5-9-2011

Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size: Understanding Efficient Types of Adaptation

Gregory P. Levin
University of Washington, glevin11@uw.edu

Sarah C. Emerson
Oregon State University, emersosa@stat.oregonstate.edu

Scott S. Emerson
University of Washington, semerson@u.washington.edu

Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size: Understanding Efficient Types of Adaptation

Gregory P. Levin[1*], Sarah C. Emerson[2], and Scott S. Emerson[1]

[1] Department of Biostatistics, University of Washington,
Seattle, WA 98195, USA
[2] Department of Statistics, Oregon State University,
Corvallis, OR 97331, USA

[*] *email*: glevin11@uw.edu

ABSTRACT

Methods allowing unplanned adaptations to the sample size based on the interim estimate of treatment effect do not base inference on the minimal sufficient statistic and suffer losses in efficiency when compared to group sequential designs [1, 2, 3]. However, when adaptive sampling plans are completely pre-specified at the design stage of the trial, investigators can proceed with frequentist inference based on the minimal sufficient statistic at the analysis stage. In the context of two general settings where different optimality criteria govern the choice of clinical trial design, we quantify the relative costs and benefits of a variety of fixed sample, group sequential, and pre-specified adaptive designs with respect to standard operating characteristics. We find pre-specified symmetric adaptive designs that are "optimal" in the sense that they minimize the expected sample size at the design alternatives. Our results build on others' prior research [1, 4, 5, 6] by demonstrating in realistic settings that simple and easily implemented pre-specified adaptive designs provide only very small efficiency gains over group sequential designs with the same number of analyses.

In addition, we describe optimal rules for modifying the sample size, providing efficient adaptation boundaries on a variety of scales for the interim test statistic for adaptation analyses occurring at several different stages of the trial. These findings provide insight into what are good and bad choices of adaptive sampling plans and suggest that adaptive designs proposed in the literature are often based on inefficient rules for modifying the sample size.

KEY WORDS:

Adaptive designs; Clinical trials; Efficiency; Group sequential tests; Sample size modification; Sufficiency

1 Introduction

Adaptive clinical trial designs have been proposed as a promising new approach that may help improve the drug discovery process. A number of statistical papers have introduced methods to allow unplanned interim modifications to the study design while preserving the type I error rate of the clinical trial [7, 8, 9, 10, 11]. In particular, there is a large body of literature exploring designs with unplanned modifications to the sample size based on interim estimates of the treatment effect. It is unclear, however, if such adaptive designs provide any clear benefits over more standard group sequential and fixed sample approaches. Tsiatis and Mehta [3], and Jennison and Turnbull [1, 2], have demonstrated that methods allowing unplanned adaptations to the sample size do not base inference on the minimal sufficient statistic and come with costs in efficiency when compared to group sequential designs.

It is possible, however, to completely pre-specify adaptive sampling plans at the design stage of the trial, so that investigators can proceed with frequentist inference based on the minimal sufficient statistic at the analysis stage. Such designs are examples of the "sequentially planned decision procedures" proposed by Schmitz [12]. An important reason to critically evaluate the class of pre-specified adaptive designs is the lack of regulatory support, in the setting of adequate and well-controlled phase III effectiveness trials, for designs allowing unplanned modifications to the sample size. The recent FDA draft guidance on adaptive trials discussed the interpretability challenges of approaches allowing unplanned design changes and asserted that "changes in study design occurring after an interim analysis of unblinded study data and that were not prospectively planned are not within the scope of this guidance" [13].

Since group sequential designs are just one subgroup of the more flexible broader class of pre-specified adaptive designs, one would expect that some efficiency gains can be made by incorporating the opportunity for sample size adaptations into the sampling plan. A few recent papers

have derived optimal pre-specified adaptive designs under a Bayesian decision problem framework and then shown that these adaptive designs can attain only minor efficiency gains over alternative group sequential designs [1, 4, 5, 6]. In particular, Jennison and Turnbull concluded that standard group sequential designs are nearly as efficient as any optimized adaptive design. They also noted that "it will be quite a challenge to find simply defined adaptive procedures with such robustly efficient performance" and that they have "observed the sampling rules of optimal adaptive tests to be qualitatively different from rules based on conditional power commonly used in adaptive designs" [4]. It is these two comments that we hope to illuminate clearly and build on in this paper.

There is a need for studies that exactly quantify and discuss the relative costs and benefits of simple and easily implemented pre-specified adaptive designs as compared to alternative designs in realistic settings. This includes settings in which efficiency is the primary concern, and also settings in which other scientific issues govern the choice of clinical trial design. In addition, since adaptive trials are being proposed and carried out in actual clinical research, there is a need for a detailed description of the sampling rules that lead to efficient adaptive designs. It is not clear what are good and bad choices of rules for modifying the sample size on different scales for the interim test statistic, nor is it well understood at what time it is best to perform such an adaptation. We find that many of the adaptive designs proposed in the literature consist of suboptimal modification rules based on poorly understood scales (such as conditional power) and carried out at poorly chosen stages of the trial. For example, one manuscript in the press cites two actual clinical trial settings in which adaptive designs were ultimately chosen [14]. Under the authors' recommended adaptive designs, the "promising region" inside which modifications may occur and the rules for sample size adaptation are based on desired levels of conditional power. The manuscript does not evaluate alternative choices of adaptive rules, so it is unclear if the chosen designs were efficient in terms of standard operating characteristics given the scientific constraints of the two particular design settings.

In general, there is a gap in the adaptive trial literature with respect to the critical evaluation of different types of adaptation rules. There is a need for simple studies that explore a range of adaptation rules, allowing the number of adaptation regions and the timing of the adaptation analysis to vary, while also computing the efficient adaptation boundaries on a variety of different scales for the interim test statistic. By carrying out such a study, our goal is to help trial investigators better understand different types of adaptation rules and what to expect with respect to their impact on standard operating characteristics. We also hope that our findings suggest where it may be best to dedicate future research efforts in the study of adaptive trial designs.

In summary, the goal of this paper is to illustrate the effects of different adaptation rules on standard measures of efficiency when simple and easily implemented adaptive sampling plans are completely pre-specified. We first describe some notation and a general setting for our comparisons. We write out the sampling density for a completely pre-specified adaptive design and show that inference can be based on the minimal sufficient statistic. We then clearly define the optimality criteria governing the choice of randomized clinical trial design in two simple, realistic settings, describe in detail the sampling plan of the optimal adaptive designs, and compare the operating characteristics of these adaptive designs to those of alternative group sequential and fixed sample designs in these scenarios. In the final sections, we discuss the implications of our results and other issues inherent to adaptive designs. All computations were performed using the R package RCTdesign built from the S-Plus module S+SeqTrial [15].

2 Setting and Notation

Consider the following simple setting of a balanced two-sample comparison, which is easily generalized [16]. Potential observations X_{Ai} on treatment A and X_{Bi} on treatment B, for $i = 1, 2, ...,$ are independently distributed, with means μ_A and μ_B, respectively, and common variance σ^2. The

parameter of interest is the difference in mean treatment effects, $\theta = \mu_A - \mu_B$. We assume the variance is known and without loss of generality, let $\sigma^2 = 0.5$. There will be up to J interim analyses conducted with sample sizes $N_1, N_2, N_3, ..., N_J$ accrued on each arm (both J and the $N_j s$ may be random variables). At the jth analysis, let $S_j = \sum_{i=1}^{N_j}(X_{Ai} - X_{Bi})$ denote the partial sum of the first N_j paired observations, and define

$$\hat{\theta}_j = \frac{1}{N_j} S_j = \overline{X}_{A,j} - \overline{X}_{B,j}$$

as the estimate of the treatment effect of interest θ based on the cumulative data available at that time. The normalized Z statistic and upper one-sided fixed sample P-value are transformations of that statistic: $Z_j = \sqrt{N_j}(\hat{\theta}_j - \theta_0)$ and $P_j = 1 - \Phi(Z_j)$. We represent any random variable (e.g. N_j) with an upper-case letter and any realized value of a random variable (e.g. $N_j = n_j$) or fixed quantity with a lower-case letter. We additionally use a * to denote incremental data. We define N_j^* as the sample size accrued between the $(j-1)$th and jth analyses, with $N_0 = 0$ and $N_j^* = N_j - N_{j-1}$. Similarly, the partial sum statistic and estimate of treatment effect based on the incremental data accrued between the $(j-1)$th and jth analyses are $S_j^* = \sum_{i=N_{j-1}+1}^{N_j}(X_{Ai} - X_{Bi})$ and $\hat{\theta}_j^* = \frac{1}{N_j^*} S_j^*$, respectively.

Assume that the potential outcomes are immediately observed. Without loss of generality, assume that positive values of θ indicate superiority of the new treatment. It is desired to test the null hypothesis $H_0 : \theta = \theta_0 = 0$ against the one-sided alternative $\theta > 0$ with type I error probability $\alpha = 0.025$. First consider a fixed sample design. In order to detect the alternative $\theta = \Delta$ with power β, the trial requires a fixed sample size on each treatment arm of

$$n = \frac{(z_{1-\alpha} + z_\beta)^2}{\Delta^2}$$

where $z_p = \Phi^{-1}(p)$. Alternatively, consider a group sequential design. We use the following

general framework [17] for such candidate sequential designs. At the jth interim analysis, we compute some statistic $T_j = T(X_1, ..., X_{N_j})$ based on the first N_j observations. Then, for specified stopping boundaries $a_j \leq d_j$, we will stop with a decision of non-superiority of the new treatment if $T_j \leq a_j$, stop with a decision of superiority of the new treatment if $T_j \geq d_j$, or continue the study if $a_j < T_j < d_j$. We restrict attention to families of stopping rules described by the extended Wang and Tsiatis unified family [18], in which the P parameter reflects the early conservatism of the stopping boundaries.

In order to reduce the dimensionality of the space of candidate clinical trial designs, we only consider symmetric designs. Symmetric sequential sampling plans consist of continuation and stopping sets that treat the null and alternative hypotheses symmetrically with respect to early stopping. We consider the one-parameter family of symmetric one-sided designs described by Emerson and Fleming [19] and shown to be nearly as efficient as the larger class introduced by Jennison [20]. Symmetric designs attain power $1 - \alpha$ at the alternative hypothesis and therefore reject the two design hypotheses with the same level of confidence. With $\alpha = 0.025$, these designs thus have the desired characteristic that a 95% confidence interval for the estimated treatment effect computed at the end of the trial will discriminate between the null and alternative hypotheses. In such a setting, we assume that the alternative hypothesis $\theta = \Delta$ is based on the therapeutic index, and thus represents an effect size that would be considered clinically meaningful when weighed against such treatment characteristics as toxicity, side effects, and cost. We note that any design with 97.5% power at $\theta = \Delta$ will obtain 80% and 90% power at some intermediate treatment effects $\theta < \Delta$, and thus, these symmetric designs can be used to target one of these common desired levels of power at an important alternative hypothesis.

3 The Completely Pre-specified Adaptive Design

In the spirit of the "sequentially planned decision procedures" proposed by Schmitz [12] and discussed further by Jennison and Turnbull [1], we can completely pre-specify adaptation sets at the design stage of the trial. With a pre-specified adaptive sampling plan, we can easily write out the distribution of the minimal sufficient statistic. In order to develop a better understanding of these designs in the simplest of settings, we restrict our attention to adaptive designs with only as many as two possible analyses. Consider the following simple example.

Suppose that we will base inference on the estimate of treatment effect equal to the difference in sample means: $\hat{\theta}_j = \overline{X}_{A,j} - \overline{X}_{B,j}$. At the first analysis, with sample size n_1 accrued on each arm, we stop early for efficacy if $\hat{\theta}_1 \geq d_1$ or futility if $\hat{\theta}_1 \leq a_1$. Now suppose that we additionally want to add a single adaptation region inside the continuation set (a_1, d_1) at the first analysis. Conceptually, the idea is that we have observed results sufficiently far from our expectations and from both stopping boundaries such that additional data (a larger sample size) might be desired. Denote this adaptation region $C_1 = [A, D]$ where $a_1 \leq A \leq D \leq d_1$. Denote the rest of the continuation region $C_2 = (a_1, A) \cup (D, d_1)$. The sampling plan proceeds as follows:

- if $\hat{\theta}_1 \leq a_1$, stop with a decision of non-superiority
- if $\hat{\theta}_1 \geq d_1$, stop with a decision of superiority
- if $\hat{\theta}_1 \in C_1$, continue the study, proceeding to sample size $n_2^{(1)}$, at which stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^{(1)}$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^{(1)}) = \frac{1}{n_2^{(1)}} \sum_{i=1}^{n_2^{(1)}} (X_{Ai} - X_{Bi})$
- if $\hat{\theta}_1 \in C_2$, continue the study, proceeding to sample size $n_2^{(2)}$, at which stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^{(2)}$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^{(2)}) = \frac{1}{n_2^{(2)}} \sum_{i=1}^{n_2^{(2)}} (X_{Ai} - X_{Bi})$

Figure 1 illustrates the stopping and continuation boundaries for one such sequential sampling plan, in which the design is symmetric so that $d_2^{(1)} = d_2^{(2)} = d_2$ (on the sample mean scale). We

can choose values of $n_2^{(1)}, n_2^{(2)}, d_2^{(1)}$, and $d_2^{(2)}$ so that the type I error rate is α, i.e.,

$$P_{\theta_0}(\hat{\theta}_1 \geq d_1) + P_{\theta_0}(\hat{\theta}_1 \in C_1, \hat{\theta}_2 \equiv \hat{\theta}(n_2^{(1)}) \geq d_2^{(1)}) + P_{\theta_0}(\hat{\theta}_1 \in C_2, \hat{\theta}_2 \equiv \hat{\theta}(n_2^{(2)}) \geq d_2^{(2)}) = \alpha.$$

We will discuss strategies for selecting these values and give optimal choices in simple settings later in the paper. The key point is that, with a completely pre-specified adaptive sampling plan like this one, we can specify exactly the distribution of the sufficient statistic and proceed with inference in a manner analogous to a standard group sequential design.

It is important to note that, by considering only completely pre-specified sampling plans with inference based on the minimal sufficient statistic, we are evaluating adaptive designs in their best possible light. Methods allowing unplanned adaptations to the sample size violate the sufficiency principle and subsequently come with costs in efficiency when compared to group sequential designs [1, 2, 3]. The following simple example demonstrates how substantial these losses in efficiency can be. We start with a symmetric O'Brien Fleming group sequential design with two equally spaced analyses, a type I error of $\alpha = 0.025$ at $\theta = 0$, and power equal to 0.975 at $\theta = \Delta$. The critical final efficacy boundary is $a_2 = d_2 = 0.5\Delta$ on the sample mean scale. Next, we optimally add one adaptation region at the first analysis, using methods which will be described in detail below. We compare two candidate adaptive designs. The first design $Adap1$ is the optimal adaptive design, with two continuation regions at the first analysis, and the preservation of the original final efficacy boundary $a_2 = d_2 = 0.5\Delta$. The second design $Adap2$ consists of the same two continuation regions at the first analysis and the same optimal choices of N_2 corresponding to those regions. However, with this design, we use the interim estimate of treatment effect at the first analysis to change the critical efficacy boundary at the final analysis (it now ranges from 0.21Δ to 0.95Δ on the sample mean scale) in order to preserve the conditional type I error under the original group sequential design. This is a common proposed adaptive method for preserving the type I

error rate and is equivalent to a variety of other adaptive methods in this simple 2-stage setting [2]. Both of our candidate designs have the same type I error and expected sample size. However, the second design $Adap2$ suffers a substantial loss in power as a result of its failure to base inference on the minimal sufficient statistic, which is the final sample size and the estimate of treatment effect based on the cumulative data at the time of stopping. $Adap2$ violates the sufficiency principle since the final critical boundary is a function of the first-stage estimate of treatment effect, and it is possible for the same value of the minimal sufficient statistic to lead to opposite decisions at the end of the trial. Based on the results of 100,000 simulations, under the design alternative $\theta = \Delta$, designs $Adap1$ and $Adap2$ have power 0.975 and 0.845, respectively. Under the intermediate alternative $\theta = \Delta/2$, $Adap1$ attains power equal to 0.500, as compared to 0.468 for $Adap2$. This simple example demonstrates that the loss in efficiency resulting from the violation of the sufficiency principle can, in some cases, be quite large. In this paper, we evaluate the most efficient adaptive designs, in which the sampling plan is completely pre-specified and inference is based on the minimal sufficient statistic.

3.1 The Sampling Density of the Minimal Sufficient Statistic

In the simple two-stage setting, first consider the joint density of the incremental partial sum statistics S_1^* and S_2^*. By appealing to the central limit theorem, we have approximate distributions $S_1^* \sim N(n_1\theta, n_1)$ and $S_2^* \,|\, S_1^* \sim N(n_2^*\theta, n_2^*)$, since $N_2^* = n_2^*$ is fixed conditional on $S_1^* = s_1^*$. Therefore,

$$\begin{aligned} f(s_1^*, s_2^*; \theta) &= f(s_2^* \,|\, s_1^*; \theta)\, f(s_1^*; \theta) \\ &= \frac{1}{\sqrt{n_2^*}}\, \phi\left(\frac{s_2^* - n_2^*\theta}{\sqrt{n_2^*}}\right)\, \frac{1}{\sqrt{n_1}}\, \phi\left(\frac{s_1^* - n_1\theta}{\sqrt{n_1}}\right) \\ &= \frac{1}{2\pi\sqrt{n_1 n_2^*}} \exp\left(-\frac{(n_2^* s_1^{*2} + n_1 s_2^{*2})}{2n_1 n_2^*}\right) \exp\left((s_1^* + s_2^*)\theta - \frac{(n_1 + n_2^*)}{2}\theta^2\right) \end{aligned}$$

Since $S_2 = S_1^* + S_2^*$ and n_1 is fixed, it is clear from the factorization criterion and a result of Lehmann and Scheffé that (N_2^*, S_2) is minimal sufficient for θ. Equivalently, $(N_2, \hat{\theta}_2)$ is minimal sufficient since $N_2 = n_1 + N_2^*$ and $\hat{\theta}_2 = \frac{1}{N_2} S_2$.

Therefore, define the test statistic $(N, \hat{\theta})$, where N is the sample size when the trial is stopped, and $\hat{\theta} \equiv \hat{\theta}(N)$ is the sample mean statistic computed on the cumulative data at the time of stopping. For the simple two-stage setting presented in the previous section, following Armitage, McPherson, and Rowe [21], the sampling density for observation $(N = n, \hat{\theta} = t)$ is then defined as

$$p(n, t; \theta) = \begin{cases} f(n, t; \theta) & \text{if } t \text{ falls in a stopping set} \\ 0 & \text{otherwise} \end{cases}$$

where the (sub)density is recursively defined as

$$\begin{aligned} f(n_1, t; \theta) &= \sqrt{n_1}\, \phi(\sqrt{n_1}\,(t - \theta)) \\ f(n_2^{(1)}, t; \theta) &= \int_{C_1} \sqrt{n_2^{(1)} - n_1}\, \phi\left(\sqrt{n_2^{(1)} - n_1}\,(t - u - \theta)\right) f(n_1, u; \theta)\, du \\ f(n_2^{(2)}, t; \theta) &= \int_{C_2} \sqrt{n_2^{(2)} - n_1}\, \phi\left(\sqrt{n_2^{(2)} - n_1}\,(t - u - \theta)\right) f(n_1, u; \theta)\, du \end{aligned}$$

and $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the standard normal density.

Since we can write out exactly the sampling density of the minimal sufficient statistic, we can numerically compute any of the standard operating characteristics used to evaluate group sequential designs. For example, the power and expected sample size (ASN) of this design at a particular alternative θ are easily calculated as

$$\text{Power}_\theta = P_\theta(\hat{\theta}_1 \geq d_1) + P_\theta(\hat{\theta}_1 \in C_1, \hat{\theta}_2 \equiv \hat{\theta}(n_2^{(1)}) \geq d_2^{(1)}) + P_\theta(\hat{\theta}_1 \in C_2, \hat{\theta}_2 \equiv \hat{\theta}(n_2^{(2)}) \geq d_2^{(2)})$$

$$\text{ASN}_\theta = n_1\left[P_\theta(\hat{\theta}_1 \geq d_1) + P_\theta(\hat{\theta}_1 \leq a_1)\right] + n_2^{(1)} P_\theta(\hat{\theta}_1 \in C_1) + n_2^{(2)} P_\theta(\hat{\theta}_1 \in C_2)$$

Thus, we can carefully evaluate the important operating characteristics of the adaptive sampling plan using standard software for group sequential designs [22], and then can compare these characteristics to those of alternative group sequential designs at the planning stage of the clinical trial. The ability to compute the sampling density and standard operating characteristics across the range of the parameter space easily generalizes to pre-specified adaptive designs with more than two analyses and continuation regions.

3.2 Finding the Optimal Adaptive Design

We follow a similar strategy for finding the "optimal" adaptive design in each of two general settings that will be described below. In each setting, we first clearly enumerate the optimality criteria governing the choice of randomized clinical trial (RCT) design. These optimality criteria include a particular null hypothesis and associated type I error and a design alternative at which there is a desired level of statistical power, along with constraints limiting the number of analyses desired and/or the minimal sample size at which early stopping is permitted.

In the class of symmetric adaptive designs with up to two analyses, the following parameters need to be specified. We need to choose the number m of continuation regions at the first analysis. For each of these regions, we must specify one of the boundaries (A or D in the example above). Due to symmetry, the other boundary is then determined, since these boundaries are symmetric about the midpoint between the null and alternative hypotheses on the sample mean scale. Finally, we must choose the maximal sample size $N_2 = n_2^{(\ell)}$ to which the study will proceed if the estimate of treatment effect falls in each respective continuation region C_ℓ, for $\ell = 1, 2, ..., m$. We note that the *a priori* specification of desired type I and type II errors restricts the range of these parameters and ensures that the specification of the first $\ell - 1$ possible values for N_2 determines the final possible maximal sample size. We also note that the stopping boundary d_2 at the final analysis is determined by symmetry and is equal to the midpoint between the design alternatives on the

sample mean scale.

Given these free parameters, our optimization procedure proceeds as follows:

- Start with $\ell = 2$ continuation regions. Holding constant the desired type I error and power, choose C_1 and $n_2^{(1)}$ to minimize the average sample size at the design alternatives (the ASN is the same at the null and alternative hypotheses due to symmetry). We perform a numerical grid search to minimize ASN over these two free parameters.

- Proceed to $\ell = 3$ continuation regions by holding C_1 constant and finding an optimal split of C_2 into 2 continuation regions (to minimize the ASN).

- Proceed to $\ell = 4$ continuation regions by finding an optimal split of C_1 (holding the other regions constant).

- Proceed with this method of increasing the number of continuation regions until there is evidence of approximate convergence to a minimum achieved ASN.

This optimization procedure conditions on all but one of the continuation regions, and the corresponding selected maximal sample sizes, that were chosen at the previous step. Therefore, it is not guaranteed that for $\ell > 2$ continuation regions, we have actually achieved the minimum ASN possible for this class of designs. However, sensitivity procedures iterating back and forth between adjacent regions do not provide further reduction in the expected sample size. In fact, most of the potential gain in efficiency is achieved with the first step, as will be discussed further in the examples below. It is also important to note that minimizing the average sample size at other values of the treatment effect (e.g. moderate effect sizes), or minimizing the expected ASN with respect to some prior distribution on the parameter space, would produce different "optimal" adaptive designs.

4 Comparing Adaptive and Alternative Designs

4.1 Setting #1

In this setting, we are interested in finding the most efficient clinical trial design given a constraint on the number of analyses that can be conducted. We acknowledge that statistical efficiency should never be the sole factor leading to a particular choice of clinical trial design, due to the numerous ethical, economic, and scientific issues that must be considered first at the design stage. However, it is still important to describe the optimal rules for making interim adaptations to the sample size and to discuss the gains that can be attained by the use of an adaptive design in a setting where efficiency is the primary concern. Suppose the following optimality criteria govern the choice of RCT design:

- The number of analyses is constrained to a maximum of two, which in our experience is the typical proposed setting for an adaptive design.
- The desired type I error is $\alpha = 0.025$ and power is $\beta = 0.975$ at the design alternative $\theta = \Delta$. The initial candidate design is a fixed sample design with $n = \frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$ subjects required to meet these operating characteristics.
- The primary interest is in finding the most efficient design meeting these constraints. Efficiency is measured by the average sample size in the presence of a truly ineffective (under the null hypothesis) or effective (under the alternative hypothesis) treatment.

The first alternative design is a standard group sequential design (GSD). Given the above constraints, we consider all symmetric group sequential sampling plans in the unified family with a maximum of $J = 2$ analyses. We choose values for P (degree of early conservatism) and N_1 (spacing of the two analyses) in order to maintain the desired α and β while minimizing the ASN at the design alternatives. This yields a 2-analysis GSD with $P = 0.542$ (close to a Pocock design,

which corresponds to $P = 0.5$) and analyses at 50% and 118% of the original fixed sample size n. The stopping boundaries for futility and efficacy at the first analysis are 0.21Δ and 0.79Δ on the sample mean scale, respectively. These boundaries correspond to $(0.57, 2.21)$ on the Z-scale, $(4.9\%, 95.1\%)$ on the conditional power scale assuming the interim maximum likelihood estimate $\hat{\theta}_1$ is the true treatment effect, and $(81.8\%, 99.0\%)$ on the conditional power scale assuming the design alternative Δ is the true treatment effect. This choice of GSD achieves an average sample size of 68.54% of the fixed sample size n at the design alternatives.

Next we consider optimal adaptive designs. We hold constant the timing and stopping boundaries of the first analysis of the optimal GSD and search for optimal adaptive designs over the different possible divisions of the continuation region at $n_1 = 0.5n$ (using the optimization routine described previously). Table 1 displays the average and maximal sample sizes of optimal adaptive designs with an increasing number of continuation regions (displayed in units of the original fixed sample size n), as well as the corresponding percent reduction in ASN as compared to the optimal GSD.

All of the designs displayed in Table 1 have power $\beta = 0.975$ at the design alternative $\theta = \Delta$. The adaptive design with one continuation region is the reference GSD. The efficiency gain achieved by the optimal adaptive design is minimal (less than 0.5% per treatment arm) and is largely produced by the first split of the GSD's single continuation set into two regions. The ASN decreases 0.34%, from $0.6854n$ to $0.6831n$, at this first split. Allowing more than four continuation regions leads to negligible decreases in the ASN, and approximate convergence to a minimum ASN is achieved by a design with 8 different regions. It is interesting that increasing the number of continuation regions of the optimal adaptive design only marginally increases the maximal N, to as large as $1.28n$ with 8 continuation regions. This result suggests that adaptive designs which include the possibility of very large increases in the sample size, to as much as twice or more the original n, are not efficient designs in terms of the expected sample size. Figure 2 displays the

optimal rule for N_2 as a function of the interim test statistic, computed on four commonly used scales, for a symmetric adaptive design with 8 continuation regions.

The gains in efficiency at the design alternatives ($\theta = 0$ and $\theta = \Delta$) are offset by losses in efficiency at intermediate values of the treatment effect. Figure 3 displays de-trended power and ASN curves, comparing two representative optimal adaptive designs with the original group sequential design. We can see that the adaptive designs suffer efficiency losses for values of θ between 0.25Δ and 0.75Δ, with worst-case behavior relative to the group sequential design at $\theta = 0.5\Delta$. Worst-case efficiency losses are nearly the same magnitude as efficiency gains at the design alternatives. While the addition of continuation regions modestly increases efficiency gains at the design alternatives, it also increases efficiency losses at intermediate values of the treatment effect. Our optimality criteria and optimization procedure force the group sequential and adaptive designs to have equal power at the design alternatives, but Figure 3 demonstrates that there are some slight differences in the power of these designs at other possible values of the treatment effect. However, differences in power are less than 0.001 and thus are negligible.

It is also important to note that adding an additional analysis to a group sequential design leads to a much larger efficiency gain than allowing adaptive modifications to the sample size. For example, if we hold constant the stopping boundaries at the first analysis and choose two additional analysis times from among the 8 adaptive values for N_2 shown in Figure 2, we can decrease the ASN at the design alternatives to as low as $0.643n$. Thus, a 3-analysis group sequential design is able to reduce the average sample size of the optimal 2-analysis GSD by 6.2%, as compared to the less than 0.5% reduction achieved by the optimal 2-analysis adaptive design. This is an important result in considering the tradeoffs between the cost of carrying out additional analyses and the costs of enrolling additional patients and increasing study duration.

4.2 Setting #2

In this second setting, we are interested in the possible gains in efficiency that can be made by using an early analysis to help determine the optimal sample size for the analysis at which inference will be carried out. Consider a scenario in which the following optimality criteria govern the choice of RCT design:

- There will be only one analysis at which an efficacy decision can be made. Another analysis is permitted to help determine the optimal sample size for the final analysis.

- The desired type I error is $\alpha = 0.025$ and power is $\beta = 0.975$ at the design alternative $\theta = \Delta$. The initial candidate design is a fixed sample design with $n = \frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$ subjects required to meet these operating characteristics.

- A minimum sample size for early stopping of $n_{min} < n$ is required so that an adequate safety profile for the new treatment can be developed. We assume that the minimal sample size for early stopping is $n_{min} = 0.75n$. Similar patterns to those described below were observed when n_{min} was set at different proportions of the fixed sample size.

- The primary interest is in finding the most efficient design satisfying these constraints, where efficiency is measured by the ASN at the design alternatives.

Given the above constraints, we consider a range of adaptive designs. The "adaptation" analysis, at which the estimate of treatment effect will be used to determine the sample size for the final analysis, may occur at a range of time points n_{adap} prior to the accrual of n_{min} subjects. Let $n_{adap} = R * n_{min}$, and consider $R \in \{0.1, 0.2, ..., 0.9, 1.0\}$. The adaptive design with $R = 1.0$ is the only one of the 10 candidate adaptive designs that allows stopping for futility and efficacy both at the analysis used to determine the final sample size and at the final analysis. Thus, as R approaches 1, these adaptive designs are essentially progressing from an adaptive design with one analysis for

early stopping towards an adaptive design with two such analyses. Each adaptive design described in Table 2 includes four continuation regions, since adding additional regions had negligible effects on the ASN. We display the average and maximal sample sizes, in units of the fixed sample size n, for these optimal adaptive designs. Table 2 also provides the probabilities that the sample size will exceed n, $1.1n$, and $1.2n$ for each of the candidate designs.

These results demonstrate several interesting characteristics of adaptive designs in this setting. First, it is clear that adapting the sample size on the basis of minimal statistical information is not a good idea. Adaptations at 10% and 20% of n_{min}, for example, provide very small efficiency gains (1% and 3% reductions in the ASN), while more substantially increasing the maximal N. Reductions in the ASN achieved by the adaptive designs grow larger as the quantity of accrued statistical information at the adaptation increases. The largest efficiency gain is attained when the adaptation occurs at an analysis which also allows early stopping ($R = 1.0$). In addition, in this setting, the designs that adapt the sample size at $1/2$ to $2/3$ of n_{min} provide worse behavior than designs with later-stage adaptations, with respect to both the maximal N and the probabilities of exceeding important sample size thresholds. These results suggest that the frequently proposed adaptive sampling plans that allow modifications to the sample size at or around one half of the minimal stopping sample size may not represent efficient choices for an RCT design.

Our results do in fact show that adding an interim analysis to modify the sample size leads to meaningful efficiency gains relative to a fixed sample test, reducing the ASN at the design alternatives by as much as approximately 20%. However, just as in the first setting, it is clear that the largest efficiency gain is attained by adding an analysis at which stopping for futility and efficacy can occur. These results suggest that, if an RCT sampling plan is to include the possibility of interim modifications to the sample size, such an adaptation should occur at an analysis that also permits early stopping. Finally, we note that these optimal adaptive designs lead to maximal increases in the sample size of only about 20%, much less than the 50% or twofold increases often

proposed in the literature.

Figure 4 displays the optimally chosen adaptation boundaries on commonly used scales, along with the corresponding choices of N, for three representative values of R. Taking into account the different ranges of values plotted on the x-axes of these three plots, we can see that the boundaries outside of which the optimal adaptive designs proceed only to accrue n_{min} subjects grow tighter as the timing of adaptation gets later (as R and thus n_{adap} increase). It is interesting to examine the chosen boundaries on the conditional power scale assuming the interim maximum likelihood estimate is the true treatment effect. When $R = 0.5$ for example, so that the adaptation occurs at half the minimum sample size, the adaptive design proceeds to the smallest possible sample size n_{min} only if the conditional power is as low as 3% or as high as 97%. This choice deviates greatly from adaptive designs that have been proposed in the literature, which have set this lower threshold for proceeding to only the minimal sample size to as high as 36% [14].

The optimal adaptive designs attain smaller efficiency gains over the original fixed sample design at treatment effects falling in between the design alternatives. Figure 5 displays de-trended power and ASN curves, comparing four representative optimal adaptive designs to the original fixed sample design. We can see that, as long as the adaptation analysis occurs after the accrual of at least $0.5n_{min}$ subjects, the adaptive design is uniformly superior to the group sequential design with respect to the expected sample size. However, it is clear that the adaptive design with $R = 1.0$ is most efficient, again suggesting that a larger efficiency gain can be attained by adding an analysis at which early stopping is permitted than by adding an adaptation analysis. Differences in power are small with no clear benefit for either the fixed sample or adaptive design.

5 Other Issues with Adaptive Designs

The evaluation of adaptive designs should not be limited to type I error control and efficiency considerations. It is also critical that any confirmatory phase III clinical trial design is able to produce results that are interpretable, and that investigators are able to provide reliable inference on the treatment effect of interest. Regulatory agencies must balance estimates of efficacy against safety concerns in deciding whether to approve a new drug, and also need reliable estimates and intervals for the development of new drug labeling. In addition, clinicians may use estimates of treatment effect to compare one treatment option to alternative interventions for a particular patient and indication. Confirmatory trial designs should allow the computation of unbiased and sufficiently precise estimates of treatment effect and the construction of confidence intervals with the correct frequentist coverage probabilities, for both efficacy and safety endpoints.

Such inference is possible in the context of a completely pre-specified design with adaptations to the maximal sample size. The ability to compute the sampling density of the sufficient statistic allows the extension of methods for parameter estimation and confidence interval construction developed for the analysis of group sequential trials. For example, we could perform a numerical search to compute the bias-adjusted mean or median-unbiased estimate of the treatment effect. It should also be possible to construct intervals based on different orderings of the outcome space, although further research in this area is needed. It may be challenging to compute confidence intervals in the case of non-symmetric adaptive designs. Completely pre-specified adaptive designs with interim modifications to the number or spacing of analyses, or to the choice of stopping boundaries, should also allow enumeration of the sampling density and thus, the computation of reliable estimates and intervals. The use of the previously described orderings of the outcome space [23, 24, 25, 26, 27] with adaptive designs may exhibit different and less desirable behavior than is observed with group sequential tests with respect to the generation of convex confidence intervals,

inclusion of point estimates, agreement with test decisions, and width of confidence intervals.

We note, however, that the ability to provide inference is even more difficult if adaptations to the sample size are based on unplanned rules, or if modifications are made to scientific aspects of the study design. For example, adaptive designs with unplanned changes to the study population (e.g. "adaptive enrichment"), the primary endpoint, or the treatment strategy based on interim estimates of treatment effect alter the scientific hypotheses being tested in the different stages of the trial, and compromise the investigators' ability to provide reliable inference on a particular treatment indication at the study's end.

There are also a number of logistical issues inherent in adaptive designs. Pre-specifying exact rules for interim modifications to aspects of the study design leads to a more complicated protocol, and thus may extend the design stage of the trial. Such an increase in the duration of the design stage could outweigh small efficiency gains made during the conduct of the trial. In addition, the pre-specification of rules for adapting the maximal sample size could cause interim analyses to essentially reveal the current estimate of treatment effect to study investigators who should remain blinded until the trial's conclusion. If the choice of maximal sample size is a continuous function of the interim estimate, such as in the case of designs based on increasing the conditional power to a fixed target, investigators would be able to use the new final sample size target to infer the current estimate of treatment effect. It is well-known that in certain settings, unblinding of investigators or patients to treatment results can introduce multiple sources of bias and can compromise the validity of a confirmatory trial. This issue becomes less of a concern in the case of an adaptive design containing only a few adaptation regions and corresponding possible choices for the maximal sample size. With such a design, investigators would only be able to identify a region containing the interim estimate and may be less likely to change trial conduct.

Finally, we note that adaptive designs may have additional benefits in the setting of longitudinal settings, such as time-to-event survival studies. In the presence of delayed ascertainment of the

primary outcome, it is important to consider both the number of subjects accrued and the trial duration. Emerson et. al found some benefits of pre-specified adaptive designs over group sequential designs in certain time-to-event settings [22].

6 Conclusions and Additional Comments

The goal of this paper was to critically evaluate a range of simple and easily implemented pre-specified adaptive sampling plans in order to contribute to the understanding of adaptive designs with interim modifications to the maximal sample size. To that end, we focused on the efficiency of adaptation, although we acknowledge that a great many other considerations go into the selection of a sequential monitoring plan, including the need to be able to address long-term safety and secondary endpoints.

In the context of two general clinical trial settings, where different optimality criteria govern the choice of RCT design, we compared a variety of fixed sample, group sequential, and adaptive designs with respect to standard operating characteristics. We found simple and easily implemented symmetric adaptive designs with completely pre-specified stopping and continuation boundaries and inference based on the minimal sufficient statistic that were "optimal" in the sense that they minimized the expected sample size at the design alternatives. Our comparisons of alternative designs provide a commentary on the efficiency gains that can be attained with the use of adaptive designs in simple and realistic settings, as well as some insight into what are efficient rules for adapting the sample size at an interim stage of the trial.

Our results from the first setting are consistent with those discussed in several previous works [1, 4, 5, 6] in demonstrating that optimal completely pre-specified adaptive designs with inference based on the minimal sufficient statistic can only lead to very small efficiency gains over optimal group sequential designs with the same number of analyses. Our study builds on previous research

by quantifying precisely the efficiency gains that can be attained with the use of simple and easily implemented adaptive sampling plans in realistic RCT design settings. Constraining the RCT design to a maximum of 2 analyses, we found adaptive designs that attained an ASN at the design alternatives of nearly 0.5% lower than an efficient group sequential design. However, these gains were offset by losses in efficiency at intermediate values of the treatment effect. In addition, adding a third analysis to the group sequential design decreased the ASN by more than 6%, suggesting that the addition of stopping analyses provides more substantial efficiency gains than adding analyses used to adapt the sample size.

In addition to quantifying efficiency gains, the results of our study provide important insight into what are good and bad types of adaptation rules. In particular, we found that dividing the original group sequential continuation boundary into more than four or five adaptation regions leads to negligible efficiency gains. In fact, most of the efficiency gain obtained through adaptation was achieved by adding the first adaptation region (allowing two different potential maximal sample sizes). We have found this to be true for asymmetric adaptive procedures as well. Briefly, we investigated the use of an adaptive rule based on the procedure of Gao, Ware, and Mehta [28], which is designed to achieve a specified conditional power, conditional on the estimated effect size, by modifying both the critical value and the maximal sample size. While the procedure proposed by Gao, Ware, and Mehta uses a continuous function of the interim estimate to determine the maximal sample size, we modified this approach by discretizing: dividing the set of possible interim effect sizes into disjoint regions of values inside of which the same future boundary and sample size will be used. We computed the future boundary and sample size for each region by carrying out these computations at the region's midpoint using the formulae of Gao, Ware, and Mehta, an approach that preserves the type I error while boosting the conditional power to 90%. For this procedure, the goal of adaptation was to increase power (rather than decrease ASN). We found a negligible difference between the use of only a few adaptation regions and the use of what essentially is a

continuous function to determine the final sample size and boundary: A design with 101 adaptation regions achieved a maximal power increase of less than 0.04% over a design with five adaptation regions. These findings suggest that the frequently proposed use of a continuous function of the interim estimate to determine the maximal sample size (e.g. to raise the conditional power to a desired level) may be unnecessary. This is especially noteworthy considering the logistical issues that accompany such continuous rules. On the other hand, it is straight-forward to implement and compute the operating characteristics of simple adaptive designs that contain only a few adaptation regions using standard group sequential software.

Our results also provide interesting commentary on the merit of other characteristics of adaptation rules. The findings from our second RCT design setting demonstrate that adding an interim analysis to modify the sample size leads to meaningful efficiency gains relative to a fixed sample test, reducing the ASN at the design alternatives by as much as approximately 20%. However, just as in the first setting, our results demonstrate that a greater efficiency gain can be attained by adding an analysis at which stopping for futility and efficacy can occur. These findings suggest that, if an RCT sampling plan is to include the possibility of interim modifications to the sample size, such an adaptation should occur at an analysis that also permits early stopping. In short, it does not seem worthwhile to adapt before having the ability to stop early.

Our results also suggest that adaptive designs frequently proposed in the literature do not include optimal timing for the adaptation analyses or optimal rules for modifying the sample size. It is common for such designs to include interim modifications after the accrual of one half the original fixed sample size, which may be inefficiently early. We also note that the optimal adaptive sampling plans found for the RCT design settings considered in this paper only lead to maximal increases in the sample size of about 20% to 30%, much less than the 50% or twofold increases often proposed in the literature. Designs with the possibility of such large increases in the maximal sample likely do not result in sufficient efficiency gains to offset the huge potential investment required

of the sponsor. It is also common for proposed adaptive designs to use suboptimal thresholds and rules on the conditional power scale for modifying the sample size. Our results provide optimally chosen boundaries on several scales for different design settings, and should help investigators better understand what are good and bad choices for adaptive sampling plans. At a minimum, we hope that these results cause investigators to more rigorously evaluate candidate alternative group sequential and adaptive designs.

As with any evaluation of alternative group sequential and adaptive designs, it is very challenging to carry out a fair and reasonable comparison. There are many parameters that can vary, such as the number and timing of analyses, the family of stopping boundaries, and the operating characteristics used to determine efficiency, as well as possible scientific constraints on the conservatism of early boundaries or the minimal sample size for early stopping. In this paper, we address only a small fraction of this large space of designs. We focus our investigation on symmetric designs in two simple settings, and define "efficiency" and find "optimal" designs based on the expected sample size at the design alternatives. However, we believe that these simple comparisons should provide insight into the broader class of adaptive and group sequential designs. Many recent papers have failed to come up with fair comparisons of adaptive and alternative designs, and thus lead to results that are challenging or impossible to interpret. Although our evaluation of candidate designs is limited to two simple settings, we believe that these cases present a reasonable and fair comparison of alternative RCT designs. In each setting, we first clearly describe a set of realistic optimality criteria governing the choice of RCT design, and then find candidate fixed sample, group sequential, and adaptive designs that meet these constraints.

In summary, we have evaluated adaptive designs with pre-specified rules for modifying the maximal sample size based on interim estimates of treatment effect and inference based on the minimal sufficient statistic. Completely pre-specified adaptive designs allow the computation of the sample density of the sufficient statistic and thus should allow investigators to carry out full

frequentist inference at the end of the clinical trial. Our results suggest that simple and easily implemented pre-specified adaptive sampling plans achieve only small efficiency gains over alternative group sequential designs with the same number of analyses in realistic settings. We would argue that these very small efficiency gains are often not worth the additional logistical challenges that come with adaptive designs, and that standard group sequential designs best address the complex ethical, scientific, and efficiency issues inherent in most, if not all, RCT settings. However, adaptive designs continue to be proposed in actual clinical research, and thus, it is important to critically evaluate such sampling plans so that investigators have the tools to choose an efficient design that satisfies the scientific constraints of a specific RCT setting. To this end, our findings provide optimal adaptation rules in simple design settings and thus provide some insight into what are efficient choices of adaptive sampling plans, and where it may be best to dedicate future research efforts. In particular, our results suggest that in searching for adaptive sample plans, it is likely adequate to restrict attention to simple designs with only a few adaptation regions.

References

[1] Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. Biometrika 2006; **93**(1):1–21.

[2] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine 2003; **22**:971–993.

[3] Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika 2003; **90**(2):367–378.

[4] Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. Statistics in Medicine 2006; **25**:917–932.

[5] Benerjee A, Tsiatis AA. Adaptive two-stage designs in phase ii clinical trials. Statistics in Medicine 2006; **25**:3382–3395.

[6] Lokhnygina Y, Tsiatis AA. Optimal two-stage group-sequential designs. Journal of Statistical Planning and Inference 2008; **138**:489–499.

[7] Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994; **50**(4):1029–1041.

[8] Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. Biometrics 1995; **51**(4):1315–1324.

[9] Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. International Biometric Society 2001; **57**(3):886–891.

[10] Fisher LD. Self-designing clinical trials. Statistics in Medicine 1998; **17**:1551–1562.

[11] Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. Biometrics 1999; **55**:853–857.

[12] Schmitz N. Optimal Sequentially Planned Decision Procedures. Springer-Verlag New York, Inc., 1993.

[13] Food and Drug Administration. Guidance for industry: Adaptive design clinical trials for drugs and biologics 2010.

[14] Mehta C, Pocock S. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine. 2010; In press.

[15] S+SeqTrial. Insightful corporation 2002. Seattle, Washington.

[16] Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC, 2000.

[17] Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics 1999; **55**:874–882.

[18] Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987; **43**:193–199.

[19] Emerson SS, Fleming TR. Symmetric group sequential test designs. Biometrics 1989; **45**(3):905–923.

[20] Jennison C. Efficient group sequential tests with unpredictable group sizes. Biometrika 1987; **74**:155–165.

[21] Armitage P, McPherson C, Rowe BC. Repeated significance tests on accumulating data. Journal of the Royal Statistical Society: Series A 1969; **132**:235–244.

[22] Emerson SC, Rudser KD, Emerson SS. Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. Statistics in Medicine. 2010; In press.

[23] Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. Biometrics 1984; **40**(3):797–803.

[24] Chang MN, O'Brien PC. Confidence intervals following group sequential tests. Controlled Clinical Trials 1986; **7**:18–26.

[25] Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential test: A comparison of methods. Biometrika 1988; **75**:723–729.

[26] Chang MN. Confidence intervals for a normal mean following a group sequential test. Biometrics 1989; **45**:247–254.

[27] Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. Biometrika 1990; **77**(4):875–892.

[28] Gao P, Ware J, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. Journal of Biopharmaceutical Statistics 2008; **18**(6):1184–1196.

Tables and Figures

Table 1: Average and Maximal Sample Sizes of Adaptive Designs in Setting #1

	Number of Continuation Regions								
	0[a]	1[b]	2	3	4	5	6	7	8
$ASN_{\theta=0,\Delta}$	1	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Difference	+45.9%	*Ref*	-0.34%	-0.38%	-0.42%	-0.43%	-0.43%	-0.44%	-0.44%
Maximal N	1	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

a. Fixed Sample Design

b. Group Sequential Design (*Reference* design)

Table 2: Characteristics of the Sample Size Distribution of Adaptive Designs in Setting #2

	R (Proportion of n_{min} at which adaptation occurs)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$ASN_{\theta=0,\Delta}$	0.99	0.97	0.94	0.91	0.88	0.86	0.84	0.82	0.80	0.78
Maximal N	1.07	1.12	1.16	1.18	1.20	1.21	1.21	1.20	1.18	1.17
$P_{\theta=0,\Delta}(N > 1.0)$	0.61	0.68	0.55	0.45	0.36	0.29	0.23	0.18	0.14	0.09
$P_{\theta=0,\Delta}(N > 1.1)$	0	0.38	0.36	0.28	0.23	0.18	0.14	0.11	0.09	0.06
$P_{\theta=0,\Delta}(N > 1.2)$	0	0	0	0	0	0.11	0.09	0.07	0	0

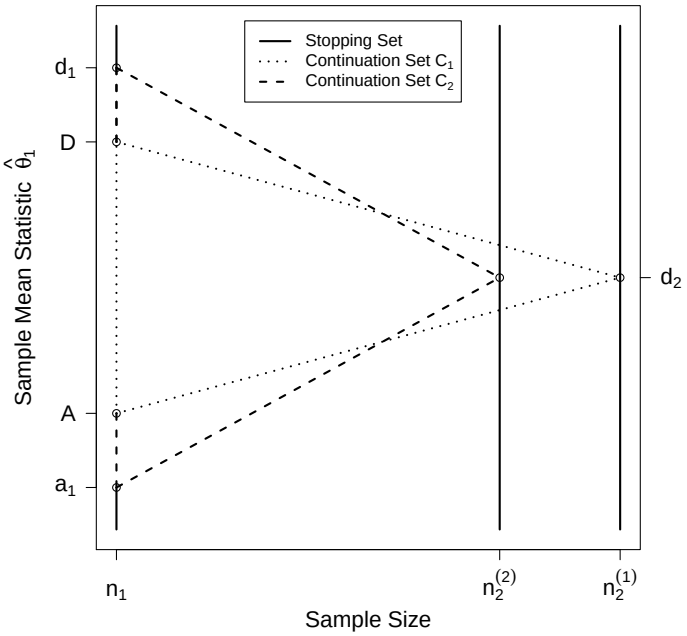


Figure 1: An illustration of possible continuation and stopping boundaries on the sample mean scale for a completely pre-specified adaptive design

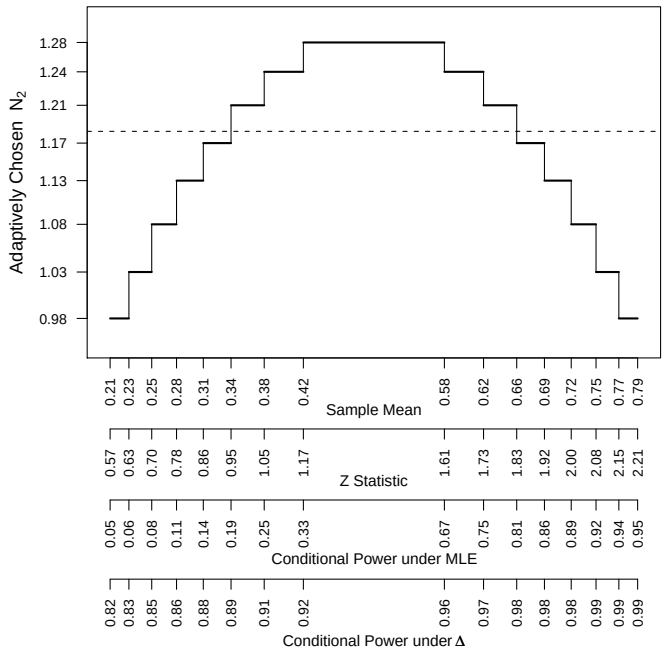
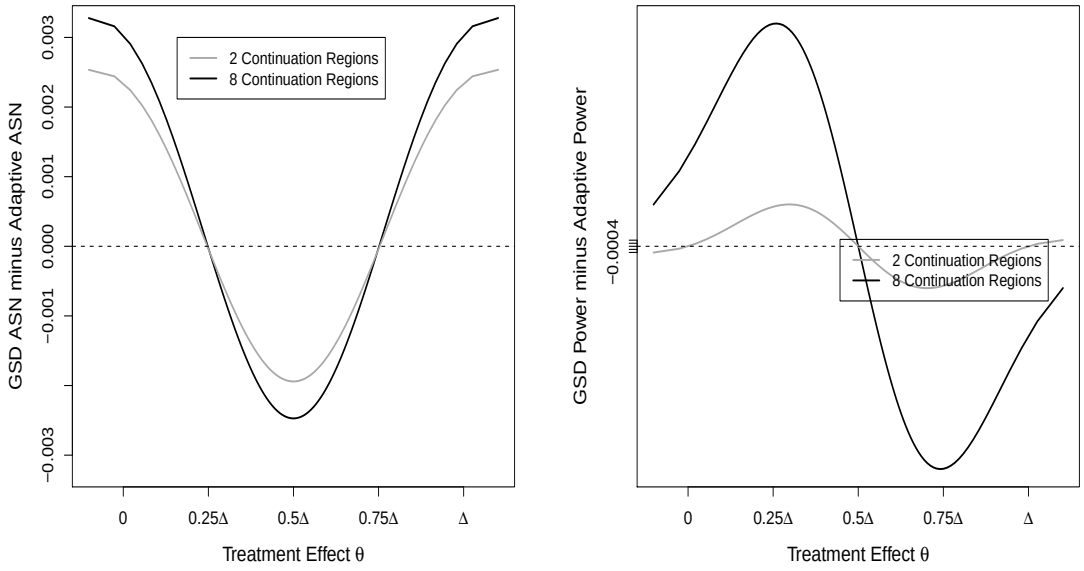


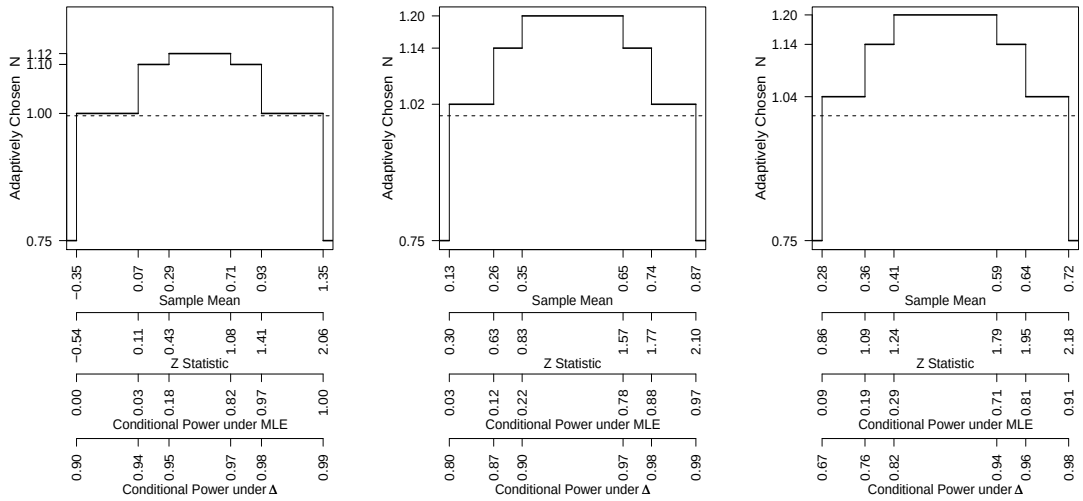
Figure 2: The optimal choice of N_2, in units of the fixed sample size n, as a function of the test statistic computed at the first analysis for a symmetric adaptive design with 8 continuation regions. The interim test statistic is displayed on the following scales: the crude estimate of treatment effect, or sample mean, scale (in units of the design alternative Δ), the normalized Z statistic scale, the conditional power scale under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), and the conditional power scale under the alternative ($\theta = \Delta$). The dotted line represents n_2 under the optimal group sequential design. The adaptive design stops early for efficacy or futility (at $n_1 = 0.5n$) if the sample mean at the first analysis is greater than 0.79Δ or less than 0.21Δ, respectively.



(a) Differences in Expected Sample Size

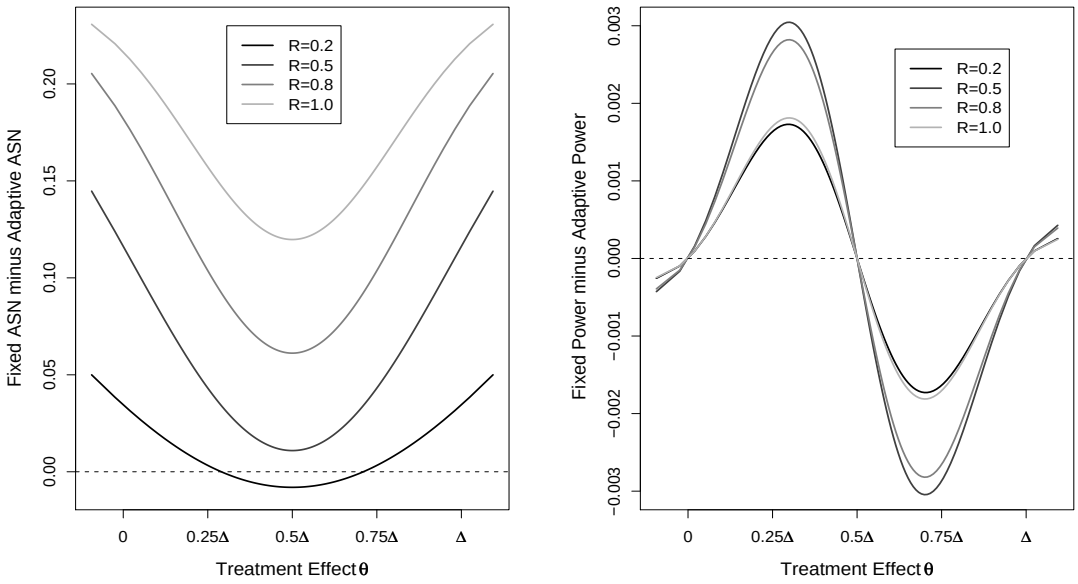
(b) Differences in Power

Figure 3: Comparison of the group sequential design to two representative optimal adaptive designs with respect to power and ASN across a range of plausible treatment effects. Differences between the group sequential and adaptive operating characteristics are shown on the y-axes. ASN differences are in units of the fixed sample size n. The dotted line indicates equality.



(a) R = 0.2 (b) R = 0.5 (c) R = 0.8

Figure 4: Optimal adaptive rules for the choice of N when the adaptation occurs at different stages of the trial. Adaptive designs select the final sample size N based on the test statistic computed after accrual of $n_{adap} = R * n_{min}$ subjects. The interim test statistic is displayed on the following scales: the crude estimate of treatment effect, or sample mean, scale (in units of the design alternative Δ), the normalized Z statistic scale, the conditional power scale under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), and the conditional power scale under the alternative ($\theta = \Delta$). Adaptively chosen values of N are displayed in units of the fixed sample size n. All designs proceed to accrual of a total of $n_{min} = 0.75n$ subjects if the estimate at n_{adap} falls outside the outermost boundaries.



(a) Differences in Expected Sample Size

(b) Differences in Power

Figure 5: Comparison of the fixed sample design to four representative optimal adaptive designs with respect to power and ASN across a range of plausible treatment effects. Differences between the fixed sample and adaptive operating characteristics are shown on the y-axes. ASN differences are in units of the fixed sample size n. The dotted line indicates equality.