Published in final edited form as:

Stat Biopharm Res. 2019 ; 11(4): 375–386. doi:10.1080/19466315.2019.1610044.

Using Adaptive Designs to Avoid Selecting the Wrong Arms in Multiarm Comparative Effectiveness Trials

Byron J Gajewski,

Department of Biostatistics & Data Science, University of Kansas Medical Center, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, USA

Jeffrey Statland,

Department of Neurology, University of Kansas Medical Center, Mail Stop 2012, 3901 Rainbow Blvd., Kansas City, KS 66160, USA

Richard Barohn

Department of Neurology, University of Kansas Medical Center, Mail Stop 2012, 3901 Rainbow Blvd., Kansas City, KS 66160, USA

Abstract

Limited resources are a challenge when planning comparative effectiveness studies of multiple promising treatments, often prompting study planners to reduce the sample size to meet the financial constraints. The practical solution is often to increase the efficiency of this sample size by selecting a pair of treatments among the pool of promising treatments before the clinical trial begins. The problem with this approach is that the investigator may inadvertently leave out the most beneficial treatment. This paper demonstrates a possible solution to this problem by using Bayesian adaptive designs. We use a planned comparative effectiveness clinical trial of treatments for sialorrhea in amyotrophic lateral sclerosis as an example of the approach. Rather than having to guess at the two best treatments to compare based on limited data, we suggest putting more arms in the trial and letting response adaptive randomization (RAR) determine better arms. To ground this study relative to previous literature we first compare RAR, adaptive equal randomization (ER), arm(s) dropping, and a fixed design. Given the goals of this trial we demonstrate that we may avoid ‘type III errors’ - inadvertently leaving out the best treatment - with little loss in power compared to a two-arm design, even when choosing the correct two arms for the two-armed design. There are appreciable gains in power when the two arms are prescreened at random.

Keywords

Response adaptive randomization; Bayesian methods; clinical trials; equal randomization; adaptive designs

Corresponding Author: Byron J Gajewski, Department of Biostatistics, University of Kansas Medical Center, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, USA, bgajewski@kumc.edu.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1. INTRODUCTION

Limited resources are a challenge when planning comparative effectiveness studies of multiple promising treatments, often prompting study planners to reduce the sample size to meet the financial constraints. For example, consider a disease in which four treatments (A, B, C, and D) are commonly prescribed in practice but little is known regarding their relative effect. Let θ_{0A} , θ_{0B} , θ_{0C} , and θ_{0D} be the true but unknown response rates for the respective treatment arms. Then power is defined as the probability of selecting a single best treatment arm in a comparative effectiveness trial. The type I error is specifically defined as the probability of selecting a single best treatment arm under the null hypothesis ($H_0: \theta_{0A} = \theta_{0B} = \theta_{0C} = \theta_{0D}$). Given a fixed budget, the investigators have two choices, keep all the four treatments in a comparative effectiveness trial (strategy 1), or in order to increase the average information per treatment arm, reduce the number of arms in the study down to two treatments, say A and B (strategy 2). A positive of strategy 1 is that the trial includes all four treatments, thus each has a chance to be demonstrated at being the best. A negative of strategy 1 is that with so many treatment arms it may have less desirable power. A positive of strategy 2 is that the increase in sample size per treatment arm will increase the power of the study. However, a negative with strategy 2 is that the investigator may inadvertently leave out the most beneficial treatment (Figure 1). Suppose among the four treatments only one of them is best, and it is treatment A. Given that the investigators have little information, a priori, about which treatment is the best, there is a 0.5 probability that treatment A will be selected using the prescreening in strategy 2. A further problem with strategy 2 is it only compares A and B (say) to each other and so one may still question if treatments C and D are better since they would not be placed in the trial for investigation.¹

This paper demonstrates a possible solution to this problem by using Bayesian adaptive designs to efficiently conduct strategy 1. We use a planned comparative effectiveness clinical trial of treatments for sialorrhea in amyotrophic lateral sclerosis as an example of the approach. Rather than having to guess at the two best treatments to compare based on limited data, we suggest putting more arms in the trial and letting response adaptive randomization (RAR) determine better arms.

There are several general tutorials on adaptive designs (e.g. Jennison & Turnbull, 2000; Bhatt & Mehta, 2016; Pallmann et al, 2018). Bayesian approaches for comparative effectiveness research allow for cumulative learning as data are collected (Berry, 2012). Using simulation, the relative benefits of Bayesian adaptive designs for comparative effectiveness trials were recently studied in ‘proof of concept’ studies in status epilepticus (Connor et al., 2013A; Connor et al., 2013B) and antihypertensive and lipid lowering treatment study (Luce et al., 2016). In the latter study the authors performed a re-execution of the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) study and found that the Bayesian approach yielded similar conclusions to the original study but placed more patients on the better arm. These studies use simulation, an

¹Note that in the case of a lot of prior knowledge, the investigators could balance the impact of what we can learn from the trial in terms of improved decision making, with the cost of generating information, see for example Welton, Maden, & Ades (2011). A value of information approach could be used to determine which treatments should be compared in a two-arm trial or whether there is value in performing a multi-arm trial.

advocated and accepted process (Mawocha et al., 2017), to study the merits of adaptive designs relative to other designs. It is noted that there is a long history, over the last 20 years (Xu et al., in press), of Bayesian adaptive designs with response adaptive randomization. Therefore, there are lots of different models and approaches. Some examples of other studies that use simulation for investigating the relative merits of response adaptive randomization in multi-arm clinical trials are by the authors Wathen & Thall (2017) and Lee et al. (2012).

Other papers have compared adaptive and balanced designs, for example Trippa et al. (2012) and Wason & Trippa (2013), but always fix the number of treatment arm comparisons across comparative designs, for example at three arms and five arms in these respective studies. An exception would be a design that drops arms as it moves along, but the studies do begin with a full set of treatment arms at initial study enrollment. We take a different approach since we add comparative designs in which investigators screen treatment arms *before* executing the trial, which is common practice in academic medicine. Given the goals of comparative effectiveness trials, we demonstrate that we may avoid ‘type III errors’ - inadvertently leaving out the best treatment - with little loss in power compared to a two-arm design. We will show in this paper that we are better off using Bayesian adaptive designs with all of the arms because it is still competitive with designs that prescreen arms correctly, but at the same time is not as risky in case the screening goes badly, that is the best treatment arm is wrongly screened out.

While highly successful in practice, response adaptive randomization is not free of limitations. For example, Korn & Freidlin (2011) discuss some of these limitations. Further, a recent study (Hey & Kimmelman, 2015) challenged the ethics of Bayesian response adaptive randomization, for two armed studies, stating that they do not stand up to careful scrutiny. While some of the responses in the commentaries highlight potential challenges in response adaptive randomization (Buyse, 2015; Korn & Freidlin, 2015), most of the responses (Saxman, 2015; Lee, 2015; Joffe & Ellenberg, 2015; and Berry, 2015) counter the conclusions raised by Hey & Kimmelman (2015), for example “Suffice to say that I disagree with essentially all their points, sometimes because they are factually wrong (Berry, 2015, p. 107).” In the case of two arms a fixed design with equal allocation has approximately optimal power, even when testing two proportions (Azriel et al., 2012). However, we are of the opinion that response adaptive randomization has some advantages over the fixed randomization even in two-armed designs if one is concerned with power plus other characteristics of the trial, for example treatment benefit to the patients in the trial (Wick et al., 2017). Indeed, moving to more than three arms Hey & Kimmelman (2015) admit that there are benefits to Bayesian response adaptive randomization. Oftentimes three or more arms are possible in comparative effectiveness studies.

There are many design choices for Bayesian adaptive designs for comparative effectiveness studies. Some of the choices are the burn-in sample size, allocation formulas, stopping criteria, and accrual rates. In fact, one of the big issues in adaptive designs is the patient accrual rate relative to how long it takes to collect the endpoint (Gajewski et al., 2015). If one accrues too fast one cannot adapt; if one accrues too slowly one cannot finish the study; therefore, the rate of accrual can be optimized with simulation to balance trial speed and sample size. We are aware of these design choices and report design parameters that are

reasonable knowing that they can be optimized. But across these choices we emphasize that the features attractive to the Bayesian adaptive designs is that they have many benefits since they can put more patients on better treatment arms and can result in trials that are smaller in sample size, more powerful, and faster than fixed trials (Berry et al., 2015). Further, as long as we pre-specify all of the adaptive possibilities, we can calculate operating characteristics that support rigorous trial designs, including controlling the overall type I error rate, but at the same time enjoying flexible forms, much like the Greek god Proteus who was capable of assuming many forms, adapting to the changing nature of the sea (Sellaturay et al., 2012).

2. Methods

We frame our simulations around a specific comparative effectiveness trial design. We propose to test which of four commonly prescribed medications for drooling (sialorrhea) would be most effective for patients with amyotrophic lateral sclerosis (ALS), a debilitating and fatal neurodegenerative disease (Jackson et al., 2015). Difficulty swallowing can lead to excess saliva and drooling in patients with ALS. This symptom causes both a social and medical burden to the patient and their family, and drooling can lead to choking episodes which can cause aspiration pneumonia. While there are many commonly utilized medications for managing drooling in ALS, the best medication is unknown. The aim of this clinical study is to determine which of the four most commonly prescribed medications is best in controlling drooling. Patients will be randomized to receive one of the following: 1) scopolamine patch (1 mg) every 72 hours; 2) glycopyrrolate 1 mg three times a day; 3) amitriptyline 25 mg at bedtime; or 4) atropine 1% sublingual drops 2 drops three times a day. The endpoint is defined as a positive response to medication if the patient says their drooling is 'slightly better' or 'markedly better' 4-weeks after randomization.

A maximum sample size of 200 patients will be enrolled in this study and the goal is to have high power to identify a single arm as best. For all of the trial designs described there are predefined rules. For example this includes the adaptive designs in which allocation ratios may change based on their responses, designed to improve the power of the study and steer more patients towards the two best arms in order to increase power. The adaptive trial designs also include interim analyses for stopping early for success.

2.1 Statistical model

Bayesian quantities evaluate the final determination of which intervention is best (e.g. significance) as well as provides allocation probabilities for response adaptive randomization (RAR). We refer to the 'best' arm as the arm of maximum response rate. The j^{th} arm runs from $j=1, \dots, J$ where J is the number of total arms in the trial design. For example, strategy 1 uses all four arms so $J=4$ and strategy 2 will have less arms after taking some of the arms out before the trial begins, for example $J=2$ or 3. n_j is the number of patients in study arm j that have outcome data. S_j is the number of responses and is modeled as a binomial distribution $S_j \sim \text{Bin}(n_j, \theta_j)$, where θ_j is the true response rate for arm j . In addition, we provide 'weakly informative' priors, $\text{logit}(\theta_j) \sim \mathcal{N}(0, 1.82^2)$. Using the endpoint data and the prior probabilities, we then use Markov Chain Monte Carlo (MCMC) computations to obtain the Bayesian posterior distributions of θ_j .

The medication with the highest response rate is labeled E_{max} . As the trial is going on we will not know which medication is E_{max} so we will estimate, via MCMC draws, the probability each of the medications is the best, denoted by $P(j=E_{max})$. Specifically the probability that treatment j is the best treatment is defined as $P(j=E_{max}) = \Pr(\theta_j > \theta_X; \theta_j > \theta_Y; \text{ and } \theta_j > \theta_Z)$ where X , Y , and Z represent the three treatments other than treatment j . At the end of the trial, that is after all patients enrolled are followed up, we will have identified a best treatment arm if there exists an arm such that $P(j=E_{max})$ exceeds some value. This value is pre-specified depending on the trial considered in order to achieve an appropriate type I error rate (less than or equal to 0.05).

We also denote $Var(\theta_j)$ as the posterior variance of the response rate for medication j . This quantity will be used explicitly in the response adaptive randomization (RAR) along with $P(j=E_{max})$ and the sample size n_j .

2.2 Several types of designs

In recent years we have designed several Bayesian adaptive clinical trials for comparative effectiveness research (e.g. Gajewski et al., 2015). We find in most of the comparative effectiveness clinical trial designs that adaptive trials' benefit outweighs the increased cost relative to a fixed trial design. Further, this process is evolutionary as we are interested in identifying strategies towards an optimal design. In theory our definition of optimal design would be an unbiased test (i.e. type I error .05) that has largest power, smallest sample size, finishes the fastest, and places the most patients on the best arm, among a broad range of alternative hypotheses (e.g. response rates scenarios). We acknowledge that proving optimality is impractical, so we aim to step forward towards optimality. In our first version of the design we proposed a Bayesian adaptive trial design with RAR and we present that methodology first. Then we move to equal randomization (ER) and fixed trial designs. A third comparison in one that is also adapted with interim analysis but rather than using RAR we allow for the dropping of arms for futility. We describe all of the design options using $J=4$ arms and modify for scenario 2 (less arms). See Figure 2 for a visual representation of all of the designs.

2.2.1 Bayesian adaptive design with RAR—In the beginning of the trial we have what looks like a fixed trial design. The first 80 subjects are randomized 1:1:1:1, then response adaptive randomized (RAR) is implemented up to a total maximum number of subjects $n_{max}=200$. At each interim analysis a decision is made to either continue enrolling subjects or to stop the trial for conclusive results.

- The first interim analysis happens after 80 patients are randomized then they occur every 13 weeks. We could have chosen 100 patients or some other value. But we felt the operating characteristics with 80 patients was reasonable.
- Early success stopping criteria is if at an interim analysis after 100 patients enrolled there is a medication arm with a high probability of being the best, that is if there is an arm such that $P(j=E_{max}) > 0.88$, then we will stop the trial for early success. This criteria was chosen to preserve a portion of the 5% Type I error but allow a relatively high probability of early stopping. The other portion of the 5%

Type I error is saved for the final success criteria. The current choice is calibrated in order to have roughly $\frac{3}{4}$ of the 5% error taken early. A less aggressive cut-point, say 0.95, could have been chosen but would have less chance of stopping the trial early.

- Response adaptive randomization RAR is conducted if stopping criteria is not achieved, then we allocate the probability the next patient to be enrolled in the j^{th} arm to be proportional to $V_j^* = \sqrt{\frac{P(j=E_{\max})\text{Var}(\theta_j)}{n_j+1}}$. We could have chosen to allocate on just the probability of the arm being the best (e.g. $P(j=E_{\max})$), however, the allocation we chose, called *information weighting*, favors better medications, however in the case of two arms being equally effective, the arm with fewer participants will be favored (e.g. Connor et al., 2013). The justification for the extra weighting is the reduction in variance from adding an extra patient to that arm. For example, with n_j patients the variance is $\text{Var}(\theta_j)$ but adding one patient to this arm the variance becomes approximately $\frac{n_j\text{Var}(\theta_j)}{n_j+1}$, thus the reduction in variance is $\text{Var}(\theta_j) - \frac{n_j\text{Var}(\theta_j)}{n_j+1} = \frac{\text{Var}(\theta_j)}{n_j+1}$. Other forms of this information weighting have been studied (e.g. Meinzer, Martin, and Suarez, 2017) and used in other proposed clinical trials (e.g. Satlin et al., 2016).
- Final success criteria is established (significance) if we have identified a best medication arm with fairly high probability, that is if there is at least one arm such that $P(j=E_{\max}) > 0.865$. This cutoff was chosen to preserve 5% Type I error, given the early success cut-point. It also is less aggressive than early because it is a verification criteria when stopping early and new data that are followed-up will likely lead to lower probabilities. This allows a lower chance of flip-flop results, for example early stopping criteria met but not the final criteria met.

2.2.2 Bayesian adaptive design with equal randomization (ER)—Previous research (Wathen & Thall, 2017) suggests always comparing RAR to equal randomization (ER). Thus we also consider fixed allocation of 1:1:1:1 through the entire adaptive design. Regarding the ER design, it should be noted that this is still an “adaptive” design even though it maintains equal allocation probabilities throughout. This is because we can still have early stopping for success which may adapt the final sample size. This design has all the features of the RAR design except it does not use the RAR allocation formula. In the ER design early stopping and final success criteria are if we have identified an arm such that $P(j=E_{\max}) > 0.90$ and $P(j=E_{\max}) > 0.880$, respectively.

2.2.3 Bayesian adaptive design with arm(s) dropping—This design is similar to the Bayesian adaptive design with equal randomization (ER) except that it allows to drop arms if they exhibit a high probability of being futile. Specifically, if at any point during the interim analysis there are arms such that $P(j=E_{\max}) < .15$ then we drop those arms and then continue with equal randomization. The choice of the futility rule is somewhat arbitrary and other choices could be explored. However, this futility rule coupled with the success criteria

achieve 5% type I error making it calibrated with the other designs. In the arm dropping design early stopping and final success criteria are if we have identified an arm such that $P(j=E_{max})>0.865$ and $P(j=E_{max})>0.825$, respectively.

2.2.4 Bayesian fixed trial design—We also consider a fixed trial design for comparison purposes. In this trial design all 200 subjects are assumed to be randomized 1:1:1:1. Since this design is required to accrue 200 and the other designs have a maximum sample size of 200 but also have options for early stopping, they will always beat the fixed sample design with respect to expected sample size. The question becomes will there be any loss of power? The final analysis used for the fixed design defines trial success if there exists a treatment arm such that $P(j=E_{max})>0.875$.

2.2.5 Criterion for designs with less than 4 arms (2 or 3)—In the previous sections we presented early stopping and final success criteria across several study designs for four arms. The criteria were specified to provide a Type I error rate of 5%. Later in this paper we report designs with a smaller number of arms (e.g. two and three arms). These latter designs require adjustments of the early stopping and final success criteria to maintain the Type I error rate of 5%. The criteria for all trial designs are presented in Table 1.

2.3 Comparing specific strategies

Our goal is to provide benefits and risks to two overall design strategies outlined in Figure 1. Recall from Figure 1 that our overall approach is to first decide a design under strategy 1 (don't leave out arms). This requires us to tweak the study design through a comparison of several types of designs, specifically, RAR, ER, arms (s) dropping, and fixed. Then the best approach from strategy 1 requires a comparison to strategy 2. Recall that strategy 2 is to choose only two arms to be in the trial design with the hope that the correct two arms are chosen. Before we can compare designs and strategies we will need to hypothesize different plausible scenarios of how we think the participants will vary in their response to the medications.

2.3.1 Strategy 1: don't leave arms out—For the purposes of this investigation we look at several scenarios for responses to determine the power, sample size and time (duration) needed for our study. We have several scenarios reflecting response rates of six patterns (Table 2).

2.3.2 Strategy 2: Choosing trials with only 3- and 2-arm alternatives (deciding on arms)—We also look at the scenarios when we left out arm 1 (Table 3), which in all scenarios is the correct choice since its response rate is always the lowest among medications. We also investigate the scenarios where the two worst arms are left out before conducting the trial (Table 4).

2.4 Comparing trial designs with operating characteristics

For all trial designs we wish to calculate operating characteristics including power, type I error, sample size, duration, and the percentage of patients expected in each arm. This is accomplished through simulation. Using the true response rates for each of the medications

we run simulations through each of the trial designs which will include the adaptations where appropriate. For each trial iteration we will track whether we have identified a medication arm with the best response rate (e.g. power) as well as under the null hypothesis (e.g. no difference in response rates across arms) if we have incorrectly identified an arm with the best response rate (e.g. type I error). For each trial we will also track the total sample size, trial duration, and the number assigned to each medication arm. We will repeat this 1,000 times per scenario and design. For each combination of scenario we will compare the operating characteristics of the designs under RAR, ER, arm(s) dropping, and fixed. Note that when we are prescreening arms before starting the trial that we appropriately adjust the early stopping rules and final success such that we have similar early and late stopping probabilities under the no difference scenario, essentially changing the stopping and success criteria defined by the Bayesian quantity $P(j=E_{max})$.

In addition to scenarios of response rates we will need to make an assumption about how fast or slow we accrue the patients. Accrual patterns are important to Bayesian adaptive designs and refer to how rapidly each site enrolls patients in the trial. We assume that the distribution of the accrual patterns follows a Poisson distribution with a mean number of patients accrued per week. We expect to accrue an average of 2 patients/week. Specifically, the number of patients enrolled/week (n_t) is expected to follow a Poisson distribution, $n_t \sim \text{Poisson}(2)$. This accrual rate assumption is important for identifying how long the trial will last. Note that since the recruitment rate is on average 2 patients per week and there is a 4-week delay in response, there will be on average 8 patients in the pipeline at each interim analysis (who have commenced treatment but are yet to be followed-up for their primary response). If the trial stops early for success at an interim analysis then the patients in the pipeline will be followed up and the final success criteria will confirm early success. Later in the paper we are going to consider additional rates and discuss its impact on trial design.

2.5 Code for computing operating characteristics for trial designs

Computations are performed using software: Fixed and Adaptive Clinical Trial Simulator (FACTS) (Berry & Sanil, 2010), a commercially available program designed to rapidly design, compare, and simulate both fixed and adaptive trial designs. Because FACTS is unavailable to the general user, we wrote code in R (see Appendix A) that provides calculation for the probability each of the arms is maximum as well as allocation probabilities of each arm using RAR. The program uses a Bayesian bootstrap (Gelman et al., 2014) approach to gain posterior draws of size S in each arm. The code requires the user to input responses, sample size for each arm, and the number of draws.

3. Results

The following results are based on R code in the appendix as well as using the FACTS software. For the R code we examined time series plots of the MCMC draws and found quick convergence since it is using a Bayesian bootstrap it converged quickly. We also investigated similar plots for the draws provided from a few simulation results from FACTS, these had similar convergence.

3.1 The way the adaptive design works

Table 5 shows an example, for RAR, with interim analyses occurring every 13 weeks after the first 80 patients are randomized, this is off a bit by the expected 2 patients/week*39 weeks=78 patients because accrual is a Poisson process. This example simulated trial was generated from ‘best and second best’ scenario. This trial does not stop early for success at the first interim analysis despite the 4th arm meeting the stopping criteria

$P(j=E_{max})=0.97>0.88$ because the trial had not yet met the minimum sample size of 100 enrolled patients. This trial stops early for success at the second interim analysis since the 4th arm meets the stopping criteria of both $P(j=E_{max})=1.00>0.88$ and having enrolled 107 patients which is larger than the 100 patients required. Further, after all of the patients are followed up the trial achieves final success at 64 weeks since the 4th arm meets the final success criteria of $P(j=E_{max})=0.97>0.865$. Example calculations for this case are shown, using our R-code in the appendix. This trial took about ½ of the maximum sample size to complete. Also note that after the first interim analysis if we had not met the stopping criteria then the RAR probability of the next patient being randomized to the best arm is large, 0.81.

This example trial shown in Table 5 allows investigators the opportunity to see what is happening and reacting in a positive way to make the trial more efficient for the investigative team and the patients. But at the same time, this trial design does not ignore important statistical properties, such as controlling the familywise type I error. As will be shown in the next section there are no negative implications for type I error of funneling more patients into the ‘winning’ arm at interim analysis because of our carefully constructed stopping rules.

3.2 Comparison of designs

By replicating through simulations we are able to get frequentist properties of the different trial designs (e.g. type I error and power). Table 6 shows a comparison between adaptive (both RAR, ER, and arm(s) dropping) and fixed clinical trial designs across several scenario profiles for strategy 1. We are comparing the designs in a fair way since they all have an overall ‘power’ (type I error), calculated using the ‘no difference’ case, at a rate 5%. Similar to the references in the introduction, in all the other scenarios the adaptive approaches are superior as they produce trials that have lower expected sample size (E(N)), are more powerful, and have a lower expected time (E(T)) to be completed. Further, as is typical in trials with three or more arms, the RAR allows the trial to operationally target more patients towards the better arms making the resources more efficient in use, leading to trials that are both smaller in sample sizes and better in power, with better targets to the better therapies. If we average all of the scenarios outside of ‘no differences’ the comparisons in operating characteristics for RAR, ER, arm(s) dropping, and fixed is: 131, 138, 131, and 200 patients (subjects), 94%, 88%, 92%, and 84% power, and 69, 73, 69, and 104 weeks to operate (Table 7). Also illustrated in Table 7 is the average percentage of patients placed on the best arm is 38%, 25%, 34%, and 25% for RAR, ER, arm(s) dropping, and fixed respectively. Therefore, in addition to power and sample size, this is a benefit of RAR, from a patient’s perspective, in that the RAR approach places a higher percentage of

patients on the best arm. The property of the RAR to increase the percentage of patients receiving the best therapy is also very desirable to clinicians, and helps reassure patients.

Note that for the procedures allowing early stopping, we tracked how often we trigger termination anticipating selection of one treatment as the best, for example with $P(j = E_{\max}) > 0.880$ for the RAR design, only to see $P(j = E_{\max}) < 0.865$ once all recruited patients have been followed up for their primary outcome. This rate is 0% across these studies, because we always require early stopping to be more stringent than the final success (see Table 1).

We have added more extensive simulation studies and report them in Appendix B. Specifically, we look at the operating characteristics as the treatment difference between the best and the rest gets increasingly narrow while also looking at what happens when the best treatment is at the lower variance part of the binary distribution (e.g. treatment 1 has $p=0.5$ and treatment 4 has $p=0.9$). A further analysis investigates what happens after we do a sensitivity analysis to the stopping rule. In the end, RAR still holds an advantage over the other designs.

Now that we have optimized strategy 1 (all arms), we need to consider strategy 2 (deciding on arms) and compare the two strategies. To this end we explore the situation in which the investigator is considering designs in which we take out one arm and then two arms to explore power differences to the full armed RAR. The justification for this change might be to raise the overall number of patients per arm, and possibly increase study power. However, one can see in Table 7 that going from 4 arms to 3 arms has very little change in sample size, trial duration, and power. The only real advantage is that we would be eliminating the investigation of a bad arm and so more patients are allocated to the best arm quickly. In the two armed case the difference in these parameters is not great. Although with two arms the fixed case has better power than the adaptive case. But adaptive outperforms fixed in all other cases. The Bayesian procedure permitting arm dropping performs far worse than the other procedures when we begin comparing by 2 arms, but it is very competitive when we start by comparing 3 or 4 arms. The explanation for this dramatic change is because once we get to 2 arms and we drop one of the arms (for futility) we are only enrolling in a single arm until we meet one of the success criteria, offering very inefficient use of enrollments.

The only advantage strategy 2 would have over strategy 1 is if we left out arms correctly, more overall patients would be placed on the best arm. However, this is a big assumption of having this done correctly. Is it worth it to decide on how many arms to use? We don't think so, in this example. The power differences between 4 arm RAR and 2 arm fixed are just not big enough to justify the risk of deciding on the right arms. One should caution, however, among the strategy 1 options, RAR is the only one that enjoys similar power properties to strategy 2 (deciding on arms to use).

A final comparison is between RAR using strategy 1 but assuming various accrual rates. The original rate was 2 patients/week, so we investigate a slower rate (1.5 patients/week) as well as a faster rate (3 patients/week). This is compared to a fixed design using strategy 2. However, in strategy 2 with fixed arms we investigate all arms having an equal chance of being selected in the prescreening. Table 8 shows the average sample size as well as the

average probability of selecting arm 4 for both scenarios. For scenario 1, we can see that average sample size does get larger as the accrual rate increases as we have less information for interim decision making and RAR, at the same time the probability of selecting the best arm remains stable. This reaction to various accrual rates is similar to the results in Gajewski et al. (2015). When switching to strategy 2, the total sample size is larger and the average probability of selecting the best arm drops substantially, this is because the prescreening causes 50% of the time the best arm is not selected at all for the trial.

4. Discussion

There are limitations of the Bayesian adaptive designs, for example, they require careful planning and execution (Brown et al., 2016). Indeed, a key reason to deviate from the adaptive approaches is if the investigators are unwilling to budget for proper statistical support for several interim analyses. Yes, the adaptive designs are more statistical effort - but in the scenario of a 'best' treatment they end up saving money due to shorter study time, and smaller number of participants.

Also, one may argue that our general design strategy has many features, such as the stopping criterion, the criterion for unbalancing the randomization, accrual rate assumptions, and the decision to pursue equal randomization for 80 of 200 patients. One could extend the general design strategy by optimizing these features, for example investigating a probability allocation formula (i.e. $P(j=E_{max})$) for RAR instead of using the information allocation

formula (i.e. $\sqrt{\frac{P(j=E_{max})Var(\theta_j)}{n_j+1}}$). Our presentation is a march towards optimization. We did

mention in the introduction that the ratio of accrual rate and the endpoint assessment time matters in adaptive designs. One way to handle the case of fast accrual with relative long assessment time window would be to use longitudinal modeling to predict the final endpoint (e.g. Brown et al., 2016). Related to prediction, not addressed in this paper but certainly could be studied, is the use of a predictive distribution (for example the predictive probability of reaching a conclusion with the next set of patients).

Another limitation of this paper is we do not study how the various designs would perform if our objective was to understand not only which was the best treatment, but also what the response probability is on that treatment (and by what margin this exceeds the response rates on other studied treatments). This estimation question is relevant since we would not wish to select a treatment for further use or study which was likely to be inferior to the current standard of care or superior by an irrelevant difference. This is an important question but one we are unlikely to have any prior knowledge about.

A fourth limitation of this study is that we compare the competing designs under configurations of response probabilities where there is quite a large degree of separation between the best treatment and the rest. The different methods would be quite comparable, however, when several therapies are highly competitive.

Here we show that a 4-arm study utilizing our RAR design has very favorable operating characteristics compared to studies with fewer arms. This was a bit of a surprise to

us. Although there may be increased costs to fund the interim analysis, in the setting of a 'best' treatment, or 'best and second best' this cost can be offset by shorter trial duration and smaller numbers of participants – and more importantly this reduces the study-burden on the patients. Further, think about one possible risk of a fixed 2-arm design. What if we left out arm 4 instead? Then there is a probability 0 we correctly choose that arm, even though it is in truth the 'best' treatment. This is a 'type III' error, or the error of the 'arm not studied.'

We emphasize in this paper hypothesis testing in the presence of selection. We note that "comparative effectiveness" can also include a pure selection procedure, without statements of statistical significance. That is, the goal of the study could be to simply select the most promising treatment for further confirmatory testing in a separate study. Pure selection designs can achieve their goals much more quickly than the ones we discussed, because there may be no need to control the type I error rate (because no statements of significance are made).

In conclusion, rather than having to guess at the two best treatments to compare based on limited data, we propose clinical trialists consider to utilize a RAR design to improve trial efficiency, effectively test more therapeutic options, and avoid making type III errors. By careful planning, and pre-specifying the approach to adaptation, we can possibly learn from the example of Proteus, and allow our trials to take on many forms during its cycle of operation, so that we get smaller, faster, and more powerful trials for with better therapies for more of the patients. Additionally, as we have shown here, RAR enjoys similar power to leaving out arms without the risk of leaving out the wrong arms!

Acknowledgements

We would like to thank Chuanwu Zhang for assistance in formatting this article.

Funding

This study was supported in part by a NIH Clinical and Translational Science Award (UL1TR002366) to the University of Kansas.

Appendix A: R Code

```

#### R-code for calculating the probability of each arm being the best and
the new RAR ####allocation probabilities
####Drooling_RAR

#####This function calculates the response adaptive randomization
#####probabilities and probability of success for the Drooling trial

##### 4/17/2018 (rev 9/19/18, fixed bug on prior)
##### Byron J. Gajewski

#####Inputs:
#y=4-vector each arm number of responses Arms
#n=4-vector each arm number of patients

#S=Number of simulation iterations

#####Outputs:
#Pbest=4-vector each arm is best
#RAR= 4-vector each arm is best new randomization probability

#####Function

Drooling_RAR<-function(y,n,S)
{
  #####Proportional posterior (last part is the Jacobian)
  logpost<-function(y,n,P)
  {
    dbinom(y,n,P,log=TRUE)+dnorm(log(P/(1-P)),0,1.82,log=TRUE)-log(P)-log(1-P)
  }

  #####Each
  q=matrix(NA,S,4)
  theta=matrix(NA,S,4)
  for (j in 1:4)
  {
    P=runif(S)
    q[,j]=exp(logpost(y[j],n[j],P))/dunif(P)
    q[,j]=q[,j]/sum(q[,j])
    #####Draws from control parameter
    theta[,j]=sample(P,S,replace=TRUE,prob=q[,j])
  }

  #####Best arm
  id=matrix(NA,S,1)
  Pbest =matrix(NA,1,4)

  for (i in 1:S) {id[i]=which(theta[i,]==max(theta[i,]))}
  for (j in 1:4) {Pbest[j]=mean(id==j)}

  #####Variance of parameters
  Vtheta=matrix(NA,1,4)
  for (j in 1:4) {
    Vtheta[j]=var(theta[,j])
  }

  #####Response adaptive randomization parameters
  V=matrix(NA,1,4)
  RAR=matrix(NA,1,4)

  for (j in 1:4) {
    V[j]=(Pbest[j]*Vtheta[j]/(n[j]+1))^(1/2)
  }
  for (j in 1:4) {
    RAR[j]=V[j]/sum(V)
  }

  return(c(Pbest,RAR))
}

#### Results for a simulated trial under 'best and second best' scenario.
####Example (completed)
>set.seed(1234)

>Drooling_RAR(y=c(3,2,3,9),n=c(15,15,15,15),S=1000000)
[1] 0.01196200 0.00328900 0.01164100 0.97310800 0.07575826 0.03510421 0.07468
233 0.81445519
>Drooling_RAR(y=c(3,2,3,13),n=c(20,20,20,21),S=1000000)
[1] 0.000932000 0.000201000 0.000891000 0.997976000 0.023125577 0.009420269 0
.022551381 0.944902774
>Drooling_RAR(y=c(3,2,6,21),n=c(22,20,23,42),S=1000000)
[1] 0.00154100 0.00072200 0.02999300 0.96774400 0.03882413 0.02637997 0.20230
052 0.73249538

```

Appendix B: Supplemental Operating Characteristics

Table B1.

Response rate scenarios when all arms are in the trial design when the best treatment is at the lower variance and the treatment difference between best and the rests gets narrower.

Scenarios	Arm 1	Arm 2	Arm 3	Arm 4
	<i>Response Rates</i>			
One Best, Small	0.50	0.50	0.50	0.63

Scenarios	Arm 1	Arm 2	Arm 3	Arm 4
One Best, Medium	0.50	0.50	0.50	0.77
One Best, Large	0.50	0.50	0.50	0.90

Table B2.

Simulated trial operating characteristics for 4-arms comparing adaptive to fixed trials.

Scenarios	Design	E(N)	Proportion Early Success	Proportion Late Success	Power	E(T) (weeks)	% Arm 4	Probability Select Arm 4	Probability Select Inferior Arm
1 Best, Small	RAR	182	0.32	0.05	0.37	95	32%	0.35	0.02
	ER	184	0.27	0.06	0.33	96	25%	0.31	0.02
	Arm(s) Drop	177	0.36	0.05	0.41	93	34%	0.39	0.02
	Fixed	200	0.00	0.26	0.26	104	25%	0.25	0.01
1 Best, Medium	RAR	132	0.92	0.03	0.95	70	35%	0.95	0.00
	ER	140	0.85	0.05	0.90	74	25%	0.90	0.00
	Arm(s) Drop	132	0.90	0.04	0.95	70	34%	0.94	0.00
	Fixed	200	0.00	0.87	0.87	104	25%	0.87	0.00
1 Best, Large	RAR	110	1.00	0.00	1.00	59	38%	1.00	0.00
	ER	112	1.00	0.00	1.00	60	25%	1.00	0.00
	Arm(s) Drop	110	1.00	0.00	1.00	59	32%	1.00	0.00
	Fixed	200	0.00	1.00	1.00	104	25%	1.00	0.00
Average	RAR	141	0.75	0.03	0.78	75	35%	0.76	0.01
	ER	145	0.71	0.04	0.75	77	25%	0.74	0.01
	Arm(s) Drop	140	0.75	0.03	0.78	74	33%	0.77	0.01
	Fixed	200	0.00	0.71	0.71	104	25%	0.71	0.00

RAR=full response adaptive randomization; ER=adaptive but equal randomization; Arm(s) Drop=arm dropping and early stopping; Fixed=full sample size

Table B3.

Simulated trial operating characteristics for 4-arms comparing adaptive to fixed trials, flipping the allocation of Type I error to be later, 1% early success and 4% late success, instead of 4% early success and 1% late success.

Scenarios	Design	E(N)	Proportion Early Success	Proportion Late Success	Power	E(T) (weeks)	% Arm 4	Probability Select Arm 4	Probability Select Inferior Arm
1 Best, Small	RAR	190	0.18	0.17	0.35	99	32%	0.35	0.00
	ER	193	0.13	0.18	0.31	100	25%	0.30	0.01

Scenarios	Design	E(N)	Proportion Early Success	Proportion Late Success	Power	E(T) (weeks)	% Arm 4	Probability Select Arm 4	Probability Select Inferior Arm
1 Best, Medium	Arm(s) Drop	193	0.13	0.18	0.31	100	34%	0.30	0.01
	Fixed	200	0.00	0.26	0.26	104	25%	0.25	0.01
	RAR	144	0.84	0.11	0.95	76	37%	0.95	0.00
	ER	154	0.72	0.18	0.90	81	25%	0.90	0.00
	Arm(s) Drop	154	0.72	0.18	0.90	81	36%	0.90	0.00
1 Best, Large	Fixed	200	0.00	0.87	0.87	104	25%	0.87	0.00
	RAR	111	1.00	0.00	1.00	59	39%	1.00	0.00
	ER	115	1.00	0.00	1.00	61	25%	1.00	0.00
	Arm(s) Drop	115	0.99	0.01	1.00	61	33%	1.00	0.00
	Fixed	200	0.00	1.00	1.00	104	25%	1.00	0.00
Average	RAR	148	0.67	0.09	0.76	78	36%	0.77	0.00
	ER	154	0.62	0.12	0.74	81	25%	0.73	0.00
	Arm(s) Drop	154	0.63	0.12	0.74	81	34%	0.73	0.00
	Fixed	200	0.00	0.71	0.71	104	25%	0.71	0.00

RAR=full response adaptive randomization; ER=adaptive but equal randomization; Arm(s) Drop=arm dropping and early stopping; Fixed=full sample size

References

1. Azriel D, Mandel M, Rinott Y (2012), "Optimal allocation to maximize the power of two-sample tests for binary response," *Biometrika*, 99, (1), 101–113.
2. Berry S, Sanil A (2010), *FACTS™ Dose finding: single endpoint engine specification*, Tessella, Newton, MA.
3. Berry SM, Carlin BP, Lee JJ, and Mueller P (2010), *Bayesian adaptive methods for clinical trials*, Chapman & Hall.
4. Berry DA (2012), "Bayesian approaches for comparative effectiveness research," *Clinical Trials*, 9(1): 37–47. [PubMed: 21878446]
5. Berry DA. (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 107–109. [PubMed: 25649111]
6. Bhatt DL, Mehta C (2016), "Adaptive designs for clinical trials," *New England Journal of Medicine*, 375, 65–74. [PubMed: 27406349]
7. Brown AR, Gajewski BJ, Aaronson LS, Mudaranthakam DP, Hunt SL, Berry SM, Quintana M, Pasnoor M, Dimachkie MM Jawdat O, Herbelin L, Barohn RJ (2016), "A Bayesian comparative effectiveness trial in action: developing a platform for multi-site study adaptive randomization," *Trials*, 17 (428). [PubMed: 27577191]
8. Buyse M (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 119–121. [PubMed: 25649112]
9. Connor JT, Elm JJ, Broglio KR, ADAPT-IT Investigators (2013), "Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus," *Journal of Clinical Epidemiology*, 66(8S), S130–137. [PubMed: 23849147]

10. Connor JT, Luce BR, Broglio KR, Ishak KJ, Mullins D, Vanness DJ, Fleurence R, Saunders E, Davis BR (2013), "Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study," *Clinical Trials*, 10, 807–827. [PubMed: 23983160]
11. Gajewski BJ, Berry SM, Quintana M, Pasnoor M, Dimachkie M, Herbelin L, Barohn R (2015), "Building efficient comparative effectiveness trials through adaptive designs, utility functions, and accrual rate optimization: finding the sweet spot," *Statistics in Medicine*, 34(7), 1134–1149. [PubMed: 25640114]
12. Gelman A, Carlin JB, Stern HS, Rubin DB (2014), *Bayesian data analysis*, Chapman & Hall/CRC Boca Raton, FL, USA.
13. Hey S, Kimmelman J (2015), "Are outcome-adaptive allocation trials ethical?," *Clinical Trials*, 12 (2), 102–106. [PubMed: 25649106]
14. Hu F, Rosenberger (2006), *The theory of response adaptive randomization in clinical trials*, Wiley, 2006.
15. Jackson CE, McVey AL, Rudnicki S, Dimachkie MM, & Barohn RJ (2015), "Symptom management and end-of-life care in amyotrophic lateral sclerosis," *Neurol Clin*, 33(4), 889–908. [PubMed: 26515628]
16. Jennison C, Turnbull BW (2000), *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRS, New York.
17. Joffe S, Ellenberg SS (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 116–118. [PubMed: 25649113]
18. Korn EL, Freidlin B (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 122–124. [PubMed: 25649108]
19. Korn EL, Freidlin B, (2011), "Outcome-adaptive randomization: is it useful?" *Journal of Clinical Oncology*, 29(6), 771–776. [PubMed: 21172882]
20. Lee JJ, Chen N, Yin G (2012), "Worth adapting? Revisiting the usefulness of outcome-adaptive randomization," *Clinical Cancer Research*, 18(17): 1–10.
21. Lee JJ (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 110–112. [PubMed: 25649110]
22. Luce BR, Connor JT, Broglio KR, Mullins CD, Ishak KJ, Saunders E, Davis BR. READAPT (REsearch in ADaptive methods for Pragmatic Trials) Investigators (2016), "Using Bayesian adaptive trial designs for comparative effectiveness research: a virtual trial execution," *Annals of Internal Medicine*, 20, 165(6), 431–8. [PubMed: 27273013]
23. Mawocha SC, Feters MD, Legocki LJ, Guetterman TC, Frederiksen S, Barsan WG, Lewis RJ, Berry DA, Meurer WJ (2017), "A conceptual model for the development process of confirmatory adaptive clinical trials within an emergency research network," *Clinical Trials*, 14(3), 246–254. [PubMed: 28135827]
24. Meinzer C, Martin R, and Suarez JI (2017), "Bayesian dose selection design for a binary outcome using restricted response adaptive randomization," *Trials*, 18, 420. [PubMed: 28886745]
25. Pallmann P, Bedding AW, Choodari-Oskoei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T (2018), "Adaptive designs in clinical trials: why use them, and how to run and report them," *BMC Medicine*, 16(1), 29. [PubMed: 29490655]
26. Satlin A, Wang J, Logovinsky V, Berry S, Swanson C, Dhadda S, and Berry D (2016), "Design of a Bayesian adaptive phase 2 proof-of-concept trial for BAN2401, a putative disease-modifying monoclonal antibody for the treatment of Alzheimer's disease," *Alzheimer's & Dementia*, 2, 1–12.
27. Saxman SB (2015), "Commentary on Hey, Kimmelman," *Clinical Trials*, 12 (2), 113–115. [PubMed: 25649109]
28. Sellaturay SV, Nair R, Dickinson IK, & Sriprasad S (2012), "Proteus: Mythology to modern times," *Indian Journal of Urology*, 28(4): 388–391. [PubMed: 23450503]
29. Trippa L, Lee EQ, Wen PY, Batchelor TT, Cloughesy T, Parmigiani G, & Alexander BM (2012), "Bayesian adaptive randomized trial design for patients with recurrent Glioblastoma," *J. of Clinical Oncology*, 30 (26), 3258–3263.
30. Wason JMS & Trippa L (2013) "A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials," *Statistics in Medicine*, 33, 2206–2221.

31. Wathen JK, Thall PF (2017), “A simulation study of outcome adaptive randomization in multi-arm clinical trials,” *Clinical Trials*, 14(5), 432–440. [PubMed: 28982263]
32. Welton NJ, Madan J, Ades AE. (2011), “Are head-to-head trials of biologics needed? The role of value of information methods in arthritis research,” *Rheumatology*, 50, 19–25.
33. Wick J, Berry SM, Yeh H, Choi W, Pacheco CM, Daley C, Gajewski BJ (2017), “A Novel Evaluation of Optimality for Randomized Controlled Trials,” *Journal of Biopharmaceutical Statistics*, 27 (4), 659–672 (PMCID: PMC5154788). [PubMed: 27295566]
34. Xu W, Hu F, Cheung SH (in press), “Adaptive designs for non-inferiority trials with multiple experimental treatments,” *Statistical Methods in Medical Research*.

Figure 1.

Two strategies for designing comparative effectiveness research (CER) with multiple arms and an approach for deciding which one to use. The examples provide possible situations. Every study is unique and this approach can assist trialists in deciding which strategy to implement.

RAR

ER

Arm(s) Dropping

Fixed

Figure 2.

Flow charts for study designs RAR, ER, Arm(s) Dropping, and Fixed designs. The rectangles are action points, parallelograms are decision points with a yes (Y) or no (N) routes, and the octagons are highlighted actions such as stopping early.

Table 1.

Early stopping and final success criteria by number arms included in the design as well as the Bayesian design. All of these trial designs have 5% Type I error rates.

# of Arms	Bayesian Design	Early Stopping ($P(j=E_{max})> .15$)	Final Success ($P(j=E_{max})> .15$)
4	adaptive with RAR	0.880	0.865
	adaptive with equal randomization (ER)	0.900	0.880
	adaptive with arm(s) dropping*	0.865	0.825
	fixed	--	0.875
3	adaptive with RAR	0.950	0.920
	adaptive with equal randomization (ER)	0.950	0.930
	adaptive with arm(s) dropping*	0.930	0.910
	fixed	---	0.900
2	adaptive with RAR	0.990	0.980
	adaptive with equal randomization (ER)	0.995	0.990
	adaptive with arm(s) dropping*	0.975	0.955
	fixed	---	0.975

The * means that an arm was dropped in interim analysis if $P(j=E_{max})< .15$, except for two arms where the adaptive arm dropping is the same as the adaptive with RAR.

Table 2.

Response rate scenarios when all arms are in the trial design.

Scenarios	Arm 1	Arm 2	Arm 3	Arm 4	
		<i>Response Rates</i>			
No Difference	0.10	0.10	0.10	0.10	all arms are equal
Best and 2 nd Best	0.10	0.10	0.20	0.40	one arm is best, one is 2 nd best
All Different	0.10	0.15	0.20	0.40	all different
One Strong Best	0.10	0.10	0.10	0.40	one arm is much better
One Modest Best	0.10	0.10	0.10	0.30	one arm is modestly better

Table 3.

Response rate scenarios when only three arms are in the trial design.

Scenarios	Arm 1	Arm 2	Arm 3	Arm 4
	<i>Response Rates</i>			
No Difference	0.10	0.10	0.10	all arms are equal
Best and 2 nd Best	0.10	0.20	0.40	one arm is best, one is 2 nd best
All Different	0.15	0.20	0.40	all different
One Strong Best	0.10	0.10	0.40	one arm is much better
One Modest Best	0.10	0.10	0.30	one arm is modestly better

Table 4.

Response rate scenarios when only two arms are in the trial design. Note that the scenarios “Best and 2nd Best” and “All Different” now have the same responses after prescreening arms 1 and 2.

Scenarios	Arm 1	Arm 2	Arm 3	Arm 4	
		<i>Response Rates</i>			
No Difference			0.10	0.10	all arms are equal
Best and 2 nd Best			0.20	0.40	one arm is best, one is 2 nd best
All Different			0.20	0.40	all different
One Strong Best			0.10	0.40	one arm is much better
One Modest Best			0.10	0.30	one arm is modestly better

Table 5.

Example trial demonstrating data gathered at each interim analysis, probability each treatment arm is the best (Emax) and randomization probabilities for the next patients enrolled. N is the number enrolled. The number with complete follow-up is a sum of “completers.”

Time (weeks)	N	Observed Responses/Completers				Probability Emax				Randomization Probs for Next 13-weeks			
		Arm1	Arm2	Arm3	Arm4	Arm1	Arm2	Arm3	Arm4	Arm1	Arm2	Arm3	Arm4
39	80	3/15 20%	2/15 13%	3/15 20%	9/15 60%	0.012	0.003	0.012	0.973	0.076	0.035	0.075	0.814
52	107	3/20 15%	2/20 10%	3/20 15%	13/21 62%	0.001	0.000	0.001	0.998	-	-	-	-
64	107	3/22 14%	2/20 10%	6/23 26%	21/42 50%	0.002	0.001	0.030	0.968	-	-	-	-

Table 6.

Simulated trial operating characteristics for 4-arms comparing adaptive to fixed trials.

Scenarios	Design	E(N)	std(N)	Proportion Early Success	Proportion Late Success	Power	E(T) (weeks)	% Arm 4	Probability Select Arm 4	Probability Select Inferior Arm
One Strong Best	RAR	116	21	0.99	0.01	1.00	62	39%	1.00	0.00
	ER	120	26	0.97	0.02	0.98	64	25%	0.98	0.00
	Arm(s) Drop	118	22	0.97	0.01	0.98	63	33%	0.98	0.00
A Best & 2 nd Best	Fixed	200	0	0.00	0.97	0.97	104	25%	0.97	0.00
	RAR	132	33	0.91	0.03	0.94	66	38%	0.94	0.00
	ER	137	36	0.85	0.03	0.88	72	25%	0.89	0.00
All Different	Arm(s) Drop	129	32	0.91	0.03	0.94	68	34%	0.94	0.00
	Fixed	200	0	0.00	0.81	0.81	104	25%	0.81	0.00
	RAR	135	34	0.90	0.03	0.93	71	38%	0.92	0.01
One Modest Best	ER	144	38	0.80	0.06	0.86	76	25%	0.85	0.01
	Arm(s) Drop	134	34	0.88	0.03	0.91	71	34%	0.91	0.00
	Fixed	200	0	0.00	0.79	0.79	104	25%	0.79	0.00
No Differences	RAR	141	36	0.86	0.04	0.90	75	38%	0.89	0.01
	ER	149	39	0.75	0.05	0.80	78	25%	0.78	0.02
	Arm(s) Drop	143	39	0.79	0.05	0.84	75	35%	0.84	0.00
No Differences	Fixed	200	0	0.00	0.78	0.78	104	25%	0.78	0.00
	RAR	198	14	0.03	0.01	0.04	103	25%	0.01	0.03
	ER	198	13	0.04	0.01	0.05	103	25%	0.01	0.04
No Differences	Arm(s) Drop	197	16	0.04	0.01	0.05	102	27%	0.02	0.03
	Fixed	200	0	0.00	0.04	0.04	104	25%	0.01	0.03

RAR=full response adaptive randomization; ER=adaptive but equal randomization; Arm(s) Drop=arm dropping and early stopping; Fixed=full sample size

Table 7.

Simulated trial operating characteristics for averaged across all alternative hypotheses.

Strategy	Design	Average Sample Size	Average Time (weeks)	Average Power	%Arm 4	Probability Select Arm 4
Strategy 1	4 arms: RAR	131	69	0.94	38%	0.94
	4 arms: ER	138	73	0.88	25%	0.88
	4 arms: Arm(s) Drop	131	69	0.92	34%	0.92
	4 arms: Fixed	200	104	0.84	25%	0.84
Strategy 2	3 arms: RAR	137	72	0.94	53%	0.94
	3 arms: ER	136	72	0.91	33%	0.91
	3 arms: Arm(s) Drop	141	74	0.82	64%	0.82
	3 arms: Fixed	200	104	0.92	33%	0.92
	2 arms: RAR	141	74	0.90	71%	0.90
	2 arms: ER	144	75	0.88	50%	0.88
	2 arms: Arm(s) Drop	148	77	0.77	80%	0.77
	2 arms: Fixed	200	104	0.93	50%	0.93

AR=full response adaptive randomization; ER=adaptive but equal randomization; Arm(s) Drop=arm dropping and early stopping; Fixed=full sample size

Table 8.

Simulated trial operating characteristics for averaged across all alternative hypotheses, focusing on RAR and Fixed for Strategy 1 and Fixed for Strategy 2, after slowing the accrual rate from 2 patients/week down to 1.5 patients/week or speeding up to 3 patients/week. For Strategy 2 we average across all combinations of choices.

Strategy	Design	Average Sample Size	Average Probability Select Arm 4
Strategy 1	4 arms: RAR (accrual 1.5/week)	131	0.94
	4 arms: RAR (accrual 2/week)	131	0.94
	4 arms: RAR (accrual 3/week)	140	0.93
Strategy 2	2 arms: Fixed	200	0.49

AR=full response adaptive randomization; Fixed=full sample size