Comparing Three Regularization Methods to Avoid Extreme Allocation Probability in Response-Adaptive Randomization

Yining Du[a], John D. Cook[b], and J. Jack Lee[c,*]

[a] *Department of Biostatistics, Incyte Corporation, 1801 Augustine Cut-Off, Wilmington, DE 19803, USA*

[b] *Singular Value Consulting*

[c] *Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA*

Abstract

We examine three variations of the regularization methods for response adaptive randomization (AR) and compare their operating characteristics. A power transformation is applied to refine the randomization probability. The clip method is used to bound the randomization probability within specified limits. A burn-in period of equal randomization (ER) can be added before AR. For each method, more patients are assigned to the superior arm and overall response rate increase as the scheme approximates simple AR, while statistical power increases as it approximates ER. We evaluate the performance of the three methods by varying the tuning parameter to control the extent of AR to achieve the same statistical power. When there is no early stopping rule, PT method generally performed the best in yielding higher proportion to the superior arm and higher overall response rate but with larger variability. The burn-in method showed smallest variability compared with the clip method and the PT method. With the efficacy early stopping rule, all three methods performed more similarly. The PT and clip

*Corresponding author. Tel.:(1)-713-794-4158. E-mail address:jjlee@mdanderson.org.

methods are better than the burn-in method in achieving higher proportion randomized to the superior arm and higher overall response rate but burn-in method required fewer patients in the trial. By carefully choosing the method and the tuning parameter, RAR methods can be tailored to strike a balance between achieving the desired statistical power and enhancing the overall response rate.

Abbreviations: SAR, simple response-adaptive randomization; RAR, response-adaptive randomization; ER, equal randomization; PT, power transformation

Keywords: Bayesian methods, clinical trial design, early stopping, operating characteristics

1 Introduction

In clinical research, randomization plays an important role. Usually, the ultimate goal of a clinical trial is to evaluate the effect (e.g., efficacy and safety) of a tested treatment compared to that of a control (e.g., a placebo or a standard therapy). The appropriate use of a randomization procedure is important, not only to provide an unbiased assessment regarding the efficacy and safety of the tested treatments, but also to form the theoretical basis of valid hypothesis testing. Detailed discussions about randomization in clinical trials can be found in Zelen [1], Rosenberger and Lachin [2], and Hu and Rosenberger [3]. An increased trend of randomization in phase II trials has also been reported [4].

Adaptive randomization designs are methods that change the randomization probability over time depending upon the cumulative information obtained from the previously enrolled patients. The application of adaptive randomization designs in clinical trials has received more attention in recent years and response-adaptive randomization (RAR) methods have been developed in different settings [5]. Adaptive randomization designs can have advantages over conventional randomization schemes because they can adaptively assign more patients to the better treatment in the trial based on the accrued data. Moreover, the randomized probability within the RAR can be constructed

to balance the goals of maximizing the statistical power for testing the treatment efficacy (group ethics) and treating each patient with the best possible treatment (individual ethics). A detailed review of adaptive randomization methods in clinical trials can be found in Chow and Chang [6]. In 2010, a guidance draft issued by the U.S. Food and Drug Administration on adaptive designs in clinical trials highlighted the importance of such designs, including RAR designs. The guidance discussed the need for a clear definition of the scope of the term "adaptive design" in clinical trials, and considered adaptive designs as a critical component of a development program in clinical trials [7].

In randomized trials, patients are assigned to the different treatment arms with certain probabilities. Equal randomization (ER) assigns patients to the treatment arms with equal probability throughout the trial. In RAR, the probability of treatment assignment changes according to the outcome of patients who have already been treated in the trial, with the goal of assigning more patients to the more effective treatment(s) based on the observed interim data. The original concept of RAR can be traced back to Thompson [8], who was devoted to the theoretical derivation and calculation of randomization probability. Incorporating the sequentially accruing outcome data into future randomized probabilities, RAR procedures have been successfully applied in clinical trials [2], [9]. One of the earliest applications of RAR for treatment assignment was Zelen's play-the-winner design [10]. Subsequently, a generalization of that design, the randomized play-the-winner design, was developed by Wei and Durham [11]. A further development of this type of design was the doubly-biased-coin-design, proposed by Hu and Zhang [12]. RAR methods have more recently been extended to address the issue of delayed response [13], [14], incorporate multiple treatment arms in a trial [15], and to implement covariates [16]. Zhang and Rosenberger developed RAR designs with continuous endpoints [17] and later discussed the application of RAR on the basis of survival outcomes [18]. In this paper, we focus on the use of RAR to compare two treatment arms with binary outcomes based on the Bayesian framework. Since the characteristics of a Bayesian adaptive trial will depend on the definition of the design parameters and the randomization probabilities,

the operating characteristics presented in this paper are not trying to match the entire class of adaptive design. A detailed evaluation of the RAR design for binary outcomes can be found in Hu and Rosenberger [19]. In addition, the value of RAR has been debated in Korn and Freidlin [20] and Lee et al. [21]. Recently, Thall et al. [22], [23] simulated some response adaptive randomization designs and pointed out that the designs could have large variability in the allocation ratio and a non-ignorable percentage of patients randomized into the inferior arms. Furthermore, the ethical implications of the response adaptive designs have been reviewed and discussed in recent articles [24], [25], and [26].

In this paper, we compare three variations of response-adaptive randomization under a Bayesian framework. We vary the tuning parameters in the allocation probabilities to approximate either simple response-adaptive randomization (SAR) or ER. The paper is constructed as follows: In Section 2, we introduce the three regularization methods as compromises between SAR and ER. We describe the simulation settings in Section 3. In Section 4, we report the simulation results for trials both with and without an efficacy early stopping rule. We evaluate the operating characteristics of the three regularization methods with respect to the mean proportion of patients assigned to the superior treatment arm, mean overall response rate, statistical power, and mean total number of patients required in the trial when using an efficacy early stopping rule. We also study the variability of the three methods by reporting the 10% and 90% qunatiles of these measures. We provide some recommendations for designing adaptive randomization designs in Section 5. We present additional plots and detailed simulation results in tabular form in the Appendix. Note that the characteristics of Bayesian adaptive designs depend on the definition of the design parameters and the randomization probabilities etc., the operating characteristics presented in this paper are not trying to match the entire class of adaptive designs.

2 Methods

A primary goal of clinical trials is to compare efficacy among different treatment arms. In addition, it is desirable to treat more patients with the better treatment, if any, in the trial. Data accumulate sequentially as the clinical trial moves along. If one treatment seems to be more efficacious than another, it is ethically appealing to consider treating more patients with the superior treatment and fewer patients with the less promising treatments until a definitive conclusion in comparing the treatment efficacy can be drawn. We can use the available patients' outcomes observed at the time of randomization to skew the allocation probability of a new patient toward the treatment(s) that show a better results. This is the RAR scheme. ER is commonly used as the conventional randomization design in clinical trials. It provides a scientifically valid inference for comparing treatment effects. However, either from prior knowledge or based on interim results, when there is compelling evidence that one treatment is likely to be superior to the other(s), continuing to use ER in the trial may raise ethical concerns. ER holds a constant probability of randomization throughout the trial, which may result in treating more patients than necessary with a treatment that is inferior.

Clinical trials often have multiple and sometimes competing objectives. Maximizing the statistical power and maximizing the benefit to the individual patient within the trial can be considered as competing goals. Several approaches of RAR that incorporate different preferences or objectives in a trial design have been proposed to explore more efficient and ethical approaches in the development of clinical trials. Thall and Wather [27] introduced the method using the tuning parameter $n/2N$, which varies with the number of patients that have been previously randomized in the trial. In this paper, we examine the operating characteristics of three adaptive randomization methods by varying the tuning parameter in the allocation probabilities. The tuning parameter can be chosen such that the randomization scheme leans toward ER or toward SAR.

We consider a two-arm trial with a binary outcome. Let θ_i be the probability of response on

arm i, where i is 1 or 2. We assume a beta prior for θ_i with the distribution:

$$\theta_i \sim Beta(\alpha_i, \beta_i).$$

Therefore, the SAR design under the Bayesian framework assigns treatment i with probability

$$p = P_r(\theta_i > \theta_j | data), \quad i \neq j. \tag{1}$$

To avoid extreme allocation of the randomization probability under SAR, we propose three regularization methods, the power transformation method, clip method, and burn-in method.

2.1 Power Transformation (PT) Method

In the first method, we adjust the posterior probability in equation 1, and apply a power transformation to that probability

$$f(p, \lambda) = \frac{p^\lambda}{p^\lambda + (1-p)^\lambda}$$

for a given $\lambda \geq 0$. That is, rather than assigning treatment i with probability p, we first calculate p and assign treatment i with probability $f(p, \lambda)$. We represent this approach as PT(t), where t denotes the tuning parameter. Note that $f(p, 1) = p$; thus, the PT method reverts to the SAR design when $t = 1$. Also, $f(p, 0) = 1/2$ for all $0 < p < 1$; therefore, when $t = 0$, the PT randomization scheme reduces to the ER scheme. We show that as we vary t between 0 and 1, the operating characteristics change continuously between those of ER and SAR.

Therefore,

$$\begin{cases} PT(0) = ER, \\ PT(1) = SAR. \end{cases}$$

2.2 Clip Method

The second approach, which also represents a compromise between ER and SAR, is to simply clip the randomization probabilities by setting a minimum randomization probability that $r \leq 1/2$. If

p falls below r, treatment i is assigned with probability r. By symmetry, this implies that if the probability p rises above $1-r$ then the treatment is assigned with probability $1-r$. For moderate values of p, those probabilities are between r and $1-r$, and thus are left unchanged. Explicitly, a patient is assigned treatment i with probability

$$g(p,r) = \max\left(r, \min(p, 1-r)\right).$$

We note that when $r = 0$, the clip method is reduced to the SAR scheme; it is equivalent to the ER scheme when $r = 0.5$. We refer to this approach as Clip(t) where t is the tuning parameter. To be consistent in notation and the extent of AR, we let $t = 1 - 2r$.

Therefore,

$$\begin{cases} Clip(0) = ER, \\ Clip(1) = SAR. \end{cases}$$

2.3 Burn-in Method

The third variation on RAR we consider is the adoption of a burn-in period of ER. Let N be the maximum accrual of a trial in which the first n patients are assigned to the treatments using blocked randomization (i.e., $n/2$ patients randomized to each treatment), and the remaining $N-n$ patients are randomized to treatment i with probability p as defined in equation 1. We denote this design as Burn(t), where $t = 1 - n/N$ is the tuning parameter.

We note that when $t = 0, n = N$ the burn-in method is the ER scheme, and when $t = 1, n = 0$ it corresponds to the SAR scheme.

Hence,

$$\begin{cases} Burn(0) = ER, \\ Burn(1) = SAR. \end{cases}$$

2.4 An Illustrative Example

Next, we demonstrate how these three approaches perform when applied to a specific example. For a two-arm trial with 80 patients, we simulated 20 realizations of the trial to illustrate the underlying variabilities. The prior distribution is assumed that $\theta_i \sim Beta(0.6, 1.4)$. Setting the power to 80%, tuning parameters are 0.43, 0.58, and 0.65 for the PT, clip and burn-in methods, respectively. Under the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$, we plot their performances of the allocation probability in Figure 1. The x-axis represents the total patients in the trial; the y-axis represents the probability of allocation to the superior arm (arm 2). For the PT method, the change in the allocation probability as the trial evolved is illustrated in the top panel. In several trials, the probabilities of randomization to arm 2 dropped to below 0.4. Such trends reverted as the trial evolved. As the sample size reached 80, the probability of allocation to arm 2 is increased to greater than 0.8 for most realizations. We can see the variablity of the process is larger in the beginning of the trial but gets smaller toward the end of the trial. Regarding the change in the allocation probability of the clip method (the middle panel), the highest probabilities are 0.79 according to the threshold assigned. We can see that the allocation probability is bounded between 0.21 and 0.79. The variability is larger in the beginning of the trial and diminishes toward the end of the trial with the randomization probability reaches to 0.79. For the burn-in method (the bottom panel), given $t = 0.65$, there were 28 patients for burn-in period with equal randomization probability to each treatment. After 28 patients, we see a large variability in the randomization probability when AR begins. As the trial evolved, the allocation probabilities are larger than 0.9 for most trials.

3 Simulation Setting

In this section, we describe the simulation settings we use in this work. In consideration of striking a balance between group ethics and individual ethics, researchers may use different criteria when evaluating a trial. It is desirable that clinical trials have a high probability of being able to identify

effective treatments being evaluated (statistical power) and also have a high proportion of patients assigned to the superior treatments, which is associated with a high overall response rate. We examine the proportion of patients assigned to the superior arm, overall response rate, statistical power, and total number of patients required in the trial when adding an efficacy early stopping rule among these three randomization regularization approaches.

In clinical trial design, there usually exists a trade-off between correctly inferring the results at the end of the trial and assigning more patients to the superior treatment arm during the trial. Generally, ER designs provide higher statistical power (with the exception of the Neyman allocation ratio, which attains the highest power). Adaptive randomization designs generally assign a larger proportion of patients to the better treatment arm and thus obtain a higher favorable response rate at the end of the trial. By varying the tuning parameters of the three regularization methods to achieve the same power, we compare the operating characteristics between the two extremes (ER or SAR), and provide recommendations regarding their use.

To explore and compare the operating characteristics of the three regularization approaches, we simulated a two-arm trial with a maximum of 80 patients. We assumed the prior for the response probabilities θ_i distributed as Beta(0.6, 1.4), and examined the following scenarios. In each we assumed the true probability of response on arm 1 was 0.2. The probabilities of response on arm 2 were $\{0.2, 0.3, 0.4, 0.5\}$. We also denoted the observed data as x. We calibrated the threshold, denoted by C_T, so that the trial could fulfill the frequentist operating characteristics of controlling the type I error rate. Without an early stopping rule, one of the following decisions would be made at the end of the trial:

1. If $P_r(\theta_2 > \theta_1|x) \geq C_T$, we claim that treatment 2 is superior.

2. If $P_r(\theta_1 > \theta_2|x) \geq C_T$, we claim that treatment 1 is superior.

3. Otherwise, we claim that the two treatments are equally effective.

For each method, we performed 100,000 simulations for each value of the tuning parameter t. We

controlled the type I error rate at $\alpha = 0.10$, and set each value of t between 0 and 1 using increments of 0.1. Note that except for the boundaries cases (t=0 or 1), the same tuning parameter t in different regularization methods could result in different randomization probabilities. To provide a fair comparison, we choose different t's in the three methods yielding the same statistical power in testing the treatment effect. We present the average performance and the percentiles in some cases.

Similarly, implementing an efficacy early stopping rule, we stop the trial early if at any point

$$P(\theta_i > \theta_j | data) > C_S \tag{2}$$

where C_S is the threshold to control the type I error rate at 0.10. It was calibrated under the null hypothesis ($\theta_1 = \theta_2 = 0.2$) with 100,000 simulations and chose the value of C_S such that the proportion of claiming treatment 1 or 2 better equalled to 0.05, respectively. If we reach the maximum number of patients without either arm satisfying the inequality above, we declare that the two treatments are equally effective.

We first discuss the results for the three regularization methods without using an early stopping rule, accruing the maximum enrollment of 80 patients in each simulation. Then we consider the scenarios with the addition of an efficacy early stopping rule.

4 Results

4.1 Simulation Results Without An Early Stopping Rule

We present the simulation results in figures and provide detailed information in tabular form in the online Supplemental Tables 1-3 summarizing the simulation results for the three methods (PT, clip, and burn-in) by varing the tuning parameters from 0 to 1 without adding an efficacy early stopping rule. In the tables, we present the mean proportions of patients assigned to the superior arm, mean overall response rates, and probabilities of determining which treatment is superior (or that the treatments are equally effective) under 4 scenarios with $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.3, 0.4$ and

0.5, respectively. We provide the thresholds for the decision rules in the tables.

In Figure 2, we compare the performance of the three methods without early stopping. The distributions of the overall response rate and proportion of patients assigned to the superior arm yielding the same power are shown in the box plots. We chose the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$, and plotted these distributions resulting in 70%, 80%, and 85% power, respectively. From the figures, we see that in terms of having higher overall response rate and more patients assigned to the better arm, the PT method performs the best among all three methods. The burn-in method had the smallest variability in the percentage of patients assigned to better arms among all three methods in all settings. The clip method assigns more patients to the better arm at 75% power but fewer at 80% and 85% power compared with the burn-in method.

Also, in Supplemental Figure 1 we present the performances of mean proportion of patients assigned to the superior arm (arm 2), mean overall response rate, and probability of declaring that arm 2 is better (statistical power) under 4 scenarios where $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$ as the tuning parameter t varies between 0 and 1. When $t = 0$, the randomization approach represents the ER design, and when $t = 1$, it represents the SAR design. For example, when $\theta_1 = 0.2$ and $\theta_2 = 0.5$, the mean proportion randomizing to Arm 2 increased from 0.5 to 0.85 but the power reduced from 0.89 to 0.69 as t increase from 0 to 1.

In addition, we provide the comparisons of the variability of the three regularization methods in Supplemental Figure 2 showing the performance of the 10% and 90% quantile estimates of the proportion to the superior arm (arm 2) and overall response rate for the 3 methods under 3 scenarios, when $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.4, 0.5\}$. In each figure, the bottom three lines represent the 10% quantile estimates and the upper three lines represent the 90% quantile estimates. Under the null hypothesis ($\theta_1 = \theta_2 = 0.2$), the 10% and 90% quantile estimates for the proportion to the superior arm were around 40% and 60% respectively for ER design. As the tuning parameter increased, the difference between the 10% and 90% quantile estimates for the proportion on arm 2 widened. The three regularization methods performed similarly. For SAR, the proportion on arm 2 was 0.15 for

the 10% quantile and 0.85 for the 90% quantile, and the overall response rate was close to 0.15 for the 10% quantile and more than 26% for the 90% quantile. For the scenarios $\theta_2 = 0.4$ and $\theta_2 = 0.5$, in terms of the proportions to the superior arm, the PT method had the largest variability followed by the clip method. The burn-in method was least variable. In terms of the overall response rate, the differences in variability among all methods were less pronounced. Generally speaking, the PT method yielded the highest overall response rate for both the 90% and 10% quantiles compared to the clip or the burn-in method. The differences in the overall response rate among the three methods were not as large as that in the proportion to the better arm. One caveat in examining the figures by varying the tuning parameter t is that the same t could result in different adaptive randomization proportions in different regularization methods.

In general, both the mean proportion of patients assigned to the superior arm and the mean overall response rate increased when the methods were closer to SAR; whereas the statistical power increased when the methods were closer to ER. SAR obtained both the higher mean proportion of patients assigned to the better treatment arm and higher mean overall response rate, while the ER design provided larger power. However, when the difference in the true probability of response between the two arms increased, SAR designs can also achieve sufficiently high power.

4.2 Simulation Results With An Early Stopping Rule

In this section, we discuss the simulation results when we add the efficacy early stopping rule given in equation 2. We present the results in Figure 3, Supplemental Figures 3, 4 and 5, and provide detailed information in online Supplemental Tables 4-6. Similarly to the scenario without early stopping, we present the characteristic performances for $\theta_1 = 0.2$ and $\theta_2 = 0.5$ in Figure 3. With the early stopping rule applied, we provide the distributions of the overall response rate, proportion of patients assigned to the better arm, and total patients enrolled in the trial for three methods resulting in 72%, 75%, and 80% power, respectively. PT method and clip method performed similarly and better than the burn-in method in terms of the overall response rate and proportion

of patients assigned to the better arm, but burn-in method enrolled slightly fewer patients in the trial than the other two methods.

In addition to the mean proportion of patients assigned to the superior arm, mean overall response rate, and power, Supplemental Figure 3 also shows the mean total number of patients required in the trial for the 3 methods. The addition of the early stopping rule leads to somewhat different operating characteristic curves. In Supplemental Figure 3, we also observed an increasing trend of the mean proportion of patients assigned to the superior arm and mean overall response rate; a decreasing power trend when the three regularization methods were close to the SAR design; and larger power when the three methods were close to the ER design. Furthermore, since the SAR design required more patients than the ER design, the curves of the mean total patients required in the trial were decreasing when the tuning parameter was reduced to closer to the ER design.

Supplemental Figure 4 presents the performance of the 10% and 90% quantile estimates of the 3 methods. Similar to the scenario without early stopping, when $\theta_1 = \theta_2 = 0.2$, SAR had larger variability than ER, providing smaller 10% quantile estimates and larger 90% quantile estimates in terms of the proportion of patients assigned to the superior arm. For the scenarios $\theta_2 = 0.4$ and $\theta_2 = 0.5$, the patterns of the 10% and 90% quantile estimates were consistent with the mean estimates. Considering the total patients in the trial, all methods had very similar 90% quantiles across the range of the tuning parameters (i.e., from ER to SAR). Regarding the 10% quantile, the curves of the total sample size required were also largely flat when the tuning parameter ranges from 0 to 1 for all three methods except that the burn-in method had smaller sample size especially when tuning parameter is around 0.9.

Lastly, we provide the distributions of total patients enrolled in the trial with the histograms in Supplemental Figure 5. The plots are from the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$ given power to 72%, 75%, and 80%, respectively. From the histograms, we can find that the required sample sizes for the three methods were very similar. About 20% to 30% of trials were not stopped earlier. In summary, with the early stopping rule applied, PT method and clip method performed similarly

and better than the burn-in method in terms of the overall response rate and proportion of patients assigned to the superior arm, but burn-in method required fewest patients enrolled in the trial. Similar to the scenarios without early stopping, SAR design shows higher overall response rate and larger proportion of patients to the better arm, but ER design presents higher power. Considering the variability, ER design has smaller variability than SAR design.

4.3 Sensitivity Analysis of the Choice of the Prior Distribution of the Response Rate

We selected some representative parameters to perform the sensitivity analysis to the choice of the prior for both scenarios without and with early stopping. The simulation results are from the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$ with the tuning parameter $t = 0.5$. From the results below, we see that the design is not very sensitive to the choice of the prior. The results are shown in the Supplemental Table 7.

5 Discussion

The PT, clip, and burn-in methods are three generalizations of the RAR design under the Bayesian framework. Each contains a parameter that can be used to create designs with operating characteristics that are intermediate between those of ER and SAR. For each method, the mean proportion of patients assigned to the superior arm and mean overall response rate increase as the tuning parameter changes from 0 to 1 (ER to SAR). The probability of correctly selecting the better arm (power) at the end of the trial decreases as the method approximates SAR. However, when the difference in the true probability of response between the two arms increases, the methods that approximate SAR can also achieve sufficiently large power. When the difference in the true response rate between the two arms increases, the gains achieved by SAR over ER are more obvious. One may use any of these methods when designing an adaptively randomized trial with properties somewhere between those

of ER and SAR, according to the trade-off between obtaining higher statistical power and treating patients in the trial most effectively. One drawback of the SAR method is that the variability of the trial design measures (such as the allocation proportion, overall response rate, and sample size, etc.) is larger than the ER method which can result in undesirable operating characteristics [23]. The variability can be mitigated by reducing the tuning parameter and applying it to the three regularization methods to avoid extreme allocation probability. We also explored the sensitivity to the choice of the prior. When the prior is non-informative, the measurements are not sensitive to the chioce of the priors.

Without an early stopping rule for efficacy, the simulation results suggest that of these 3 methods, the PT method assigns the most patients to the superior arm and achieves higher overall response rate for a given statistical power. Therefore, we recommend the PT method. With the early stopping rule, the PT and clip methods appear to perform equally well, and the burn-in method performs the worst. Considering the mean total number of patients enrolled in the trial when adding an early stopping rule, the PT and clip methods require more patients in the trial than the burn-in method. We examined only designs of the form PT(t) with $t \leq 1$, though designs with $t > 1$ are possible.

We suggest statisticians begin exploring clinical trial designs by simulating PT(0) and PT(1) methods. The PT(0) approach will achieve the most statistical power. The PT(1) approach will provide an idea of how many more patients can be assigned to the superior arm, and at what loss of power, under given scenarios. Alternatively, if one wants a combination of power and treatment imbalance somewhere between that of PT(0) and PT(1), interpolation of the tuning parameter t will provide an initial guess at an acceptable value of t. One may also consider hybrid designs, for example, using a burn-in period initially following with a PT(t) method. By carefully choosing the regularization method and the tuning parameter, response-adaptive randomization methods can be tailored to strike a balance between achieving the desired statistical power and enhancing the overall response rate in computing the required sample size in clinical trials.

6 Conflict of Interest

There are no conflicts of interest to declare.

7 Acknowledgment

JJL's work was supported in part by grant CA016672 from the National Cancer Institute.

References

[1] M. Zelen. The randomization and stratification of patients to clinical trials. Journal of Chronic Diseases, 27: 365-375, 1974.

[2] W.F. Rosenberger and J.M. Lachin. Randomization in Clinical Trials: Theory and Practice. New York: John Wiley and Sons, 2002.

[3] F.F. Hu, and W.F. Rosenberger. The theory of response-adaptive randomization in clinical trials. John Wiley and Sons, 2006.

[4] J.J. Lee, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. J Clin Oncol, 23(19): 4450-7, 2005.

[5] W.F. Rosenberger, and S. Oleksandr, and F.F. Hu. Adaptive randomization for clinical trials. Journal of Biopharmaceutical Statistics, 22: 719-736, 2012.

[6] S.C. Chow and M.Chang. Adaptive design methods in clinical trials - A review. Orphanet Journal of Rare Diseases, 3: 11, 2008.

[7] T. Cook and D.L. DeMets. Review of Draft FDA Adaptive Design Guidance. Journal of Biopharmaceutical Statistics, 20: 1132-1142, 2010.

[8] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25: 275-294, 1933.

[9] E.S. Kim, R.S. Herbst, Wistuba II, J.J. Lee, G.R. Blumenschein, A. Tsao, D.J. Stewart, M.E. Hicks, J. Erasmus, S. Gupta, C.M. Alden, S. Liu, X. Tang, F.R. Khuri, H.T. Tran, B.E. Johnson, J.V. Heymach, L. Mao, F. Fossella, M.S. Kies, V. Papadimitrakopoulou, S.E. Davis, S.M. Lippman, W.K. Hong. The BATTLE Trial: Personalizing Therapy for Lung Cancer. Cancer Discov 1(1):44-53, 2011.

[10] M. Zelen. Play the winner rule and the controlled clinical trial. Journal of the American Statistical Association, 64: 131-146, 1969.

[11] L.J. Wei and S. Durham. The randomized play-the-winner rule in medical trial. Biometrics, 73: 840-843, 1978.

[12] F.F. Hu and L.X. Zhang. Asymptotic properties of double adaptive biased coin designs for multi-treatment clinical trials. Annals of Statistics, 32: 268-301, 2004.

[13] L.X. Zhang and W.S. Chan and F.F. Hu. A generalized drop-the-loser urn for clinical trials with delayed responses. Statistica Sinica, 17: 387-409, 2007.

[14] Z.D. Bai and F.F. Hu and W.F. Rosenberger. Asymptotic properties of adaptive designs for clinical trials with delayed response. Annals of Statistics, 30: 122-139, 2002.

[15] J. Wason and L. Trippa. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. Statistics in Medicine, 33: 2206-2221, 2014.

[16] W.F. Rosenberger and A.N. Vidyashankar and D.K. Agarwal. Covariate-adjusted response-adaptive designs for binary response. Journal of Biopharmaceutical Statistics, 11: 227-236, 2001.

[17] L. Zhang and W.F. Rosenberger. Response-adaptive randomization for clinical trials with continuous outcomes. Journal of the American Statistical Association, 62: 562-569, 2006.

[18] L. Zhang and W.F. Rosenberger. Response-adaptive randomization for survival trials: the parametric approach. Journal of the Royal Statistical Society: Series C (Applied Statistics), 56: 153-165, 2007.

[19] F.F. Hu and W.F. Rosenberger. Optinality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. Journal of the American Statistical Association, 98: 671-678, 2003.

[20] E.L. Korn and B. Freidlin. Outcome-adaptive randomization: is it useful? Journal of Clinical Oncology, 29: 771-776, 2011.

[21] J.J. Lee and N. Chen and G.S. Yin. Worth adapting? Revising the usefulness of outcome-adaptive randomization. Clinical Cancer Research, 18: 4498-4507, 2012.

[22] P. Thall and J. Wathen. Statistical Controversies in Clinical Research: Scientific and Ethical Problems with Adaptive Randomization in Comparative Clinical Trials. Annals of Oncology, May 15, 2015.

[23] P.F. Thall, P. Fox and J. Wathen. Scientific and ethical problems with adaptive randomization in comparative clinical trials. Annals of Oncology, 2015.

[24] S.P. Hey and J. Kimmelman. Are outcome-adaptive allocation trials ethical? Clinical Trials, 12(2): 102-106, 2015.

[25] J.J. Lee. Commentary on Hey and Kimmelman. Clinical Trials, 12(2): 110-112, 2015.

[26] S.P. Hey and J. Kimmelman. Rejoinder. Clinical Trials, 12(2): 125-127, 2015.

[27] P.F. Thall and J.K. Wathen. Practical Bayesian adaptive randomization in clinical trials. European Journal of Cancer, 43: 859-866, 2007.

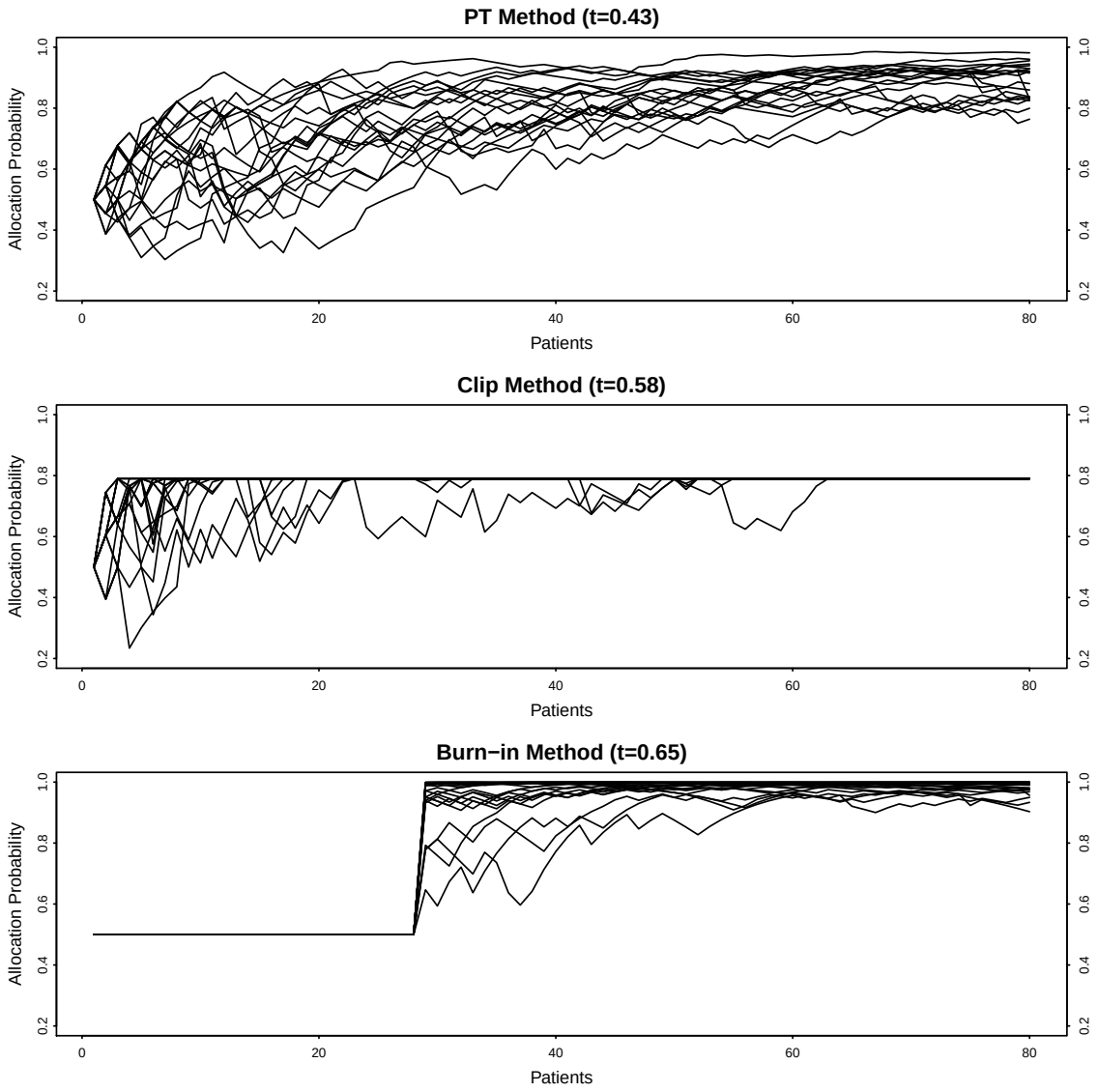


Figure 1: Performances of the allocation probability during the trial for the three regularization methods for adaptive randomization yielding 80% power. A sample of 20 realizations of the trial are shown. The x-axis represents the total number of patients; the y-axis represents the probability of allocation to arm 2, the superior treatment. The true response rates are $\theta_1 = 0.2$ and $\theta_2 = 0.5$.

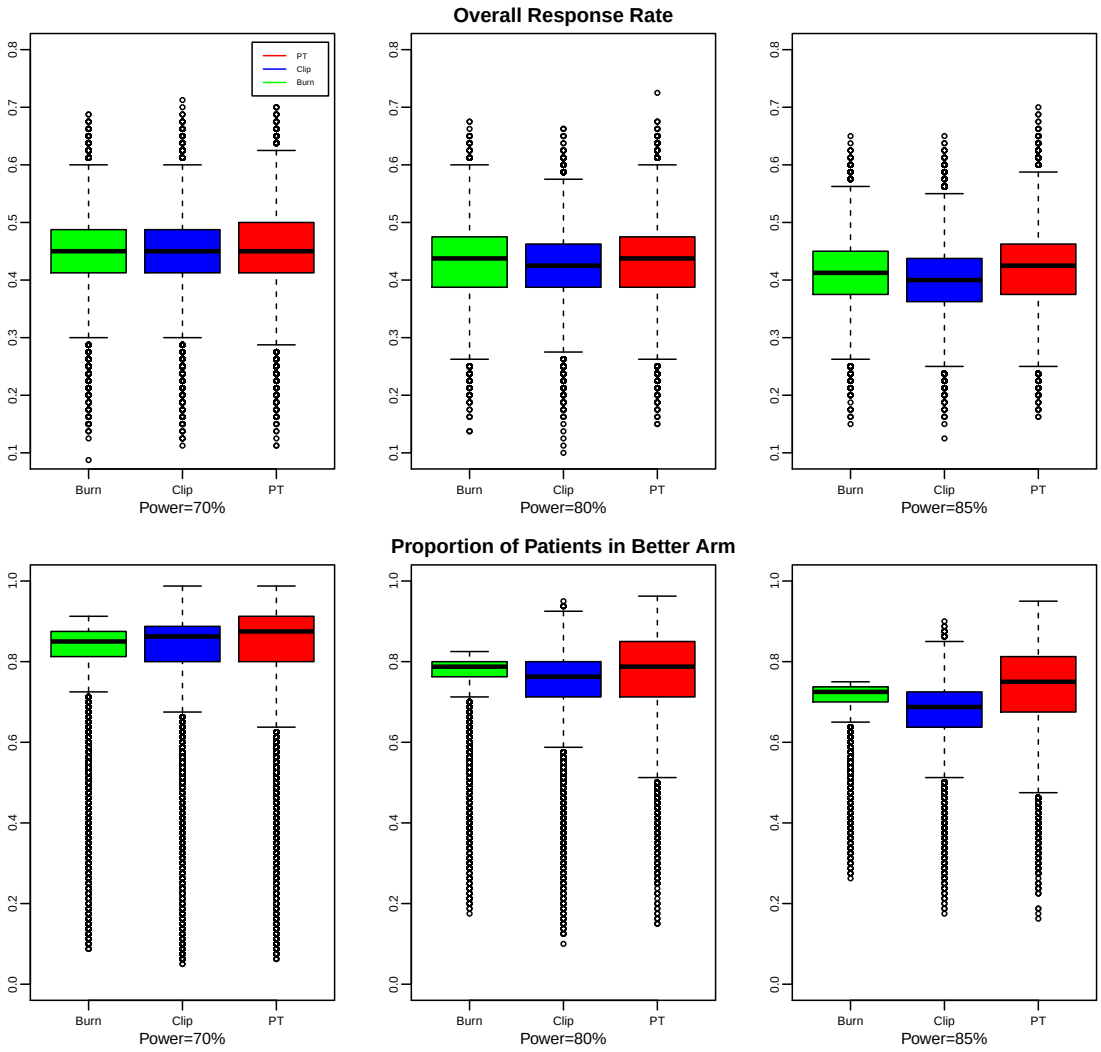


Figure 2: Box plots for the distributions of the overall response rate and proportion of patients assigned to the better arm (arm 2) with no early stopping rule implemented. The left, middle, and right panels show the results at power equals to 75%, 80%, and 85%, respectively. The red plots represent the PT method; blue plots represent the clip method; green plots represent the burn-in method. All the plots are from the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$. Each plot was derived from 100,000 simulations.

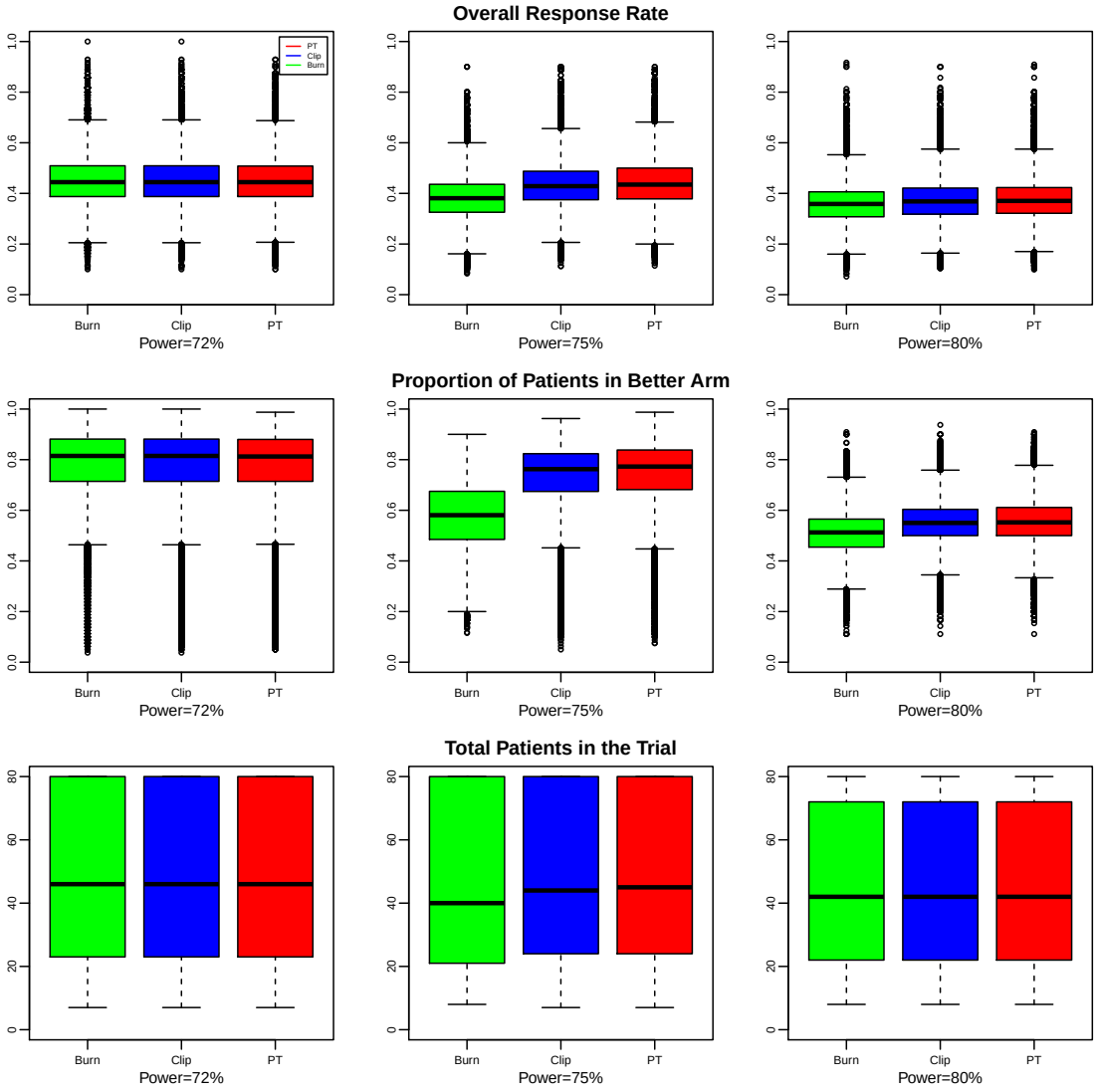
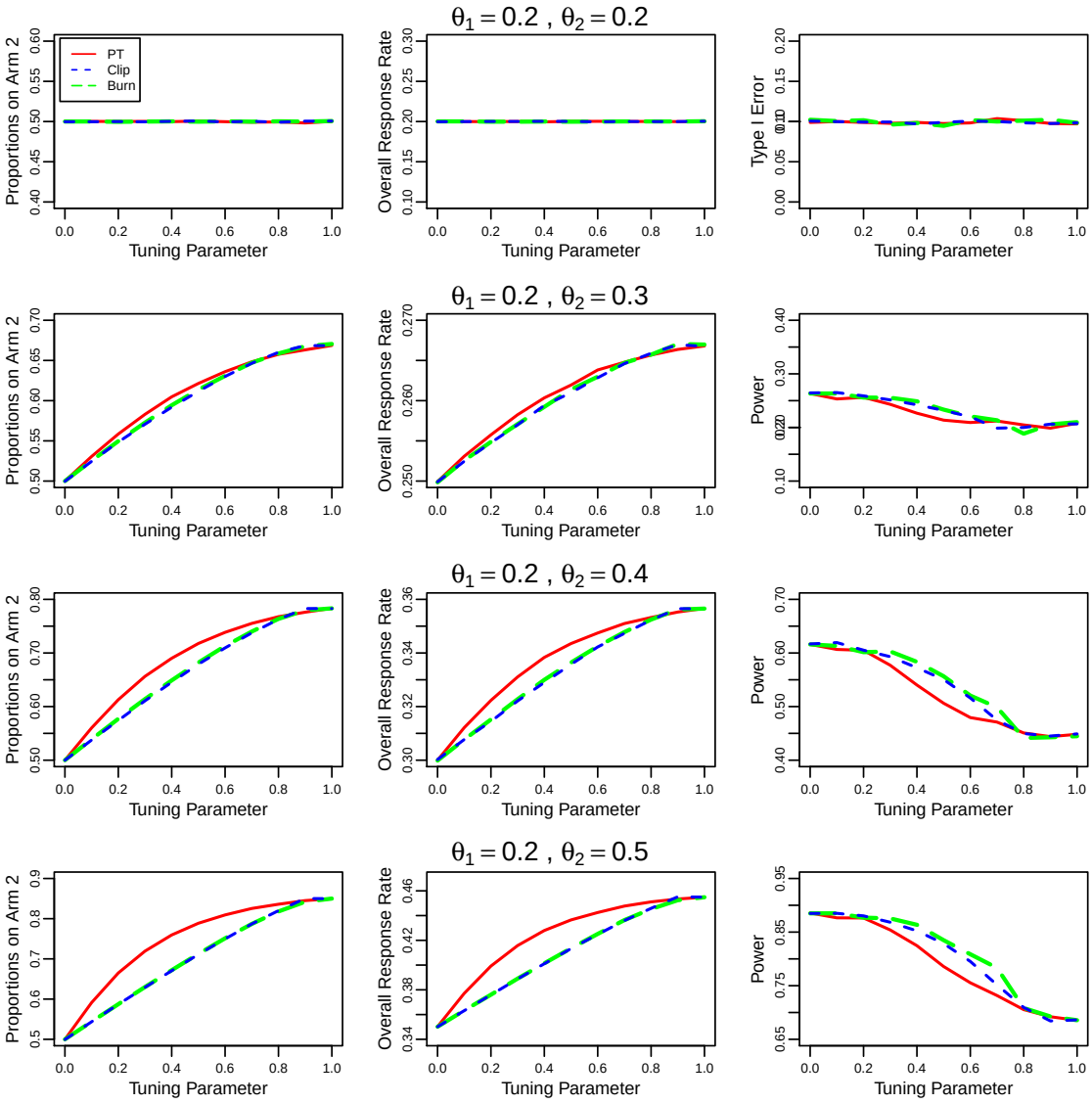
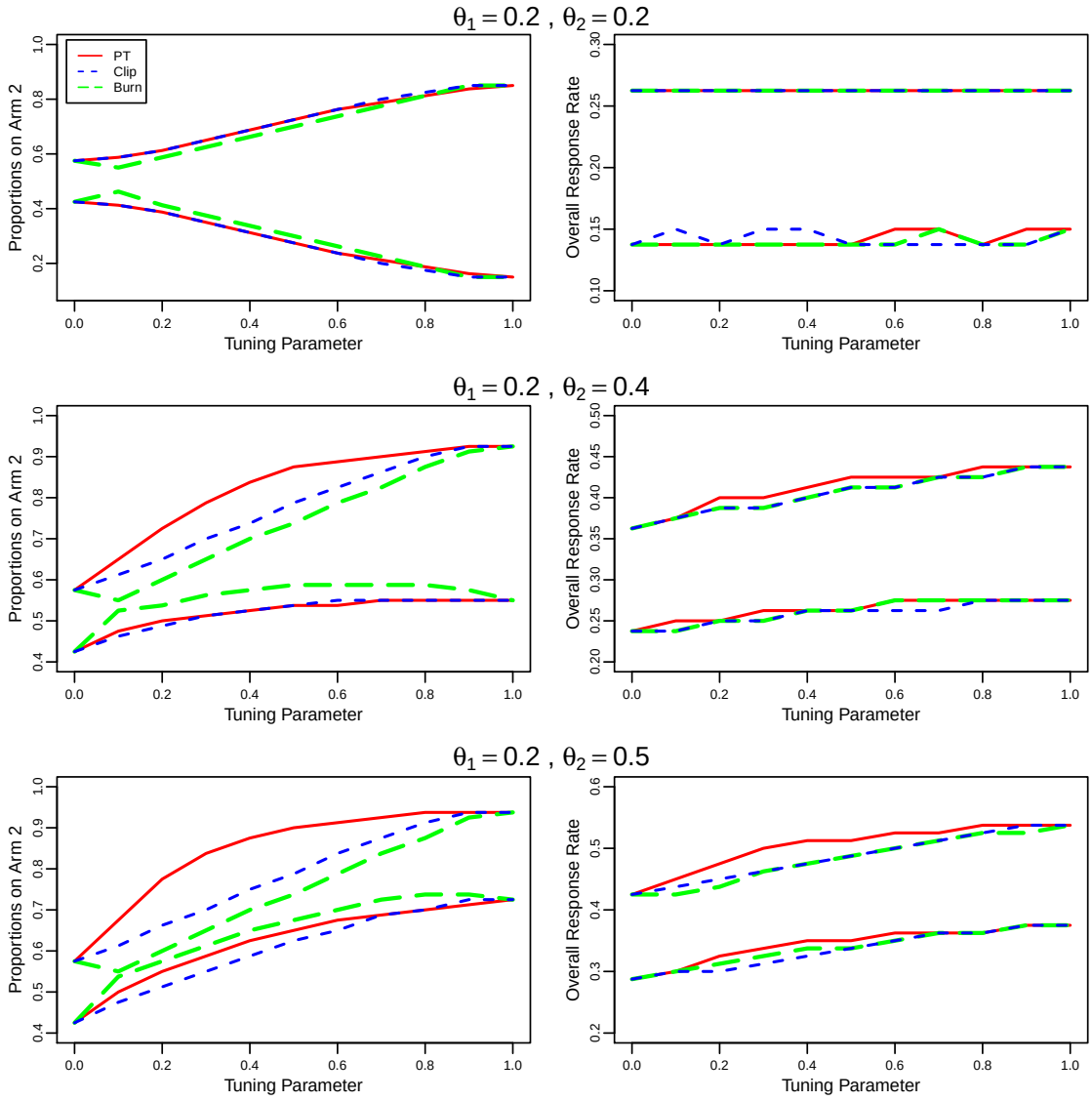


Figure 3: Box plots for the distributions of the overall response rate, proportion of patients assigned to the better arm (arm 2), and total patients in the trial with early stopping rule implemented. The left, middle, and right panels show the result at power equals to 72%, 75%, and 80%, respectively. The red plots represent the PT method; blue plots represent the clip method; green plots represent the burn-in method. All the plots are from the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$. Each plot was derived from 100,000 simulations.

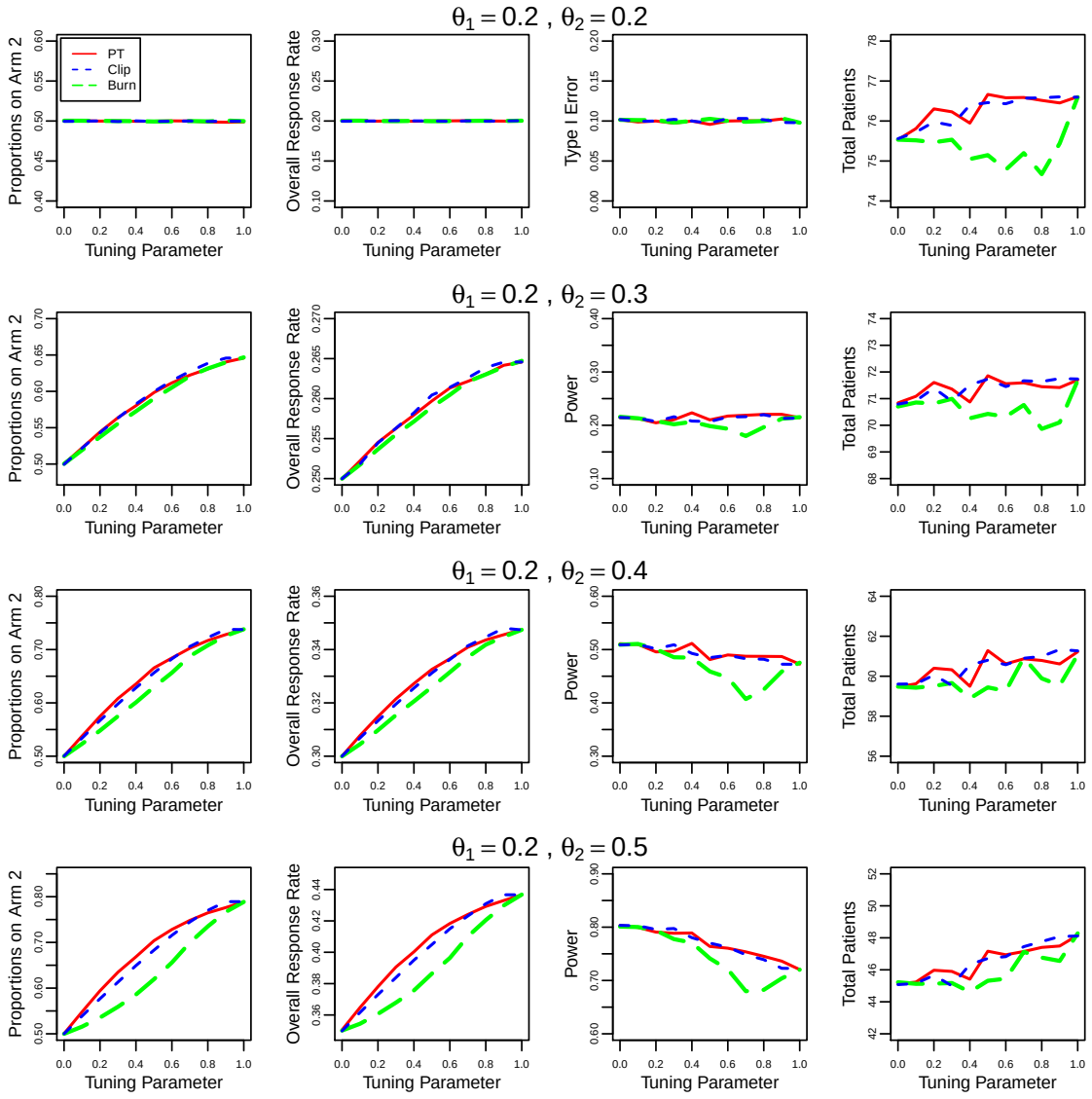
8 Appendix



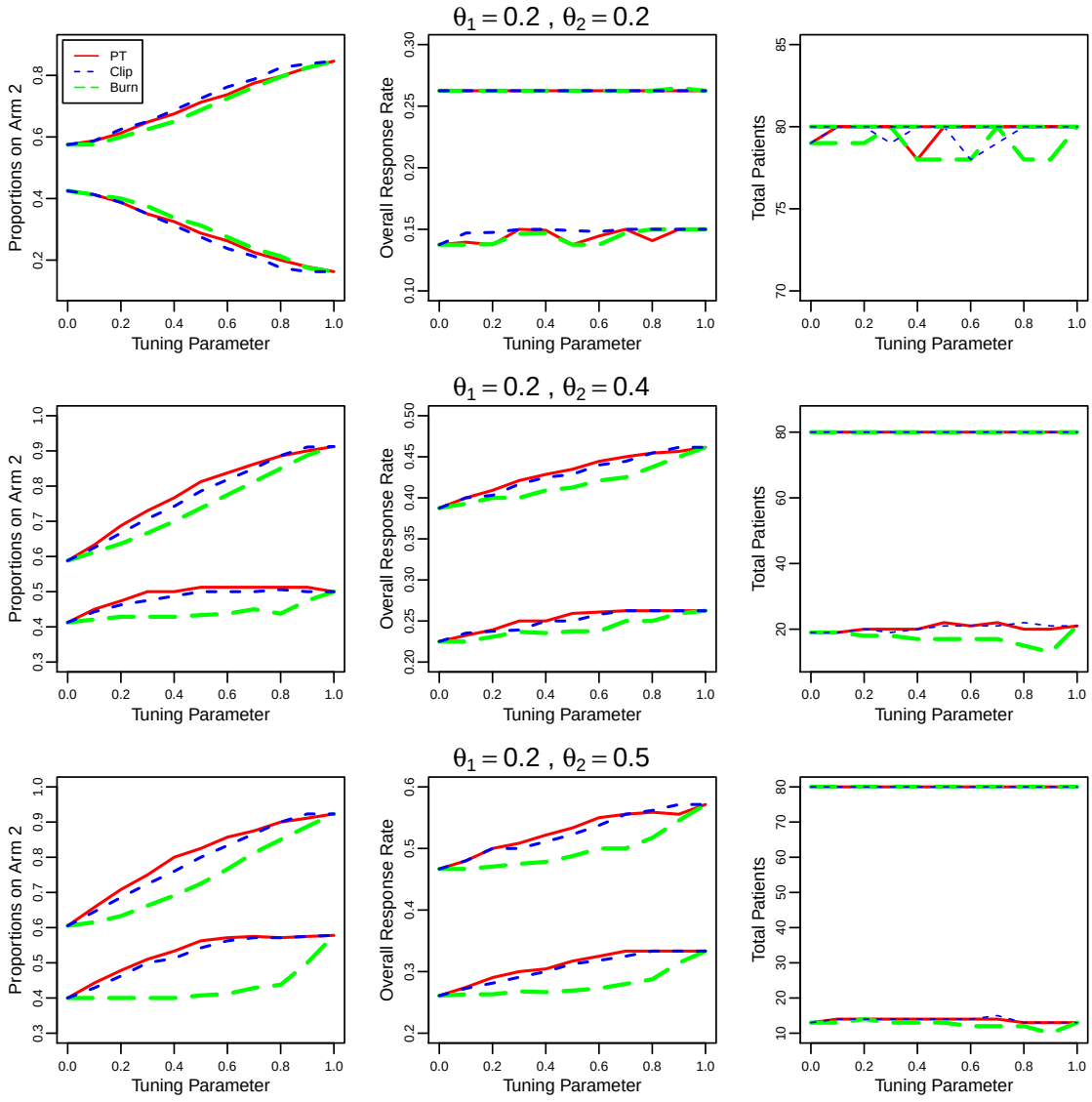
Supplemental Figure 1: Operating characteristics without early stopping for efficacy. Proportion of patients assigned to arm 2 (the superior treatment), overall response rate, and statistical power are shown for the 3 methods under 4 scenarios, where $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$ as t varies from 0 to 1 in increments of 0.1. The red solid lines represent the PT method; blue dashed lines represent the clip method; green dashed lines represent the burn-in method. Each line was derived from 100,000 simulations.



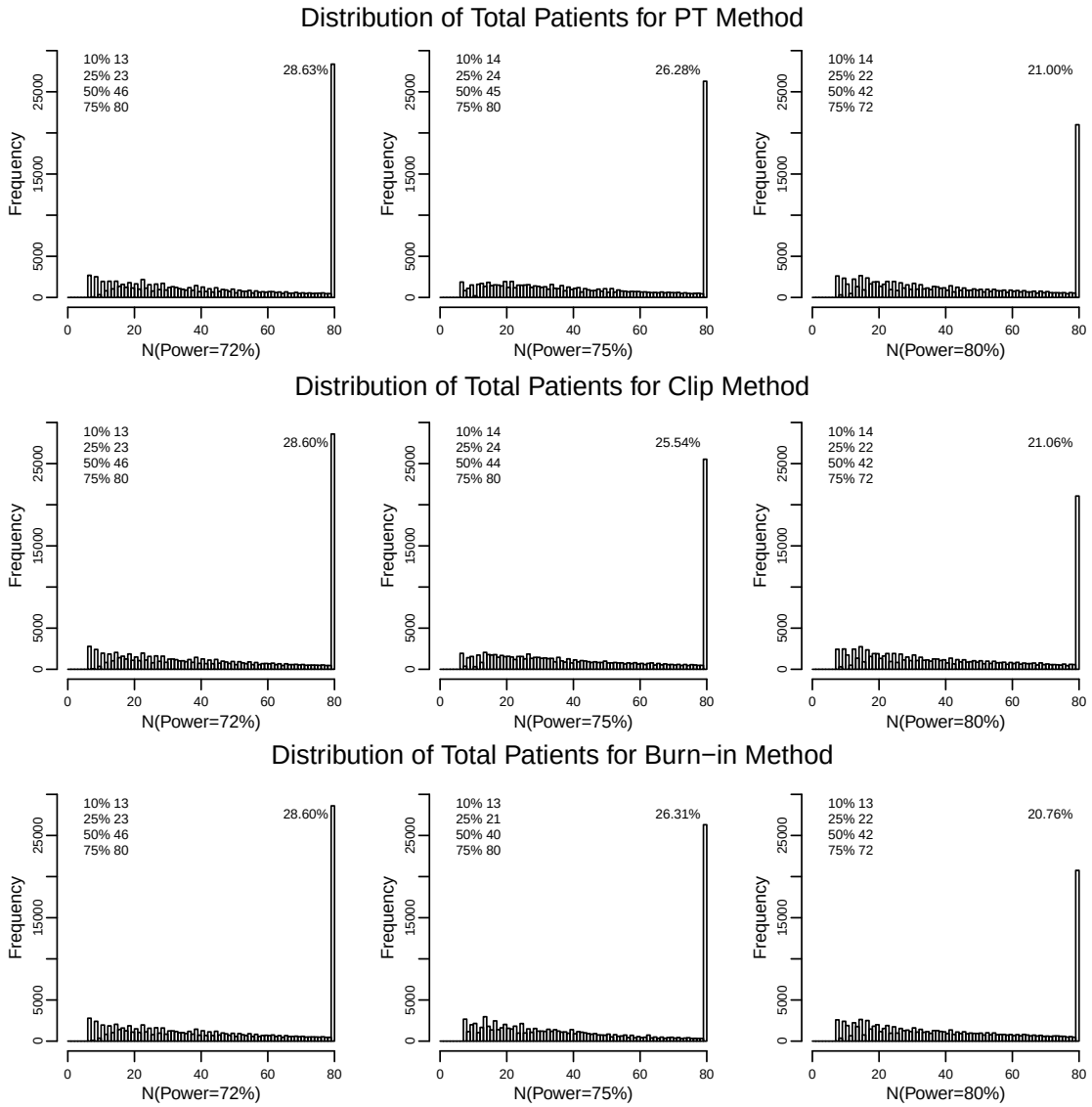
Supplemental Figure 2: Operating characteristics with early stopping for efficacy. Proportion of patients assigned to arm 2 (the superior treatment), overall response rate, statistical power, and total patients are shown for the 3 methods under 4 scenarios, where $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$ as t varies from 0 to 1 in increments of 0.1. The red solid lines represent the PT method; blue dashed lines represent the clip method; green dashed lines represent the burn-in method. Each line was derived from 100,000 simulations.



Supplemental Figure 3: Performance without early stopping as assessed by the 10% and 90% quantile estimates of the proportion to arm 2 (the superior treatment) and overall response rate for the 3 methods under 3 scenarios, where $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.4, 0.5\}$. The red solid lines represent the PT method; blue dashed lines represent the clip method; green dashed lines represent the burn-in method. Each one was derived from 100,000 simulations.



Supplemental Figure 4: Performance with early stopping as assessed by the 10% and 90% quantile estimates of the proportion to arm 2 (the superior treatment), overall response rate, and total patients in the trial for the 3 methods under 3 scenarios, where $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.4, 0.5\}$. The red solid lines represent the PT method; blue dashed lines represent the clip method; green dashed lines represent the burn-in method. Each one was derived from 100,000 simulations.



Supplemental Figure 5: Distributions of total patients in the trial for the 3 methods with early stopping under the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$ setting power to 72%, 75%, and 80%, respectively. Each plot was derived from 100,000 simulations.

Supplemental Table 1: Simulation results for the PT method without an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for the decision rule was C_T, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value of t.

	$N = 80$				
$\theta_2 = 0.2$ (C_T)	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$(0.952)	0.200	0.500	0.050	0.901	0.049
$t = 0.1$(0.955)	0.200	0.500	0.050	0.900	0.050
$t = 0.2$(0.955)	0.200	0.500	0.050	0.901	0.049
$t = 0.3$(0.959)	0.200	0.500	0.049	0.902	0.049
$t = 0.4$(0.963)	0.200	0.500	0.050	0.901	0.049
$t = 0.5$(0.968)	0.200	0.501	0.050	0.902	0.048
$t = 0.6$(0.969)	0.200	0.500	0.049	0.902	0.050
$t = 0.7$(0.968)	0.200	0.499	0.050	0.900	0.050
$t = 0.8$(0.967)	0.200	0.500	0.050	0.899	0.051
$t = 0.9$(0.965)	0.200	0.498	0.048	0.902	0.049
$t = 1.0$(0.961)	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.250	0.500	0.263	0.733	0.004
$t = 0.1$	0.253	0.531	0.253	0.743	0.004
$t = 0.2$	0.256	0.559	0.256	0.740	0.004
$t = 0.3$	0.258	0.583	0.243	0.752	0.005
$t = 0.4$	0.260	0.605	0.227	0.767	0.006
$t = 0.5$	0.262	0.621	0.214	0.779	0.007
$t = 0.6$	0.264	0.636	0.209	0.783	0.008
$t = 0.7$	0.265	0.648	0.212	0.778	0.010
$t = 0.8$	0.266	0.658	0.205	0.784	0.011
$t = 0.9$	0.266	0.663	0.199	0.790	0.011
$t = 1.0$	0.267	0.669	0.208	0.780	0.012

	$N = 80$				
$\theta_2 = 0.4$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.300	0.500	0.616	0.384	0.000
$t = 0.1$	0.312	0.560	0.607	0.393	0.000
$t = 0.2$	0.322	0.613	0.605	0.395	0.000
$t = 0.3$	0.331	0.656	0.577	0.423	0.000
$t = 0.4$	0.338	0.690	0.540	0.459	0.001
$t = 0.5$	0.344	0.718	0.506	0.493	0.001
$t = 0.6$	0.347	0.738	0.479	0.519	0.001
$t = 0.7$	0.351	0.755	0.471	0.527	0.002
$t = 0.8$	0.353	0.768	0.451	0.547	0.002
$t = 0.9$	0.355	0.776	0.444	0.554	0.002
$t = 1.0$	0.357	0.783	0.448	0.549	0.002
$\theta_2 = 0.5$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.350	0.500	0.886	0.114	0.000
$t = 0.1$	0.377	0.591	0.877	0.123	0.000
$t = 0.2$	0.399	0.665	0.876	0.124	0.000
$t = 0.3$	0.416	0.719	0.854	0.146	0.000
$t = 0.4$	0.428	0.759	0.825	0.175	0.000
$t = 0.5$	0.436	0.789	0.786	0.214	0.000
$t = 0.6$	0.442	0.809	0.755	0.244	0.000
$t = 0.7$	0.448	0.825	0.732	0.268	0.000
$t = 0.8$	0.451	0.836	0.705	0.294	0.000
$t = 0.9$	0.453	0.845	0.692	0.307	0.000
$t = 1.0$	0.455	0.850	0.685	0.314	0.001

Supplemental Table 2: Simulation results for the clip method without an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for the decision rule was C_T, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value of t.

	$N = 80$				
$\theta_2 = 0.2$ (C_T)	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$(0.952)	0.200	0.500	0.050	0.901	0.049
$t = 0.1$(0.952)	0.200	0.500	0.049	0.900	0.051
$t = 0.2$(0.953)	0.200	0.500	0.050	0.901	0.049
$t = 0.3$(0.954)	0.200	0.500	0.049	0.901	0.050
$t = 0.4$(0.956)	0.200	0.500	0.049	0.903	0.048
$t = 0.5$(0.957)	0.200	0.501	0.050	0.901	0.049
$t = 0.6$(0.959)	0.200	0.499	0.050	0.899	0.050
$t = 0.7$(0.961)	0.200	0.500	0.048	0.904	0.048
$t = 0.8$(0.961)	0.200	0.499	0.049	0.902	0.050
$t = 0.9$(0.961)	0.200	0.500	0.049	0.902	0.048
$t = 1.0$(0.961)	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.250	0.500	0.263	0.733	0.004
$t = 0.1$	0.252	0.524	0.265	0.731	0.004
$t = 0.2$	0.255	0.549	0.259	0.737	0.004
$t = 0.3$	0.257	0.571	0.251	0.744	0.004
$t = 0.4$	0.259	0.592	0.242	0.753	0.005
$t = 0.5$	0.261	0.612	0.232	0.762	0.006
$t = 0.6$	0.263	0.630	0.220	0.773	0.007
$t = 0.7$	0.265	0.647	0.199	0.794	0.007
$t = 0.8$	0.266	0.660	0.200	0.790	0.010
$t = 0.9$	0.267	0.669	0.206	0.782	0.011
$t = 1.0$	0.267	0.669	0.208	0.780	0.012

	$N = 80$				
$\theta_2 = 0.4$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.300	0.500	0.616	0.384	0.000
$t = 0.1$	0.308	0.538	0.619	0.380	0.000
$t = 0.2$	0.314	0.574	0.605	0.395	0.000
$t = 0.3$	0.322	0.611	0.593	0.407	0.000
$t = 0.4$	0.329	0.645	0.572	0.427	0.000
$t = 0.5$	0.336	0.679	0.550	0.449	0.000
$t = 0.6$	0.342	0.709	0.516	0.483	0.001
$t = 0.7$	0.347	0.737	0.475	0.524	0.001
$t = 0.8$	0.352	0.763	0.450	0.549	0.002
$t = 0.9$	0.357	0.783	0.445	0.552	0.003
$t = 1.0$	0.357	0.783	0.448	0.549	0.002
$\theta_2 = 0.5$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.350	0.500	0.886	0.114	0.000
$t = 0.1$	0.363	0.544	0.885	0.115	0.000
$t = 0.2$	0.376	0.587	0.880	0.120	0.000
$t = 0.3$	0.389	0.629	0.868	0.132	0.000
$t = 0.4$	0.401	0.671	0.852	0.148	0.000
$t = 0.5$	0.413	0.711	0.829	0.171	0.000
$t = 0.6$	0.425	0.750	0.795	0.205	0.000
$t = 0.7$	0.436	0.787	0.752	0.248	0.000
$t = 0.8$	0.446	0.820	0.709	0.290	0.000
$t = 0.9$	0.455	0.850	0.684	0.315	0.001
$t = 1.0$	0.455	0.850	0.685	0.314	0.001

Supplemental Table 3: Simulation results for the burn-in method without an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case, $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for the decision rule was C_T, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value t.

	$N = 80$				
$\theta_2 = 0.2$ (C_T)	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$(0.952)	0.200	0.500	0.050	0.901	0.049
$t = 0.1$(0.953)	0.200	0.500	0.051	0.899	0.050
$t = 0.2$(0.954)	0.200	0.500	0.050	0.898	0.051
$t = 0.3$(0.955)	0.200	0.500	0.048	0.904	0.048
$t = 0.4$(0.956)	0.200	0.500	0.049	0.902	0.050
$t = 0.5$(0.958)	0.200	0.500	0.047	0.905	0.047
$t = 0.6$(0.960)	0.200	0.500	0.052	0.898	0.050
$t = 0.7$(0.960)	0.200	0.500	0.053	0.892	0.055
$t = 0.8$(0.966)	0.200	0.500	0.051	0.899	0.050
$t = 0.9$(0.962)	0.200	0.500	0.051	0.898	0.051
$t = 1.0$(0.961)	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.250	0.500	0.263	0.733	0.004
$t = 0.1$	0.253	0.525	0.263	0.733	0.004
$t = 0.2$	0.255	0.550	0.256	0.740	0.004
$t = 0.3$	0.257	0.573	0.256	0.740	0.004
$t = 0.4$	0.259	0.594	0.249	0.746	0.005
$t = 0.5$	0.261	0.614	0.233	0.761	0.006
$t = 0.6$	0.263	0.631	0.221	0.772	0.007
$t = 0.7$	0.265	0.647	0.213	0.778	0.008
$t = 0.8$	0.266	0.659	0.188	0.801	0.010
$t = 0.9$	0.267	0.668	0.206	0.783	0.012
$t = 1.0$	0.267	0.669	0.208	0.780	0.012

	$N = 80$				
$\theta_2 = 0.4$	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.300	0.500	0.616	0.384	0.000
$t = 0.1$	0.308	0.539	0.613	0.387	0.000
$t = 0.2$	0.315	0.577	0.601	0.398	0.000
$t = 0.3$	0.323	0.614	0.603	0.397	0.000
$t = 0.4$	0.330	0.649	0.583	0.416	0.000
$t = 0.5$	0.336	0.682	0.557	0.443	0.000
$t = 0.6$	0.342	0.713	0.521	0.479	0.001
$t = 0.7$	0.348	0.740	0.499	0.500	0.001
$t = 0.8$	0.352	0.764	0.441	0.557	0.002
$t = 0.9$	0.356	0.780	0.443	0.555	0.003
$t = 1.0$	0.357	0.783	0.448	0.549	0.002
$\theta_2 = 0.5$	Mean Response Rate	Mean Patients on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	0.350	0.500	0.886	0.114	0.000
$t = 0.1$	0.363	0.544	0.885	0.115	0.000
$t = 0.2$	0.376	0.587	0.877	0.123	0.000
$t = 0.3$	0.389	0.630	0.876	0.124	0.000
$t = 0.4$	0.401	0.672	0.864	0.136	0.000
$t = 0.5$	0.414	0.713	0.835	0.165	0.000
$t = 0.6$	0.425	0.751	0.809	0.191	0.000
$t = 0.7$	0.436	0.787	0.783	0.217	0.000
$t = 0.8$	0.445	0.818	0.708	0.291	0.000
$t = 0.9$	0.452	0.842	0.693	0.307	0.001
$t = 1.0$	0.455	0.850	0.685	0.314	0.001

Supplemental Table 4: Simulation results for the PT method with an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case, $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for early stopping was C_S, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value of t.

	$N_{MAX} = 80$					
$\theta_2 = 0.2$ (C_S)	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0(0.991)$	75.534	0.200	0.500	0.051	0.898	0.051
$t = 0.1(0.991)$	75.810	0.200	0.500	0.050	0.901	0.049
$t = 0.2(0.991)$	76.304	0.200	0.500	0.050	0.900	0.050
$t = 0.3(0.990)$	76.230	0.200	0.500	0.049	0.902	0.049
$t = 0.4(0.988)$	75.947	0.200	0.500	0.050	0.900	0.050
$t = 0.5(0.988)$	76.665	0.200	0.500	0.049	0.902	0.049
$t = 0.6(0.986)$	76.581	0.200	0.500	0.049	0.900	0.051
$t = 0.7(0.984)$	76.588	0.200	0.500	0.049	0.900	0.051
$t = 0.8(0.982)$	76.515	0.200	0.499	0.051	0.899	0.051
$t = 0.9(0.980)$	76.452	0.200	0.499	0.050	0.898	0.052
$t = 1.0(0.979)$	76.607	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	70.840	0.250	0.500	0.214	0.775	0.011
$t = 0.1$	71.087	0.252	0.522	0.213	0.777	0.011
$t = 0.2$	71.606	0.255	0.544	0.205	0.785	0.010
$t = 0.3$	71.359	0.256	0.564	0.210	0.778	0.011
$t = 0.4$	70.872	0.258	0.580	0.223	0.763	0.014
$t = 0.5$	71.857	0.260	0.598	0.210	0.778	0.012
$t = 0.6$	71.559	0.261	0.612	0.217	0.769	0.014
$t = 0.7$	71.593	0.262	0.622	0.219	0.767	0.014
$t = 0.8$	71.446	0.263	0.631	0.220	0.764	0.016
$t = 0.9$	71.415	0.264	0.640	0.221	0.764	0.016
$t = 1.0$	71.697	0.265	0.646	0.214	0.770	0.016

	$N = 80$					
$\theta_2 = 0.4$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	59.467	0.300	0.500	0.511	0.486	0.003
$t = 0.1$	59.620	0.308	0.538	0.510	0.487	0.003
$t = 0.2$	60.400	0.315	0.575	0.496	0.502	0.003
$t = 0.3$	60.323	0.321	0.608	0.497	0.500	0.003
$t = 0.4$	59.497	0.327	0.635	0.512	0.485	0.004
$t = 0.5$	61.283	0.333	0.665	0.481	0.515	0.004
$t = 0.6$	60.616	0.337	0.684	0.490	0.506	0.004
$t = 0.7$	60.869	0.341	0.702	0.487	0.508	0.004
$t = 0.8$	60.788	0.344	0.717	0.487	0.508	0.005
$t = 0.9$	60.613	0.346	0.728	0.487	0.508	0.005
$t = 1.0$	61.223	0.347	0.738	0.473	0.522	0.005
$\theta_2 = 0.5$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	45.116	0.350	0.500	0.803	0.196	0.001
$t = 0.1$	45.235	0.365	0.548	0.799	0.200	0.001
$t = 0.2$	45.974	0.378	0.594	0.791	0.209	0.001
$t = 0.3$	45.901	0.391	0.635	0.789	0.210	0.001
$t = 0.4$	45.415	0.400	0.668	0.789	0.210	0.001
$t = 0.5$	47.164	0.411	0.704	0.764	0.235	0.001
$t = 0.6$	46.941	0.418	0.728	0.761	0.238	0.001
$t = 0.7$	47.145	0.424	0.748	0.754	0.245	0.001
$t = 0.8$	47.402	0.429	0.765	0.745	0.253	0.002
$t = 0.9$	47.488	0.433	0.776	0.736	0.262	0.002
$t = 1.0$	48.163	0.437	0.789	0.720	0.278	0.002

Supplemental Table 5: Simulation results for the clip method with an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case, $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for early stopping was C_S, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value of t.

	$N_{MAX} = 80$					
$\theta_2 = 0.2$ (C_S)	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0(0.991)$	75.534	0.200	0.500	0.051	0.898	0.051
$t = 0.1(0.991)$	75.722	0.200	0.500	0.048	0.900	0.051
$t = 0.2(0.991)$	75.977	0.200	0.500	0.050	0.900	0.050
$t = 0.3(0.990)$	75.890	0.200	0.499	0.051	0.898	0.051
$t = 0.4(0.990)$	76.398	0.200	0.500	0.050	0.900	0.050
$t = 0.5(0.989)$	76.460	0.200	0.499	0.050	0.900	0.050
$t = 0.6(0.987)$	76.433	0.200	0.500	0.050	0.900	0.050
$t = 0.7(0.985)$	76.571	0.200	0.500	0.050	0.900	0.050
$t = 0.8(0.982)$	76.580	0.200	0.499	0.050	0.898	0.052
$t = 0.9(0.979)$	76.604	0.200	0.500	0.049	0.902	0.049
$t = 1.0(0.979)$	76.607	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	70.840	0.250	0.500	0.214	0.775	0.011
$t = 0.1$	70.920	0.252	0.522	0.213	0.776	0.011
$t = 0.2$	71.395	0.254	0.543	0.207	0.783	0.010
$t = 0.3$	70.904	0.256	0.563	0.216	0.772	0.012
$t = 0.4$	71.515	0.258	0.583	0.208	0.781	0.011
$t = 0.5$	71.731	0.260	0.600	0.207	0.781	0.012
$t = 0.6$	71.452	0.261	0.615	0.215	0.772	0.013
$t = 0.7$	71.670	0.263	0.628	0.216	0.770	0.014
$t = 0.8$	71.641	0.264	0.639	0.220	0.766	0.014
$t = 0.9$	71.752	0.265	0.646	0.213	0.772	0.015
$t = 1.0$	71.697	0.265	0.646	0.214	0.770	0.016

	$N = 80$					
$\theta_2 = 0.4$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	59.467	0.300	0.500	0.511	0.486	0.003
$t = 0.1$	59.617	0.307	0.534	0.510	0.488	0.003
$t = 0.2$	60.059	0.313	0.567	0.501	0.496	0.003
$t = 0.3$	59.540	0.320	0.598	0.509	0.488	0.003
$t = 0.4$	60.559	0.326	0.628	0.493	0.504	0.003
$t = 0.5$	60.802	0.331	0.656	0.485	0.512	0.003
$t = 0.6$	60.575	0.336	0.683	0.489	0.507	0.004
$t = 0.7$	60.898	0.341	0.706	0.483	0.513	0.004
$t = 0.8$	60.990	0.345	0.723	0.482	0.513	0.005
$t = 0.9$	61.336	0.348	0.738	0.473	0.522	0.005
$t = 1.0$	61.223	0.347	0.738	0.473	0.522	0.005
$\theta_2 = 0.5$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	45.116	0.350	0.500	0.803	0.196	0.001
$t = 0.1$	45.145	0.362	0.539	0.803	0.196	0.001
$t = 0.2$	45.661	0.373	0.577	0.796	0.204	0.001
$t = 0.3$	45.001	0.384	0.613	0.797	0.202	0.001
$t = 0.4$	46.335	0.395	0.649	0.780	0.219	0.001
$t = 0.5$	46.712	0.405	0.683	0.770	0.229	0.001
$t = 0.6$	46.825	0.415	0.715	0.762	0.236	0.001
$t = 0.7$	47.447	0.423	0.745	0.748	0.250	0.001
$t = 0.8$	47.767	0.431	0.770	0.739	0.259	0.002
$t = 0.9$	48.081	0.437	0.789	0.723	0.275	0.002
$t = 1.0$	48.163	0.437	0.789	0.720	0.278	0.002

Supplemental Table 6: Simulation results for the burn-in method with an early stopping rule. "Mean proportions on arm 2" is the mean proportion of patients assigned to treatment arm 2; "Arm 2 better" is the probability that treatment arm 2 is declared better at the end of the trial; "Equal" is the probability that the trial declares that the two treatments are equally effective; "Arm 1 better" is the probability that treatment arm 1 is declared better at the end of the trial. In each case, $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.3, 0.4, 0.5\}$. For each scenario, the trial was simulated $100,000$ times, the threshold for early stopping was C_S, and the type I error rate was controlled at 0.10. The tuning parameter is specified by the value of t.

	$N_{MAX} = 80$					
$\theta_2 = 0.2$ (C_S)	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0(0.991)$	75.534	0.200	0.500	0.051	0.898	0.051
$t = 0.1(0.991)$	75.517	0.200	0.500	0.051	0.899	0.050
$t = 0.2(0.991)$	75.468	0.200	0.500	0.050	0.899	0.051
$t = 0.3(0.991)$	75.532	0.200	0.500	0.048	0.902	0.050
$t = 0.4(0.990)$	75.048	0.200	0.500	0.052	0.894	0.053
$t = 0.5(0.990)$	75.147	0.200	0.499	0.051	0.897	0.051
$t = 0.6(0.989)$	74.782	0.200	0.500	0.053	0.893	0.054
$t = 0.7(0.989)$	75.197	0.200	0.500	0.049	0.901	0.050
$t = 0.8(0.986)$	74.671	0.200	0.500	0.054	0.891	0.055
$t = 0.9(0.982)$	75.441	0.200	0.501	0.052	0.897	0.051
$t = 1.0(0.979)$	76.607	0.200	0.500	0.049	0.902	0.049
$\theta_2 = 0.3$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	70.840	0.250	0.500	0.214	0.775	0.011
$t = 0.1$	70.854	0.252	0.519	0.213	0.776	0.011
$t = 0.2$	70.819	0.254	0.537	0.208	0.781	0.011
$t = 0.3$	70.992	0.256	0.556	0.202	0.788	0.010
$t = 0.4$	70.263	0.257	0.573	0.207	0.781	0.013
$t = 0.5$	70.427	0.259	0.590	0.198	0.790	0.012
$t = 0.6$	70.323	0.260	0.605	0.193	0.793	0.014
$t = 0.7$	70.762	0.262	0.621	0.180	0.806	0.014
$t = 0.8$	69.864	0.263	0.631	0.196	0.785	0.018
$t = 0.9$	70.109	0.264	0.640	0.212	0.769	0.019
$t = 1.0$	71.697	0.265	0.646	0.214	0.770	0.016

	$N = 80$					
$\theta_2 = 0.4$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	59.467	0.300	0.500	0.511	0.486	0.003
$t = 0.1$	59.427	0.304	0.523	0.511	0.487	0.003
$t = 0.2$	59.511	0.310	0.548	0.501	0.496	0.003
$t = 0.3$	59.660	0.315	0.575	0.486	0.511	0.003
$t = 0.4$	58.888	0.321	0.601	0.485	0.512	0.003
$t = 0.5$	59.443	0.326	0.630	0.459	0.538	0.003
$t = 0.6$	59.260	0.332	0.656	0.447	0.549	0.004
$t = 0.7$	60.925	0.337	0.688	0.407	0.589	0.004
$t = 0.8$	59.891	0.342	0.708	0.425	0.569	0.006
$t = 0.9$	59.541	0.345	0.726	0.459	0.534	0.007
$t = 1.0$	61.223	0.347	0.738	0.473	0.522	0.005
$\theta_2 = 0.5$	Mean Total Patients	Mean Response Rate	Mean Proportions on Arm 2	Arm 2 Better	Equal	Arm 1 Better
$t = 0.0$	45.116	0.350	0.500	0.803	0.196	0.001
$t = 0.1$	45.124	0.354	0.516	0.800	0.199	0.001
$t = 0.2$	45.136	0.361	0.535	0.794	0.205	0.001
$t = 0.3$	45.175	0.368	0.559	0.778	0.221	0.001
$t = 0.4$	44.615	0.376	0.586	0.771	0.228	0.001
$t = 0.5$	45.316	0.386	0.619	0.741	0.258	0.001
$t = 0.6$	45.444	0.396	0.656	0.721	0.278	0.001
$t = 0.7$	47.122	0.410	0.700	0.680	0.319	0.001
$t = 0.8$	46.764	0.421	0.735	0.682	0.316	0.002
$t = 0.9$	46.557	0.430	0.767	0.704	0.293	0.003
$t = 1.0$	48.163	0.437	0.789	0.720	0.278	0.002

Supplemental Table 7: Sensitivity Analysis of the Choice of the Prior Distribution of the Response Rate. The results are from 100,000 simulation under the scenario $\theta_1 = 0.2$ and $\theta_2 = 0.5$ with the tuning parameter $t = 0.5$.

	Without Early Stopping		
PT	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	43.6%	78.9%	78.6%
$Beta(0.5, 0.5)$	43.6%	78.6%	76.9%
$Beta(1, 3)$	43.6%	78.5%	80.1%
Clip	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	41.3%	71.1%	82.9%
$Beta(0.5, 0.5)$	41.4%	71.1%	81.7%
$Beta(1, 3)$	41.4%	71.1%	83.9%
Burn-in	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	41.4%	71.3%	83.8%
$Beta(0.5, 0.5)$	41.4%	71.2%	82.5%
$Beta(1, 3)$	41.4%	71.2%	84.4%
	With Early Stopping		
PT	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	41.1%	70.4%	76.4%
$Beta(0.5, 0.5)$	41.1%	70.2%	75.5%
$Beta(1, 3)$	41.1%	70.5%	75.5%
Clip	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	40.5%	68.3%	77.0%
$Beta(0.5, 0.5)$	40.5%	68.3%	76.5%
$Beta(1, 3)$	40.7%	68.7%	75.9%
Burn-in	Overall Response Rate	Proportion on Arm 2	Power
$Beta(0.6, 1.4)$	38.6%	61.9%	74.1%
$Beta(0.5, 0.5)$	38.6%	61.9%	74.7%
$Beta(1, 3)$	38.8%	62.4%	72.5%