Published in final edited form as:

Stat Med. 2015 March 30; 34(7): 1134–1149. doi:10.1002/sim.6403.

Building Efficient Comparative Effectiveness Trials through Adaptive Designs, Utility Functions, and Accrual Rate Optimization: Finding the Sweet Spot

Byron J. Gajewski¹, Scott M. Berry², Melanie Quintana², Mamatha Pasnoor³, Mazen Dimachkie³, Laura Herbelin³, and Richard Barohn³

¹Department of Biostatistics, The University of Kansas Medical Center, Kansas City, KS 66160

²Berry Consultants, 4301 Westbank Drive, Suite 140, Bldg B, Austin, TX 78746

³Department of Neurology, The University of Kansas Medical Center, Kansas City, KS 66160

Abstract

The time is right for the use of Bayesian Adaptive Designs (BAD) in comparative effectiveness trials. For example, PCORI has joined the FDA and NIH in adopting policies/guidelines encouraging their use. There are multiple aspects to BAD that need to be considered when designing a comparative effectiveness design. First, the adaptation rules can determine the expected size of the trial. Second, a utility function can be used to combine extremely important co-endpoints (e.g. efficacy and tolerability), and is a valuable tool for incorporating clinical expertise and potentially patient preference. Third, accrual rate is also very, very important. Specifically, there is a juxtaposition related to accrual and BAD. If accrual rate is too fast we never gain efficient information for adapting. If accrual rate is too slow we never finish the clinical trial. We propose methodology for finding the “sweet spot” for BAD that addresses these as design parameters. We demonstrate the methodology on a comparative effectiveness BAD of pharmaceutical agents in cryptogenic sensory polyneuropathy (CSPN). The study has five arms with two endpoints that are combined with a utility function. The accrual rate is assumed to stem from multiple sites. We perform simulations from which the composite accrual rates across sites results in various piecewise Poisson distributions as parameter inputs. We balance both average number of patients needed and average length of time to finish the study.

Keywords

comparative effectiveness; Bayesian adaptive design; accrual; Group sequential monitoring; Patient Centered Outcomes

1. INTRODUCTION

Bayesian Adaptive Designs (BAD) are no longer “a dream for statisticians only [1].” BAD clinical trials have been published not only in biostatistics journals but also clinical

[†]Address for correspondence: The University of Kansas Medical Center, School of Medicine, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, Phone: 913-588-1603, Fax: 913-588-0252, bgajewski@kumc.edu.

epidemiology, drug discovery [2], and medical journals [3], including in the form of reviews [4,5,6] and clinical trial reports [1]. This broad acceptance with the property that BAD save time and money and lean towards more ethical studies [7], indicates that the time is right for the use of BAD in comparative effectiveness clinical trials. In fact, the Patient Centered Outcomes Research Institute (PCORI), a leader in comparative effectiveness research, has adopted policies/guidelines encouraging their use [8]. A recent example is a trial for treating epileptic seizures that uses comparative effectiveness research and BAD [9].

When applying for funding through the PCORI comparative effectiveness grant request for proposals, researchers are asked to explain why a particular study design and statistical methods were chosen. We recently proposed a BAD to achieve an overall objective to determine which pharmaceutical therapy is most effective and causes the fewest side effects in cryptogenic sensory polyneuropathy (CSPN). We proposed a comparative effectiveness study using BAD. We built the BAD with efficiency in mind--the employment of adaptive randomization allows for substantially smaller sample sizes and better conclusions about what treatments are the most effective. As we detail in this paper, we conducted extensive trial simulations comparing different designs measuring the resources (time and patients) and the ability to draw important conclusions about relative efficacy and selected the proposed design as the most effective and efficient. What we found in the process of these simulations and the development of the trial is the importance of three key trial aspects: (1) the Bayesian adaptive design parameters, (2) the utility function for weighing endpoints, and (3) the patient accrual rate. These three developmental parameters are vital for building adaptive, cost-effective comparative effectiveness designs.

The BAD is discussed extensively [7], focusing on Phase I-III clinical trials, dose finding studies, and the assessment of safety and efficacy in the presence of historical prior information. In many cases these studies have a functional form that is unique to classical pharmaceutical clinical trials (e.g. control group or dose). A different challenge in comparative effectiveness trials is the fact that there is typically no control group since we are investigating the relative effectiveness of active treatments. Nor is there typically a dose structure to the problem, hence the treatments are not ordered in any explicit way. We discuss the unique framework of BAD under this setting.

One of the challenges we face in our BAD is that we wish to combine two co-endpoints: efficacy and tolerability. One extremely useful and intuitive approach is to combine the endpoints using a utility function [10]. Many papers have appeared that present ways of incorporating utilities into clinical trial design. For example, Stallard, Thall, and Whitehead [11] consider utility functions that combine multiple outcomes, albeit in single-arm studies. Thall and Cook [12] discuss a utility function that combines efficacy and toxicity. Rossell et al. [13] consider the situation where one is screening a large number of drugs, adaptively deciding which drugs meet pre-specified criteria of activity and which do not. Incorporation of a utility function into a clinical trial's design occurs rarely in comparative effectiveness research so in this paper we demonstrate how we elicited utility functions from the clinical team. We do so in the context of our CSPN example. This extends the Berry et al. paper [10] since that step-by-step procedure was not detailed until now.

The promise of an adaptive design is that important things are learned about the treatments that were not known before the trial started. This learning depends on getting good empirical information from the accrual of the trial. There are two types of accrual: accrual rate or accrual duration. The accrual rate is the number of patients entered to the trial during a fixed time. The accrual duration is the time it takes to enter all the patients. The accrual rate is an important aspect of a BAD, as it can dictate the speed of learning. Modeling patient accrual has recently been studied extensively e.g. [14, 15, 16, & 17]. There are two common themes in the work cited above. First, they use some form of exponential waiting times or Poisson model for estimating accrual rate. Second, their primary aim is to gain either prediction (with intervals) of accrual duration, either as time to finish a study or sample size given a fixed timeframe. The work of Gajewski, Simon, and Carlson [15] is based on an exponential waiting model. An equivalent model is based on a Poisson counting process [14]. A more complex accrual model [16] is not always necessary [18]. One model [17] allows for event appearance (e.g., recurrence, death). This is important for trials that recruit for a time interval and then stop recruitment and wait until a specific number of events occur.

BAD modeling has been applied to a plethora of types of trials and a variety of endpoints [7] that might be efficacy, safety, or a combination of efficacy and safety [10]. The importance of accrual rate in BAD can be emphasized through logical thought. If the accrual rate is too fast we never gain information for adapting. If accrual rate is slow we likely will never finish the clinical trial (or will take too long). We propose methodology for finding the “sweet spot” for accrual rate in relation to BAD and demonstrate the methodology on our comparative effectiveness BAD design.

In Section 2 we present two analytic examples and show how simulation quickly becomes the tool of choice because of its flexibility. In Section 3 we provide details about the comparative effectiveness trial including endpoints, models, and the simulations. We perform multiple simulations in which the composite accrual rates result in various piecewise Poisson distributions, as parameter inputs. We find the “sweet spot” balancing both the average number of patients needed and average length of time to finish the study, and the cost of the study. In Sections 4 and 5 we present results and discussion.

2. BASIC ANALYTIC EXAMPLES

The next two examples focus on the BAD for the simple cases and focuses on the costs involved with the accrual duration schedule. For now we only discuss one endpoint and use the combination of efficacy and tolerability later.

Example 1

Consider an immediately measured endpoint for the CSPN pharmaceutical therapy study, with one arm. Suppose we have two accrual duration periods of time lengths T_1 and T_2 , each respectively n_1 and n_2 patients ($n=n_1+n_2=85$). Patient i shows efficacy ($E_i=1$) or not ($E_i=0$). We assume $E_i \sim \text{Bern}(\theta)$, where θ is unknown, and γ is a pre-specified threshold for declaring success. So we decide the drug is successful if $P(\theta > \gamma | S_E) > \gamma$, where $S_E = \sum E_i$. We stop the trial *early at period 1* if $P(\theta > \gamma | S_E) > \gamma$, otherwise we move to *period 2*. The operating characteristics of this trial can be calculated by assuming the true underlying sampling

distribution is $E_i|\theta_0 \sim \text{Bern}(\theta_0)$, where θ_0 is the true efficacy rate. The role of θ_0 is to provide the necessary parameter for defining the virtual observed data for calculating the trial design's operating characteristics. The role of θ is to provide a distribution for driving the decision making in the trial and is informed at first by a prior and updated with the observed data. With a uniform prior on θ the probability of stopping the trial early, is

$$P_1 = \sum_{S_E=0}^{n_1} \left\{ I \left(\int_{\Delta} \frac{\Gamma(n_1+2)}{\Gamma(S_E+1)\Gamma(n_1-S_E+1)} \theta^{S_E} (1-\theta)^{n_1-S_E} d\theta > \gamma \right) \binom{n_1}{S_E} \theta_0^{S_E} (1-\theta_0)^{n_1-S_E} \right\}$$

, where $I(x>y)$ is 1 if $x>y$ and 0 otherwise. Thus the expected time (T) of the trial is $E(T) = P_1 T_1 + (1-P_1)(T_1+T_2)$ and the expected sample size (N) of the trial is $E(N) = P_1 n_1 + (1-P_1)85$.

Let $\theta_0=0.45$, $\gamma=0.3$, $T_1=T_2=28$ days, and $\gamma=0.9$. We then inspect various allocation of the total sample size ($n=85$) to the first period, $n_1=10, 30, 50$, & 70 . The probability of stopping early (P_1) linearly increases from 0.4956 for $n_1=10$ to 0.9262 for $n_1=70$ resulting in a linear *decrease* for expected time and a quadratic *increase* for expected sample size (Figure 1). The more resources we place in period 1 the larger the study, but will finish in a shorter time because the study will have higher power to stop earlier, by T_1 , rather than stopping later at time T_1+T_2 . We would like to balance the cost between the size and duration of the study. For illustrative purposes, let the per participant cost be 1/patient and the cost of time be 1/day, after combining the linear equations we get $E(\text{Cost}) = 0.0077n_1^2 - 0.4331n_1 + 93.5240$. Taking the derivative of the expected cost with respect to n_1 , setting to zero and solving we get the optimal accrual rate to be $n_1 \hat{=} 28.1$ in the first 28 days.

Example 2

Next consider the most trivial comparative effectiveness trial, two arms. The evaluation would then be the posterior probability that one drug (rate θ_1) is better than the other (rate θ_2). We stop early for efficacy if the posterior probability that one drug is better than the other is high: $P(\theta_1 > \theta_2 | S_{E1}, S_{E2}) > \gamma$ or if $P(\theta_1 < \theta_2 | S_{E1}, S_{E2}) > \gamma$. For this example, under Beta priors, a closed posterior distribution is available [19]. Using a double sum across S_{E1} and S_{E2} provides closed formulas for the expected time of the trial and the expected sample size ($E(T)$ and $E(N)$) as done in Example 1. However, in the full comparative effectiveness study we wish to have five arms and two endpoints, rather than one. Further, the accrual rate patterns tend to be random and not fixed, and their patterns tend to be complicated by the fact that sites enter into a study in a staggered fashion. All of these uncertainties together quickly complicate things. Therefore, as advocated in the past[7], we utilize simulations of the clinical trials to gain properties of the general effect of accrual rate on the expected duration and sizes of studies utilizing adaptive randomization and stopping rules.

3. METHODS

3.1. Details of Comparative Effectiveness Trial

Peripheral Neuropathy affects over 20 million Americans. Some are secondary to readily identifiable causes, for example, diabetic sensory polyneuropathy is the most common cause in the United States. Once known causes are excluded, approximately 25% remain

idiopathic and are referred to as cryptogenic sensory polyneuropathy (CSPN)[20]. Currently there have been no prospective trials performed in the large CSPN group.

The primary aim of this study is to determine which drug is most effective in producing pain relief that is tolerable. We will perform a comparative effectiveness adaptive design prospective randomized study in those who do not have diabetes and in whom no other cause for neuropathy has been found. Five unnamed drugs (for this paper) are proposed.

Subjects will be randomized to one of five treatment arms with a maximum number of subjects $n_{max} = 600$. Using a Bayesian Adaptive Design (BAD), at each interim analysis a decision will be made to either continue enrolling subjects or to stop the trial for unequivocal results. If patient enrollment continues the randomization structure will also be updated. Two endpoints are used to drive the adaptive randomization and stopping criteria. The first endpoint is a responder definition of efficacy (if the patient drops visual analog scale (VAS) pain score by 50% in 12 weeks then the subject is deemed a responder). The second endpoint is an indicator of whether a subject discontinues treatment due to an adverse event. After several patients have been randomized, all current data will be analyzed and an updated randomization schedule will be created. We create an allocation probability for each arm proportional to the probability the arm is the maximally therapeutic arm. This adaptation is repeated every several weeks until the trial is stopped. The success stopping criteria is if the probability of the maximally effective arm being identified (measured by the combined utility of the endpoints) is larger than a pre-specified value. After all enrolled patients are evaluated on their co-endpoints, we deem an arm the best if it has the highest probability of being the maximally effective arm.

3.2. Combining Endpoints

We propose combining the two endpoints using a utility function framework, which is discussed from a BAD point of view in a previous paper [10], although the details of the utility function development are left out. Therefore, we detail the building of a utility function here.

Suppose Drug B has better efficacy than Drug A but has a higher drug discontinue rate due to tolerability, than Drug A. What would that “discontinue rate” have to be in order for Drug B to be clinically the same as Drug A? The first two authors created an elicitation tool that looks at various comparisons (seven of them). Then we asked the clinical experts on this paper to provide the point of equivalency (see Table 1).

Specifically, we create two component utility functions. The first for the efficacy endpoint and the second for the tolerability endpoint, these components are then summed to form a joint utility measure. We assume that E is the rate of response and Q is the rate of discontinuation due to an adverse event for an arm. We use a linear component utility function for efficacy reflecting a utility of 1 for 100% efficacy and utility of 0 for 0% efficacy. For the response rate we use a linear utility of parameter b and add utility of quit rate. Then we sum these to form a joint utility of the form $U(E, Q) = bE + (1 - Q)$, and estimate b from the expert data in Table 1. Labeling the s^{th} scenario and Drugs “A” and “B”

respectively and taking the average of calculated ratios our estimate of b is

$$\sum_{s=1}^7 (Q_{As} - Q_{Bs}) / (E_{As} - E_{Bs}) / 7 \sim 0.75.$$

From the estimate of b , we plotted the points that were considered "equal" (Figure 2); they are common solid lines in the graph and plotted lines (dashed) of constant utility (4 points efficacy = 3 points drop rate). We then presented the results with implied preferences to the clinical team. For example, we did notice a slight lack-of-fit for very small drop rates (lowest solid line), but the utility function was accepted by the clinical team. Therefore the final utility function is $U(E, Q) = 0.75E + (1 - Q)$.

3.3. Statistical details

An overview of the design:

- Accrual rate: this will vary and be optimized;
- Minimum sample size: 200;
- Maximum sample size: 600;
- Stopping/evaluation rule: posterior probability an arm has the maximum utility is greater than 0.90;
- Number of comparative arms: 5;
- New randomization schedule: every four weeks after 100 patients are randomized; and
- Distribution of endpoints: independent Binomial distributions.

We will provide details of each of these design parameters as well as a sensitivity analysis showing the sensitivity to these assumptions after optimizing the accrual rate.

3.3.1. Virtual subject response for five arms—For the purposes of this investigation we look at several virtual response assumptions to investigate the operating characteristics of our trial design. We label θ_{0j} the true efficacy rate for the j^{th} drug where $j=1,2,3,4,5$. Similarly, θ_{0j} is the true discontinuation rate for the j^{th} drug. The sampling distributions for the two endpoints are, respectively, $S_{EjT}|n_{jT} \sim \text{Bino}(n_{jT}, \theta_{0j})$ and $S_{QjT}|n_{jT} \sim \text{Bino}(n_{jT}, \theta_{0j})$, and independent, where n_{jT} represents the cumulative number of patients randomized to the j^{th} arm at week T (we could extend the methodology for a correlation between these endpoints – i.e. side effects and efficacy - but we do not do that here). The total number of patients

accrued at time T is then $N_T = \sum_{j=1}^5 n_{jT}$. Conditional on a time T , N_T is random based on the accrual rate patterns which we model below. For the purposes of this study we focus on two scenarios for treatment arm effects. For the first (alternative scenario, H1) we assume that the true probability of efficacy responses are $\theta_0^e = [.3, .3, .3, .4, .5]$ and the probability of discontinuation are $\theta_0^q = [.3, .3, .3, .25, .15]$. We modeled many other patient profiles but these are not reported here. This virtual response represents a case where there is a clear winner with a distinguished second. The best drug has the highest efficacy and the lowest

discontinuation rate. We also consider a scenario where all the drugs are equivalent (null scenario, H0), so that $\theta_0^e = [.3, .3, .3, .3, .3]$ and $\theta_0^q = [.2, .2, .2, .2, .2]$.

3.3.2 Accrual rate patterns—We assume that the distribution of the accrual rate patterns follows a Poisson distribution with a mean number of accrued patients per week, call it Λ_T . The distribution is $\{N_T - N_{T-1}\} | \Lambda_T \sim \text{Poisson}(\Lambda_T)$, where $T=1, 2, 3, \dots$, and $N_0=0$. The patterns of Λ_T depend on two factors: (1) the number of sites actively enrolling patients into the study and (2) how fast the sites can enroll, which we assume is a constant $\lambda_0/2$ for each. For example, in our case we have 20 sites, with two sites activating every two weeks, thus the mean of the Poisson accrual rate parameter is:

$$\Lambda_T = \begin{cases} \lambda_0, & 0 \leq T < 2 \\ 2\lambda_0, & 2 \leq T < 4 \\ 3\lambda_0, & 4 \leq T < 6 \\ \vdots & \\ 10\lambda_0, & 20 \leq T \end{cases}$$

This assumption says that once a site is recruited they will continue enrolling at the same rate throughout the study. Notice that when time is at 20 weeks, $T=20$, the accrual rate is steady at $\Lambda_{20}=10 \lambda_0$.

3.3.3. Design—The likelihood for the two endpoints are respectively $S_{EjT} | n_{jT} \sim \text{Bino}(n_{jT}, \theta_j^e)$ and $S_{QjT} | n_{jT} \sim \text{Bino}(n_{jT}, \theta_j^q)$. Further, we provide vague priors for each arm, $\text{logit}(\theta_j^e) \sim N(0, 100^2)$ and $\text{logit}(\theta_j^q) \sim N(0, 100^2)$. At each interim we collect the data on all subjects who have been followed for 12 weeks and determine the number of responders and the number of subjects who quit under each treatment (S_{EjT} and S_{QjT}). Given S_{EjT} and S_{QjT} , we obtain the posterior distributions, of $\{\theta_j^e | S_{EjT}\}$ and $\{\theta_j^q | S_{QjT}\}$ respectively, using Markov Chain Monte Carlo (MCMC). We then use the posterior probabilities under each arm to determine if we should stop the trial early for success. Furthermore, if we have not shown sufficient evidence to stop early we use the posterior probabilities to adaptive randomize more patients to the more promising arms.

Our predefined stopping criteria for determining success is restricted to be when at least 200 subjects are randomized. Specifically, we will stop the trial if the posterior probability an arm has the maximum utility is greater than 0.90. The ‘strength of evidence’ of 0.9 was chosen in order to calibrate the Type I error to an acceptable level which was between 5 and 10% depending on how many interim analyses were conducted. The utility for an arm is $U_{jT} = 0.75\theta_j^e + (1 - \theta_j^q)$. The maximum utility arm is defined as $j_{\max} = \arg \max(U_{1T}, U_{2T}, U_{3T}, U_{4T}, U_{5T})$. We stop the trial if there is an arm a that satisfies $\Pr(a = j_{\max}) > 0.9$.

If the trial continues, the next round of patients are randomized using updated randomization probabilities. The probability for each arm is proportional to

$$V_j^* = \sqrt{\frac{\Pr(j = j_{\max, T}) \text{Var}(U_{jT})}{n_{jT} + 1}}. \text{ The approach balances the randomization based on the}$$

posterior probability that the arm is the maximum utility arm and the expected change in variance (a proxy for information gained) by allocating to the arm. There are many choices of the formula of the randomization probability, for example one choice is proportional to only the posterior probability that the arm is the maximum utility arm (i.e. $\Pr(j=j_{max}, T)$). However, the information formula (V_j^*) randomizes more patients to the arm with higher standard error if all posterior probabilities that the arm is the maximum are the same. Strict allocation based on probability only does not have this advantage.

3.3.4. Simulation Algorithm—In order to evaluate the design and measure the effects of accrual rate we simulate the design. We use the simulation results to measure the time of the trial and the efficacy of the trial in order to find the accrual rate “sweet spot.” Repeating 1000 times we simulate the number of patients that have had an efficacious outcome as well as the number that have discontinued the treatment using the 12-week endpoint. Depending on the accrual rate assumptions, the time to enroll subjects per group would vary. We label that time point T . At the time of the interim analysis there are patients in the trial without reaching the 12-week endpoint. Although these patients are not included in calculating the posterior probability that each arm is the highest utility arm they do count towards the subject count in determining the trial size. Then we repeat the methods outlined above using the different BAD parameters, which means we can stop early, change randomization, etc. We repeat the approach for different accrual rate patterns and investigate the trial size and duration for each. We call this optimization search and algorithm the sweet spot algorithm (SSA).

The choice of 1000 simulations in the SSA algorithm is because the maximum 95% margin of error is $1.96 \sqrt{.5(.5)/1000} < 0.05$. However, under a Type I error of .05 or power of .95 the margin of error is much smaller, $\sqrt{.05*.95/1000} = 0.01$. This particular algorithm can be implemented in several different software platforms, for example, R, SAS, STATA, Matlab, Java, C+, etc. We implemented it in a software package called FACTS™ (Fixed and Adaptive Clinical Trials Simulator) [10]. FACTS™ is a simulation platform allowing the simulation of Bayesian and frequentist adaptive designs. We believe it is desirable to make such code freely available. Therefore we have constructed a version of the algorithm in an R program shown in the Appendix section.

3.3.5. BAD versus More Frequent Designs—We compare the key operating characteristics for the BAD with other designs to show that BAD is better in terms of requiring less number of patients, less time, better decision, and better treatment outcomes. First, we perform the BAD procedure with no response adaptive randomization (RAR), so in this design the first stage of 200 patients is equally randomized among the five treatment arms, then, a group sequential design can be applied with interim monitoring and early stopping rules just as in the full BAD. The maximum sample size is 600. Second, we perform a design with no adaptive randomization and fixed at sample size of 600. The gain of BAD is compared with both designs: dropping RAR and fixed. All designs here were calibrated to Type I error of 6% by adjusting their stopping rules (i.e. posterior probability an arm has the maximum utility).

We also construct a more frequent design with *group sequential monitoring* with multiple arms, see a text on the subject by Jennison and Turnbull [21, chapter 16]. We investigate a trial that has two stages ($K=2$), with the first stage of 200 patients equally randomized among the 5 treatment arms, then, we apply a group sequential design with interim monitoring and early stopping rules. The maximum sample size was still 600. Using the Jennison and Turnbull notation we also then compare the gain of BAD to this group sequential design. Let $k=1, 2$ be the two stages. We decide a treatment is better than another treatment if the pairwise Z-statistic $> c_k$, where $c_k = 1.2816 \sqrt{2/k}$. In this trial design we stop only if there is a treatment better than all other treatments, otherwise we continue to the maximum sample size and perform the next pairwise calculation. The 1.2816 is the 10%-tile for the standard normal. This “liberal” cut-point results in a trial having a Type I error of 6% because of the strict rule of a treatment having to be better than all others in order to be the best. We repeated this design for $K=11$ stages ($k=1,2,3,\dots,11$) to better match the average number of interim analyses for the BAD and calibrated the c_k to achieve a Type I error of 6%.

3.3.6. Other Key Operating Characteristics—The ‘sweet spot accrual rate’ (SSA) algorithm is limited by focusing on size and duration. For example, we find the optimal solution to maximize the resource utilization but we do not include other important criteria in this algorithm (such as efficacy). While optimizing the resource utilization is a desirable goal, other very important criteria such as the ability to make an accurate inference on the treatment efficacy and to yield the highest overall success rate by better aligning patients to effective treatments need to be investigated. Therefore, we also investigate key operating characteristics number (and percentage) of patients treated with the best treatment, number of patients having adverse events (AEs), and the treatment response rate, and we look at how these vary by accrual rate.

3.3.7. Sensitivity Analyses—We will conduct a sensitivity analysis by varying accrual rate. If the design’s accrual rate deviates from the ‘sweet spot’, the design may suffer in different ways. We will investigate how much this varies if the accrual rate goes slower and faster than optimal accrual rate. A second sensitivity analysis takes the basic design and systematically changes several of the design parameters including minimum and maximum sample sizes, stopping rule, number of arms, accrual rate ramp up, frequency of interim analyses, and the binomial independent assumption for the endpoints. We investigate how these changes impact the success rate (power), expected size, and duration.

4. RESULTS

For the Peripheral Neuropathy comparative effectiveness trial we perform two sets of simulations based on the two scenarios (described in Section 3.3.1). Further, we vary the accrual rate as $\Lambda_{20} = [1,2,4,6,8,10]$, which is the mean weekly accrual rate after 20 weeks into the study (all sites up and enrolling).

To focus on one scenario, consider the case where the accrual rate is on average 8 patients per week and the virtual profile with an effect. Figure 3 presents the distribution of the

maximum probability of the best arm being the maximal utility arm, the sample size, and the time in weeks of the trial. Recall that the probability of success is estimated through the proportion of simulations that achieve $P(a > j_{max,T}) > .9$, broken up into early (proportion of trials with total sample size less than N_{max} $N_b's < N_{max}$) and late (proportion of trials with total sample size equal to N_{max} $N_b's = N_{max}$), with the rest considered inconclusive, where N_b is the sample size of the b^{th} iteration of a simulation. In the alternative hypothesis scenario, in the simulations 99% of the trials had early success, 0% late success (trial goes to the maximum sample size of 600), and 1% were inconclusive. The mean sample size of this trial is 281, thus substantially smaller than the maximum of 600. The average length of the trial is 58 weeks.

In the null hypothesis scenario, in the simulations 6.0% of the trials had early success, 0.0% late success (Type I error~5%). The mean sample size of this trial is 585 subjects. The average length of the trial is 96 weeks.

In Figure 4 we contrast the mean sample size and the mean time of the trial in weeks, as a function of the accrual rate. We then fit least squares functions to the mean time and sample size, for allowing analytic solutions for finding the “sweet spot” as defined by a cost function over sample size and time of the trial. The former can be adequately summarized with a linear function and the latter with a power function. We model the cost to administer a patient through the clinical trial (provide the drugs, take the measurements, etc.) as much more than it costs to keep a trial operational for one week. We assume that the cost to operate a clinical trial for one week is 1.29 times that per subject.

Using the formulas in Figure 4 on the time and number of subjects in the trial, the expected total cost is $E(Cost) = 5.3428 \Lambda_{20} + 240.32 + 1.29(247.87 \Lambda_{20}^{-0.693})$, taking the derivative of the expected cost with respect to Λ_{20} , the steady accrual rate after 20 weeks of site recruitment, setting to zero and solving we get the optimal accrual rate to be $\hat{\Lambda}_{20} = [(1.29 * 247.87 * .693) / 5.3428]^{1/1.693} = 9.03$, which means between 8 and 10 patients per week would be an optimal maximization of resources. The average optimal cost using these equations is 358.2. For a BAD without response adaptive randomization (RAR) and a fixed design at 8 patients per week the cost is higher than the BAD cost since they are 439 and 726 respectively. We are not including efficacy in the cost of the trial – there is no reflection of the strength of the trial. We find that across all the accrual rate patterns the proportion of times it is successful (i.e. proportion $P(a = j_{max}) > .9$) is between 0.99 and 1.00. Thus success does not vary much over the scale of accrual rates.

Table 2 compares the Bayesian Adaptive Design (BAD) versus BAD without response adaptive randomization (RAR), versus a fixed design, versus a group sequential model (GSM) on all the parameters described in Figure 4 assuming the accrual rate is at 8 patients per week. The BAD is superior to the fixed design and better than the BAD without RAR since it has lowest sample size, fastest duration, lowest cost, highest power, lowest number of AEs, highest percentage of patients on best arm, highest percentage efficacious. On all measures of operating characteristics the GSM ($K=2$) is poorer than all designs except the Fixed, although GSM ($K=11$) is better as it is closer (in all operating characteristics) to the BAD no RAR.

In Figure 5 we show the tradeoff between expected time and expected sample size for the null scenario (all equal). In this case the expected cost is completely determined by the accrual rate and the faster we can enroll, the faster we will get to the maximum sample size of 600 and determine that we cannot find the best drug (which is correct). But given that we don't know the true answer until we start the study, an accrual rate of $\Lambda_{20}=8$ works fine, we would not want to go below 4 as we really start to see a dramatic increase in overall costs.

Back to the alternative scenario, to investigate the sensitivity we look at some of the key operating characteristics if the accrual rate is slower than optimal ($\Lambda_{20}=6$) versus faster than optimal ($\Lambda_{20}=10$). Relative to optimal the expected cost for faster accrual rate is higher and for slower accrual rate is higher. From that point of view favors faster accrual rate, among these two. However, when looking at expected number of adverse events the faster accrual rate has higher adverse events and the slower has lower adverse events. Similar gains in slower than optimal accrual rate occurs for number of patients randomized to the best arm. The slight gain in relative cost may justify the relative big gain in more patients receiving a better treatment during the trial.

The results of the second sensitivity analysis are shown in Table 3. Obviously dropping the accrual rate drops size and increases duration. Decreasing minimum sample size decreases size and duration. Decreasing the maximum size obviously decreases success rate, size, and duration. Increasing the stopping rule lowers success rate but increases size and duration. Dropping the bad arm (first one) from consideration improves all three. Shortening the ramp up of accrual rate has little impact. Lengthening the gap between interim analyses slightly lowers the success rate, increases the sample size, but shortens the duration. A large incremental impact is going from an independent model to a dependent model since it lowers the success rate and makes the trial bigger and longer.

5. DISCUSSION

Typically what is done in a classical framework is to determine the sample size of a study based on an initial guess of the effect that the treatment will have. A major limitation of this approach is that the information gained (i.e. getting the right answer) depends on the quality of our guess. This limitation is partially addressed here because we “flip” the approach to clinical trials design. The effect we learn is fixed while the sample size used for a particular clinical trial can vary depending on the data. This is a very attractive aspect of the BAD approach in that we have a much better chance of getting the right answer as the sample size is a proxy for the scientific knowledge. This means that the cost-benefit of the trial differs from the fixed approach.

The SSA algorithm can be used in different ways for the Peripheral Neuropathy comparative effectiveness trial. We can vary the number arms (say 2, 6, or more), we can use one endpoint instead of two, change the maximum sample size from 600 to higher. We can change the rules to stop the trial in different ways, say with a minimal efficacy or a futility stopping rule. For example, we have a stopping rule for the whole study that is based on an arm satisfying the probability based criterion of being associated with the maximum utility. It might be, however, that two or more arms have combined probability greater than 90% of

associated utility being the maximum (i.e., if there is a tie for the maximum utility). We might not want to stop then. Another way of stating this is if all 5 utilities in our example were equal, then $\Pr(a=j_{max}) = 20\%$ and we may want to stop.

Another way the SSA algorithm can be used in different ways is the accrual rate pattern could change. For example, we could assume more clinical sites are in the trial and the spacing of their joining the trial could vary. We could learn about the accrual rate patterns and then adapt how we recruit sites with that new information. However, the current study that we performed here has properties to it that are very similarly found in other studies [17] where the accrual rate starts off small at the beginning of the study before getting larger and larger and achieving its steady state where the average is essentially flat. However, to do this kind of adaptive accrual rate one would need to incorporate some of the accrual prediction models in the literature [14, 15, 16, & 17]) so that one may have updates of what the true accrual rate patterns are in real time. For example, the investigative team may have under projected what the true accrual rate is, which might happen considering that have a documented tendency to overpromise and under deliver on patient accrual rate they plan for their research studies[22].

The SSA algorithm can also be extended to a range of different Bayesian adaptive clinical trials designs. For example, we could extend the model to include time to event endpoints, which would bring in another random component to the balance. Rather than binary endpoints we could include models that are ordinal or continuous or a mix of the two endpoints. For example, rather than an efficacy dichotomous endpoint, we could treat the VAS score as continuous. Along with the binary discontinuation endpoint we would have co-primary endpoints where one is continuous and the other is binary. The methodology would be especially valuable in enrichment designs where subsets of the population may be discontinued due to poor effect size – the remaining “enriched” population provides the appropriate population to study. These studies would have the possibility that the design changes enrollment and thus changes the inherent accrual rate.

One point that can be straightforwardly extended is the fact that each arm might have a different cost associated with the drug. Thus the sweet spot for the balance of sample size might not be the same for all drugs but can be incorporated into the SSA algorithm which can allow the design team to balance potential efficacy knowledge with the overall cost of the trial. Another opportunity to minimize costs would be to design the trial to have less arms in the beginning (say three). Then we can drop arms that are doing poorly and add new ones.

In our main analysis we present the marginal probability of response and the marginal probability of discontinuation as independent endpoints. In some instances this may not be a good approximation because the treatment’s efficacy and toxicity are correlated. In the case of non-independence it might be better to model the four categories via a multinomial distribution and then provide the posterior distribution of the marginal distributions. Or one could define all patients who drop from the study to be non-efficacious resulting in a trinomial distribution. In either of these latter cases we would make the straightforward

extension in the algorithm to include these models. We see in the sensitivity analysis this is a very important assumption.

While not immediately generalizable, the principles we provide for finding the optimal accrual rate pattern via the SSA algorithm can be used in all of the different types of studies and these principles should be considered when designing Bayesian adaptive clinical trials, specifically the adaptation rule, utility function, and the accrual rate should all be parameters considered for optimizing the design of comparative effectiveness research.

ACKNOWLEDGEMENTS

This work was supported by an Institutional Clinical and Translational Science Award, NIH/NCATS Grant Number UL1TR000001 and the PAIN-CONTRoLS Study #CER-1306-02496. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or PCORI.

REFERENCES

1. Chevret S. Bayesian adaptive clinical trials: a dream for statisticians only? *Statistics in Medicine*. 2012; 31:1002–1013. [PubMed: 21905067]
2. Berry DA. A guide to drug discovery: Bayesian clinical trials (2006) *Nature Reviews Drug Discovery*. 2006; 5(1):27–36.
3. Berry DA. Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology*. 2012; 9(4):199–207.
4. Lee JJ, Chu CT. Bayesian clinical trials in action. *Statistics in Medicine*. 2012; 31(25):2955–2972. [PubMed: 22711340]
5. Bonangelino P, Irony T, Liang S, Li X, Mukhi V, Ruan S, Xu Y, Yang X, Wang C. Bayesian approaches in medical device clinical trials: A discussion with examples in the regulatory setting. *Journal of Biopharmaceutical Statistics*. 2011; 21(5):938–953. [PubMed: 21830924]
6. Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clinical Trials*. 2009; 6(3):205–216. [PubMed: 19528130]
7. Berry, SM.; Carlin, BP.; Lee, JJ.; Muller, P. *Bayesian Adaptive Methods for Clinical Trials*. New York: CRC Press; 2011.
8. Patient Centered Outcomes Research Institute (PCORI). [accessed August 13, 2014] PCORI Methodology Standards. 2012. <http://www.pcori.org/assets/PCORI-Methodology-Standards.pdf>
9. Connor JT, Elm JJ, Broglio KR. ADAPT-IT Investigators. Bayesian Adaptive Trials offer Advantages in Comparative Effectiveness Trials: An Example in Status Epilepticus. *Journal of Clinical Epidemiology*. 2013; 66(8S):S130–S137. [PubMed: 23849147]
10. Berry SM, Spinelli W, Littman GS, Liang JZ, Fardipour P, Berry DA, Lewis RL, Krams M. A Bayesian dose-finding trial with adaptive dose expansion to flexibly assess efficacy and safety of an investigational drug. *Clinical Trials*. 2010; 7:121–135. [PubMed: 20338905]
11. Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*. 1999; 55:971–977. [PubMed: 11315037]
12. Thall PF, Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*. 2004; 60:684–693. [PubMed: 15339291]
13. Rossell D, Müller P, Rosner GL. Screening designs for drug development. *Biostatistics*. 2007; 8:595–608. [PubMed: 17030589]
14. Anisimov VV, Fedorov VV. Modelling prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine*. 2007; 26(27):4958–4975. [PubMed: 17639505]
15. Gajewski BJ, Simon SD, Carlson SE. Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in Medicine*. 2008; 27(13):2328–2340. [PubMed: 17979152]
16. Zhang X, Long Q. Stochastic modeling and prediction for accrual in clinical Trials. *Statistics in Medicine*. 2010; 29:649–658. [PubMed: 20082363]

17. Anisimov VV. Effects of unstratified and centre-stratified randomization in multi-centre clinical trials. *Pharmaceut. Statist.* 2011; 10:50–59.
18. Gajewski BJ, Simon SD, Carlson SE. On the existence of constant accrual rates in clinical trials and direction for future research. *International Journal of Statistics and Probability.* 2012; 1(2):43–46.
19. Kawasaki Y, Miyaoka E. A Bayesian inference of $P(\pi_1 > \pi_2)$ for two proportions. *Journal of Biopharmaceutical statistics.* 2012; 22:425–437. [PubMed: 22416833]
20. Pasnoor M, Dimachkie MM, Barohn RJ. Cryptogenic sensory polyneuropathy. *Neurologic Clinics.* 2013; 31:463–476. [PubMed: 23642719]
21. Jennison, C.; Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman & Hall/CRS; 2000.
22. Breau RH, Carnat TA, Gaboury I. Inadequate statistical power of negative clinical trials in urological literature. *J. Urol.* 2006; 176(1):263–266. [PubMed: 16753417]

APPENDIX: MULTINOMIAL APPROACH AND R CODE

The two endpoints are respectively S_{EjT} for efficacy and S_{QjT} for quitting, and with not efficacy, Sn_{EjT} is modeled, the dependent case, as a multinomial distribution $(S_{QjT}, S_{EjT}, Sn_{EjT}) \sim \text{MultNomial}(\theta_j^q, \theta_j^e, \theta_j^{ne})$. In addition, we provide uniform priors, $(\theta_j^q, \theta_j^e, \theta_j^{ne}) \sim \text{MultNomial}(1/3, 1/3, 1/3)$. Using the endpoint data and the prior probabilities, we then use Markov Chain Monte Carlo computations to obtain the Bayesian posterior distributions of $\{\theta_j^e | S_{EjT}\}$ and $\{\theta_j^q | S_{QjT}\}$ respectively for each endpoint (i.e., efficacy and quitting), and clearly the endpoints are no longer independent.

```
#
*****
*****

# Copyright 2014 Berry Consultants, LLC
# Berry Consultants Intellectual Property
# *****
# PROGRAM NAME: PcoriPain.R
# PROJECT NAME: Pcori Pain
# PROGRAM PURPOSE: Functions for Simulating Bayesian Adaptive Design for
# trial with utility function created from safety and eff. endpoints
# PROGRAM AUTHOR: Melanie Quintana
# Gajewski, Quintana, Berry, Pasnoor, Dimachkie, Herbelin, & Barohn (in
# review)
# *****

## This function generates an accrual time (in weeks) from a Poisson process
# for a specified number of patients.#####
accrual.fun = function(Nmax,accrual,ramp){
  ## Nmax: Maximum sample size
  ## accrual: Accrual rate (patients/week) at peak enrollment
  ## ramp: Ramp up period (weeks until peak enrollment)
  ## Vector of average patients enrolled each week
  ppw = c(seq(accrual/ramp, accrual, length=ramp),rep.int(accrual, Nmax))
```



```

weekin = NULL
ptsin = 0 ## Counter for number of patients currently enrolled
for(i in 1:length(ppw)){ ## Loop through the weeks
  temp = cumsum(rexp(Nmax-ptsin, ppw[i])) ## Cumulative time to enroll for all
  remaining patients based on rate in ith
  week
  temp = temp[temp<=1] ## Keep only the times for patients enrolled in ith week
  weekin = c(weekin, (i-1)+temp)
  ptsin = length(weekin)
  if(length(ptsin) >= Nmax){ ## Stop once we have reached Nmax patients
    weekin = weekin[1:Nmax]
    break}
}
return(weekin) ## weekin = vector of enrollment times (week) per patient
}

## Function to sample from Dirichlet
distribution#####
rdir = function(n,alpha){
  ##n: number of vectors to generate from dirichlet distribution
  ##alpha: parameters of dirichlet distribution
  l = length(alpha)
  x = matrix(rgamma(l * n, alpha), ncol = l, byrow = TRUE)
  sm = x %%% rep(1, l)
  return(x/as.vector(sm))}

##Calculate posterior probability drug is the max under Mult. likelihood and
Dir. Prior and corresponding AR#####
PrBest = function(X,alpha=c(1/3,1/3,1/3),iter=1000,droptresh = .05){
  ##X: where each row corresponds to the number of Q, the number of E & NQ,
  and the number of NE & NQ for a given
  treatment
  ##alpha: vector of prior probabilities of Q, E & NQ and NE & NQ
  ##Sample iter values from posterior distribution of p1 (pr(Q)), p2 (pr(E &
  NQ)) and p3 (pr(NE & NQ)) under each treatment
  ##And calculate utility under each sampled vector (p1,p2,p3)
  ## 3/4p2 + (1-p1)?
  util = function(x,alpha,iter){
    p = rdir(iter,x+alpha)
    u = (3/4)*p[,2] + (1-p[,1])
    U = apply(X,1,util,alpha=alpha,iter=iter)
    ##Caluclate the probability that each treatment has best utility
    maxU = apply(U,1,which.max)
    numTreat = dim(X)[1]
    PrMax = apply(as.matrix(c(1:numTreat)),1,function(x){mean(maxU==x)})
    ##Calculate Adaptive randomiztion probabilties

```

```

numSubj = apply(X,1,sum)
AR = sqrt(PrMax*apply(U,2,var)/(numSubj+1))
AR = AR/sum(AR)
##Assign 0 probability to any treatment with AR<=droptresh
AR = ifelse(AR<=droptresh,0,AR)
AR = AR/sum(AR)
##Return list of PrMax and AR
results = list(PrMax,AR)
names(results) = c("PrMax","AR")
return(results)
}

##Function to Simulate one
trial#####
PCORI.Sim = function(theta,Nmax=400,accrual=4.2,ramp=8,successThresh=.
95,dropThresh=.05){
##theta: rows of theta are correspond to (p1,p2,p3) for each treatment
##p1 = pr(Q), p2 = pr(E & NQ), and p3 = pr(NE & NQ)
##Get enrollment time of all 400 patients
Time.Enroll = accrual.fun(Nmax,accrual,ramp)
Time.Data = Time.Enroll + 12
##Time we get final data on each patient
##Burnin data (20 patients on each arm)
numTreat = dim(theta)[1]
Treat = rep(1:numTreat,20)[sample(1:(20*numTreat),(20*numTreat))]
##Treatment assignment for first 80 patients
response = apply(theta[Treat,],1,function(x){sample(c(1:3),1,prob=x)})
##Response/patient Q=1, E|NQ=2, NE|NQ=3
##After 20 patients on each arm start performing interims every 13 weeks
time.interims = seq(Time.Enroll[80],Time.Enroll[Nmax]+12,13)
##Make sure to include Final analysis when all 400 have 12 week data
if(time.interims[length(time.interims)]<(Time.Enroll[Nmax]+12))
{time.interims = c(time.interims,(Time.Enroll[Nmax]+12))}
##Make matrix of results for each interim
interim.results = matrix(NA,nrow=length(time.interims),ncol=5+2*numTreat)
colnames(interim.results) =
c("Interim.Time","StoppedSuccess","N.Enrolled","N.
12Week","Best",paste("PrMax",c(1:numTreat),sep=""),paste("N.",c(1:numTreat),
sep=""))
for(i in 1:length(time.interims)){
time.i = time.interims[i]
##Use data for all patients that have been enrolled for atleast 12 weeks
treat.i = Treat[Time.Data<=time.i]
response.i = response[Time.Data<=time.i]
X = t(apply(as.matrix(1:numTreat),1,function(x){c(sum(treat.i==x &

```

```

response.i==1),sum(treat.i==x
& response.i==2),sum(treat.i==x & response.i==3))}))
##Get posterior prob. each treatment is max and adaptive randomization prob.
results.i = PrBest(X,droptresh=dropThresh)
Best = c(1:numTreat)[results.i$PrMax==max(results.i$PrMax)]
interim.results[i,] =
c(time.i,0,sum(Time.Enroll<=time.i),length(treat.i),which.max(results.i
$PrMax),results.i$PrMax,apply(as.matrix(1:numTreat),1,function
n(x){sum(Treat[Time.Enroll<=time.i]==x)}))
##If we have 100 subjects complete & PrBest> .95 for some treatment stop for
success.
if(sum(Time.Data<=time.i)>=130 & max(results.i$PrMax)>=.95){
if(i< length(time.interims)){interim.results[i,2]=1}
if(i== length(time.interims)){interim.results[i,2]=2}
break
}
##If we have not stopped for success or reached max patients then enroll
patients for the next 13 weeks
##Assign treatment based on AR prob
if(sum(Time.Enroll<=time.i)<Nmax){
N.new = sum(Time.Enroll>time.interims[i] & Time.Enroll<=time.interims[i+1])
Treat.new = sample(c(1:numTreat),N.new,replace=TRUE,prob=results.i$AR)
response.new =
apply(matrix(theta[Treat.new,],ncol=3,nrow=N.new),1,function(x)
{sample(c(1:3),1,prob=x)})
Treat = c(Treat,Treat.new)
response = c(response,response.new)}
}
return(interim.results)}
##Simulate many trials under specified theta and report some operating
characteristics#####
#####
PCORI.MultSims =
function(theta,sims,Nmax=400,accrual=4.2,ramp=8,successThresh=.
95,dropThresh=.05){
numTreat = dim(theta)[1]
Full.Results = matrix(NA,nrow=sims,ncol=5+2*numTreat)
colnames(Full.Results) =
c("Interim.Time","StoppedSuccess","N.Enrolled","N.
12Week","Best",paste("PrMax",c(1:numTreat),sep=""),paste("N.",c(1:numTreat),
sep=""))
for(i in 1:sims){
cat("Processing Trial",i,"\n")
results.i =

```

```

PCORI.Sim(theta,Nmax=Nmax,accrual=accrual,ramp=ramp,successThresh=successthresh,dropThresh=dropThresh)
final.interim.i = results.i[is.na(results.i[,1])==FALSE,]
Full.Results[i,] = final.interim.i[dim(final.interim.i)[1],]
}
results = round(c(mean(Full.Results[,2]!=0),mean(Full.Results[,2]==1),apply(Full.Results[,c(1,3,10:13)],2,mean),
apply(as.matrix(1:numTreat),1,function(x){mean(Full.Results[,5+x]>=.95)}),apply(as.matrix(1:numTreat),1,function(x){mean(Full.Results[,5+x]<=.01)})),3)
names(results) =
c("ProbSuccess","EarlySuccess","MeanDuration","MeanSubj",paste("MeanSubj",c(1:dim(theta)[1]),sep=""),paste("ProbWin",c(1:dim(theta)[1]),sep=""),paste("ProbLose",c(1:dim(theta)[1]),sep=""))
return(results)
}
Full.Results = NULL
theta = matrix(NA,ncol=3,nrow=4)
colnames(theta) = c("Pr(Q)","Pr(Eff & NQ)","Pr(NotEff & NQ)")
rownames(theta) = c("Treat1","Treat2","Treat3","Treat4")
theta[,1] = c(.2,.2,.2,.2)
theta[,2] = c(.3,.3,.3,.3)
theta[,3] = 1 - (theta[,1] + theta[,2])
Full.Results = rbind(Full.Results,PCORI.MultSims(theta,1000))
theta[,1] = c(.3,.25,.15,.3)
theta[,2] = c(.3,.4,.5,.3)
theta[,3] = 1 - (theta[,1] + theta[,2])
Full.Results = rbind(Full.Results,PCORI.MultSims(theta,1000))
Full.Results

```

Figure 1.

As a function of the accrual rate (n_1) – patients enrolled in 28 days -displayed are the expected study sample size ($E(N)$), expected trial duration in days ($E(T)$), and the expected cost ($E(\text{Cost})$).

Figure 2. Trade-off between the efficacy and discontinue combinations as elicited from the clinical team (each solid line represents a pair). The efficacy and tolerability tradeoff was determined by taking the average of ratio of change in discontinue (tolerability) and change in efficacy across scenarios. This resulted in 0.7595 so we decided on 0.75 (4 points efficacy=3 points discontinue).

Figure 3.

The probability of maximum utility (P), sample size (N), duration (T), percentage of patients responding ($\%Efficacy$), and percentage of patients who discontinue ($\%Discontinue$) predictive distributions when the accrual rate is $\Lambda_{20} = 8$ patients/week; with the scenario having a best and second best arm (top) and under the null hypothesis (bottom). The first and fourth plots represent arm 5 which is the best arm for H1 (best and second best scenario).

Figure 4.

For the alternative scenario (best and second best arm) we display the expected sample size ($E(N)$) and the expected duration ($E(T)$). We also display the expected cost ($E(\text{Cost})$). Other plots include expected number ($E(n_5)$) of patients treated with best treatment (5), and number of patients having adverse events ($E(AEs)$). . The $E(\%n_5)$ was approximately 37% and the treatment response rate $E(\%Effic)$ was approximately 40% regardless of accrual rate. Λ_{20} is the mean weekly accrual rate after 20 weeks into the study.

Figure 5.

For the null scenario we display the expected sample size ($E(N)$) and the expected duration ($E(T)$). We also display the expected cost ($E(\text{Cost})$). Other plots include expected number ($E(n_5)$) of patients treated with best treatment (5), and number of patients having adverse events ($E(\text{AEs})$). Λ_{20} is the mean weekly accrual after 20 weeks into the study.

Table 1

This table was presented to clinical experts and they were asked to fill in the underlined values, which are filled-in here, to indicate the value that makes Drug A equal to Drug B for combined efficacy and discontinue rates.

Scenario	Drug A		Drug B	
	Efficacy (E_A)	Discontinue (Q_A)	Efficacy (E_B)	Discontinue (Q_B)
#1	30%	15%	40%	<u>30%</u>
#2	30%	5%	<u>40%</u>	10%
#3	50%	20%	40%	<u>15%</u>
#4	50%	30%	<u>40%</u>	20%
#5	30%	15%	<u>50%</u>	30%
#6	20%	2%	<u>50%</u>	10%
#7	25%	15%	50%	<u>35%</u>

Table 2

Outcomes of operating characteristics for the scenario having a best and second best arm for full Bayesian Adaptive Design (BAD) versus BAD without response adaptive randomization (RAR) versus a fixed design as well as group sequential monitoring (GSM) with $K=2$ stages. We compare average size ($E(N)$), duration ($E(T)$), cost ($E(\text{Cost})$), power, expected number of patients having adverse events ($E(\text{AEs})$), expected percentage of patients treated with best treatment ($E(\%n_5)$), and the expected treatment efficacy rate ($E(\% \text{Efficacy})$). $\%n_5$ is the percentage assigned to the best treatment. The BAD is best by all measures. The accrual rate is 8 patients/week. Note: time and duration are calculated directly from simulations not from smoothed estimates.

Design	Type I Error	E(N)	E(T)	E(Cost)	Power	E(AEs)	E($\%n_5$)	E($\% \text{Efficacy}$)
BAD	6%	281	58	359	99%	66	37%	40%
BAD no RAR	6%	353	67	439	90%	92	20%	36%
Fixed	6%	600	98	726	94%	156	20%	36%
GSM ($K=2$)	6%	488	84	596	89%	127	20%	36%
GSM ($K=11$)	6%	403	72	496	91%	105	20%	36%

Table 3

Sensitivity analysis after changing several of the design parameters investigating its effect on success rates, size, and duration of the trial.

Design	Change	%Early Success	%Late Success	%Total Success	E(N)	E(T) weeks
1	None	99%	0%	99%	281.3	57.5
2	Accrual Rate=from 8 to 4.2	99%	0%	99%	264.0	85.1
3	Nmin=from 200 to 130	99%	0%	99%	224.9	75.9
4	Nmax=from 600 to 400	94%	0%	94%	219.3	74.6
5	Stopping Rule=from .90 to .95	88%	0%	88%	254.8	83.2
6	Arms=from 5 to 4	92%	0%	92%	226.1	76.2
7	Ramp up=from 20 to 8 weeks	91%	0%	91%	232.3	71.6
8	Interim Analyses=from 4 to 13 weeks	89%	1%	90%	253.5	76.7
9	Distribution=from BI to MD	69%	5%	74%	320.4	83.7

BI=Binomial Independent

MD=Multinomial Dependent

Design 1: accrual rate=8 pts/week, Nmin=200, Nmax=600, Stopping Rule =.9, Arms=5, Ramp Up 20 weeks, Interim Analysis every 4 weeks, Distribution=BI