Published in final edited form as:

Contemp Clin Trials. 2015 January ; 40: 15–25. doi:10.1016/j.cct.2014.11.006.

Simulation Study for Evaluating the Performance of Response-Adaptive Randomization

Yining Du^{a,b}, Xuan Wang^c, and J. Jack Lee^{a,*}

^aDepartment of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA

^bDivision of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler Street, Houston, TX 77030, USA

^cBaylor Institute for Immunology Research, Baylor Research Institute, Baylor Health Care System, Dallas, TX 75246, USA

Abstract

A response-adaptive randomization (RAR) design refers to the method in which the probability of treatment assignment changes according to how well the treatments are performing in the trial. Holding the promise of treating more patients with the better treatments, RARs have been successfully implemented in clinical trials. We compared equal randomization (ER) with three RARs: Bayesian adaptive randomization, sequential maximum likelihood, and sequential posterior mean. We fixed the total number of patients, considering as patient horizon, but varied the number of patients in the trial. Among the designs, we compared the proportion of patients assigned to the superior arm, overall response rate, statistical power, and total patients enrolled in the trial if adding an efficacy early stopping rule. Without early stopping, ER is preferred when the number of patients beyond the trial is much larger than the number of patients in the trial. RAR is favored for large treatment difference or when the number of patients beyond the trial is small. With early stopping, the difference between these two types of designs was reduced. By carefully choosing the design parameters, both RAR and ER methods can achieve the desirable statistical properties. Within three RAR methods, we recommend SPM considering the larger proportion in the better arm and higher overall response rate than BAR and similar power and trial size with ER. The ultimate choice of RAR or ER methods depends on the investigator's preference, the trade-off between group ethics and individual ethics, and logistic considerations in the trial conduct, etc.

© 2014 Elsevier Inc. All rights reserved.

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

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest

There are no conflicts of interest to declare.

Keywords

allocation probability; Bayesian adaptive design; efficacy early stopping; operating characteristics; patient horizon

1 Introduction

When multiple treatment arms are involved in clinical trials, randomization is commonly applied to avoid allocation bias and to yield a valid statistical inference. It also has ethical implications regarding treatment efficacy. In the beginning of a trial, the equipoise principle implies that treatment effects are equal among the treatment arms. Hence, equal randomization (ER) which randomizes patients equally among treatment arms can be justified. However, as the information accrues in the trial, the treatment effects may no longer be equal among treatments. Response-adaptive randomization (RAR) dynamically assigns patients to treatments based on the accumulating clinical responses, where ‘response’ generically refers to treatment outcomes. One of the RAR’s appealing features is that it can assign more patients to better treatments based on available data.

In two-arm trials, RAR assigns more patients to the superior treatment and exposes fewer patients to the inferior treatment. The challenge is that we know very little about the relative treatment effect initially. As a trial proceeds, data in the trial give RAR higher probability of assigning more patients to the superior treatment. The relative effectiveness of treatments can be estimated by some sequential estimation rule on a meaningful outcome measure. An important issue is that many of the RAR procedures may not be optimal based on the specific measures [1]. Since they are derived heuristically, we need to define criteria for comparing and evaluating them. A properly chosen RAR method can strike a balance in maximizing the statistical power for testing the treatment efficacy (group ethics) and giving each patient the best treatment (individual ethics). Key desirable features of the RAR designs include retaining randomizations advantage (eliminate bias), maintaining the required statistical power, and treating more patients with superior treatments, etc. The play-the-winner design ([2]) was a deterministic precursor of RAR. Other approaches use urn models ([3]; [4]), and consider multi-arm trials and covariates ([5]; [6]), extend RAR to trials with delayed response ([7]; [8]), and compare RAR for frequentist group sequential binary response trials ([9]).

We use the setting of a binary outcome, but RAR can be generalized to other settings. We evaluate only treatment efficacy and assume comparability in safety, cost, and feasibility. We compare ER to three RAR methods: Bayesian adaptive randomization (BAR), derived from [10]; sequential maximum likelihood (SML), with randomization probability calculated based on frequentist sequential maximum likelihood estimators (Section 10.4.1 in [4]); and sequential posterior mean (SPM), which replaces the role of the sequential maximum likelihood in the allocation probability.

The use of RAR versus ER is yet debated in statistical and clinical trial communities ([11]; [12]). Traditionally, the primary goal of clinical trials is to determine better treatment options to benefit future patients. Providing the best possible treatment to the patients

enrolled in the trial is rarely the main purpose of the trial. This view, however, has major shortcomings. For patients enrolled in clinical trials, will there be a single patient who does not want to receive the best possible treatment? Although equal randomization can be well justified based on the equipoise principle in the beginning of the trial, as the information accrues, RAR can be applied to assign more patients to better treatments based on available data. Two recent advances in medicine compel us to critically examine the traditional paradigm. (1) New treatment modalities are changing rapidly. Even if a trial is positive and the standard practice is changed, its impact may not be long lasting because new treatment modalities could change again. (2) In the genomic era, considering all possible markers such as mutations from deep sequencing (next generation sequencing), copy number variation, mRNA expression, microRNA expression, and protein expression, etc., no two patients are alike. This N-of-1 trial concept challenges the traditional paradigm that patients are homogeneous and the results are generalizable to all alike patients in the future. Therefore, we take one step further to evaluate designs which can strike a balance between the individual ethics and the group ethics. The concept of the patient horizon ([13]) is used to evaluate such balance. The patient horizon is the total number of patients with the particular disease that is relevant to the treatments being investigated. It includes patients enrolled in the trial and the future patient population beyond the trial who will benefit from what is learned in the trial. For some diseases the future patient population can be quite large compared to the size of the trial. In that setting, making the right decision about the treatments studied (which depends on the power of the test) is more important. On the other hand, in the case of some rare cancers, most of the patients with the disease of interest will be enrolled in the trial and there will be few future patients beyond the trial. Under that scenario, giving the best treatment possible to each patient in the trial is ethically the right thing to do. We fix the patient horizon size but varying the trial size, then, compare the proportion of patients assigned to the superior treatment, overall response rate, and statistical power between RAR and ER designs. We evaluate the operating characteristics for trials without and with early stopping rules for futility and efficacy.

2 Methods

2.1 Bayesian Adaptive Randomization Method

Bayesian adaptive randomization (BAR) is an allocation scheme based on the posterior probability of one treatment being more effective than the other(s). This posterior probability can be obtained from a binomial likelihood and a beta prior for a binary outcome:

$$x_{ij} \sim \text{Bin}(1, \theta_i), \theta_i \sim \text{Beta}(\alpha_i, \beta_i) \quad (1)$$

where x_{ij} is the outcome of patient j on treatment i and θ_i is the probability of a response for treatment i . We consider the setting of two treatments $i \in \{1, 2\}$. The probability that treatment 2 is better than treatment 1 is given by $\Pr(\theta_2 > \theta_1 | x)$. Because of the consistency of the posterior, this probability approaches either 0 or 1 as more data are collected for both treatments unless the two treatments are exactly the same. The basic idea of BAR can be traced to [10], although a good part of his paper was devoted to the computation of $\Pr(\theta_1 >$

$\theta_2|x$), which was then a big hurdle. A detailed review on BAR is provided by Thall and Wathen [14], and an application of BAR in the development of targeted agent can be found in [15]. The main goal of BAR is to assign more patients to the better treatment arms with higher probability. A patient is adaptively randomized to treatment i with probability $Pr(\theta_i > \theta_j | data)$. However, the probability $Pr(\theta_i > \theta_j | data)$ can be highly variable in the beginning of the trial when the number of patients is small. By adding a tuning parameter, the randomization probability can be stabilized

$$\rho(\lambda) = \frac{\Pr(\theta_2 > \theta_1 | x)^\lambda}{\Pr(\theta_1 > \theta_2 | x)^\lambda + \Pr(\theta_2 > \theta_1 | x)^\lambda} \quad (2)$$

where λ is a tuning parameter and $\lambda \geq 0$. An introduction about the tuning parameter could be found in [16]. Furthermore, [14] recommended using the tuning parameter $n/2N$ instead of a fixed value. Note that $\rho(1) = \Pr(\theta_2 > \theta_1 | x)$, $\rho(0) = \frac{1}{2}$ and $\rho(\infty)$ behaves like the play-the-winner rule. Therefore, λ controls the level of imbalance in the allocation probability.

As is evident, BAR may lead to an extreme preference for a certain treatment arm. One way to avoid such extreme allocation probability is to set bounds on the allocation probability; thus, it does not converge to 0 or 1. For example, we may constrain the randomization probability to be bounded within 0.05 to 0.95, or 0.1 to 0.9 ([17]) to allow for continued randomization to both arms to gather information for further assessment of the treatment effects.

An appealing advantage of the Bayesian approach is that the information continues to be updated naturally as the trial moves along. Early stopping rules can be incorporated based on the probability of relevant clinical events (e.g., [18]). This is important because the trial should be stopped when there is sufficient information to declare that one treatment is better than the other (so that we cease to randomize patients to the inferior arm) or when there is strong evidence that the treatments are equivalent. Thall and Simon [18] considered stopping rules for declaring efficacy based on the probabilities $\Pr(\theta_1 > \theta_2 | x)$ or $\Pr(\theta_2 > \theta_1 | x)$. It is common practice to calibrate the thresholds so that the trial fulfills certain criteria regarding the frequentist operating characteristics (such as controlling the type I error). In this paper, we denote C_I as the threshold on $\Pr(\theta_1 > \theta_2 | x)$ or $\Pr(\theta_2 > \theta_1 | x)$ for the efficacy early stopping rules, and C_F as the corresponding threshold for the final decision rules. For our work, we perform simulations to determine the cutoff values for both C_I and C_F to control the type I error rate at $\alpha = 0.05$. If the trial is not stopped early, one of the three following decisions can be reached:

1. If $P_r(\theta_2 > \theta_1 | x) \geq C_F$, we claim that treatment 2 is superior.
2. If $P_r(\theta_1 > \theta_2 | x) \geq C_F$, we claim that treatment 1 is superior.
3. Otherwise, we claim that the two treatments are equally effective.

For the efficacy early stopping rule, we will stop if at any point we have

$$P_r(\theta_i > \theta_j | x) \geq C_I$$

and claim that treatment i is superior (i.e., treatment j is inferior), where C_I represents the corresponding threshold for early stopping and we also calibrate it by simulations to control type I error rate.

2.2 Sequential Maximum Likelihood Method

Alternatively, if the treatment allocation is targeted based on the maximum likelihood estimates which are sequentially estimated, we call such RAR designs the sequential maximum likelihood (SML) method. Section 10.4.1 in [4] introduced the SML, and an early discussion about the application of the SML was presented in [19]. SML takes the allocation probability

$$\rho = \frac{\theta_2}{\theta_1 + \theta_2} \quad (3)$$

to assign patients to arm 2. Since θ_1 and θ_2 are unknown, they have to be estimated. One way to proceed is to use the corresponding maximum likelihood estimates

$$\hat{\rho} = \frac{\hat{\theta}_2}{\hat{\theta}_1 + \hat{\theta}_2}.$$

Similar to the BAR method, a generalization to mitigate the problem of extreme allocation probability can be obtained by introducing a tuning parameter λ to the allocation probability:

$$\hat{\rho}(\lambda) = \frac{\hat{\theta}_2^\lambda}{\hat{\theta}_1^\lambda + \hat{\theta}_2^\lambda}, \lambda \geq 0. \quad (4)$$

To determine which treatment is better at the end of the trial, a test is performed. The null hypothesis is $H_0: \theta_1 = \theta_2$ and the alternative is $H_1: \theta_1 \neq \theta_2$. The test can be performed by constructing the following confidence interval:

$$\hat{\theta}_2 - \hat{\theta}_1 \pm Z_{\alpha/2} \sqrt{\frac{\hat{\theta}_1(1 - \hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1 - \hat{\theta}_2)}{n_2}},$$

where $Z_{\alpha/2}$ represents the upper $\alpha/2$ percentile of the standard normal distribution. For comparing treatment efficacy between the two groups, one of the following three decisions can be made:

1. If the lower bound is greater than 0, we claim that treatment 2 is superior.
2. If the upper bound is less than 0, we claim that treatment 1 is superior.
3. Otherwise, we claim that the two treatments are equally effective.

Note that the SML method is based on frequentist estimation and hypothesis testing, and the main reason we apply it is that we can compare the methods from the frequentist framework.

When adding an efficacy early stopping rule, we calibrate the value of $Z_{\alpha/2}^*$ to control the type I error rate at 0.05.

2.3 Sequential Posterior Mean Method

Last, we propose a method in the Bayesian framework that is similar to SML, which we call the sequential posterior mean (SPM) method. In SPM, we replace the MLE by the posterior mean. The main motivation we propose this method is that we can compare the Bayesian framework with SML which is based on frequentist framework. Also, SPM can be further improved by applying the informative priors. For Expressions 2 and 4, we set the probability of allocation to arm 2 as

$$\tilde{\rho}(\lambda) = \frac{\tilde{\theta}_2^\lambda}{\tilde{\theta}_1^\lambda + \tilde{\theta}_2^\lambda}, \lambda \geq 0.$$

Here, $\tilde{\theta}_i$ is the Bayesian estimator (posterior mean) for θ_i . With the choice of non-informative priors, SPM and SML should perform in a similar way. To test treatment efficacy in both the interim and final analyses, we apply the same decision rules as in BAR.

Next, we demonstrate how the three RAR methods work by conducting the following simulation: The true probabilities of response were $\theta_1 = 0.4$ and $\theta_2 = 0.7$. First, we performed an equal randomization as burn-in with 20 patients, i.e., 10 randomized to each treatment. For BAR, we assumed the model given by expression (1) with the prior distribution $\theta_i \sim \text{Beta}(0.5, 0.5)$. The next 80 patients were assigned to treatments according to the probabilities given by equation (2), with $\lambda = 1$. The changes in the allocation probabilities as the trial evolved are illustrated in Figure 1. We considered 5 realizations of the randomization process in order to illustrate the underlying variability (each realization identified by a different color). For patients enrollments $n = 1$ to 20, they are assigned to the treatments using the blocked randomization during the ER burn-in period. The performance of BAR is shown in the top panel. In one trial, the probability of randomization to arm 2 dropped to 0.42. This trend reverted as the trial evolved. As the sample size reached 100 under BAR, the probability of allocation to arm 2 approached 1 for all of the realizations. Despite the variability of the process, there was little difference in the posteriors by the end of the trial. This illustrates the essence of learning throughout the trial under the Bayesian framework. Similar plots illustrate the simulated performances of SML and SPM (middle and lower panels, respectively) in Figure 1. Note that BAR behaved differently from SML and SPM mainly because of different target allocation probabilities. The limit of the allocation probability for BAR is 1; whereas it is 0.636 for SML. Also, we found that SPM and SML performed similarly because we assumed a non-informative prior for SPM.

3 Simulation Setting

To illustrate how BAR, SML, SPM and ER behave and evaluate their relative performance, we perform the following simulations: A burn-in of the first $2k$ patients is carried out using blocked randomization; $k = 10$ patients are assigned to each of treatment 1 and treatment 2. The response rates for treatment i are denoted by θ_i with the prior distribution $\theta_i \sim \text{Beta}(0.5,$

0.5). We analyze three scenarios, for which the response rates are $\theta_1 = 0.2$, and $\theta_2 \in \{0.2, 0.4, 0.6\}$. We allocate $m - 2k$ patients to the treatments according to either BAR, SPM, SML, or ER. Here, $m \in \{50, 100, 200, 300, 400, 500\}$ denotes the size of the trial. We consider a horizon of $N = 1, 000, 5, 000$ or $50, 000$ patients corresponding to small, medium, and large patient population beyond the trial. The decision rule depends on the randomization scheme used: for SML, we base the decision rule on confidence intervals; for BAR, SPM, and ER, we compare the Bayesian posterior probabilities of responses for the two treatments with the threshold. When no early stopping is applied, the decision rule is evaluated after all m patients enrolled in the trial. On the other hand, when early stopping is applied, starting from the $2k+1$ patient, we evaluate the decision rule after the outcome of every patient is observed. For simplicity, we assume the outcome is instantaneously observed. The trial is stopped early if the stopping boundaries crossed. After the trial is stopped, we assign the remaining patients (N minus the number of patients in the trial) to the better treatment as determined by the decision rule. If the trial declares that the two treatment effects are not different, the $N - m$ patients are assigned evenly to treatment 1 and treatment 2 (via blocked randomization). We allow m to vary in different simulation settings, and set $\lambda = 1$ in this paper. For each scenario, we simulate the trial 100,000 times. The parameters of the decision rules are chosen to control the type I error rate at 0.05, and we also constrain the randomization probability to be bounded within 0.05 and 0.95. Moreover, we will add an efficacy early stopping rule to the RAR and ER designs. At the end of the trial, we compare the mean proportion of patients allocating to the better arm, the mean overall response rate, statistical power, and the mean total patients enrolled in the trial. Note that the proportion of patients in the better arm and overall response rate were measured at the patient horizon level while the statistical power and mean total patients in the trial were evaluated at the trial level.

4 Simulation Results

We present the simulation results for the patient horizon size $N = 1, 000$ and $N = 50, 000$ in Figures 2–5, and provide detailed information in Supplemental Tables 1–6 and Supplemental Figures 1–6 for size $N = 5, 000$ in the Appendix. Each plot presents the performance of the four randomization methods. With $\theta_1 = 0.2$ and $\theta_2 \in \{0.2, 0.4, 0.6\}$, Figures 2 and 3 depict the trial performance without early stopping; Figures 4 and 5 depict the trial performance with an efficacy early stopping rule. The vertical axes respectively represent the mean proportion in the better arm (treatment 2), mean overall response rate, power, and mean trial size with early stopping taken over the simulations. The power and mean trial size are based on the trial level and the mean proportion in the better arm and mean overall response rate are based on the horizon level. The horizontal axes represent the trial size. The four trial designs are depicted by lines with different colors; the red line represents the BAR design, and orange, green, and black lines represent the SML, SPM, and ER designs, respectively. Each line was derived from 100,000 simulations. The tables in the Appendix provide detailed simulation results for the four randomization methods for $N = 5, 000$ when $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4$ and 0.6 . In the tables, we present the mean response rate, mean proportion in the better treatment, and the probabilities that claim which treatment is superior (trial level), and present the mean overall response rate and mean

proportion in the better treatment (horizon level). Also, we provide the thresholds for the decision rules and boundaries of the allocation probability in the tables.

For all RAR designs without early stopping, when θ_2 was greater than θ_1 , in the trial (first stage), a larger trial size was associated with a higher mean response rate and larger mean proportion in arm 2, the better treatment. For example, as $N = 5,000$, at the trial level, the mean response rate of the BAR method increased from 0.336 to 0.382 and mean proportion in arm 2 increased from 0.682 to 0.908 as the trial size increased from 50 to 500 when $\theta_1 = 0.2$ and $\theta_2 = 0.4$ (Supplemental Table 2). In contrast, for the ER design without early stopping, the mean response rate and mean proportion in the better arm remained the same throughout the first stage (Supplemental Table 2). This indicates an advantage of the RAR design, which is that it continues to learn throughout the trial. As the sample size increases, more information accumulates; hence, more patients are assigned to the superior arm and the mean estimated response rate in the trial increases.

To achieve the highest estimated response rate for the patient horizon (total patient population), there is a trade-off between the size of the trial (in the first stage) and the amount of information gained in the trial that can be applied to future patients. Assuming the size of patient horizon is fixed, if the trial size is too small in the first stage, inadequate information is gathered and the remaining patients may not be assigned to the better arm. On the other hand, if the trial size is too large in the first stage, the power will be high. However, there will not be many patients remaining who can benefit from the trial. Thus, the benefit from learning in the trial will be limited. The maximal estimated response rate for the patient horizon can be achieved by enrolling a sufficient number of patients but not too many in the first stage. For example, when $N = 5,000$ and $\theta_2 = 0.4$, for the BAR method without adding the early stopping rule, the highest mean overall response rate (0.381) for the patient horizon was achieved at $m = 500$ patients in the first stage (Supplemental Table 2); however, when $\theta_2 = 0.6$, the highest mean overall response rates (0.588) for the patient horizon were achieved at $m = 300$ (Supplemental Table 3). Similarly, for both the SML and SPM methods without early stopping, the highest mean overall response rates for the total patient horizon were achieved for $\theta_2 = 0.4, 0.6$ when the first stage sample sizes were $m = 300$ and 100 , respectively (Supplemental Tables 2 and 3). In comparison, the corresponding highest mean overall response rates for the total patient horizon for the ER design without early stopping were achieved for $\theta_2 = 0.4$ and 0.6 at first stage sample sizes of $m = 200$ and 100 , respectively (Supplemental Tables 2 and 3).

Figure 2 presents the average performance without early stopping for the scenarios when $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4$ and 0.6 as $N = 1000$. Compared to the ER design, the RAR designs offered larger mean proportion in the superior arm except that BAR showed the smallest mean proportion in the better arm for small trial sizes ($m < 300$) when $\theta_2 = 0.4$. However, the mean proportion in the better arm and overall response rate of BAR continued to increase as the trial size increased compared to other three methods when $\theta_2 = 0.4$. The RAR methods provided much larger mean proportion in the better arm and resulted in higher mean overall response rate than ER when the difference in the true response rate between two arms increased ($\theta_2 = 0.6$). When $m > 200$, BAR yielded the highest mean overall

response rate, and the response rate decreased slightly for BAR when $m > 300$. However, it decreased much for ER, SML, and SPM with the decrease of ER the most severe.

In terms of statistical power, without early stopping, the simulations showed that the probability of concluding that arm 2 is the better treatment in the first stage, which can be considered as the power, increased as the trial size increased. This was true under all the trial designs. For the comparisons, ER provided much higher power than BAR when the difference in the true response rate between arms is moderate ($\theta_1 = 0.2$ and $\theta_2 = 0.4$). Compared to either SML or SPM method, ER demonstrated a slightly higher power. For SML, SPM, and ER, to achieve 80% power when $\theta_1 = 0.2$ and $\theta_2 = 0.4$, they needed m of 200 or larger with ER having the highest power. But for BAR, it required $m = 500$. Furthermore, for SML, SPM, and ER, the optimal sample size to yield the highest overall response rate is to have sufficient power (around 80%–90%). More than 90% power is not necessary. The overall response rate can be lower as there is no need to put too many patients on trials. When the difference in the true response rate between two arms increased ($\theta_2 = 0.6$), the BAR method can also achieve sufficiently higher power. To reach 80% power, BAR needed $m > 100$, and SML, SPM, and ER needed m between 50 and 100. Generally, ER showed higher power than BAR, while SML and SPM outperformed ER when $m > 200$ considering the larger proportion in the better arm and higher overall response rate and the similar power. For the moderate treatment difference, SML and SPM performed well and struck a good balance between obtaining high power with small size and overall high response rate. Since we assumed the non-informative prior for SPM, it performed very similarly with SML. ER suffered with low overall response rate most when putting too many patients on trials. BAR was the best for the large treatment difference.

Figure 3 presents the average performance of four methods without early stopping for the large patient horizon $N = 50,000$. When we increased the patient size to 50,000, BAR showed consistently smallest mean proportion in the superior treatment and lowest mean overall response rate for moderate difference in the true response rates between two arms ($\theta_2 = 0.4$) for all m between 50 and 500. Also, BAR generated the lowest power compared to other three methods. All three other methods performed similarly but ER was slightly better than SML and SPM for $m < 300$ in terms of power, and slightly better considering the proportion in the better arm and overall response rate. When the difference in response rates increased ($\theta_2 = 0.6$), SML and SPM methods showed a little larger mean proportion in the better treatment than the ER method for $m > 200$ and BAR is the best for $m = 500$. Power of SML, SPM, and ER were very similar with ER performing the best, followed by SPM and SML.

Also, we provided the simulation results for $N = 5,000$ in the Appendix. The performance was in between the two extreme patient horizons. For the moderate response difference, ER was the best in terms of the proportion in the better arm and overall response rate for $m = 100$, but SML and SPM were better for $m > 100$. Although ER was better for small m , SML and SPM yielded higher response for patients on trials. For example, when $\theta_1 = 0.2$ and $\theta_2 = 0.4$, SML and SPM provided response rates from 0.319 to 0.333 in the trial, while ER showed the response rate 0.3 in the trial. These can be found in the Supplemental Table 2.

For the large response difference, SML and SPM were the best for $m = 100$ but BAR was the best for $m = 200$.

Next, we discuss the results when the efficacy early stopping rule was added to both RAR and ER designs. In Figure 4, we present the average performance of four methods when $N = 1,000$. With early stopping, the gain from SML and SPM over ER in terms of the proportion in the better arm and overall response rate was reduced. When $\theta_1 = 0.2$ and $\theta_2 = 0.4$, BAR showed the smallest mean proportion in the superior arm and lowest mean overall response rate, and consistently the lowest power compared to the other three methods for all m . Also, in terms of the mean trial size (first stage), BAR required the most patients. When comparing the SML, SPM and ER methods, they performed very similarly with respect to the mean proportion in the better arm, mean overall response rate, and power. SPM was the best for $m > 300$, followed by SML and ER. Considering the mean trial size, ER required the smallest number of patients, and SML required more patients than SPM. Even when the originally planned trial size was at $m = 500$, with early stopping, ER and SPM only needed a mean of about 150 patients, while SML needed a mean of 200 patients and BAR needed a mean of about 310 patients. When $\theta_2 = 0.6$, SML, SPM, and ER performed similarly in terms of the proportion in the better arm, overall response rate, and power. BAR had smaller mean proportion in the better arm and worse mean overall response rate when $m < 300$ but provided the highest overall response rate for $m > 400$. It, however, continued to have lower power and larger trial size. For $m > 100$, SPM was slightly better than SML, and ER was slightly worse than SPM and SML considering the mean overall response rate. In terms of the mean trial size, ER needed less than 50, SPM about 50, SML about 57 and BAR required 90. Similar results can be found in Figure 5 when $N = 50,000$. In terms of the overall response rate, SML, SPM, and ER were all very similar. BAR showed much smaller mean proportion in the better treatment and lower overall response rate than other three methods for all m when $\theta_2 = 0.4$. When $\theta_2 = 0.6$, BAR caught up with other methods in terms of the mean proportion in the better arm when $m = 500$. It, however, consistently yielded lower power and larger trial size.

Through the previous discussion, we provide the comparisons and evaluations of operating characteristics which are given conditional on whether an efficacy early stopping rule is applied. We also compare the results between without and with implementing early stopping for each method. With early stopping, when $m = 300, 400$, and 500 for the moderate treatment difference, for SML, the actual trial size were 176.92, 192.00, and 203.34, respectively. It required fewer patients but provided higher overall response rate than without early stopping (Supplemental Tables 2 and 5). SPM and ER performed similarly with SML, where they enrolled fewer patients in the trial but provided higher overall response rates than without early stopping. For the large response difference, SML, SPM, and ER required much fewer patients in the trial and provided much higher overall response rates than without early stopping. Since BAR had lower power and required more patients in the trial than other methods, it performed similarly for both with and without early stopping (Supplemental Tables 2 and 5).

Last, we added the 10% and 90% quantile estimates to the key operating characteristics when $N = 5,000$ in the Supplemental Figures 3 to 6 in the Appendix. Generally speaking,

the patterns were consistent with the mean measurements. However, the estimates of the BAR method have larger variance than other three methods. For example, under the scenario with no difference in the true response rate between two arms ($\theta_1 = \theta_2 = 0.2$) for both without and with early stopping, BAR had the lowest 10% quantile estimates and the highest 90% quantile estimates of the proportion of patients assigned to the superior arm compared to the other three methods. For the 10% quantile estimates, under the alternative hypothesis without early stopping, BAR performed the worst in terms of the proportion of patients assigned to the better treatment and overall response rate when the difference in the true response rate between two arms was moderate ($\theta_1 = 0.2$ and $\theta_2 = 0.4$). For the large response difference, the 10% quantile estimates of BAR reflected the similar pattern with the mean measurements. However, the performances at the 90% quantile estimates were different from the mean measurements, and it performed the best with respect to the proportion in the better arm and overall response rate. By adding early stopping, the 10% quantile estimates of the proportion to the better arm and overall response rate of the four methods were consistent with the mean measurements. In terms of the trial size, SML, SPM, and ER showed similar pattern with the mean measurements. However, BAR required the fewest patients in the trial, which was different from the mean measurements. For the moderate treatment difference, the 90% quantile estimates of four methods showed the similar pattern with the mean estimates with respect to the overall response rate and trial size. For the proportion in the better arm, four methods performed very similarly. For the large treatment difference, BAR had the highest overall response rate for the 90% quantile estimate. In terms of the trial size, four methods reflected the similar pattern with the mean measurements.

Within the three RAR methods, the performance of BAR was more extreme compared to the performances of the SML and SPM designs. Without early stopping, BAR yielded the highest mean overall response rate and highest mean overall response rate when the difference in the true response rates between the two arms was large ($\theta_1 = 0.2$ and $\theta_2 = 0.6$) but power was much reduced. However, the extremely large treatment difference is seldom seen in real trials. With the moderate difference, SML and SPM struck a good balance in achieving sufficiently high power but yielded higher overall response rate. Early stopping is helpful in reducing the trial size. By calibrating the stopping boundaries, type I error was preserved but the trial can be stopped if sufficient evidence was reached in concluding the treatment difference. It should be recommended in all trials whenever possible. When an early stopping rule was added, ER had the smallest trial size. SPM needed slightly larger trial size followed by SML, and BAR needed much larger trial size. SML, SPM, and ER performed similarly. But when the patient horizon was small, SPM can yield slightly larger overall response rate than SML and ER when $m > 300$. When the patient horizon was large, no difference between the three methods was found in terms of the overall response rate. With the large treatment difference and small patient horizon, BAR yielded the highest overall response rate when $m > 300$. For the large response difference and large patient horizon, SML, SPM, and ER showed similar overall response rate with BAR performing the worst except when BAR reached over 90% power at $m = 500$, the overall response rate was similar to other three methods. Generally speaking, we recommend the SPM method

considering the larger proportion in the better arm and higher overall response rate than BAR, and similar power and trial size with ER.

5 Discussion

Overall, one important point is that the patient horizon needs to be considered in trial design. Furthermore, the standard practice of calculating the sample size needed to achieve at least 80% power is still very important to yield desirable operating characteristics within and beyond the trial. The main criterion for evaluating a clinical trial design is based on the mean number of responses in the patient horizon and statistical power in the trial level. Without early stopping, the mean overall response rate of RAR designs is higher than that of ER designs, while ER provides higher power than the RAR methods. However, a sufficiently high statistical power of RAR designs can be achieved by increasing m or when there is a large difference in the true response rate between the two treatment arms under study. When an early stopping rule was incorporated, the difference between RAR and ER was also reduced.

Within the three RAR methods, when no early stopping rule is applied, the performances of SPM and SML are very similar. When the difference in the true response rate between two treatment arms was moderate ($\theta_1 = 0.2$ and $\theta_2 = 0.4$), SML and SPM performed better than BAR in terms of the larger proportion in the better arm and much higher power, while when that difference increased ($\theta_1 = 0.2$ and $\theta_2 = 0.6$), BAR performed better than SML and SPM with respect to both the larger proportion in the better arm and higher overall response rate and also the sufficiently high power. Furthermore, BAR has large variability compared to the other three methods, which is shown in the quantile estimates in terms of the key operating characteristics. Compared to SML, when an early stopping rule is applied, SPM required fewer patients in the trial. The main reason for incorporating an early stopping rule is to reduce the size of the trial, thereby trying to minimize the number of patients exposed to the inferior treatments. In this regard, SPM is superior to SML (it produces smaller trial size on average) and SPM can also be further improved by incorporating the informative prior. Therefore, it reinforces our recommendation in favor of SPM compared to the other methods.

Early stopping for efficacy can reduce the number of patients receiving ineffective treatments. On the other hand, early stopping for efficacy can increase the study efficiency by reducing the trial size. Note that the type I error rate could inflate with early efficacy stopping because there is more chance to reject the null hypothesis. Hence, the threshold for declaring efficacy needs to be re-calibrated to control type I error rate. Also note that, in this paper, all the conclusions and comparisons were based on the use of a fixed tuning parameter, where $\lambda = 1$. Further evaluation for comparing the RAR and ER methods can be made by varying the tuning parameter in the allocation probability. In summary, the relative advantages of RAR and ER were compared under the setting of patient horizon. ER is preferred for smaller treatment difference and when the number of patients beyond the trial is much larger than the number of patients in the trial. RAR is favored for large treatment difference or when the number of patients beyond the trial is small. Early stopping rules can be applied to reduce the trial size but the stopping boundaries need to be calibrated to

control the type I error rate while retain the study power. When early stopping rule is applied, the performance difference between the RAR and ER methods is reduced. By carefully choosing the design parameters, both RAR and ER methods can achieve the desirable statistical properties. The ultimate choice of RAR or ER methods depends on the investigator's preference, the trade-off between group ethics and individual ethics, and logistic considerations in the trial conduct, etc.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors thank the contributions of earlier work on this project by Degang Wang and Simon Lunagomez. The authors thank Ms. LeeAnn Chastain for editorial assistance. JJL's work was supported in part by grant CA016672 from the National Cancer Institute.

Abbreviations

RAR	response-adaptive randomization
ER	equal randomization
BAR	Bayesian adaptive randomization
SML	sequential maximum likelihood
SPM	sequential posterior mean

References

1. Berry DA. Bayesian Statistics and the Efficiency and Ethics of Clinical Trials. *Statistical Science*. 2004; 19:175–187.
2. Zelen M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*. 1969; 64:131–146.
3. Wei LJ. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*. 1978; 73:559–563.
4. Rosenberger, WF.; Lachin, JM. *Randomization in Clinical Trials: Theory and Practice*. New York: John Wiley and Sons; 2002.
5. Hu FF, Zhang LX. Asymptotic properties of double adaptive biased coin designs for multi-treatment clinical trials. *Annals of Statistics*. 2004; 32:268–301.
6. Rosenberger WF, Vidyashankar AN, Agarwal DK. Covariate-adjusted response-adaptive designs for binary response. *Journal of Biopharmaceutical Statistics*. 2001; 11:227–236. [PubMed: 12018777]
7. Zhang LX, Chan WS, Cheung SH, Hu FF. A generalized drop-the-loser urn for clinical trials with delayed responses. *Statistica Sinica*. 2007; 17:387–409.
8. Bai ZD, Hu FF, Rosenberger WF. Asymptotic properties of adaptive designs for clinical trials with delayed response. *Annals of Statistics*. 2002; 30:122–139.
9. Morgan CC, Coad DS. A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Statistics in Medicine*. 2007; 26:1937–1954. [PubMed: 16981177]
10. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*. 1933; 25:285–294.

11. Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? *Journal of Clinical Oncology*. 2011; 29:771–776. [PubMed: 21172882]
12. Lee JJ, Chen N, Yin G. Worth adapting? Revising the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*. 2012; 18:4498–4507. [PubMed: 22753588]
13. Anscombe FJ. Sequential Medical Trials (with discussion). *Journal of the American Statistical Association*. 1963; 58:365–387.
14. Thall PF, Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *European Journal of Cancer*. 2007; 43:859–866. [PubMed: 17306975]
15. Lee JJ, Gu XM, Liu SY. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*. 2010; 7:584–596. [PubMed: 20571130]
16. Wathen JK, Cook JD. Power and bias in adaptive randomized clinical trials. Technical Report. 2006 UTMDABTR-002-06.
17. Yin G, Chen N, Lee JJ. Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2012; 61:219–235.
18. Thall PF, Simon R. Practical Bayesian guidelines for phase IIb clinical trials. *Biometrics*. 1994; 50:337–349. [PubMed: 7980801]
19. Melfi VF, Page C, Geraldles M. An adaptive randomized design with application to estimation. *Canadian Journal of Statistics*. 2001; 29:107–116.

Figure 1.

Change in the allocation probability during the trial for BAR (top panel), SML (middle panel), and SPM (bottom panel) methods, respectively. We considered five realizations of the trial. 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 $\theta_1 = 0.4$ and $\theta_2 = 0.7$.

Figure 2.

Performances of four randomization methods without early stopping for the horizon size $N = 1000$ as assessed by the mean proportion in arm 2 (the superior treatment), mean overall response rate at the horizon level, and power at the trial level, with $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4$, and 0.6 , respectively. On the horizontal axis, m represents the trial size. The red line represents the BAR design, and orange, green, and black lines represent SML, SPM, and ER, respectively. Each line was derived from 100,000 simulations. The tuning parameter λ was specified as 1.

Figure 3.

Performances of four randomization methods without early stopping for the horizon size $N = 50,000$ as assessed by the mean proportion in arm 2 (the superior treatment), mean overall response rate at the horizon level, and power at the trial level, with $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4$, and 0.6 , respectively. On the horizontal axis, m represents the trial size. The red line represents the BAR design, and orange, green, and black lines represent SML, SPM and ER, respectively. Each line was derived from 100,000 simulations. The tuning parameter λ was specified as 1.

Figure 4.

Performances of four randomization methods with early stopping for the horizon size $N = 1,000$ as assessed by the mean proportion in arm 2 (the superior treatment), mean overall response rate at the horizon level, power, and mean trial size at the trial level, with $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4, \text{ and } 0.6$, respectively. On the horizontal axis, m represents the trial size. The red line represents the BAR design, and orange, green, and black lines represent SML, SPM and ER, respectively. Each line was derived from 100,000 simulations. The tuning parameter λ was specified as 1.

Figure 5.

Performances of four randomization methods with early stopping for the horizon size $N = 50,000$ as assessed by the mean proportion in arm 2 (the superior treatment), mean overall response rate at the horizon level, power, and mean trial size at the trial level, with $\theta_1 = 0.2$ and $\theta_2 = 0.2, 0.4$, and 0.6 , respectively. On the horizontal axis, m represents the trial size. The red line represents the BAR design, and orange, green, and black lines represent SML, SPM, and ER, respectively.

SPM and ER, respectively. Each line was derived from 100,000 simulations. The tuning parameter λ was specified as 1.