjusting for covariates. Instead, we model the spatial variation in the odds of a favorable response to a particular treatment (treatment success). The splines are used in a novel approach to estimate spatially-varying treatment effects. Our objective is to compare an an extension of response-adaptive randomization that incorporates spatial information to account for SDH with existing allocation approaches.

4.2.1 LARA Model

We propose the Location-Adjusted Response-Adaptive (LARA) approach to account for spatial variability in treatment effectiveness not previously accounted for in past allocation methods. LARA is an extension of the SARA approach when more precise geographical information is available. Specifically, we incorporate the observed patient residential locations into a Bayesian low-rank kriging model [61, 62, 63]:

$$\begin{aligned}\text{logit}(p_i) &= \alpha_0 + \left(\beta_1 + \sum_{k=1}^{K} \beta_{11k} C\{||s_i - \boldsymbol{\kappa}_k||\}\right) T_{i1} + \left(\beta_2 + \sum_{k=1}^{K} \beta_{22k} C\{||s_i - \boldsymbol{\kappa}_k||\}\right) T_{i2} + \sum_{j=1}^{p} \alpha_j X_{ij} \\ &= \boldsymbol{X_i'\alpha} + \boldsymbol{z_i'\beta}.\end{aligned} \tag{4.8}$$

Here, α_0 is the overall intercept, T_{i1}, T_{i2} are treatment group 1 and 2 indicators for the i^{th} patient, $i = 1, \ldots, n$, and β_{11k}, β_{22k} are the spatially-varying joint treatment random effects for the k knots selected. Let $\boldsymbol{z} = [C\{||s_i - \boldsymbol{\kappa}_k||/\rho\}]_{1 \leq i \leq n,\ 1 \leq k \leq K}$, where the covariance function is defined as $C\{d\} = (1 + |d|)e^{-|d|}$. The spatial range parameter ρ controls the degree of smoothing and is specified under the assumption the degree of smoothing is the same for both treatment groups.

The separate random components for each treatment group allow treatment effectiveness to vary over space for both groups. Specifically, β_1, β_2 are (optional) overall treatment effects for each group while β_{11k}, β_{22k} are interpreted as spatially-varying local deviations from the overall treatment effects. The X_j term are p (optional) subject- and/or location-specific covariates, with corresponding regression coefficients α_j, that can be incorporated into the model.

The set of k knot locations $\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_k\}$ are representative of the observed patient residential locations $\{s_1, \ldots, s_n\}$. The knots were selected using a swapping algorithm developed by Johnson et al. (1990) which minimizes a geometric space-filling criterion. This space-filling algorithm is implemented in the *cover.design* {fields} function in R [66]. French and Wand (2004) used $k = 33$ knots but express interest in further simulations to assess varying the number of knots. A more smooth estimated surface may be produced with a larger knot size but requires larger matrices to be stored and inverted. Wheeler and Calder (2016) use $k = 25$ knots in their models [61, 62]. As a model is fit for each patient randomized, we selected $k = 20$ knots for each model to limit the

computation time.

The priors for the regression coefficient and variances for the spatially-varying random effects are specified as follows. The prior on the overall intercept is an informative normal: $\alpha_0 \sim N(0, \tau_{\alpha_0})$, where precision $\tau_{\alpha_0} = 1/\sigma^2 = 0.4$ and σ^2 is the variance. The priors for the joint random effects are specified as

$$\boldsymbol{\beta_{11}}, \boldsymbol{\beta_{22}} \sim N(\mathbf{0}, \tau_\beta \boldsymbol{\omega}^{-1}),$$

where τ_β is the precision and $\boldsymbol{\omega}$ is a square covariance matrix of the pairwise distances between the knot locations,

$$\boldsymbol{\omega} = [C\{||\boldsymbol{\kappa}_i - \boldsymbol{\kappa}_k||/\rho\}]_{1 \leq i,\ k \leq K}\,.$$

The priors for the precision τ_β of the spatially-varying random effects are specified using the variance $\sigma^2 = 1/\tau_\beta$, where $\sigma^2 \sim \text{Uniform}(0, 2)$. The prior for regression coefficients for any additional (optional) covariates could be specified as a vague Gaussian: $\alpha_j \sim N(0, \tau_\alpha)$, where the precision $\tau_\alpha \sim \text{Gamma}(0.1, 0.0001)$. The prior distributions for the intercept and random effects are similar to, but less vague than, those used in Wheeler and Calder (2016) and Lawson (2013) [61, 46].

To reduce the computational burden of the simulations, we decided to fix the spatial range parameter in our low-rank kriging model. A smaller ρ produces a more coarse surface effect and should be used when the true surface is expected to resemble this [62]. A simple simulation comparing different choices for ρ showed that values smaller than the maximum interpoint distance generally resulted in smaller DIC values and lower MSE, or prediction error (results omitted). Accordingly, ρ was fixed as one-quarter of the maximum interpoint distance,

$$\hat{\rho} = \frac{1}{4} \times \max\{||s_i - s_j||\}\ \forall i, j.$$

The observed distances have to be recalculated and the fixed spatial range parameter $\hat{\rho}$ updated each time a new patient is randomized to predict the treatment effects for the subsequent patient.

As each patient is recruited, the accrued data up to that point, including treatment assignments,

observed patient outcomes and residential locations, are fit using the LARA model (Eqn. 4.8) to predict the spatially-varying treatment effects at the residential location of the subsequent patient. The predicted treatment effect for the t^{th} group at the residential location for the $(m+1)^{st}$ patient is

$$\hat{\theta}_{m+1,t} = \boldsymbol{z}'_{\boldsymbol{m+1}}\hat{\boldsymbol{\beta}}. \tag{4.9}$$

The regression coefficients and predicted spatial treatment effects (Eqn. 4.9) are used to calculate the estimated success probabilities for each treatment as follows:

$$\begin{aligned} \hat{p}_i^1 &= \frac{\exp\{\hat{\alpha}_0 + \hat{\theta}_{i1} + \boldsymbol{X}'_{\boldsymbol{i}}\hat{\boldsymbol{\alpha}}\}}{1 + \exp\{\hat{\alpha}_0 + \hat{\theta}_{i1} + \boldsymbol{X}'_{\boldsymbol{i}}\hat{\boldsymbol{\alpha}}\}}, \\ \hat{p}_i^2 &= \frac{\exp\{\hat{\alpha}_0 + \hat{\theta}_{i2} + \boldsymbol{X}'_{\boldsymbol{i}}\hat{\boldsymbol{\alpha}}\}}{1 + \exp\{\hat{\alpha}_0 + \hat{\theta}_{i2} + \boldsymbol{X}'_{\boldsymbol{i}}\hat{\boldsymbol{\alpha}}\}}. \end{aligned} \tag{4.10}$$

The predicted probabilities of success account for spatial variability in treatment effectiveness. The optimal allocation weights (Eqn.2.3) are updated using these conditional probabilities to skew allocation towards the treatment group performing 'best' closest to where each patient resides. Due to our focus on recapturing the true spatially varying treatment effects, when they exist, to account for variability in treatment effectiveness, we do not estimate the overall treatment effects β_1, β_2 nor do we include covariates or additional random effects in our simulation study.

4.3 Simulation Study

A simulation was conducted to evaluate the performance of the location-adjusted response-adaptive design, which accounts for spatial variability in treatment effectiveness, relative to existing allocation designs. In a real clinical trial, patient addresses can be geocoded in R (see *geocode* {ggmap} function) to the corresponding geographic coordinates and census tract. If no residential location is available, using the centroid of the zip code is one approach to include the patient data in the randomization process. However, to simulate the patient data from a hypothetical clinical trial, a study area with spatial coordinates $u, v \in (0, 1)$ was selected and broken into g grid cells, with each grid cell assigned a treatment success probability and effect size to create spatial variation.

The scenarios considered and the corresponding sample size determinations are shown in Section 4.3.1 and were selected to be characteristic of a typical multi-center trial.

4.3.1 Scenarios Considered

In Scenario 1 (Figure 4.1), the unit square was split into $g = 100$ grid cells and the effect size ($\delta = p_1 - p_2$) fixed at 15%. The treatment 1 success rate was set at 10% ($p_1 = 0.10, p_2 = 0.25$) in the 25 grid cells in the northwest corner and at 25% elsewhere ($p_1 = 0.25, p_2 = 0.10$). In Scenarios 2-4 (Figures 4.2-4.4), this 'true' treatment success probability surface was extended to include a gradation in success rates across the study area to represent increasingly more continuous surfaces. Scenarios 5 and 6 represent a 'spatial null' hypothesis where there is no spatial variability in the success probabilities. There is no treatment effect in Scenario 5 (Figure 4.5) while there is a 15% effect size ($p_1 = 0.25, p_2 = 0.10$) set for Scenario 6 (Figure 4.6). In Scenario 7 (Figure 4.7), there is no treatment effect but the success rates are gradually increasing from 75% in the northwest corner to 90% in the southern and eastern borders of the unit square. Scenario 8 (Figure 4.8) generates a continuous surface (Scenario 4) for treatment group 1 but a low success rate ($p_2 = 0.10$) with no spatial variability for treatment 2.

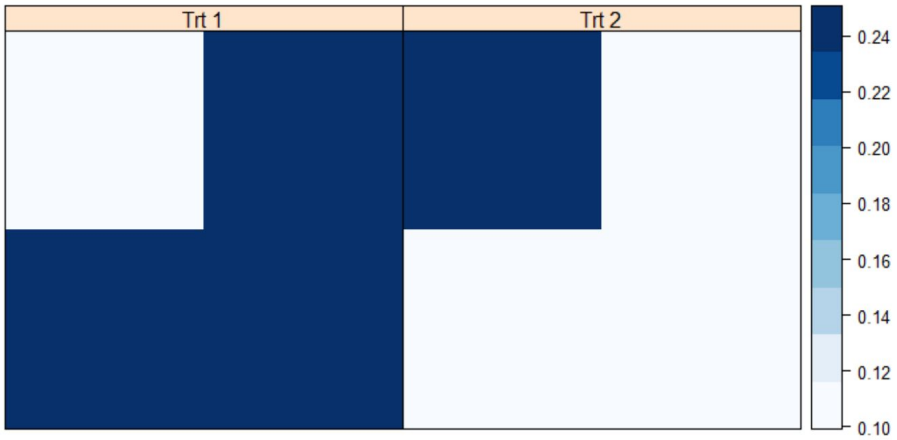


Figure 4.1: Scenario 1 True Surface - Discrete Treatment Effect, Spatial Variability

Sample Size Justification: $p_1 = 0.10, p_2 = 0.25, \delta = 0.15, n = 200$

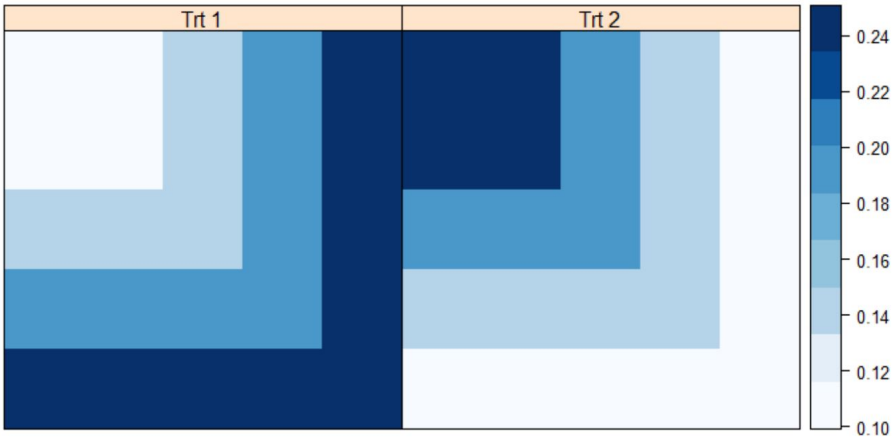


Figure 4.2: Scenario 2 True Surface - Continuous Treatment Effect, Spatial Variability

Sample Size Justification: $p_1 = 0.10, p_2 = 0.25, \delta = 0.15, n = 200$

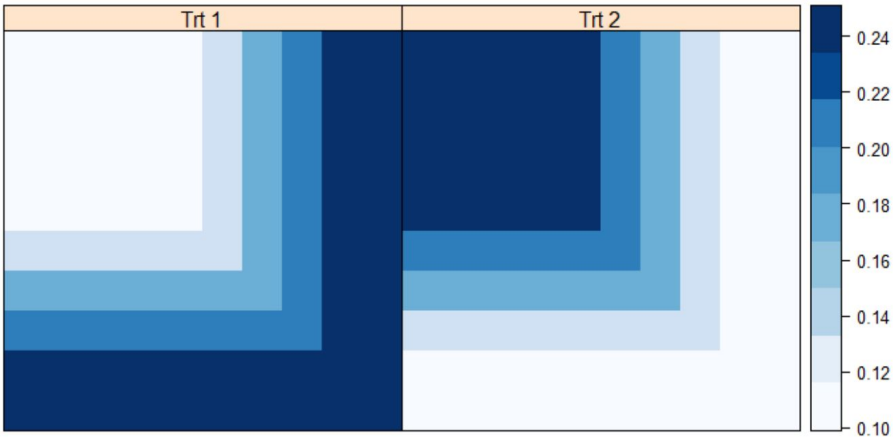


Figure 4.3: Scenario 3 True Surface - Continuous Treatment Effect, Spatial Variability

Sample Size Justification: $p_1 = 0.10, p_2 = 0.25, \delta = 0.15, n = 200$

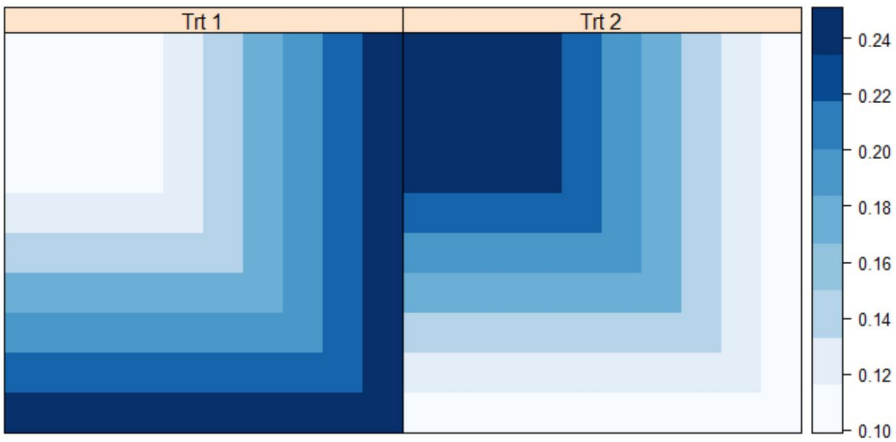


Figure 4.4: Scenario 4 True Surface - Continuous Treatment Effect, Spatial Variability

Sample Size Justification: $p_1 = 0.10, p_2 = 0.25, \delta = 0.15, n = 200$

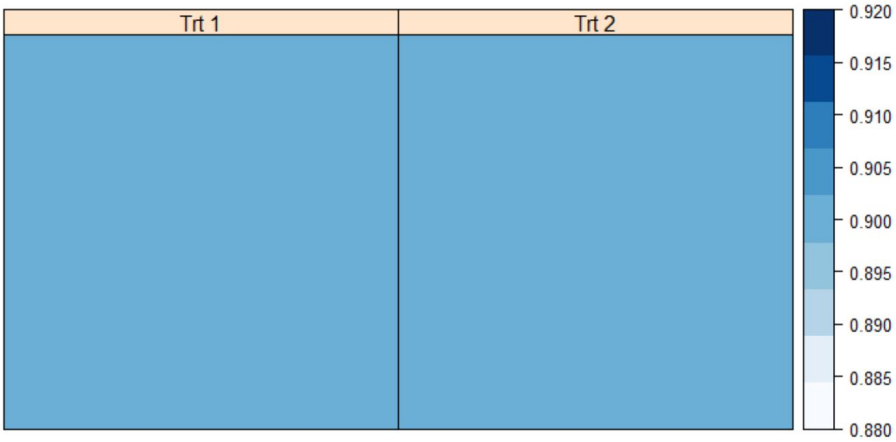


Figure 4.5: Scenario 5 True Surface - No Treatment Effect, No Spatial Variability

Sample Size Justification: $p_1 = 0.90, p_2 = 0.80, \delta = 0.1, n = 398$

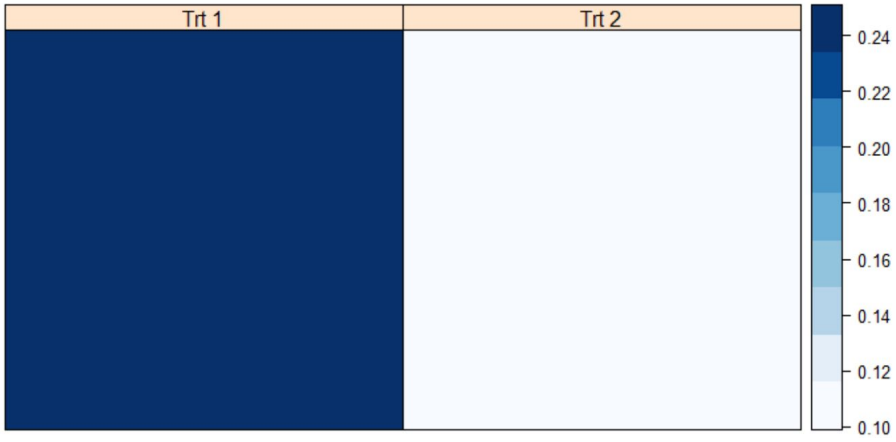


Figure 4.6: Scenario 6 True Surface - Treatment Effect, No Spatial Variability

Sample Size Justification: $p_1 = 0.25, p_2 = 0.10, \delta = 0.15, n = 200$

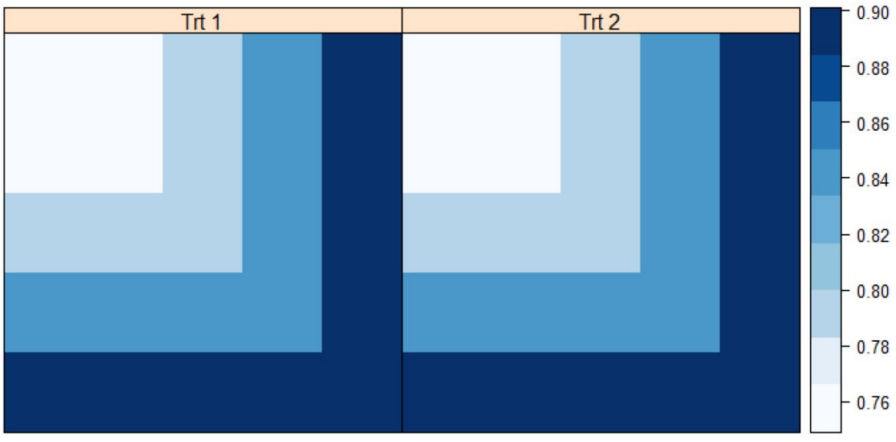


Figure 4.7: Scenario 7 True Surface - No Treatment Effect, Spatial Variability

Sample Size Justification: $p_1 = 0.9, p_2 = 0.8, \delta = 0.10, n = 398$

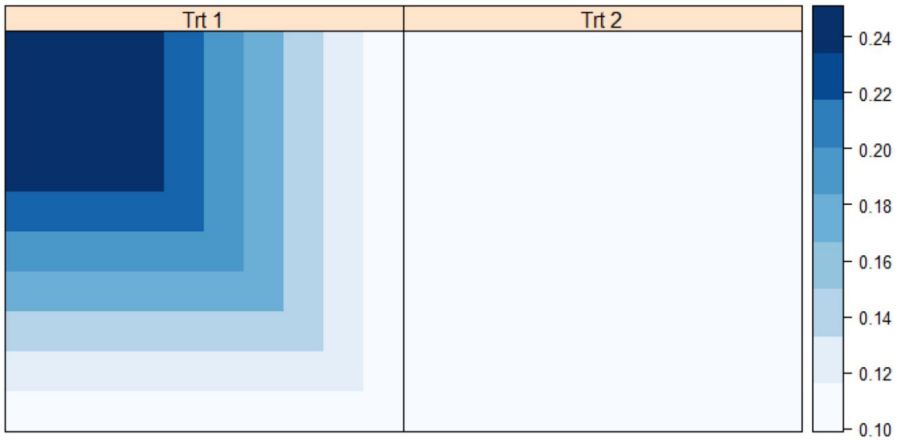


Figure 4.8: Scenario 8 True Surface - Treatment Effect, No Spatial Variability for Treatment 2

Sample Size Justification: $p_1 = 0.25, p_2 = 0.10, \delta = 0.15, n = 200$

4.3.2 Simulation Settings

Each subject was randomly assigned to a point location within one of the g grid cells using a homogeneous Poisson point process (*rpoispp* {spatstat} function [67]). Specifically, a realization of a uniform Poisson process, which assumes Complete Spatial Randomness, was generated by specifying an intensity function $n + 100$. We randomly sampled n patients without replacement from the total $n+100$ points. As each patient was recruited in the simulated trial, a $U(0, 1)$ variate was drawn and compared to the previous treatment group 1 allocation weight to determine the assignment. If the value was less than or equal to the previous weight (Eqn. 2.3), the patient was assigned to treatment group 1, otherwise they were allocated to the second treatment group. A hard-lead required that balanced allocation be performed for the first $m = 20$ patients to prevent estimator instability or convergence issues. The patient outcomes were then generated from a Bernoulli distribution using the success probability for the g^{th} grid cell corresponding to where the patient resides. After each patient is randomized to a treatment group, the spatial range parameter $\hat{\rho}$ was calculated and new knots selected. Accordingly, the distance and covariance matrices $\boldsymbol{z}, \boldsymbol{\omega}$ must be precomputed.

The simulations were powered for an analysis not accounting for location-based variability by ignoring clustering in the sample size determination (Eqn. 2.15). This is similar to previous chapters and will allow us to see how much power LARA gains back from both balanced and RA randomization. The sample sizes for the scenarios were calculated without complete information, specifically the true spatial variability in the success rates. Instead, a sample size justification was performed for the expected effect size assuming there was no spatial variation. In the case where there was no treatment effect (Scenarios 5, 7), the sample size was calculated with a small effect size ($\delta = 0.1$) to allow a reasonable sample size to be recruited.

The joint posterior distribution of the regression coefficient and joint random effects were estimated using MCMC techniques implemented through the BRugs interface in R. We estimated the posterior median of the model parameters using 4000 iterations for burn-in and 1000 iterations from two parallel Markov chains. The two Markov chains were generated without thinning because it has been shown to be unnecessary and inefficient, particularly due to the increased computation time in simulation studies [43]. The initial values were specified for the overall intercept, joint random effects ($\boldsymbol{\beta_{11}}, \boldsymbol{\beta_{22}}$) and the precision for the random effects ($\tau_{\beta_{11}}, \tau_{\beta_{22}}$) to ensure they are reasonable and within the bounds of the priors used. For the first subject randomized following the lead-in (e.g. patient 21), fixed starting values were used to uniquely initialize each of the Markov chains. For both chains, a fixed starting value of zero was used for the joint random effects. For subsequent models, the initial values are set as the posterior medians of the previous patient for the first chain and as 0.95 times the posterior estimates for the second chain.

To evaluate the proposed allocation design, we use the operating characteristics established by Rosenberger et al. (2008), namely balance (allocation proportion), efficiency (power, type I error) and ethics (total number of treatment successes) [19]. We are particularly interested in the mean total number of successes/failures ($\pm$ standard error), expected allocation proportion for treatment group 1 ($E\left[\frac{n_1}{n_1+n_2}\right]$), empirical power (95% Wald Confidence Interval) and mean squared error. In addition, we compare the computation time for the proposed approach relative to Chapter 3.

We expect that the Spatial Bayesian extensions will require even more computation time due to the added complexity of the the model relative to the Bayesian GLMMs in Chapter 2. For the LARA model, we calculate the average simulation time (seconds) per subject to compare the effect of the increasing complexity of the models on computation times. Balanced allocation and

RA randomization are not considered because these models are fit almost instantaneously for each patient. The simulation times represent the amount of time (seconds) to fit each model and update the allocation probabilities to randomize the subsequent patient in a trial. The times only account for the $n-20$ patients adaptively randomized after the initial lead-in. Each trial was run separately and the average seconds/patient along with standard error of the simulation times were reported.

We compared the performance of LARA (Eqn. 4.8) to both balanced allocation and RA randomization (Eqn. 2.1). The balanced and RA designs do not account for spatial variation in treatment effectiveness. We hypothesized that LARA, which accounts for location-based variability in treatment responses and effectiveness, will reduce the variability in randomization, thus reducing improper patient allocation. There were 1000 trials simulated separately for each scenario and the results aggregated. Due to our focus on ethics, we only considered Optimal allocation. For a single trial scenario, the following metrics were stored: sample size, number of patients assigned to each treatment, number of successes for each treatment, total number of successes, simulation time and mean squared error. After all patients were recruited in a trial, the proportions of success were compared using a Pearson's chi-squared or Fisher's Exact test to assess the treatment effect ($H_0 : p_1 = p_2$). The expected cell counts were checked (> 5) to ensure the assumptions of a chi-squared test were satisfied. Site, or specifically location here, was ignored when calculating the empirical power to be consistent with the previous chapters. This naive approach is also convention when analyzing multi-center trials.

We assessed the quality of the estimated conditional success probabilities, and ultimately the predicted treatment effects, for RA randomization and the LARA design using mean squared error. After all patients were randomized in a trial, the MSE was calculated using the true and estimated success probabilities (Eqn. 4.10) for each treatment separately. The true success probabilities for each grid cell correspond to the scenario being considered (Section 4.3.1). Since the spline model results in smooth estimates, it is important to see how well LARA performs in replicating the true success probability surface. We hypothesized that LARA would result in better predictions since RA randomization ignores the spatial variation in treatment effectiveness and instead averages the success rates across the study area. For the t^{th} treatment, $t = 1, 2$, the squared differences between the true and estimated success probabilities were aggregated across all g grid cells and n subjects into a single metric as follows:

$$MSE_t = \frac{1}{n}\sum_{g=1}^{k}\sum_{i=1}^{n}\left(\theta_g - \hat{\theta}_{ig}\right)^2 I(i \in g). \tag{4.11}$$

Here, θ_g is the true population success probability for the g^{th} grid cell and t^{th} treatment, and $\hat{\theta}_{ig}$ is the estimated success probability for the i^{th} subject located in the g^{th} grid cell. Balanced allocation does not make predictions and therefore we are unable to calculate mean squared error for this randomization approach.

We plotted the true and estimated surfaces for both RA randomization and the LARA design to visually compare how both methods are estimating the conditional probabilities of success used to update the allocation probabilities. For LARA, the posterior estimates calculated from all patient data were used to interpolate the spatially varying treatment effects (Eqn. 4.9) to a fixed 10×10 grid after all patients were randomized. The predicted treatment effects were then used to calculate the estimated success probability surface. For RA randomization, this equates to calculating the success probabilities for each treatment using the estimates of the intercept and slope, irrespective of the location. The estimated surfaces were aggregated and averaged across 1000 trials. We hypothesized that LARA would reduce the variability in allocation because it incorporates a variable success probability surface into adaptive randomization whereas the RA approach ignores the spatial information.

4.3.3 Convergence Study

In a real world application, convergence would be assessed for the fitted model after each new patient is randomized using one of the convergence diagnostic test statistics. In practice, visual assessment of the trace plots should accompany the convergence statistics. However, this is not feasible in simulation studies, particularly when the model is refitted for each new patient randomized. Accordingly, convergence was not assessed during our simulation study due to the increased computation burden from storing and processing the raw chain data necessary to calculate the diagnostic test statistic.

A small simulation was performed to examine the robustness of the inferences with and without monitoring convergence. We simulated 1000 trials and for each, estimated the model parameters

using 4000 iterations for burn-in and 1000 iterations from two parallel chains with no thinning. The two Markov chains were initialized similarly, where the previous patient's converged posterior estimates were used to start the chain close to where it left off. However, when convergence was not achieved, the previous patient's parameter estimates that converged were used to interpolate the treatment effects at the new residential location, calculate the location-specific success probabilities and initialize the subsequent chains. Gelman and Rubin's convergence diagnostic $\hat{R}$, or the potential scale reduction factor, was used as the criteria for evaluating convergence. Adaptation continued with the more strict cutoff of $\hat{R} < 1.1$ for the parameter estimates considered [39, 40], otherwise the previous patient's posterior estimates were used to interpolate the treatment effects and update the allocation probabilities. However, for the first patient after the lead-in or in the case where the previous patient's parameters did not converge, equal allocation was used until convergence was achieved. We examined the robustness of the mean total number of treatment successes (SE), expected allocation proportion1 for treatment 1 (SE), empirical power (95% confidence interval), MSE for each treatment, mean simulation time (seconds) per patient and the average number of non-convergence warnings per trial. The simulation time will represent the expected amount of additional time to fit and assess the model for convergence per patient.

4.3.4 Simulation Algorithm

The basic algorithm for our simulation study is as follows.

1. Hard lead-in for the first m patients, where individuals are allocated equally between the two treatment groups.

2. Let $\mathscr{F}_i$ be accrued information which includes treatment assignments, observed patients outcomes and geographic information. Use $\mathscr{F}_i$ for patients $i = 1, \ldots, m, m+1, \ldots, n$ to begin adaptation, starting with patient $m+1$.

3. Using the residential location for patient $m+1$, fit the Bayesian low-rank kriging model (Eqn. 4.8) to predict the treatments effects

$$\hat{\theta}_{m+1,t} = \boldsymbol{z}'_{\boldsymbol{m+1}}\hat{\boldsymbol{\beta}},$$

and calculate the conditional probability of success for the t^{th} treatment as follows:

$$\hat{p}_i^t = \frac{\exp\{\hat{\alpha}_0 + \hat{\theta}_{it} + \boldsymbol{X_i'}\hat{\boldsymbol{\alpha}}\}}{1 + \exp\{\hat{\alpha}_0 + \hat{\theta}_{it} + \boldsymbol{X_i'}\hat{\boldsymbol{\alpha}}\}},\ t = 1, 2.$$

4. Using the estimated success probabilities $\hat{p}_i^1, \hat{p}_i^2$, update the allocation weights W_1, W_2 and assign the next patient to a treatment.

$$W_1(\mathscr{F}_i) = \sqrt{\hat{p}_i^1}/\sqrt{\hat{p}_i^1 + \hat{p}_i^2},\ W_2(\mathscr{F}_i) = 1 - W_1.$$

5. Repeat Steps 2-4 to randomize patients $m + 2, \ldots, n$.

For a single scenario, this algorithm is repeated for $N = 1000$ trials.

4.4 Results

The scenarios considered are shown in Section 4.3.1.

4.4.1 Mean Total Number of Treatment Successes

The mean total number of treatment successes are displayed in Table 4.1. When there are discrete true spatially-varying surfaces with a medium effect size (Scenario 1), the adaptive designs have more expected successes, with slightly less improvement for LARA. LARA has similar variability in treatment successes to balanced allocation while RA randomization has the highest variability. When the success probability surfaces become increasing more continuous (Scenarios 2-4), all three methods perform comparably with respect to treatment successes and variability in successes, with slightly more improvement for LARA.

When there is no spatial variability and a null effect size (Scenario 5), the three methods perform similarly, with slightly lower variability in successes for LARA. For a medium effect size (Scenario 6), the two adaptive designs have a notable increase in total treatment successes, with the improvement slightly lower for LARA. However, LARA has lower variability than RA randomization, similar to balanced allocation.

When there is spatial variability but a null effect size (Scenario 7), the three approaches are comparable with respect to treatment success and variability, with slightly less improvement in both for LARA. When there is only spatial variability in success rates for treatment 1 (Scenario 8), the adaptive designs results in more treatment successes relative to the BC, but with less improvement for LARA. In this case, LARA has similar variability in treatment successes to balanced allocation while RA randomization has the highest variability.

Table 4.1: Mean Total Number of Treatment Successes (SE)

Scenario	**n**	**BC**	**RA**	**LARA**
1	200	34.8 (5.5)	36.1 (6.5)	35.9 (5.4)
2	200	34.8 (5.4)	35.0 (5.6)	35.3 (5.3)
3	200	34.9 (5.4)	34.9 (5.5)	35.4 (5.3)
4	200	34.9 (5.3)	34.9 (5.4)	35.2 (5.3)
5	398	358.3 (5.8)	358.3 (5.8)	358.3 (5.7)
6	200	34.7 (5.5)	40.4 (8.3)	36.8 (5.4)
7	398	335.2 (7.0)	335.2 (7.0)	334.9 (7.3)
8	200	26.5 (4.8)	28.0 (6.0)	27.3 (4.8)

4.4.2 Expected Allocation Proportion - Treatment Group 1

The allocation proportions are displayed in Table 4.2. When there are discrete true spatially-varyying surfaces with a medium effect size (Scenario 1), the adaptive designs have a larger allocation ratio relative to balanced case, though the improvement is lower for LARA. This explains the increase in expected treatment successes in the adaptive designs. LARA has notably lower variability in allocating relative to RA randomization and similar to the balanced case. When the true success probability surfaces becomes increasing more continuous (Scenarios 2-4), the pattern is similar where the adaptive designs have larger allocation proportions relative to balanced allocation, with less adaptation for LARA. In addition, LARA maintains lower variability in allocating than RA randomization. As the surface becomes increasingly continuous, the advantage in adapting for the RA/LARA approaches diminishes, particularly for LARA.

When there is no spatial variability and a null effect size (Scenario 5), the three methods perform similarly with equal allocation, as expected, but with nominally more variability for RA randomization. For a medium effect size (Scenario 6), the two adaptive designs have notably larger allocation ratios, with the improvement slightly lower for LARA. However, LARA has lower variability than RA randomization, similar to the BC.

When there is spatial variability but a null effect size (Scenario 7), the three approaches are comparable with equal allocation, as expected, but with slightly more variability for the RA approach. When there is only spatial variability in success rates for treatment 1 (Scenario 8), the adaptive designs results in larger allocation ratios, however, there is less improvement for LARA. LARA has similar variability in allocating relative to balanced allocation while RA randomization has the highest variability.

Table 4.2: E$[n_1/(n_1+n_2)]$ (SE)

Scenario	**n**	**BC**	**RA**	**LARA**
1	200	0.50 (0.03)	0.58 (0.24)	0.53 (0.04)
2	200	0.50 (0.03)	0.53 (0.25)	0.51 (0.05)
3	200	0.50 (0.03)	0.51 (0.25)	0.51 (0.04)
4	200	0.50 (0.03)	0.51 (0.25)	0.50 (0.04)
5	398	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)
6	200	0.50 (0.03)	0.69 (0.23)	0.56 (0.04)
7	398	0.50 (0.02)	0.50 (0.03)	0.50 (0.02)
8	200	0.50 (0.03)	0.61 (0.29)	0.53 (0.05)

4.4.3 Empirical Type I Error and Power

The empirical type I error (α) and power ($1-\beta$) results are displayed in Table 4.3. We compare the type I error rates for Scenarios 5 and 7 where there is a null effect size. The three methods perform comparably, where there is a slight inflation in the type I error for LARA irrespective of the presence (Scenario 7) or absence (Scenario 5) of spatial variability in the success rates.

Overall, there is a notable loss of power for all three methods when there is spatial variation in

the true success probabilities. When there are discrete true spatially-varying surfaces and a medium effect size (Scenario 1), LARA has slightly higher power than balanced allocation while there is a notable loss of power for RA randomization. When the true success probability surfaces becomes increasing more continuous (Scenarios 2-4), the pattern is similar where LARA outperforms the other two approaches. As the surface becomes increasingly continuous, there is a loss of power for all three approaches, particularly balanced allocation, and the advantage in power for LARA diminishes. When there is only spatial variability in success rates for treatment 1 (Scenario 8), the BC/LARA approaches have comparable power, with a slight loss of power for LARA. RA randomization has a notable loss of power relative to the other two methods. When there is no spatial variation in success rates (Scenario 6), the power for the BC/LARA approaches reach nominal levels, with more improvement for balanced allocation, and a notable loss of power for RA randomization.

Table 4.3: Empirical Type I Error and Power (95% Confidence Interval)

Scenario	**n**	**Metric[1]**	**BC**	**RA**	**LARA**
1	200	$1-\beta$	0.27 (0.24, 0.29)	0.19 (0.16, 0.21)	0.28 (0.26, 0.31)
2	200	$1-\beta$	0.09 (0.07, 0.11)	0.07 (0.05, 0.08)	0.10 (0.08, 0.11)
3	200	$1-\beta$	0.06 (0.04, 0.07)	0.06 (0.04, 0.07)	0.07 (0.05, 0.08)
4	200	$1-\beta$	0.04 (0.02, 0.05)	0.05 (0.03, 0.06)	0.06 (0.05, 0.07)
5	398	α	0.05 (0.03, 0.06)	0.05 (0.03, 0.06)	0.06 (0.04, 0.07)
6	200	$1-\beta$	0.82 (0.79, 0.84)	0.53 (0.50, 0.56)	0.80 (0.78, 0.83)
7	398	α	0.05 (0.03, 0.06)	0.05 (0.03, 0.06)	0.06 (0.04, 0.07)
8	200	$1-\beta$	0.28 (0.25, 0.31)	0.12 (0.10, 0.14)	0.27 (0.24, 0.29)

[1] Empirical Type I Error (α) or Power ($1-\beta$)

4.4.4 Mean Squared Error (MSE) by Treatment

The mean squared error results are displayed in Table 4.4. Overall, when there is spatial variability in the true success probabilities and an effect size (Scenarios 1-4, 7-8), LARA has a lower MSE for both treatments relative to RA randomization. There is less of an improvement for LARA when there is a null effect size and high success rates (Scenario 7). When there is no spatial

variability in the true success rates, the two methods are comparable for both null (Scenario 5) and medium effect sizes (Scenario 6), with slight improvement for LARA when there is a medium effect size.

Table 4.4: Mean Squared Error (MSE) by Treatment

Scenario	**Design**	**MSE 1**	**MSE 2**
1	RA	0.011	0.010
	LARA	0.008	0.007
2	RA	0.011	0.010
	LARA	0.006	0.006
3	RA	0.011	0.010
	LARA	0.006	0.006
4	RA	0.010	0.010
	LARA	0.006	0.005
5	RA	0.001	0.001
	LARA	0.001	0.001
6	RA	0.006	0.005
	LARA	0.006	0.004
7	RA	0.005	0.005
	LARA	0.004	0.004
8	RA	0.009	0.005
	LARA	0.005	0.003

4.4.5 Simulation Time per Patient

The results of the simulation times and standard errors are displayed in Table 4.5. Overall, each scenario takes a significant amount of time to implement the LARA approach for each patient. Although the times are reported per patient, thus removing the effect of different sample sizes, scenarios with larger sample sizes ($\delta = 0$) have increased simulation times. The LARA approach

has a notable increase in computation time due to the added complexity of the model compared to the SARA approaches in Chapter 3 (Tables 3.24-3.26).

In a real life application, these computation times are feasible as only one patient is adaptively randomized at a time for a single trial. However, the computation times represent a significant burden when performing a simulation study with n patients, 1000 trials and a variety of scenarios. For example, a single simulated clinical trial with a continuous success probability surface (Scenario 4) takes approximately 3.80 hours. This equates to only 76.04 seconds for each patient adaptively randomized.

Table 4.5: Computation Time

Scenario	**n**	**Seconds/patient (SE)**
1	200	67.39 (14.48)
2	200	64.03 (12.83)
3	200	67.17 (14.20)
4	200	76.04 (12.41)
5	398	147.53 (29.97)
6	200	68.93 (14.20)
7	398	153.04 (31.19)
8	200	67.66 (14.29)

4.4.6 Convergence Simulation

A convergence simulation was performed to explore the trade-off between monitoring convergence and both robustness of the inferences and computation time. The results of the convergence simulation for Scenario 4 are displayed in Table 4.6. In an average trial, there were approximately 1.79 ± 1.50 convergence warnings, where either equal allocation or the previous patient's posterior estimates were used to adapt. Overall, the total number of treatment successes, expected allocation proportions and MSE for each treatment are comparable with and without monitoring convergence of the parameter estimates. However, the simulation time per patient is 1.5 times larger when checking convergence for $n = 200$. This equates to an additional 2 hours (5.8 vs. 3.8 hours) of

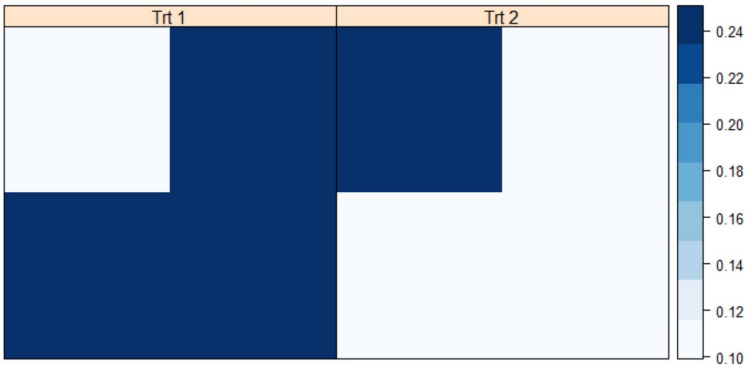
computation time when assessing convergence for a single simulated trial.

Table 4.6: Convergence Simulation: Scenario 4

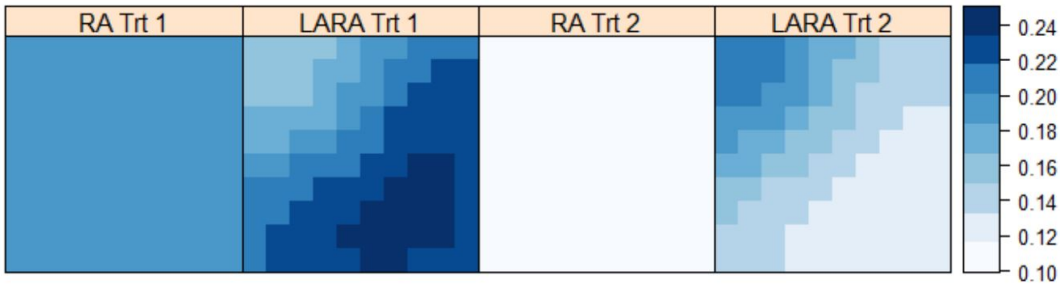
Convergence	**Successes (SE)**	**E[$n_1/(n_1+n_2)$] (SE)**	**Power (CI)**	**MSE 1**	**MSE 2**	**Seconds/Patient**
No	35.2 (5.30)	0.50 (0.04)	0.06 (0.05, 0.07)	0.006	0.005	76.04 (12.41)
Yes	35.2 (5.29)	0.50 (0.04)	0.06 (0.05, 0.07)	0.006	0.005	115.32 (20.16)

4.4.7 True and Estimated Treatment Success Surfaces

When there are discrete true spatially-varying surfaces with a medium effect size (Figure 4.9), LARA estimates a highly variable success probability surface for each treatment that resemble the true surfaces. RA randomization estimates constant success probabilities that are averaged across the true surface. This approach does not appear to recapture the true spatially-varying surfaces. When the true success probability surfaces become increasingly more continuous (Figures 4.10-4.12), the pattern is similar where LARA estimates highly variable success probabilities surfaces and does a better job at recapturing the true spatially-varying surfaces. As the true surfaces become increasing continuous, the estimated success probabilities for each treatment become more similar for RA randomization and thus do not recapture the true spatially-varying treatment effect sizes. In addition, there is less improvement for LARA in recapturing the true surfaces as the true surfaces become more continuous.

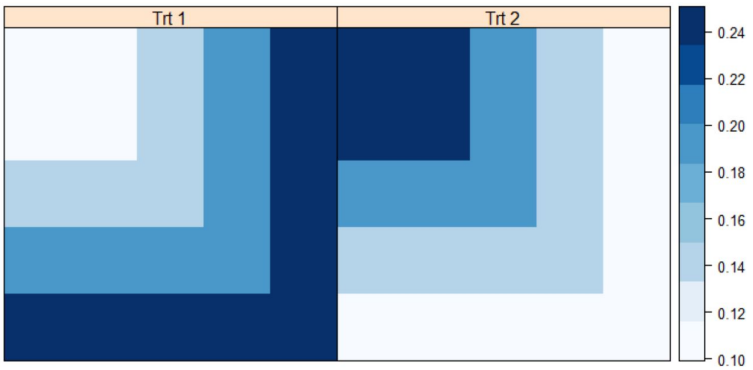


(a) True Surface

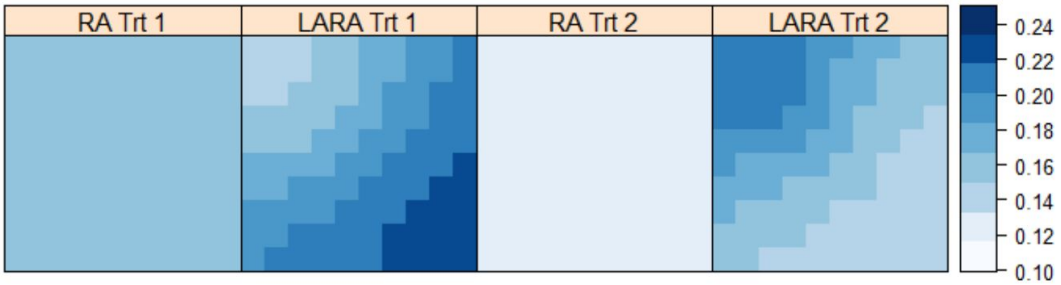


(b) Estimated Surface

Figure 4.9: Scenario 1 - Treatment Success Probabilities

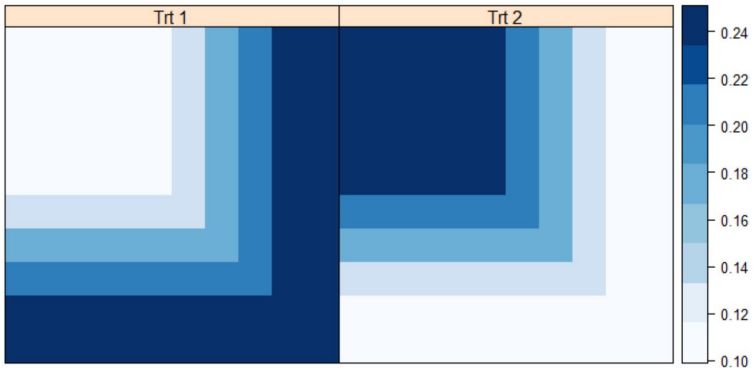


(a) True Surface

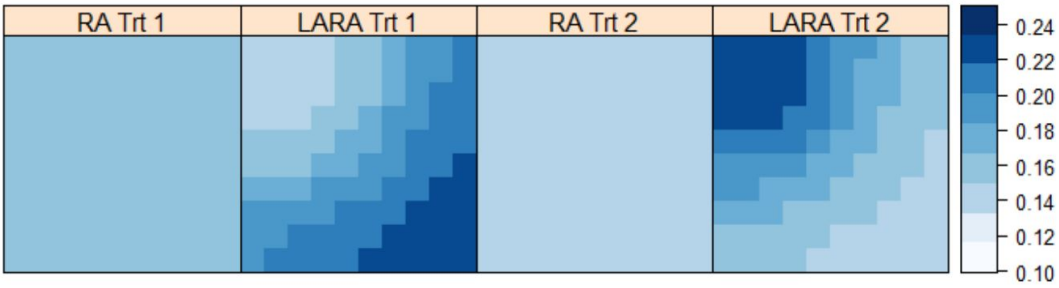


(b) Estimated Surface

Figure 4.10: Scenario 2 - Treatment Success Probabilities

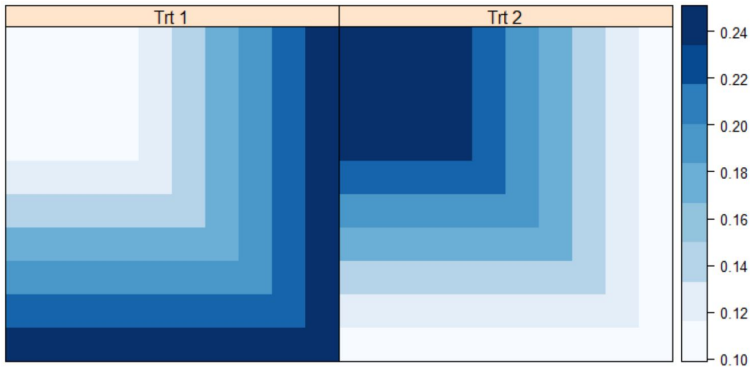


(a) True Surface

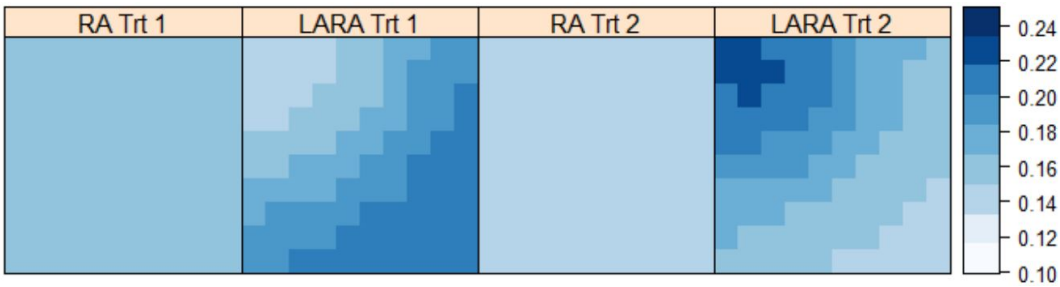


(b) Estimated Surface

Figure 4.11: Scenario 3 - Treatment Success Probabilities



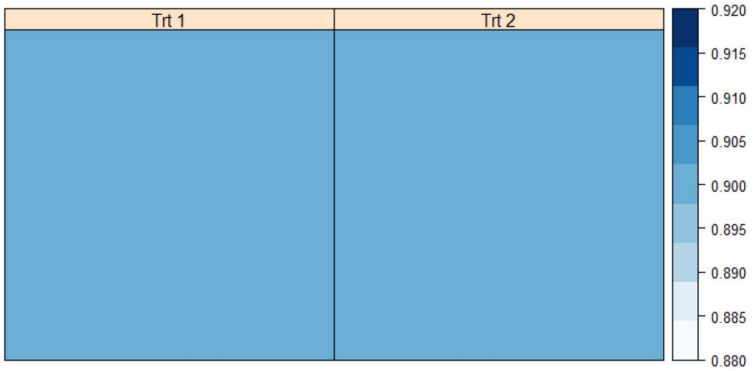
(a) True Surface



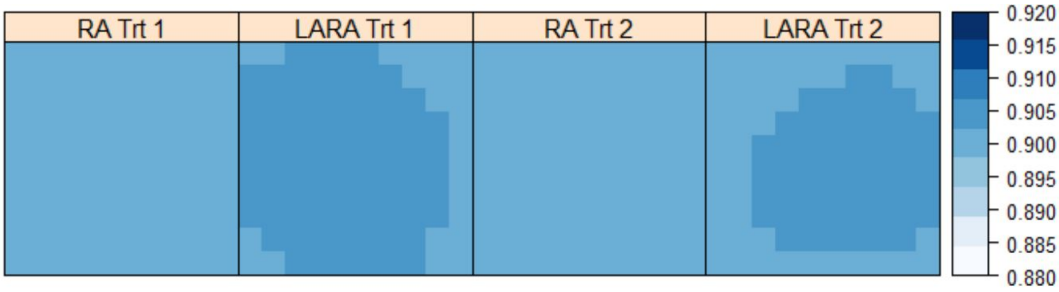
(b) Estimated Surface

Figure 4.12: Scenario 4 - Treatment Success Probabilities

When there is no spatial variability in the true surface and a null effect size (Figure 4.13), RA randomization recaptures the true success probability surfaces. However, there is some noise in estimating the true success probabilities for LARA, where there is nominal overestimation of the true success probabilities towards the center of the study area. For a medium effect size (Figure 4.14), the patter is similar where RA randomization recaptures the true surface and the estimates are slightly biased for LARA.

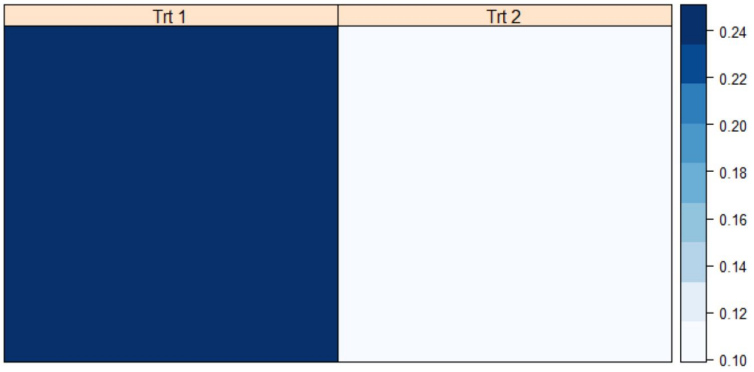


(a) True Surface

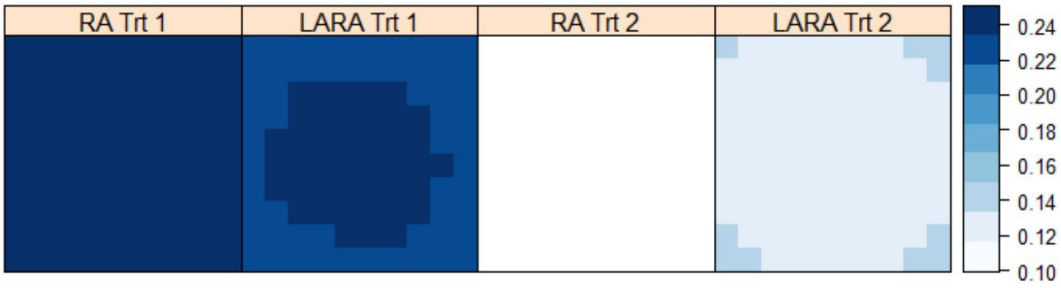


(b) Estimated Surface

Figure 4.13: Scenario 5 - Treatment Success Probabilities



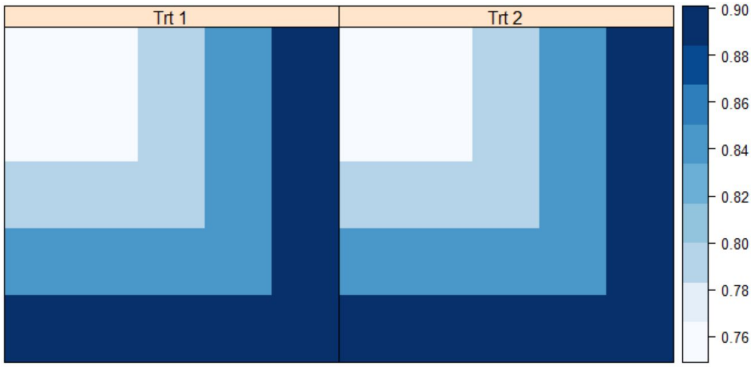
(a) True Surface



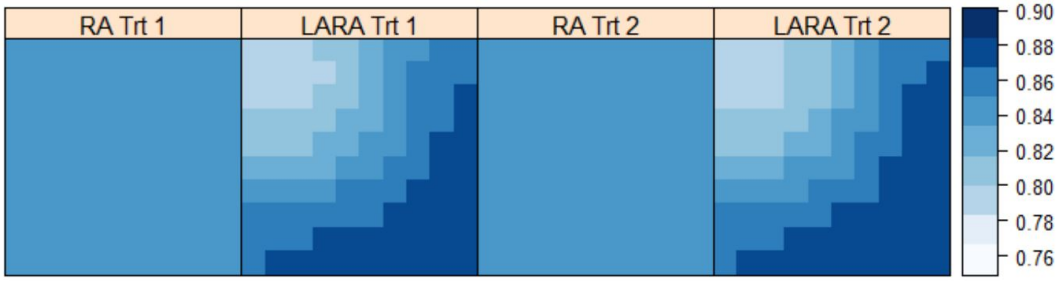
(b) Estimated Surface

Figure 4.14: Scenario 6 - Treatment Success Probabilities

When the true success probabilities vary across space and the effect size is null (Figure 4.15), LARA estimates a highly variable success probability surface that resembles the true surfaces. While RA randomization estimates constant success probabilities with no treatment difference, the estimates do not vary over space to recapture the true surfaces. When there is only spatial variability in success rates for treatment 1 (Figure 4.16), the pattern is similar where LARA outperforms RA in recapturing the true surfaces.

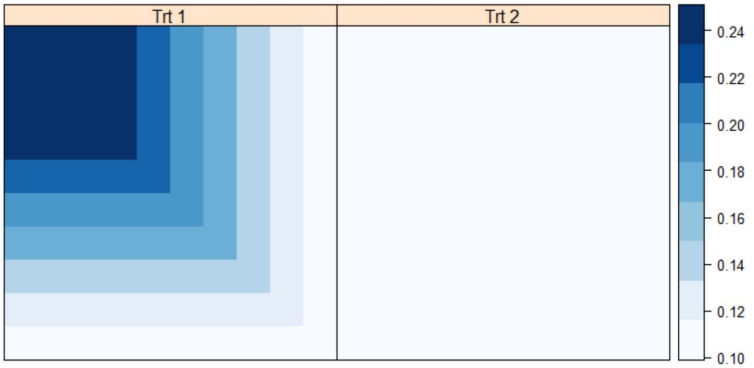


(a) True Surface

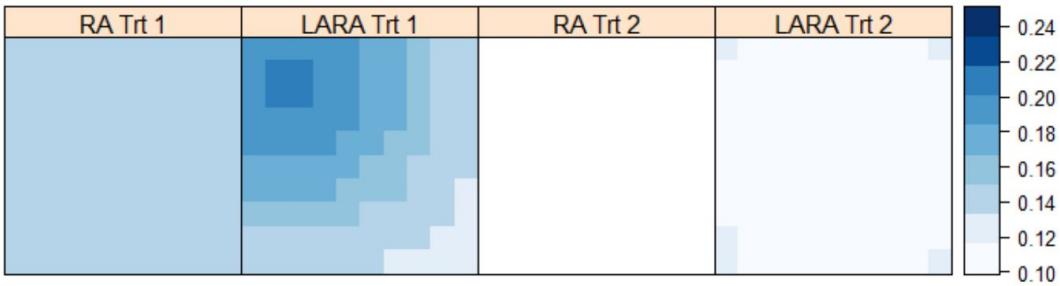


(b) Estimated Surface

Figure 4.15: Scenario 7 - Treatment Success Probabilities



(a) True Surface



(b) Estimated Surface

Figure 4.16: Scenario 8 - Treatment Success Probabilities

4.5 Discussion

We proposed the LARA approach to account for location-based variability in treatment effectiveness by interpolating spatially-varying treatment effects. The conditional success probabilities specific to each location were incorporated in a novel approach to skew the treatment assignments towards the treatment performing best at each patient's residence. We are not aware of any previous research that incorporated location into the adaptive design framework.

When there are true spatially-varying treatment effects and a medium effect size, LARA results in more treatment successes and larger allocation proportions while maintaining less variability in successes than RA randomization and, in some cases, balanced allocation. LARA achieves higher power and lower mean squared error relative to the other methods. This pattern is similar when only the success probabilities for treatment 1 vary across space. When the effect size is null, the three approaches perform similarly in terms of mean total number of successes, expected allocation proportions, type I error rates. There is nominally less improvement in expected successes and type I error but lower MSE for the LARA design. It is encouraging to note that LARA works best when there are true spatially-varying treatment differences, when implementing this approach would be most realistic.

When the true success surfaces are constant and there are no treatment differences, the three approaches perform comparably with respect to treatment successes, allocation proportions, type I error and MSE. There is slightly less improvement in the type I error rate for LARA. However, when there is a medium effect size, the adaptive designs result in more treatment successes and larger allocation ratios relative to the balanced case, with slightly less improvement for LARA. However, LARA maintains lower variability in both treatment successes and adaptation and achieves lower MSE compared to RA randomization. In addition, LARA gains back all of the power lost with the RA method relative to balanced allocation.

There is a notable loss of power for all three methods when there is spatial variation in the success probabilities. This can be explained by the fact that the sample size determination was performed for a fixed effect size, ignoring spatial variability. However, the true effect sizes vary across the study area and are often smaller than what the trial was powered for, thus the scenarios are underpowered. In addition, we are using a naive approach that ignores location to calculate

the operating characteristics. As trials would be designed and implemented without complete information, specifically the spatial variability of treatment effectiveness, this phenomena would be common in real world applications. When there was a medium effect size with spatial variability in the success rates, the power for the BC/LARA approaches were close to nominal levels. For a null effect size, there was a slight inflation in the type I error rate for LARA relative to the other methods.

By interpolating spatially-varying treatment effects, LARA is able to estimate treatment success probability surfaces that resemble the true surfaces when spatial variability exists. RA randomization ignores the spatial information and instead estimates constant success probabilities that are averages of the true surfaces. This results in more instances of under- or over- estimation of the true success probabilities and may lead to biased or even improper allocation. However, when there is no spatial variability in the true surfaces, RA randomization is able to recapture the true surface as expected, while LARA results in slightly biased estimates of the true success probabilities.

Although we use only 20 knots to represent the geographic locations, LARA may be able to better capture the true surfaces by increasing the number of knots (e.g. $k = 25$) in the low-rank kriging model. In addition, we fixed the spatial range parameter as one-quarter of the maximum interpoint distance. However, several authors have acknowledged the importance of estimating ρ as it controls the degree of smoothing in the spatially structured component [62, 46, 61]. This may lead to an improved model fit and ultimately better predictions of the spatially-varying treatment effects. Both of these approaches come at the cost of increased computation times, particularly in our adaptive randomization simulations.

The results of the computation times demonstrate that LARA is more computationally demanding relative to the SARA approaches. Although the simulation times are reported per patient, thus removing the effect of varying sample sizes, scenarios with larger sample sizes (null or small effect sizes) have increased simulation times. This may be explained by the additional computation and storing of results when there are more patient data. This problem is even more pronounced when the posterior estimates are monitored for convergence using a convergence diagnostic statistic. In our convergence simulation, we chose to use the previous patient's posterior estimates to interpolate the treatment effects necessary to update the allocation probabilities for the subsequent patient. Alternatively, we could have coded an auto-update function to stop the chains once convergence is

achieved, otherwise the chains would continue to update until convergence is attained. However, we suspect this would artificially increase the simulation times by repeatedly checking and storing the raw data. It is important to note that in a real life application the differences in computation times would be trivial, as only one patient is adaptively randomized at a time. Convergence would be assessed for each model fit after a new patient is adaptively randomized.

To ensure we were using the best interpolation approach for our LARA design, a small simulation study was performed in advance to compare the performance of a GAR model (Appendix D.1) with IDW interpolation to the low-rank kriging model (Eqn. 4.8). For the former, two independent CAR priors were specified for each treatment indicator to allow treatment effectiveness to vary over space for both groups. We calculated the success probabilities at the observed locations and used IDW interpolation to predict the success probabilities to 15 fixed unobserved locations, using 5-fold cross validation to select the inverse distance weighting power. For the LRK model, we considered varying knot sizes ($k = 15, 20, 25, 50$) to see how performance was affected. LRK requires that the number of patients be more than the number of knots when selecting candidate points for the knot locations. We simulated 100 iterations of simulated patient locations with varying sample sizes ($n = 50, 100, 200$). We compared the mean squared error and the computation times to assess the quality of the predictions and the feasibility of the method for the adaptive randomization simulation study. The results (Appendix D.1) show that the LRK model resulted in better quality predictions (lower MSE) in most cases, particularly with a moderate number of knots. The IDW approach resulted in lower computation times than LRK, with the number of knots increasing the time but without much improvement in MSE.

We investigated scenarios where there was spatial variability in the success probabilities, both discrete and more continuous patterns, as well as success rates that remained constant across space. However, the scenarios may not be realistic in real life. We could conceive even more continuous, realistic surfaces by increasing the number of grid cells and the complexity of the true success probability surfaces. In addition, we focused on recapturing the variability in treatment effectiveness by incorporating the spatially-varying treatment effects into the adaptive designs. Future work can include adding covariates and other random effects, including spatially uncorrelated frailty terms, to account for any unmeasured confounding at the subject-level. A small simulation study would be necessary to examine which, if any, covariates and/or random effects should be included in the

LARA model and candidate models compared using the goodness of fit (e.g. DIC). In addition, the two independent Gaussian priors for the spatially-varying joint treatment random effects assume independence of the two treatments. Instead, a multivariate CAR (MCAR) prior could be specified to model the cross-correlations of the spatial random effects.

An alternative to the proposed LARA model is to extend the SARA models (Eqns. 2.13, 3.1, 3.3) to incorporate a more variable site effect in a GLMM without using spatial modeling. Composite centers are created when small sites are 'pooled' together or combined with a larger site to minimize data instability and variance inflation due to small sample sizes [12]. However, this may lead to composites centers that are more heterogeneous and patients may respond differently to treatments. In a similar sense, patient residential locations can be geocoded to a census unit (e.g. tracts) and these spatial units used as a random 'site' effect. This would be appropriate for trials with sparse patient locations or in studies with a small recruitment area, to ensure that the samples sizes in each 'site' would be large enough to estimate the treatment effect.

CHAPTER 5

DISCUSSION

We have proposed in Aim 1 a method to account for site-based response variability by calculating conditional success probabilities specific to each center that can be incorporated in a novel approach to skew the treatment allocations. In Aim 2, we extended this design to also account for variability in treatment effectiveness under the Bayesian framework to avoid the limitations associated with asymptotic approximations. Finally, a proxy, specifically residential location, is used for factors that affect treatment differences to account for location-based variability in treatment responses and effectiveness in Aim 3. This allows for a more highly variable treatment success surface to be incorporated into the randomization method. These models were shown to be viable and are generally at least as efficient as comparator methods.

We are not aware of any previous research that incorporated either site or patient residential location into the response-adaptive design framework. Selvaratnam et al. (2017) proposed a GLMM approach, which includes a center effect, to analyze data from a multi-center trial after all patients are allocated using response-adaptive randomization [11]. The authors express the importance of considering differences between centers and thus follow the approach of Li et al. (2012), where a multi-center trial was analyzed using a Poisson GLMM that contained both treatment and random center effects [68]. Since there are no closed form expressions for the asymptotic variance of the MLEs, Selvaratnam et al. use the Gauss-Hermite quadrature method along with influence function approaches to approximate the integrals necessary to establish the asymptotic properties of the MLEs. The authors apply their approach to a real data set as well as perform a simulation study. The model parameters are shown to be consistent and asymptotically normally distributed [11]. However, the authors only consider their GLMM approach to analyze the trial data, not to adaptively randomize the patients as in our research.

Our approach uses conditional success probabilities specific to the subsequent patient's site or location to skew the allocation weights in a novel way. Although our research incorporates site or location information, the method is flexible enough to include additional covariates to adjust the treatment assignments. Here, we substitute the estimated site-specific success probabilities from

the GLM or GLMM/GLSM directly into the Optimal or Neyman allocation equations to update the randomization weights. While we use the relative size of the success probabilities $\hat{p}_1, \hat{p}_2$ to adapt, previous research by Bandyopadhyay et al. (2007) used a standardized test statistic of the treatment difference, or log-odds ratio, while Rosenberger et al. (2001) used an "Odds Ratio Mapping" approach to transform the adjusted odds ratio to a probability that could be plugged into one of the allocation equations [20, 21]. In addition, our approach can be easily extended to accommodate other outcome types, including continuous and time-to-event data.

The trade-off between power and treatment failures is an expected consequence of adaptive randomization. Adaptive methods skew the allocation ratio depending on the aim, thus the sample sizes for each group are often unbalanced and may result in a loss of power compared to balanced randomization. However, by accounting for an extra source of variation in the randomization process, the variability in the allocation weights and thus treatment assignments were reduced. Specifically, this loss of power for the proposed methods was not as notable as for RA randomization. Therefore, our methods may reduce variability in allocation but only at the slight cost of power compared to the balanced case. The simulations were designed with 80% power, assuming no inter-site variability, which can require a relatively large sample size. Due to the additional resources (e.g. time, money) necessary, this assumption may not be feasible or even unrealistic for some clinical trials.

Besides the statistical performance benefits, the proposed methods have several important clinical implications. First, using the optimal allocation rule reduces the number of expected treatment failures and is thus an ethical approach to randomization. Adaptive randomization may be more appealing to patients because they have a higher chance of being assigned to the 'superior' treatment group and with our approach, to the 'superior' treatment group at their recruitment center or residential location. Furthermore, by incorporating site information into the designs to account for unmeasured constructs that may affect the outcome, such as social determinants of health, the variability in the allocation process was reduced so more patients are assigned to the superior treatment. Additionally, by gaining back most of the power lost with the other adaptive methods, these approaches achieve higher power to detect a statistically significant treatment effect when one exists at the end of a study. Importantly, this method is a step closer to personalized or precision medicine, where treatment is tailored to maximize individual patient benefit. In addition to

incorporating site and location information, our proposed methods are flexible enough to allow for covariates to be included to adjust the allocation.

We investigated numerous scenarios where we controlled the amount of inter-site variability, number of sites, treatment success rates, effect sizes and the optimization criteria for the allocation, or we manually set the true treatment success surfaces. While these scenarios are finite in scope, having only considered a subset of the potential realistic trial conditions, they are representative of many possible real-world scenarios. The limited number of scenarios is in part due to the extensive computation times necessary for the simulation studies. Since the simulations were modeled as a Phase 2 efficacy trial, there are a small number of centers which requires that enough patients be recruited to each. This may result in convergence issues early on in the study or lead to unstable estimates. However, natural lead-in approaches could be used to mitigate these issues, such as those described in Thall & Wathen (2007) [22] and Sabo (2017) [69].

There is an array of logistical challenges associated with implementing any clinical trial, though there are additional issues associated with adaptive randomization designs. For covariate-adjusted response-adaptive randomization, the outcome and covariate data are required to be readily available in order to update the randomization schedule for the subsequent patient. This requires infrastructure to be in place across all centers for the input and processing of patient data necessary to update the allocation weights. Since the assignment probabilities change throughout a trial, there are ethical implications for the informed consent process. Because there is a higher chance patients may be assigned to the 'superior' treatment, particularly later in the study when enough data are available, patients may prefer to enroll later on in the study. This would introduce bias into the randomization as some patients (e.g. less severe cases) may wait to enroll under the assumption that they will be randomized to the 'better' treatment [70]. Some additional considerations include masking, use of a data safety monitoring board (DSMB) statistician, lack of standard statistical software, sample size and power, protocol approval by regulatory agencies and budget concerns [18, 70, 71, 72].

5.1 Future considerations:

Some future extensions of this work include the following conceptual ideas.

- Incorporate a natural lead in early on in a trial when the sample proportions are unstable or when there are no successes for one of the treatment groups [22], such as using a Decreasingly Informative Prior (DIP) [73],
- Comparison of different artificial lead-in lengths where balanced allocation is used,
- Development of a method to handle subject attrition but still use their information in the adaptive rule,
- Implement an approach for delayed subject responses, such as using auxiliary outcomes [8],
- Development of site- and location-adjusted designs to handle multiple treatments [74, 75], multiple responses as well as different outcome types, such as continuous data [24, 25, 26, 27],
- Examination of alternate ways to utilize data when adapting, for example, Rosenberger's (2001) "Odds Ratio Mapping" approach [20] or calculating posterior probabilities of treatment efficacy (e.g. $\Pr\{p_B < p_A|\mathscr{F}_n)\}$) [22, 32],
- Derive asymptotic properties of estimators for validity of inference from proposed designs [76, 75, 11],
- Consider methods to reduce the computational demand, such as 1) sequential Bayesian updating, where the posterior distribution is constructed so that the likelihood of the previous patient data does not need to be calculated for each subsequent patient being randomized [77], or 2) using only 1 Markov chain and assessing convergence using Geweke's (1992) diagnostic [78],
- Applying LARA and incorporating census data, such as median income or percentage of specific race/ethnicity to adjust the allocation,
- Estimate the spatial range parameter, ρ, which controls the degree of smoothing in the spatially structured component [46, 62, 61] of the LARA model,
- Comparison of Bayesian methods to test treatment effects (e.g. 95% credible interval, posterior probabilities).

APPENDIX A

R CODE RELEVANT TO CHAPTER 2

```
## Site-Adjusted Response-Adaptive Design (SARA 1) ##

##load packages
library(Matrix)
library(lme4)

##function to bound probabilities between [0.01, 0.99]
bound <- function(f){
  if (f > 0 & f < 1){
    f <- f
  } else if (f <= 0){
    f <- 0.01
  } else {
    f <- 0.99
  }
}

#generate site-specific probabilities
p1_0 <- numeric(kcent)
p2_0 <- numeric(kcent)

for (l in 1:kcent){
  p1_0[l] <- rnorm(1, mean = P1, sd = sigm) #random variate from N(P1, sigma)
  p2_0[l] <- rnorm(1, mean = P2, sd = sigm)

  p1_0[l] <- bound(p1_0[l]) #bounds probabilities [0.01, 0.99]
  p2_0[l] <- bound(p2_0[l])
}

##inner loop 1 (h) - assign site to each subject using random draw from U(0,1)
for (h in 1:n) {
  #determine site patients come from and success rates for that practice
  v <- runif(1, min = 0, max = 1) #values between (0, 1)

  for (z in 1:kcent){
    if (v > (z-1)/kcent & v <= z/kcent){
      YGS[h, 3] <- z
    }
  }
}

#delay updating of weights until at least the 20th iteration, outcome has 2 levels and all sites observed
if (i > 20 & dim(table(YGS[1:i, 1])) == 2 & dim(table(YGS[1:i, 3])) == kcent){
  #YGS = matrix(Outcome, Group, Site)
  fit2 <- try(glmer(YGS[1:i,1] ~ factor(YGS[1:i,2]) + (1 | factor(YGS[1:i,3])), family = binomial, control =
                      glmerControl(optimizer = "bobyqa"), nAGQ = 10), silent = TRUE) #Adaptive Gauss-Hermite Quadrature
                           approximation

  cond1 <- try(summary(fit2), silent = TRUE) #use try(), summary(fit2) produces error and terminates loop on cluster

```

```
  #no error, update weights if no convergence issues
  if (class(cond1) == 'summary.merMod') {
    if (is.null(summary(fit2)$optinfo$conv$lme4$messages) == TRUE & summary(fit2)$optinfo$conv$opt == 0 & class(fit2) == '
        glmerMod'){
      #no convergence warnings, successful convergence
      blups <- as.data.frame(ranef(fit2)$`factor(YGS[1:i, 3])`) #extract conditional means of random site effects

      b0 <- summary(fit2)$coefficients[1] #intercept
      b1 <- summary(fit2)$coefficients[2] #treatment effect, reference = group 1

      #site effect for next subject if not last subject in trial
      b2 <- ifelse(i != n, blups[rownames(blups) == YGS[(i + 1), 3],], blups[rownames(blups) == YGS[i, 3],])

      p1_hat <- exp(sum(b0, b2))/(1 + exp(sum(b0, b2))) #antilogit, intercept only
      p2_hat <- exp(sum(b0, b1, b2))/(1 + exp(sum(b0, b1, b2))) #intercept + b1

      if (Alloc == 'o') {
        #update optimal allocation weights
        w1 <- sqrt(p1_hat)/(sqrt(p1_hat) + sqrt(p2_hat))
        w2 <- (1 - w1)
      }

      if (Alloc == 'n') {
        #failure probabilities
        q1_hat <- (1 - p1_hat)
        q2_hat <- (1 - p2_hat)

        #update neyman allocation weights
        w1 <- sqrt(p1_hat*q1_hat)/(sqrt(p1_hat*q1_hat) + sqrt(p2_hat*q2_hat))
        w2 <- (1 - w1)
      }
    }
  }
  weight[i, 1] <- w1 #collect updated weight use to randomize subsequent patient, w2 = 1 - w1
}
```

APPENDIX B

CHAPTER 3 ADDITIONAL TABLES

Table B.1: Scenario - 3 sites, $p_1 = 0.3, 0.5, 0.7$, $\delta = 0.1$

				SARA 2			SARA 3		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
3	0.3	0.1	0	150.0 (10.2)	0.56 (0.03)	0.84	150.3 (10.3)	0.56 (0.04)	0.84
		(n=587)	0.05	150.2 (16.3)	0.55 (0.04)	0.74	151.5 (16.3)	0.56 (0.04)	0.72
			0.1	151.7 (26.5)	0.56 (0.06)	0.68	156.2 (26.5)	0.56 (0.07)	0.66
			0.15	155.7 (37.0)	0.55 (0.08)	0.69	164.3 (37.3)	0.56 (0.08)	0.64
	0.5	0.1	0	351.4 (14.0)	0.53 (0.02)	0.79	351.4 (14.0)	0.53 (0.02)	0.80
		(n=775)	0.05	350.9 (21.4)	0.53 (0.03)	0.70	351.6 (21.4)	0.53 (0.03)	0.69
			0.1	351.3 (34.4)	0.53 (0.03)	0.66	354.6 (34.5)	0.53 (0.04)	0.65
			0.15	353.3 (48.8)	0.53 (0.04)	0.70	360.7 (48.5)	0.53 (0.05)	0.69
	0.7	0.1	0	464.5 (12.3)	0.52 (0.02)	0.79	464.5 (12.3)	0.52 (0.02)	0.78
		(n=712)	0.05	464.3 (19.5)	0.52 (0.02)	0.72	464.7 (19.5)	0.52 (0.02)	0.72
			0.1	464.4 (31.4)	0.52 (0.03)	0.68	466.2 (31.3)	0.52 (0.03)	0.67
			0.15	465.3 (44.3)	0.52 (0.03)	0.71	469.3 (43.7)	0.52 (0.03)	0.71

[1] Compare to Tables 3.3, 3.9, 3.15

Table B.2: Scenario - 6 sites, $p_1 = 0.5, 0.7$, $\delta = 0.15$

				SARA 2			SARA 3		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
6	0.5	0.15	0	147.0 (9.0)	0.54 (0.04)	0.80	147.3 (9.0)	0.55 (0.04)	0.79
		(n=339)	0.05	146.3 (10.6)	0.54 (0.04)	0.76	146.8 (10.6)	0.55 (0.04)	0.76
			0.1	146.7 (13.4)	0.54 (0.04)	0.70	148.2 (13.5)	0.55 (0.04)	0.69
			0.15	147.3 (17.3)	0.54 (0.05)	0.67	150.4 (17.1)	0.55 (0.05)	0.66
	0.7	0.15	0	204.0 (8.6)	0.53 (0.03)	0.81	204.1 (8.5)	0.53 (0.03)	0.80
		(n=325)	0.05	204.1 (9.9)	0.53 (0.03)	0.77	204.4 (9.9)	0.53 (0.03)	0.78
			0.1	204.4 (12.7)	0.53 (0.03)	0.72	205.2 (12.7)	0.53 (0.03)	0.71
			0.15	204.4 (16.5)	0.53 (0.04)	0.67	206.3 (16.3)	0.53 (0.04)	0.67

[1] Compare to Tables 3.12, 3.18

Table B.3: Scenario - 3 sites, $p_1 = 0.5$, $\delta = 0$

				SARA 2			SARA 3		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
3	0.5	0	0	387.8 (13.9)	0.50 (0.02)	0.04	387.8 (13.9)	0.50 (0.02)	0.04
		(n=775)	0.05	387.3 (21.0)	0.50 (0.02)	0.19	388.0 (21.0)	0.50 (0.02)	0.19
			0.1	388.1 (33.9)	0.50 (0.03)	0.42	390.9 (33.8)	0.50 (0.03)	0.41
			0.15	389.8 (48.4)	0.50 (0.04)	0.59	396.1 (47.9)	0.50 (0.04)	0.57

[1] Compare to Table 3.9

Table B.4: Scenario - 6 sites, $p_1 = 0.7$, $\delta = 0.1$

				SARA 2			**SARA 3**		
k	$\boldsymbol{p_1}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma}$	**Successes**	**Proportion**	**Power**	**Successes**	**Proportion**	**Power**
6	0.7	0.1	0	464.5 (12.3)	0.52 (0.02)	0.78	464.5 (12.3)	0.52 (0.02)	0.79
		(n=712)	0.05	463.5 (16.5)	0.52 (0.02)	0.71	464.0 (16.5)	0.52 (0.02)	0.71
			0.1	463.6 (24.3)	0.52 (0.02)	0.65	465.7 (24.2)	0.52 (0.02)	0.65
			0.15	463.6 (33.2)	0.52 (0.03)	0.65	468.1 (32.8)	0.52 (0.03)	0.66

[1] Compare to Table 3.18

APPENDIX C

R CODE RELEVANT TO CHAPTER 3

```
## Site-Adjusted Response-Adaptive Design without interaction (SARA 2) ##

##load packages
lapply(c('Matrix', 'coda', 'BRugs', 'lme4'), require, character.only = TRUE)

#inverse logit function
inv.logit <- function(x){
  return(exp(x)/(1 + exp(x)))
}

#function for site-specific success probabilities, p-hat_ij, i = treatment group, j = site
#reference: treatment group 1
site.prob <- function(k0 = NULL, num_s = kcent, inter = 0){
  #k0 = site for next subject, num_s = total number of sites in current trial

  #no interaction
  if(inter == 0){
    p2_hat <<- as.numeric(inv.logit(b0 + b1 + get(paste('d', k0, sep = ''))))
    p1_hat <<- as.numeric(inv.logit(b0 + get(paste('d', k0, sep = ''))))
  }

  #trt-by-site interaction
  if(inter == 1){
    p2_hat <<- as.numeric(inv.logit(b0 + b1 + get(paste('d', k0, sep = '')) +
                                      get(paste('g', k0, sep = '')))) #site + interaction effects
    p1_hat <<- as.numeric(inv.logit(b0 + get(paste('d', k0, sep = '')) +
                                      get(paste('g', num_s + k0, sep = ''))))
  }
}

#bayesian GLMM with random site effect (delta)
model <- function(){
  for(i in 1:N) {

    Y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + beta1 * G[i] + delta[S[i]]

  }

  for(j in 1:nk) {

    delta[j] ~ dnorm(0.0, 1) #site-level RE, skeptical prior N(0,1)

  }

  # Choice of Priors
  beta0 ~ dnorm(0.0, 0.2) #N(0,5)
  beta1 ~ dnorm(0.0, 0.2) #N(0,5)
}

```

```
#delay updating of weights until at least the 20th patient
if (i > 20){

  YGS$group <- ifelse(YGS$G == 2, 1, 0) #dummy code for BRugs, group 1 = reference (Aim 1)
  dat <- list(N = i, nk = kcent, Y = YGS$Y[1:i], G = YGS$group[1:i], S = YGS$S[1:i]) #nk = number of sites

  if (i == 21){
    #use frequentist GLM for starting values for first subject once adaptation begins
    init_fit <- glm(YGS[1:i,1] ~ factor(YGS[1:i,2]), family = binomial) #frequentist GLM model, no site effect

    b0.init <- summary(init_fit)$coefficients[rownames(summary(init_fit)$coefficients) == '(Intercept)', 1] #intercept
    b1.init <- summary(init_fit)$coefficients[rownames(summary(init_fit)$coefficients) == 'factor(YGS[1:i, 2])2', 1] #trt
         2 effect

    l1 <- list(beta0 = b0.init, beta1 = b1.init, delta = rep(0, kcent)) #chain starting values
    l2 <- list(beta0 = b0.init + 0.5, beta1 = b1.init + 0.5, delta = rep(0, kcent))
    inits <- list(l1, l2) #list of initial values
    inits1 <- paste('init1', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
    inits2 <- paste('init2', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
    bugsInits(inits, 2, c(inits1, inits2)) #have to output inits or they will save to temp folder when fitting model

  } else {
    #starting values after 21st patient
    l1 <- list(beta0 = inits.df[1], beta1 = inits.df[2], delta = as.numeric(inits.df[2 + (1:kcent)]))
    l2 <- list(beta0 = inits.df[1] + 0.5, beta1 = inits.df[2] + 0.5, delta = as.numeric(inits.df[2 + (1:kcent)] + .05))
    inits <- list(l1, l2) #list of initial values
    inits1 <- paste('init1', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
    inits2 <- paste('init2', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
    bugsInits(inits, 2, c(inits1, inits2))
  }

  #fit SARA 2 model
  model1 <- try(BRugsFit(model, dat, c(inits1, inits2), numChains = 2,
                         parametersToSave = c('beta0', 'beta1', 'delta'),
                         nBurnin = 2000, nIter = 4000, nThin = 1, coda = TRUE, DIC = FALSE,
                         BRugsVerbose = FALSE, seed = seed0),
                silent = TRUE)

  #no error, update weights if no convergence issues
  if (class(model1) == 'mcmc.list') {

    conv <- gelman.diag(model1)$psrf[,1] #Gelman's Rhat aka. PSRF

    if (is.element(NA, conv) == FALSE & all(conv < 1.1)){
      #parameters converged (Rhat < 1.1), use posterior estimates to update weights adapt

      post <- summary(model1)$statistics #posterior estimates
      post.mean <- post[rownames(post) == 'beta1', 1:2] #posterior mean trt effect

      post.median <- apply(rbind(model1[[1]], model1[[2]]), 2, stats::median) #calculate posterior medians

      b0 <- post.median[names(post.median) == 'beta0'] #intercept
      b1 <- post.median[names(post.median) == 'beta1'] #treatment 2 effect, reference = group 1

      #extract site effects
      blups <- as.data.frame(as.numeric(post.median[names(post.median) %in% paste('delta[', 1:kcent, ']', sep = '')]))

```

```
      #output delta = site effect objects to calculate success probs
      names1 <- paste('d', 1:kcent, sep = '')
      for(s in 1:length(names1)){
        assign(names1[s], post.median[names(post.median) == paste('delta[', s, ']', sep = '')])
      }

      #if posterior estimates converged, use as initial values for next subject
      inits.df[1] <- b0
      inits.df[2] <- b1

      for(s0 in 1:kcent){
        inits.df[2 + s0] <- get(paste('d', s0, sep = ''))
      }

      inits.df[3 + kcent] <- NA #delta.var

      #site-specific success probability for next subject if not last subject in trial, otherwise use current site
      if(i != n){
        site.prob(k0 = YGS[(i + 1), 3], num_s = kcent, inter = 0) #outputs p1_hat, p2_hat objects
      } else {
        site.prob(k0 = YGS[i, 3], num_s = kcent, inter = 0)
      }

      #allocation equations
      if (Alloc == 'o') {
        w1 <- sqrt(p1_hat)/(sqrt(p1_hat) + sqrt(p2_hat)) #optimal allocation
        w2 <- (1 - w1)
      }

      if (Alloc == 'n') {
        q1_hat <- (1 - p1_hat) #failure probabilities
        q2_hat <- (1 - p2_hat)

        w1 <- sqrt(p1_hat*q1_hat)/(sqrt(p1_hat*q1_hat) + sqrt(p2_hat*q2_hat)) #neyman allocation
        w2 <- (1 - w1)
      }
    }
  }
  weight[i, 1] <- w1 #collect updated weight use to randomize subsequent
}
```

```
## Site-Adjusted Response-Adaptive Design with interaction (SARA 3) ##

#bayesian GLMM with random site effect (delta) and interaction effect (gamma)
model <- function(){
  for(i in 1:N) {

    Y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + beta1 * G[i] + delta[S[i]] + gamma[I[i]]

  }

  for(j1 in 1:nk) {

    delta[j1] ~ dnorm(0.0, 1) #site-level RE, skeptical prior N(0,1)

  }
```

```
17
18    for(j2 in 1:(2*nk)) {
19
20      gamma[j2] ~ dnorm(0.0, 1) #trt-by-site interaction RE, skeptical prior N(0,1)
21
22    }
23
24    # Choice of Priors
25    beta0 ~ dnorm(0.0, 0.2) #N(0,5)
26    beta1 ~ dnorm(0.0, 0.2) #N(0,5)
27  }
28
29  YGS$group <- ifelse(YGS$G == 2, 1, 0) #dummy code for BRugs, group 1 = reference (Aim 1)
30  YGS$I <- ifelse(YGS$G == 2, YGS$S, (YGS$S + kcent)) #interaction indicator variable
31  dat <- list(N = i, nk = kcent, Y = YGS$Y[1:i], G = YGS$group[1:i], S = YGS$S[1:i], I = YGS$I[1:i]) #nk = number of sites
32
33  if (i == 21){
34    #use frequentist GLM for starting values for first subject once adaptation begins
35    init_fit <- glm(YGS[1:i,1] ~ factor(YGS[1:i,2]), family = binomial) #frequentist GLM model, no site effect
36
37    b0.init <- summary(init_fit)$coefficients[rownames(summary(init_fit)$coefficients) == '(Intercept)', 1] #intercept
38    b1.init <- summary(init_fit)$coefficients[rownames(summary(init_fit)$coefficients) == 'factor(YGS[1:i, 2])2', 1] #trt 2
          effect
39
40    l1 <- list(beta0 = b0.init, beta1 = b1.init, delta = rep(0, kcent), gamma = rep(0, 2*kcent)) #chain starting values
41    l2 <- list(beta0 = b0.init + 0.5, beta1 = b1.init + 0.5, delta = rep(0, kcent), gamma = rep(0, 2*kcent))
42    inits <- list(l1, l2) #list of initial values
43    inits1 <- paste('init1', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
44    inits2 <- paste('init2', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
45    bugsInits(inits, 2, c(inits1, inits2)) #have to output inits or they will save to temp folder when fitting model
46
47  } else {
48    l1 <- list(beta0 = inits.df[1], beta1 = inits.df[2], delta = as.numeric(inits.df[2 + (1:kcent)]),
49               gamma = as.numeric(inits.df[3 + kcent + 1:(2*kcent)]))
50    l2 <- list(beta0 = inits.df[1] + 0.5, beta1 = inits.df[2] + 0.5, delta = as.numeric(inits.df[2 + (1:kcent)] + 0.05),
51               gamma = as.numeric(inits.df[3 + kcent + 1:(2*kcent)]) + 0.05)
52    inits <- list(l1, l2) #list of initial values
53    inits1 <- paste('init1', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
54    inits2 <- paste('init2', '_', Alloc, '_', kcent, '_', P1, '_', delt, '_', sigm, '.txt', sep = '')
55    bugsInits(inits, 2, c(inits1, inits2)) #have to output inits or they will save to temp folder when fitting model
56  }
57
58  #fit SARA 3 model
59  model1 <- try(BRugsFit(model, dat, c(inits1, inits2), numChains = 2,
60                         parametersToSave = c('beta0', 'beta1', 'delta', 'gamma'),
61                         nBurnin = 2000, nIter = 4000, nThin = 1, coda = TRUE, DIC = FALSE,
62                         BRugsVerbose = FALSE, seed = seed0),
63              silent = TRUE)
```

APPENDIX D

GEOGRAPHICALLY ADAPTIVE REGRESSION LARA MODEL

GAR LARA Model

$$logit(p_i) = \alpha + (\beta_1 + \beta_{11i})X_{1i} + (\beta_2 + \beta_{22i})X_{2i} + U_i + V_i + \boldsymbol{X^T\theta} \tag{D.1}$$

Here, α is the overall intercept, X_{1i}, X_{2i} are treatment group 1 and 2 indicators, U_i is an unstructured random effect and V_i is a spatially structured random effect for the i^{th} patient, $i = 1, \ldots, n$. Furthermore, β_1, β_2 are overall treatment effects for each group while β_{11i}, β_{22i} are interpreted as spatially-varying local deviations from the overall treatment effects. Instead of modeling the spatially-varying treatment effects with an intrinsic MCAR prior, two univariate CAR priors were considered under the assumption that the regression coefficients are independent. The separate random components for each treatment group allow treatment effectiveness to vary over space for both groups. The $\boldsymbol{X^T\theta}$ is an optional set of subject- and/or location-specific covariates. The following priors[1] are specified:

$$\begin{aligned}
\alpha \sim N(0, \tau_\alpha), \beta_1 \sim N(0, \tau_{\beta_1}), \beta_2 \sim N(0, \tau_{\beta_2}),\\
\beta_{11i}|\cdots \sim N(\mu_{\beta_{11}}, \tau_{\beta_{11}}/n_{\xi_i}),\\
\beta_{22i}|\cdots \sim N(\mu_{\beta_{22}}, \tau_{\beta_{22}}/n_{\xi_i}),\\
U_i \sim N(0, \tau_u),\\
V_i|\cdots \sim N(\bar{\nu}_i, \tau_\nu/n_{\xi_i}).
\end{aligned} \tag{D.2}$$

The hyperparameters for the CAR priors, defined as $\mu_{\beta_{11}} = \bar{\beta}_{11\xi_i}$ and $\mu_{\beta_{22}} = \bar{\beta}_{22\xi_i}$, are the neighborhood averages for neighborhood ξ_i (n_{ξ_i} neighbors) and the i^{th} area. Similarly, the hyperparameter for the independent CAR prior mean, defined as $\bar{\nu}_i = \sum_{j\in\xi_i} \nu_j/n_{\xi_i}$, is the neighborhood average. Several unstructured or spatially structured priors (e.g. exchangeable, conditionally auto-

[1]The priors are specified as $N(\mu, \tau)$, where $\tau = 1/\sigma^2$ and σ^2 is the variance.

regressive) could be used for the correlation structures of the random effects to account for unmeasured confounding at the subject-level. However, the unstructured and spatially structured random effects are optional and were removed to avoid identifiability issues.

Table D.1: Comparing Low Rank Kriging and Inverse Distance Weighted Interpolation

	LRK				
Patients	$\boldsymbol{k = 15}$	$\boldsymbol{k = 20}$	$\boldsymbol{k = 25}$	$\boldsymbol{k = 50}$	**IDW**
$\boldsymbol{n = 50}$					
Time (hours)	2.46	0.94	0.94	-	0.56
MSE 1	0.048	0.041	0.044	-	0.053
MSE 2	0.044	0.045	0.043	-	0.057
$\boldsymbol{n = 100}$					
Time (hours)	1.00	1.88	1.65	10.85	1.08
MSE 1	0.029	0.029	0.034	0.047	0.043
MSE 2	0.030	0.030	0.031	0.050	0.043
$\boldsymbol{n = 200}$					
Time (hours)	2.21	4.24	3.30	22.31	0.88
MSE 1	0.018	0.021	0.021	0.039	0.028
MSE 2	0.020	0.022	0.023	0.042	0.027

APPENDIX E

R CODE RELEVANT TO CHAPTER 4

```
## Location-Adjusted Response-Adaptive Design (LARA) ##

##load packages
lapply(c('spatstat', 'maptools', 'spdep', 'BRugs', 'coda', 'fields'), require, character.only = TRUE)

#Bayesian Low-Rank Kriging (LRK) model
model <- function(){
  for (i in 1:nknots){
    ME[i] <- 0 #mean=0 for REs
  }

  for (i in 1:nknots){
    for (j in 1:nknots){
      SOM[i, j] <- tauS * OM[i, j] #setting up precision matrix for REs, OM is pre-calculated knots cov, treatment group 1
      SOM2[i, j] <- tauT * OM[i, j] #SOM for treatment group 2 indicator
    }
  }

  for (i in 1:N){
    #Likelihood
    Y[i] ~ dbern(p[i])

    logit(p[i]) <- beta0 + beta11[i] * G1[i] + beta22[i] * G2[i]

    beta11[i] <- inprod(gam[], CDZ[i, ]) #CDZ pre-calculated cov matrix between obs/knots, treatment group 1
    beta22[i] <- inprod(gam2[], CDZ[i, ]) #treatment group 2

  }

  gam[1:nknots] ~ dmnorm(ME[], SOM[, ]) #joint random effect prior dist, treatment group 1
  gam2[1:nknots] ~ dmnorm(ME[], SOM2[, ]) #treatment group 2

  #Choice of Priors
  beta0 ~ dnorm(0, 0.4) #N(0, 2.5)

  tauS <- 1 / sigmaS #precision for trt 1
  sigmaS ~ dunif(0, 2) #var

  tauT <- 1 / sigmaT #precision for trt 2
  sigmaT ~ dunif(0, 2) #var

}

#Calculate distance matrix between observed points
Dobs <- as.matrix(dist(YGS[1:i, c('x', 'y')]))

rho_hat <- (1/4)* max(Dobs) #fixed rho

#Generate Knots - space filling knot locations
knot0 <- fields::cover.design(R = matrix(c(YGS$x, YGS$y), ncol = 2)[1:i, ], nd = n.knots,
```

```
                              nn = ifelse(i <= 100, FALSE, TRUE),
                              num.nn = min(nrow(YGS[1:i, ]) - (n.knots + 1), 100)) #only use 100 nn when i > 100
knot <- data.frame(x = knot0$design[, 1], y = knot0$design[, 2])

#Calculate distance matrix between knots
Dknot <- as.matrix(dist(knot)) #n.knots x n.knots

#Calculate Distance and Covariance Matrices between observed[1:i] & knot locations
CDZ <- matrix(nrow = i, ncol = n.knots) #covariance matrix for observed & knot locations
DOknot <- matrix(nrow = i, ncol = n.knots) #distance between observed & knot locations

for(z1 in 1:nrow(YGS[1:i, ])){
  for (z2 in 1:n.knots){
    DOknot[z1, z2] <- sqrt((YGS$x[z1] - knot$x[z2])^2 + (YGS$y[z1] - knot$y[z2])^2) #n.obs x n.knots
    CDZ[z1, z2] <- (1 + abs(DOknot[z1, z2]) / rho_hat) * exp(-abs(DOknot[z1, z2]) / rho_hat) #n.obs x n.knots
  }
}

#Omega Matrix for Random Effects with cov applied to the knots
OM <- matrix(nrow = n.knots, ncol = n.knots)

for(z1 in 1:n.knots){
  for (z2 in 1:n.knots){
    OM[z1, z2] <- (1 + abs(Dknot[z1, z2]) / rho_hat) * exp(-abs(Dknot[z1, z2]) / rho_hat) #n.knots x n.knots
  }
}

YGS$G1 <- ifelse(YGS$G == 1, 1, 0) #separate treatment indicator for each group
YGS$G2 <- ifelse(YGS$G == 2, 1, 0)
dat <- list(N = i, Y = YGS$Y[1:i], G1 = YGS$G1[1:i], G2 = YGS$G2[1:i], nknots = n.knots, CDZ = CDZ, OM = OM)

model1 <- try(BRugsFit(model, dat, c(inits1, inits2), numChains = 2,
                       parametersToSave = c('beta0', 'sigmaS', 'sigmaT', 'gam', 'gam2'),
                       nBurnin = 4000, nIter = 1000, nThin = 1, coda = FALSE, DIC = FALSE,
                       BRugsVerbose = FALSE, seed = seed0),
              silent = TRUE)

#Posterior Estimates
post.median <- as.matrix(model1$Stats)[, 'median'] #posterior median

b0 <- post.median[names(post.median) == 'beta0'] #intercept

psi1 <- as.numeric(post.median[names(post.median) %in% paste('gam[', 1:n.knots, ']', sep = '')]) #psi RE, trt 1
psi2 <- as.numeric(post.median[names(post.median) %in% paste('gam2[', 1:n.knots, ']', sep = '')]) #psi RE, trt 2

sigmaS <- post.median[names(post.median) == 'sigmaS']
sigmaT <- post.median[names(post.median) == 'sigmaT']

##Calculate Distance and Covariance Matrices between subsequent patient & knot locations for predictions
if (i != n){
  Unobs <- data.frame(x = YGS$x[i + 1], y = YGS$y[i + 1]) #subsequent patient
} else {
  Unobs <- data.frame(x = 0.5, y = 0.5) #last patient, predict to random location (0.5, 0.5)
}

CDZ2 <- matrix(nrow = nrow(Unobs), ncol = n.knots) #covariance matrix for unobserved & knot locations
DUknot <- matrix(nrow = nrow(Unobs), ncol = n.knots) #distance between unobserved & knot locations
```

```
108
109 for(z1 in 1:nrow(Unobs)){
110   for (z2 in 1:n.knots){
111     DUknot[z1, z2] <- sqrt((Unobs$x[z1] - knot$x[z2])^2 + (Unobs$y[z1] - knot$y[z2])^2) #n.Unobs x n.knots
112     CDZ2[z1, z2] <- (1 + abs(DUknot[z1, z2]) / rho_hat) * exp(-abs(DUknot[z1, z2]) / rho_hat) #n.Unobs x n.knots
113   }
114 }
115
116 b11 <- CDZ2 %*% psi1 #predicted trt 1 effect at unobserved location (subsequent patient)
117 b22 <- CDZ2 %*% psi2 #predicted trt 2 effect
118
119 #location-specific success probability for next subject
120 p1_hat <- pts$p1_hat[i + 1] <- as.numeric(inv.logit(b0 + b11))
121 p2_hat <- pts$p2_hat[i + 1] <- as.numeric(inv.logit(b0 + b22))
```

References

[1] Paul Gallo. Treatment-by-center interaction. *Wiley Encyclopedia of Clinical Trials*, 2008.

[2] Mirjam Moerbeek and Steven Teerenstra. *Power analysis of trials with multilevel data*. CRC Press, 2015.

[3] A Kuijer, J Verloop, O Visser, G Sonke, A Jager, CH van Gils, T van Dalen, and SG Elias. The influence of socio-economic status and ethnicity on adjuvant systemic treatment guideline adherence for early stage breast cancer in the netherlands. *Annals of Oncology*, page mdx204, 2017.

[4] Ira M Longini Jr, M Elizabeth Halloran, Azhar Nizam, Mark Wolff, Paul M Mendelman, Patricia E Fast, and Robert B Belshe. Estimation of the efficacy of live, attenuated influenza vaccine from a two-year, multi-center vaccine trial: implications for influenza epidemic control. *Vaccine*, 18(18):1902–1909, 2000.

[5] William F Rosenberger, Nigel Stallard, Anastasia Ivanova, Cherice N Harper, and Michelle L Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

[6] Lawrence M Friedman, Curt Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*, volume 4. Springer, 1998.

[7] CC Rout, DA Rocke, J Levin, E Gouws, and D Reddy. A reevaluation of the role of crystalloid preload in the prevention of hypotension associated with spinal anesthesia for elective cesarean section. *Anesthesiology*, 79(2):262–269, 1993.

[8] S Sinks, RT Sabo, and ND Mukhopadhyay. Outcome-adaptive allocation using auxiliary and primary outcomes. *Austin Biometrics and Biostatistics*, 2(3):e1023, 2015.

[9] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[10] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2013.

[11] Selvakkadunko Selvaratnam, Alwell J Oyet, Yanqing Yi, and Veeresh Gadag. Estimation of a generalized linear mixed model for response-adaptive designs in multi-centre clinical trials. *Canadian Journal of Statistics*, 45(3):310–325, 2017.

[12] Paul P Gallo. Center-weighting issues in multicenter clinical trials. *Journal of biopharmaceutical statistics*, 10(2):145–163, 2000.

[13] A Russell Localio, Jesse A Berlin, Thomas R Ten Have, and Stephen E Kimmel. Adjustments for center in multicenter studies: an overview. *Annals of internal medicine*, 135(2):112–123, 2001.

[14] Alan Agresti and Jonathan Hartzel. Tutorial in biostatistics: Strategies comparing treatment on binary response with multi-centre data. *Statistics in medicine*, 19:1115–1139, 2000.

[15] Feifang Hu and William F Rosenberger. *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons, 2006.

[16] Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

[17] LJ Wei. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, 73(363):559–563, 1978.

[18] William F Rosenberger and John M Lachin. The use of response-adaptive designs in clinical trials. *Controlled clinical trials*, 14(6):471–484, 1993.

[19] William F Rosenberger and Oleksandr Sverdlov. Handling covariates in the design of clinical trials. *Statistical Science*, pages 404–419, 2008.

[20] William F Rosenberger, AN Vidyashankar, and Deepak K Agarwal. Covariate-adjusted response-adaptive designs for binary response. *Journal of biopharmaceutical statistics*, 11(4):227–236, 2001.

[21] Uttam Bandyopadhyay, Atanu Biswas, and Rahul Bhattacharya. A covariate adjusted two-stage allocation design for binary responses in randomized clinical trials. *Statistics in medicine*, 26(24):4386–4399, 2007.

[22] Peter F Thall and J Kyle Wathen. Practical bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859–866, 2007.

[23] Atanu Biswas, Rahul Bhattacharya, and Eunsik Park. On a class of optimal covariate-adjusted response adaptive designs for survival outcomes. *Statistical methods in medical research*, 25(6):2444–2456, 2016.

[24] Uttam Bandyopadhyay and Atanu Biswas. Adaptive designs for normal responses with prognostic factors. *Biometrika*, pages 409–419, 2001.

[25] Uttam Bandyopadhyay and Rahul Bhattacharya. A covariate-adjusted response-adaptive allocation in clinical trials for a general class of responses. *Statistics in medicine*, 32(29):5053–5061, 2013.

[26] Atanu Biswas and Rahul Bhattacharya. A covariate-adjusted response-adaptive allocation for a general class of continuous responses. *Journal of Statistical Theory and Practice*, 10(4):852–863, 2016.

[27] Atanu Biswas and Rahul Bhattacharya. Response-adaptive designs for continuous treatment responses in phase iii clinical trials: A review. *Statistical methods in medical research*, 25(1):81–100, 2016.

[28] William F Rosenberger et al. Asymptotic inference with response-adaptive treatment allocation designs. *The Annals of Statistics*, 21(4):2098–2107, 1993.

[29] William F Rosenberger and Padmanabhan Seshaiyer. Adaptive survival trials. *Journal of biopharmaceutical statistics*, 7(4):617–624, 1997.

[30] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[32] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[33] Peter F Thall and J Kyle Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in medicine*, 24(13):1947–1964, 2005.

[34] Atanu Biswas and Jean-François Angers. A bayesian adaptive design in clinical trials for continuous responses. *Statistica neerlandica*, 56(4):400–414, 2002.

[35] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

[36] John A Lewis. Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, 18(15):1903–1942, 1999.

[37] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 3. CRC press Boca Raton, FL, 2014.

[38] David J Spiegelhalter et al. Incorporating bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174, 2004.

[39] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[40] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[41] Andrew Thomas, Bob O'Hara, Uwe Ligges, and Sibylle Sturtz. Making bugs open. *R News*, 6(1):12–17, 2006.

[42] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[43] Steven N MacEachern and L Mark Berliner. Subsampling the gibbs sampler. *The American Statistician*, 48(3):188–190, 1994.

[44] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[45] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn. Winbugs user manual, 2003.

[46] Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press, 2013.

[47] Derek G Cook and Stuart J Pocock. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, pages 361–371, 1983.

[48] MJ Gardner. Using the environment to explain and predict mortality. *Journal of the Royal Statistical Society. Series A (General)*, pages 421–440, 1973.

[49] Ross L Prentice and Lianne Sheppard. Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes & Control*, 1(1):81–97, 1990.

[50] David G Clayton, L Bernardinelli, and C Montomoli. Spatial correlation in ecological analysis. *International journal of epidemiology*, 22(6):1193–1202, 1993.

[51] Alan E Gelfand and Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15, 2003.

[52] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[53] David C Wheeler, Lance A Waller, and John O Elliott. Modeling epilepsy disparities among ethnic groups in philadelphia, pa. *Statistics in medicine*, 27(20):4069–4085, 2008.

[54] David C Wheeler and Catherine A Calder. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2):145–166, 2007.

[55] Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.

[56] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2002.

[57] David C Wheeler. Geographically weighted regression. In *Handbook of regional science*, pages 1435–1459. Springer, 2014.

[58] George Y Lu and David W Wong. An adaptive inverse-distance weighting spatial interpolation technique. *Computers & geosciences*, 34(9):1044–1055, 2008.

[59] Edzer J. Pebesma. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691, 2004.

[60] Peter J Diggle, JA Tawn, and RA Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

[61] David C Wheeler and Catherine A Calder. *Handbook of Spatial Epidemiology*, chapter Sociospatial Epidemiology: Residential History Analysis. CRC Press, 2016.

[62] Jonathan L French and Matthew P Wand. Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*, 5(2):177–191, 2004.

[63] Douglas Nychka, Perry D Haaland, Michael A O'Connell, and Stephen Ellner. Appendix a: Funfits, data analysis and statistical tools for estimating functions. *Case studies in environmental statistics*, 132:159, 1998.

[64] EE Kammann and Matthew P Wand. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18, 2003.

[65] David C Wheeler, Jenna Czarnota, and Resa M Jones. Estimating an area-level socioeconomic status index and its association with colonoscopy screening adherence. *PloS one*, 12(6):e0179272, 2017.

[66] Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. fields: Tools for spatial data, 2017. R package version 9.6.

[67] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London, 2015.

[68] Judy X Li, Daniel R Jeske, and Jeffrey A Klein. Sequential analysis methodology for a poisson glmm with applications to multicenter randomized clinical trials. *Journal of Statistical Planning and Inference*, 142(12):3225–3234, 2012.

[69] Roy T Sabo and Ghalib Bello. Optimal and lead-in adaptive allocation for binary outcomes: a comparison of bayesian methodologies. *Communications in Statistics-Theory and Methods*, 2017.

[70] Shein-Chung Chow and Ralph Corey. Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet journal of rare diseases*, 6(1):79, 2011.

[71] Derek C Angus. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA*, 314(8):767–768, 2015.

[72] Christopher S Coffey and John A Kairalla. Adaptive clinical trials. *Drugs in R & D*, 9(4):229–242, 2008.

[73] Roy T Sabo. Adaptive allocation for binary outcomes using decreasingly informative priors. *Journal of biopharmaceutical statistics*, 24(3):569–578, 2014.

[74] Atanu Biswas and Rahul Bhattacharya. A class of covariate-adjusted response-adaptive allocation designs for multitreatment binary response trials. *Journal of biopharmaceutical statistics*, pages 1–15, 2018.

[75] Li-Xin Zhang, Feifang Hu, Siu Hung Cheung, and Wai Sum Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, pages 1166–1182, 2007.

[76] Selvakkadunko Selvaratnam, Yanqing Yi, and Alwell Oyet. Maximum likelihood estimation of generalized linear models for adaptive designs: Applications and asymptotics. *Biometrical Journal*, 2018.

[77] Zita Oravecz, Matt Huentelman, and Joachim Vandekerckhove. Sequential bayesian updating for big data. *Big Data in Cognitive Science*, pages 13–33, 2016.

[78] John Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4:641–649, 1992.

VITA

Brian Sebastiano Di Pace was born on October 21, 1988, in Hartford, Connecticut, and is an American citizen. He graduated from the Connecticut International Baccalaureate Academy, East Hartford, Connecticut in 2007. He received his Bachelor of Science in Biological Sciences from the University of Connecticut, Storrs, Connecticut in 2011. He received a Master of Public Health in Epidemiology from Eastern Virginia Medical School, Norfolk, Virginia in 2013. During his time at Virginia Commonwealth University (VCU), he interned with the the VCU Department of Health Behavior and Policy, PharPoint Research, Inc. in Durham, North Carolina, and the Biostatistical Consulting Laboratory in the VCU Department of Biostatistics.