1-13-2010

Exploring the Benefits of Adaptive Sequential Designs in Time-to-Event Endpoint Settings

Sarah C. Emerson
Harvard University, scemerson@gmail.com

Kyle Rudser
University of Minnesota, rudser@umn.edu

Scott S. Emerson
University of Washington - Seattle Campus, semerson@uw.edu

Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings

Sarah C. Emerson[1], Kyle D. Rudser[2], and Scott S. Emerson[3]

[1]Department of Statistics, Stanford University
[2]Division of Biostatistics, University of Minnesota
[3]Department of Biostatistics, University of Washington

Abstract

Sequential analysis is frequently employed to address ethical and financial issues in clinical trials. Sequential analysis may be performed using standard group sequential designs, or, more recently, with adaptive designs that use estimates of treatment effect to modify the maximal statistical information to be collected. In the general setting in which statistical information and clinical trial costs are functions of the number of subjects used, it has yet to be established whether there is any major efficiency advantage to adaptive designs over traditional group sequential designs. In survival analysis, however, statistical information (and hence efficiency) is most closely related to the observed number of events, while trial costs still depend on the number of patients accrued. As the number of subjects may dominate the cost of a trial, an adaptive design that specifies a reduced maximal possible sample size when an extreme treatment effect has been observed may allow early termination of accrual and therefore a more cost-efficient trial. We investigate and compare the tradeoffs between efficiency (as measured by average number of observed events required), power, and cost (a function of the number of subjects accrued and length of observation) for standard group sequential methods and an adaptive design that allows for early termination of accrual. We find that when certain trial design parameters are constrained, an adaptive approach to terminating subject accrual may improve upon the cost efficiency of a group sequential clinical trial investigating time-to-event endpoints. However, when the spectrum of group sequential designs considered is broadened, the advantage of the adaptive designs is less clear.

1 Introduction

In designing a clinical trial, investigators typically determine the statistical information required to discriminate between a null hypothesis of no treatment effect and some alternative hypothesis representing the minimal clinically important difference. When evaluating a proposed clinical trial design, the sponsor must balance the scientific goals with the cost of the trial and the ethical issues of human experimentation.

The number of subjects involved and the duration of a clinical trial are two key factors in determining the overall cost of the trial. Requiring fewer patients will decrease the trial costs associated with screening, treatment, and follow-up. Shorter trials allow earlier profit from treatments that are effective and waste fewer resources (investigator time, cost of following patients over time, cost of money, etc.) on ineffective treatments. Trial duration also plays a role in ethical considerations: Ending a trial sooner rather than later will expose fewer patients to an ineffective or harmful treatment and allow the broader population of patients earlier access to new effective treatments. This also frees up patients for the investigation of other treatments, which in turn can speed up the process of new treatment discovery in a disease for which clinical trial participants may be in short supply.

When measurements of treatment effect are based on comparisons of means or proportions, statistical information is generally directly proportional to the number of subjects accrued. Hence, any ability to decrease calendar time is generally related to an ability to increase the number of centers recruiting patients.

However, in a survival analysis/time-to-event setting with potentially censored observations, statistical information is more directly related to the number of observed events rather than the number of subjects. Thus in such a setting there are more tradeoffs possible between sample size and calendar time, as decreasing the number of subjects accrued generally increases the calendar time required to observe the requisite number of events and vice versa.

Historically, group sequential tests have been the primary statistical method used to address the ethical and efficiency concerns in clinical trials. In a typical group sequential design, a rule is specified for determining the maximal statistical information N_J. Then, at periodic intervals during the conduct of the study, up to J interim analyses are performed to determine whether the trial should stop early. More recently there has been much interest in the statistical literature related to "adaptive designs". Such adaptation typically takes the form of decisions to extend a clinical trial beyond some previously planned maximal stopping time. The advantages of such an approach over the group sequential design have not been established in general. Tsiatis and Mehta [1] and Jennison and Turnbull [2] have not found any efficiency gains of the adaptive approach over the more standard group sequential designs. However, such explorations have focused on the setting in which statistical information was directly proportional to the number of subjects accrued to the study.

In a survival analysis/time-to-event setting, there may be a clearer advantage to adaptive designs due to the need to consider both the number of patients accrued and the calendar time of follow-up necessary to observe the desired number of events. The appeal of an adaptive design in this setting is that it offers the possibility that early trends in the estimated treatment effect may suggest a modification of the number of subjects that need to be accrued. For instance, suppose a clinical trial is designed based on a maximal statistical information of N_J. Suppose further that at the jth analysis, the estimated treatment effect were so extreme that it seemed likely that the ultimate decision for efficacy or futility could be precisely determined prior to observing all N_J events. Then it might seem advantageous to consider a re-designed trial in which the revised maximal number of observed events $N^*_{J^*}$ were strictly less than originally planned, i.e., $N^*_{J^*} < N_J$. Such a re-design might allow the number of patients accrued to the study to be similarly decreased, thereby possibly reducing the number of subjects involved. Alternatively, if at an early interim analysis of the data a less extreme treatment effect were observed than was initially anticipated, the sponsor might want to increase the maximal number of events $N^*_{J^*} > N_J$ in order to increase the conditional power of the study to attain statistical significance. In such a setting, it may be necessary to increase the number of subjects accrued to the clinical trial in order to observe the increased number of events in an acceptable interval of calendar time.

As the focus of this manuscript is the relative flexibility of standard group sequential designs and more recently described adaptive designs to meet the optimality criteria of the collaborators in a clinical trial, we restrict attention to a single hypothetical clinical trial setting. We then consider a range of stopping rules, both group sequential and adaptive, that address the types of operating characteristics most often addressed in the statistical design of a clinical trial. In particular, we consider a setting that is a slight modification of a design proposed for an industry sponsored clinical trial. In that setting, the sponsor adopted a group sequential clinical trial design to detect a specified design alternative. The initial trial design was based on estimates of subject accrual rates and event rates. In order to protect against the possibility that the observed treatment effect might be less than that indicated by the specified design alternative, the sponsor also incorporated an adaptive modification of both the maximal number of subjects to be accrued and the maximal number of events to be observed. The conditions under which the sampling rule was modified were defined based on an interim estimate of the treatment effect. In this manuscript, we model such a modification of the clinical trial design through an adaptive switch between two group sequential stopping rules. We note that this particular form of adaptive sequential design was proposed for a confirmatory clinical trial that was to be used for regulatory approval. As such, the adaptation is completely prespecified. The statistic used to define the adaptive design stopping rule is a sufficient statistic. Furthermore, because we consider a prespecified stopping rule, there is no need to address the worst case scenarios that must be considered when any adaptation is not completely prespecified. Statistical inference can be based on the distribution of the sufficient statistic under the sampling plans specified by the sequential stopping rules.

Emerson et al. [3, 4] discuss the breadth of operating characteristics commonly considered when comparing candidate sampling plans for a clinical trial. In this manuscript we presume that the stopping boundaries at the first interim analysis were chosen based on results that would be judged clinically important and statistically credible. Hence, we only consider alternative designs that agree in their stopping boundaries at the first interim analysis. In keeping with the sponsor's initial exploration of designs, we further primarily consider the possibility that logistical constraints might be the driving force in choosing the accrual rates and the schedule of interim analyses. Thus we consider in depth comparisons in which the accrual rate is the same for a group sequential design and an adaptive design, and interim analyses are to be performed at identical intervals of accrual of statistical information. We then measure the possible advantage of the adaptive approach over the group sequential approach (and vice versa) according to the expected study duration and average number of subjects accrued for a range of adaptive and group sequential designs and several combinations of accrual and event distributions. Using the setting in which we observed the greatest advantage for the adaptive approach over a group sequential approach, we explored the effect of relaxing the constraint on the schedule of interim analyses. In so doing, we found a group sequential design that incurred lower costs than the adaptive design, thus illustrating the need for careful evaluation of a broad spectrum of group sequential rules when attempting to adaptively improve design operating characteristics.

In Section 2, we describe the survival analysis setting that will be used to compare the group sequential approach to a more adaptive approach. We define the particular form of the adaptive designs considered in this manuscript, as well as the design parameters that are held constant between the group sequential and adaptive designs. The relative behavior of the group sequential and adaptive designs are then investigated in the absence of censoring in Section 3 and in the presence of censoring in Section 4. Section 5 introduces a simple cost model used to summarize the comparisons between pairs of adaptive and group sequential designs. We conclude in Section 6 with a discussion of the impact of the particular design parameters and operating characteristics that were constrained to be equal in our comparisons, and demonstrate our ability to find more efficient group sequential designs when those constraints are relaxed.

2 Background

In designing a clinical trial there are several competing concerns: Efficiency, statistical power to detect an effect of interest, and ethical considerations are key factors in assessing the suitability of a proposed design. Efficiency generally refers to the number of subjects or events required, and is often measured as the expected sample size or *average sample number* (ASN). The maximal possible sample size is also frequently considered in efficiency comparisons. Power to detect an effect is the probability of deciding to reject the null hypothesis at a given effect size. Both power and ASN are functions of the true effect size. Ethical considerations are also addressed by minimizing the number of subjects and the time required to complete the trial.

Sequential analysis is a tool that is often employed to address these tradeoffs. The basic idea is that if results are convincing early on, there is no need to increase costs and lose efficiency by continuing with more subjects. At the jth of J potential interim analyses, a test statistic T_j is computed and compared to stopping boundaries. Following Kittelson and Emerson [5], it is generally sufficient to define at each analysis up to four stopping boundaries $a_j \leq b_j \leq c_j \leq d_j$, with early termination if $T_j \leq a_j$, if $b_j < T_j < c_j$, or if $T_j \geq d_j$. Designs with fewer than four early stopping boundaries can be obtained by setting $a_j = -\infty$, $b_j = c_j$, or $d_j = \infty$, as appropriate for the setting. Ensuring that at the Jth analysis $a_J = b_J$ and $c_J = d_J$ guarantees termination of the study. Group sequential clinical trial design typically involves choosing stopping boundaries which will maintain a desired type I error, and choosing a maximal statistical information N_J such that the study will have adequate power to detect a specified design alternative under a schedule of analyses occurring when statistical information is $N_1, N_2, \ldots, N_J$.

It should be noted that N_J can be specified in units of some unknown variance of individual observations, in which case only the maximal sample size is pre-specified. Alternatively, the maximal statistical information is pre-specified and the actual maximal sample size is determined using estimates of the variance observed during the conduct of the clinical trial. Emerson [6] discusses further the scientific and statistical validity of both approaches to the pre-specification of the rule for determining maximal statistical information in the

setting of group sequential clinical trials. In the analysis of time-to-event data that are potentially right-censored, it is most common to design a clinical trial based on the maximal statistical information, which is proportional to the number of events. The number of subjects required is then determined based on accrual rate, accrual time, and the distribution of failure times.

By way of example, we will consider a group sequential design appropriate for testing for a lessened instantaneous risk of death (decreased hazard) when administering some new treatment or placebo to a population of severely ill patients. We will let $\theta(x)$ denote the hazard ratio $\frac{h_T(x)}{h_C(x)}$ where $h_T(x)$ is the hazard at time x for the treatment arm, and $h_C(x)$ is the hazard at time x for the control arm. We use a proportional hazards model, meaning that we assume that $\theta(x) \equiv \theta$ is constant for all values of x.

As noted previously, we consider a setting that is a slight modification of a design proposed for an industry sponsored clinical trial. We start with an initial simple group sequential design, Design A, with two analysis points, at 100 and 200 events. The test statistic for this design is the estimated hazard ratio, and we desire a statistical sampling plan based on a one-sided level 0.025 type I error to discriminate between a null hypothesis of no effect (a hazard ratio of 1.0) and a design hypothesis of improved survival on the treatment arm relative to the control arm (hazard ratio less than 1.0).

2.1 Specification of the initial group sequential design: Design A

In selecting a stopping rule to be used as a guideline for early termination of the clinical trial, we take the common approach of selecting an efficacy (lower) boundary that is relatively conservative at the earliest interim analyses. The motivation for such early-conservatism is that the standards of evidence for adoption of a new, unproven treatment dictate that the drawbacks of having less available data to examine safety, longer term survival, and other secondary endpoints need to be counterbalanced by a marked benefit on survival over the shorter period of observation prior to the first interim analysis. That is, with lesser follow-up, we need to be confident of a highly effective treatment, and we should focus on stopping rules that would stop early only if, say, a 95% confidence interval includes only hazard ratios that correspond to strong effect of the new treatment. We note that it is common for clinical trialists to focus instead on the criterion that early stopping should occur only if we are highly confident of an effective treatment, e.g., perhaps a 99.9% confidence interval that excludes 1.0. However, such a criterion may not be measuring the most important scientific, clinical, and ethical issues that relate to ensuring that early-occurring effects are of sufficient benefit to outweigh uncertainty about safety and long-term effects.

We also consider the specification of a futility (upper) boundary that would correspond to a decision that the treatment effect is not sufficiently beneficial to warrant continued study. The advantages of early termination of a study for futility are that it avoids continued exposure of patients to an unproven therapy that is unlikely to be adopted and that it avoids continued delay of investigating other, more promising therapies. If there is no important secondary information that can be obtained from studying a therapy that we have with high confidence determined is not associated with a clinically important effect, there is also no need to build in the early-conservatism desirable for the efficacy boundary. Hence, a futility boundary might be chosen to afford more efficiency.

In concordance with the above criteria, we consider the popular O'Brien and Fleming [7] boundary relationship for the efficacy boundary, but we choose a less conservative, more efficient Pocock [8] futility boundary. In considering the efficiency of the stopping boundary, we follow the most common approach based on examination of the ASN.

In the unified family described by Kittelson and Emerson [5], there are typically several different parameterizations that lead to nearly the same stopping boundary or, in the case of a stopping rule with a maximum of two analyses, the exact same boundary. We prefer parameterizations that maintain the same level of confidence when rejecting the null hypothesis (when setting the efficacy boundary) and rejecting the design alternative (when setting the futility boundary). Hence, in this case we choose to define the boundary to reject the design alternative with 97.5% power. We further tend to restrict attention to boundary shapes within the extended Wang and Tsiatis [9] family. In the unified family, these boundary shape functions have parameters $A = 0$ and $R = 0$, with P allowed to vary. In this setting, the P parameter measures

the early-conservatism of the boundary, with $P = \infty$ corresponding to no early stopping. Two important cases within this family are the Pocock [8] boundary shape function when $P = 0.5$ and the O'Brien and Fleming [7] boundary shape function when $P = 1$. Emerson and Fleming [10] found that boundary shape functions similar to the Pocock [8] boundaries tended to nearly minimize the ASN under the hypothesis being accepted. That is, the ASN-efficient efficacy boundary shape for rejecting the null when the design alternative is true is close to a Pocock boundary, as would be the ASN-efficient futility boundary shape for rejecting the design alternative when the null is true. For the purposes of this manuscript, we chose a futility boundary that corresponds to the Pocock $P = 0.5$ boundary.

Using S+SeqTrial [11], we compute the design boundaries to have type I error $\alpha = 0.025$, giving boundaries at the first analysis of $a_1 = 0.5792, b_1 = c_1, d_1 = 0.8645$. The study stops at the first analysis (100 observed events) if the estimated hazard ratio $T_1 \leq a_1 = 0.5792$ (in which case the null hypothesis $\theta \geq 1.0$ is rejected) or $T_1 \geq d_1 = 0.8645$ (in which case the null hypothesis $\theta \geq 1$ is not rejected). At the second (and maximal) analysis, a result corresponding to $T_2 \leq d_2 = 0.7611$ would correspond to a rejection of the null hypothesis. With these boundary shape parameters, a design allowing analyses at 100 and 200 observed events provides 97.5% power to detect a hazard ratio of $\theta_1 = 0.5596$. Hence, the futility boundary can be interpreted based on a rejection of that design alternative, just as the efficacy boundary rejects the null hypothesis of a hazard ratio of 1.0. A more detailed examination of the power curve reveals that this design has 80% power against a hypothesis $\theta_1 = 0.6646$.

As noted above, the way in which we specified the stopping boundaries allowed greater interpretation of the study design parameters. Equal statistical errors (type I under the null and type II under the alternative) are chosen in order to have the trial design indicate the hypotheses that would be discriminated by a 95% confidence interval: At the end of the study, a 95% confidence interval for the true hazard ratio has probability 0 of including both the null hazard ratio of $\theta_0 = 1.0$ and the design alternative hazard ratio of $\theta_1 = 0.5596$. Were it to be the case that larger sample sizes were not feasible, the design alternative takes on the role of the *de facto* minimal clinically important difference in the sense that any smaller difference will not be detected with the prescribed confidence. The choice of an O'Brien-Fleming efficacy boundary is recognizable as providing early conservatism in a decision to reject the null, and the choice of a futility boundary closer to a Pocock boundary is recognizable as providing greater efficiency (and therefore perhaps better addressing ethical issues) in a decision to reject the minimal clinically important difference when the null hypothesis is true. Furthermore, with the symmetric type I and type II errors, there are generalizable approximate relationships between the stopping boundaries at each analysis and the hypotheses being rejected by the respective boundaries. For instance, halfway through accrual of the maximal statistical information in a group sequential trial, the stopping boundary under an O'Brien-Fleming boundary shape is approximately equal to the hypothesis being rejected. For a Pocock boundary, this relationship holds when one-quarter the way through accrual. (These sorts of relationships hold exactly when the boundary shape functions, as well as the statistical errors, are the same for both the efficacy and futility boundaries.)

However, we could have arrived at an almost identical boundary by choosing statistical power of 80% (a type II error of 0.2), and an O'Brien-Fleming boundary shape function for the futility boundary "rejecting" (with only 60% two-sided confidence) an alternative corresponding to a hazard ratio of $\theta_1 = 0.6652$. It should be noted that under this parameterization, we do not have the same benefit of interpretation we had when we chose identical type I and type II errors. A 95% confidence interval will not necessarily discriminate between the null and alternative hypotheses. In fact, if the alternative hypothesis is true, there is a 17.5% probability that the 95% CI computed at the end of the trial will include both the null and alternative hypotheses. Furthermore, the interpretation of the "early-conservatism" of the O'Brien-Fleming is also muddled: Though the O'Brien-Fleming boundary shape function is proposed for both the efficacy and futility boundaries, the relationship between the stopping boundaries at the first analysis and the hypotheses used to define the boundaries are not comparable, because the magnitude of statistical errors varies. There is also no similarity of the boundaries on the scale of the error spending function. For instance, in terms of the error spending function, the O'Brien-Fleming efficacy boundary spends 12.5% of the type I error (or 0.0031 out of the total error of 0.025) at the first analysis. On the other hand, the O'Brien-Fleming futility boundary spends 45.0% of the type II error (or 0.09 out of the total error of 0.2) at the first analysis.

(Emerson et al. [4] further elaborate on the difficult correspondence between error spending functions and O'Brien-Fleming and Pocock boundary shape functions.)

2.2 Specification of adaptive design

Next we consider an adaptive design to increase the power of Design A at various effect sizes θ. The basic idea is that at the first interim analysis we might 1) observe results so extreme that the stopping rule defined by Design A would suggest early termination of the study, 2) observe results that did not exceed a stopping boundary, but were close enough to the boundary as to suggest the eventual decision reached at the next analysis, or 3) observe results that were sufficiently far from our expectations that additional data might be desired to increase the power to obtain a statistically significant result. The form of adaptive design that we consider throughout is a prespecified adaptation based on an interim estimate of the effect size, and the statistic used in the specification of the stopping rule is a sufficient statistic. This differs from the adaptive designs described in, for example, Proschan and Hunsberger [12] and Cui, Hung, and Wang [13], which are not necessarily prespecified and are not based on a sufficient statistic.

Let Design B be an extension of Design A, constrained to have the same boundary as Design A at the first analysis, but with three analyses at 100, 200, and 300 events. The same parameters used in construction of Design A are used to define the boundaries at the 2nd and 3rd analyses for Design B: the test should be of level $\alpha = 0.025$ with an O'Brien-Fleming efficacy boundary and a Pocock futility boundary. Such an approach is easily performed using the constrained boundary approach described by Burington and Emerson [14] and implemented in S+SeqTrial.

The adaptive design that we consider performs an analysis at 100 subjects just as in Design A. We define parameters A and D such that $a_1 \leq A \leq D \leq d_1$ to be the values that define the adaptive behavior of the design. Based on the statistic at the first analysis, T_1, the design either proceeds with Design A, or switches to Design B$^*(A, D)$, a version of Design B that is slightly modified to maintain the overall experiment-wise type I error. If $T_1 \in (a_1, A)$ or $T_1 \in (D, d_1)$ then the Design A boundary is used for the remainder of the study. If, however, $T_1 \in (A, D)$, Design B$^*(A, D)$ is used, and the maximal possible sample size is therefore increased from 200 to 300. Design B$^*(A, D)$ has the same specifications and constraints as Design B, except that the type I error $\alpha^*(A, D)$ is chosen based on the values of A and D in order to guarantee the desired experiment-wise error for the adaptive procedure. This modified type I error and design can be found using S+SeqTrial as described in the Appendix. Figure 1 illustrates the boundaries corresponding to this adaptive design procedure.

The parameters A and D which define the adaptive behavior can be chosen based on several different criteria, such as conditional power or symmetry considerations. Note that if $A = D$, the adaptive design reduces to Design A, and if $A = a_1$ and $D = d_1$, the adaptive design reduces to Design B. As shown in Figure 2, different choices of A and D move the power curves and ASN curves of the resulting adaptive designs smoothly between the power curve of Design A and the power curve of Design B. In our investigations, we consider a full range of possible A and D values to explore the space of adaptive designs defined in this way.

2.3 Specification of a nonadaptive group sequential design for comparison

Let $ASD(A, D)$ denote the adaptive sequential design resulting from a particular choice of A and D. For fixed parameters A and D we then seek to identify a comparable, nonadaptive group sequential design $GSD(A, D)$. The spectrum of group sequential designs is quite extensive and flexible, and, as noted by Tsiatis and Mehta [1] and Jennison and Turnbull [2], it is in general possible to find a group sequential design that matches the efficiency of an adaptive design. Our goal, however, is to determine the extent to which a traditional group sequential design may equally well satisfy the sponsor's constraints, which extend beyond the usual statistical power and sample size considerations. As such, these considerations, rather than error-spending functions, guide our choice of competing group sequential designs to consider.

To that end we presume that the stopping boundaries at the first interim analysis (when $N_1 = 100$ events) were chosen to guarantee the scientific and statistical credibility of results should the study be terminated early. We further presume that the timing of possible interim analyses was fixed by logistical constraints,

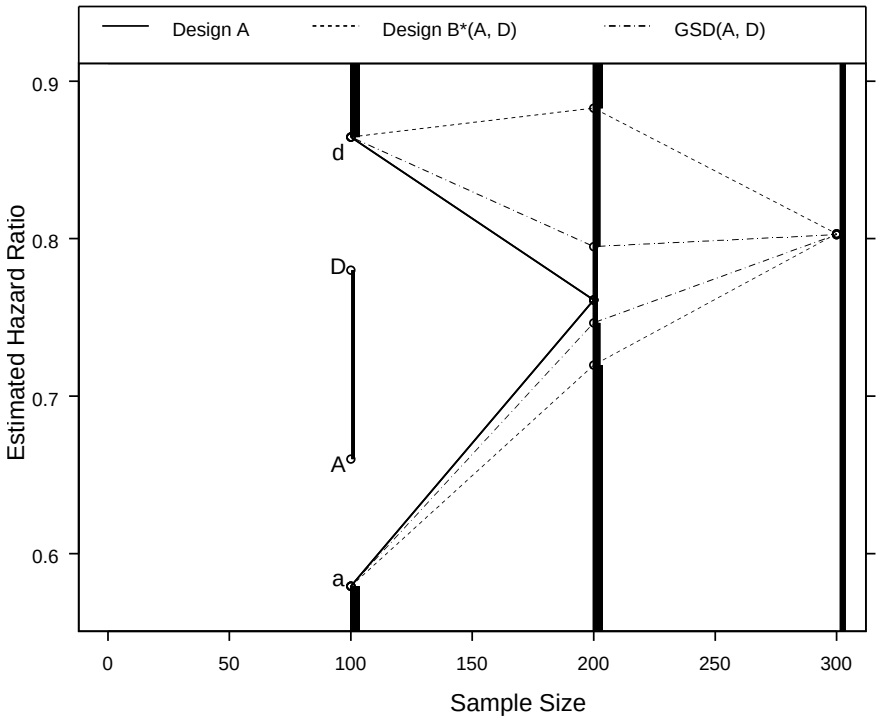


Figure 1: Stopping boundaries for the adaptive design comprised of Design A and Design B$^*(A, D)$, as well as for the group sequential design $GSD(A, D)$.

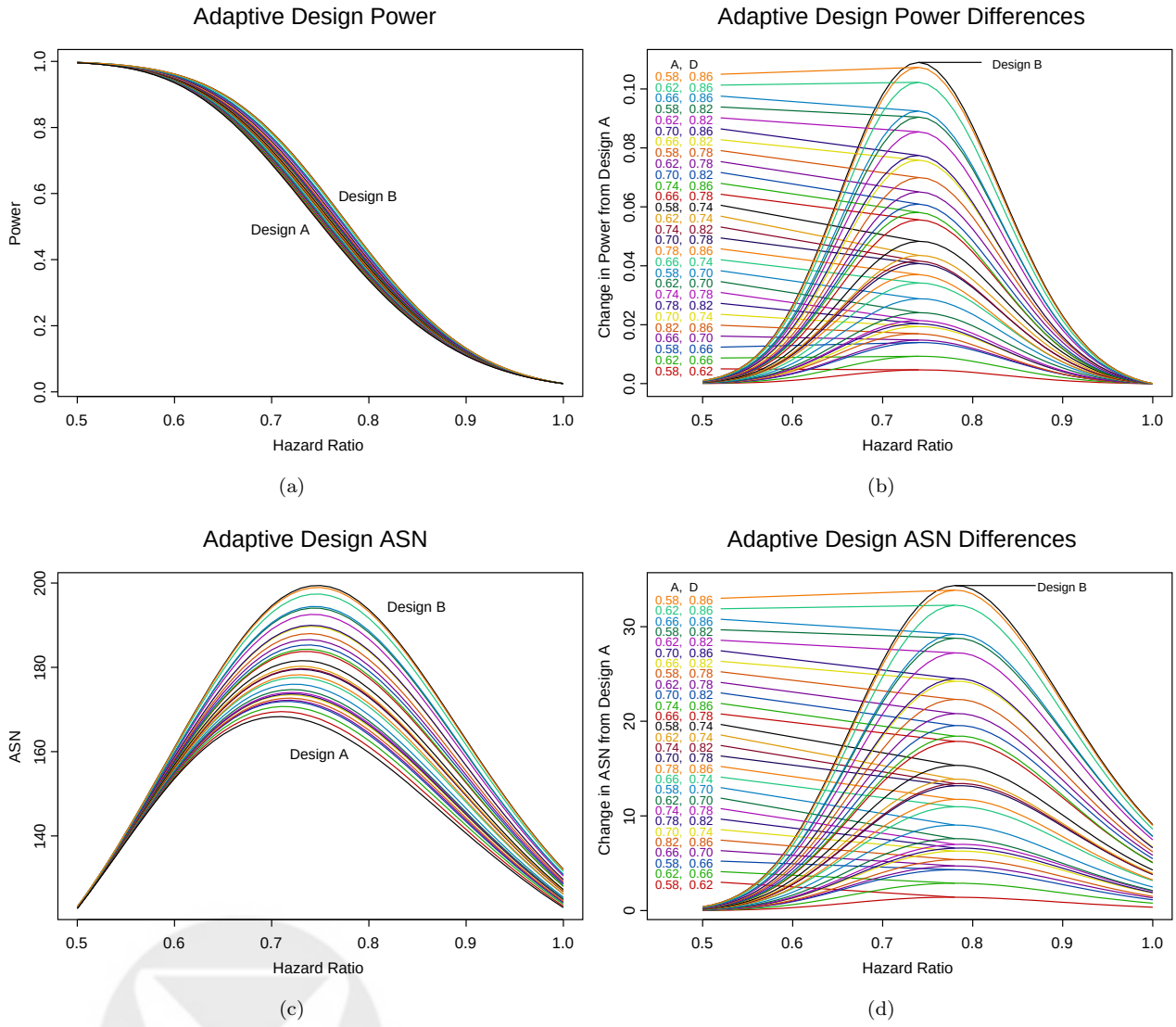


Figure 2: Operating characteristics for the range of adaptive designs considered, illustrating that the adaptive designs provide a continuous transition from the operating characteristics for Design A to those of Design B: (a) Power curves; (b) Change in power from Design A, with the corresponding A and D adaptive parameters given on the left; (c) ASN curves; (d) Change in ASN from Design A, with the corresponding A and D adaptive parameters given on the left.

and thus consider only group sequential tests having analyses at 100, 200, and 300 events. We also restrict attention to those group sequential test designs that have power curves that closely match the power curve of the corresponding adaptive design. That is, we want power curves $\beta_{ASD(A,D)}(\theta) \cong \beta_{GSD(A,D)}(\theta)$ for all θ in a range that includes the null and alternative hypotheses. Note that the adaptive design $ASD(A, D)$ and the matching group sequential design $GSD(A, D)$ have the same maximal possible sample size.

Even under the above constraints, there are likely many different group sequential stopping boundaries that could be considered. In searching for a group sequential design that matched the power curve of a given adaptive design $ASD(A, D)$, we found that a specially modified version of Design B worked remarkably well. We defined $GSD(A, D)$ to be a standard group sequential design with the same boundaries as Design B at the first and third analyses ($N_1 = 100$ events and $N_3 = 300$ events). Then we modified the boundary at the second analysis by changing the value of the P parameter for the design to be zero for the efficacy boundary ($P_a = 0$), and to be some appropriately chosen positive number (in the range 0.05 - 1.25 for the examples we consider) for the futility boundary ($P_d = P^*(A, D)$). This modification effectively shrinks the stopping boundary at the second analysis in from that of Design B toward the boundary of Design A, and the value of $P^*(A, D)$ controls the degree of shrinkage: smaller values of $P^*(A, D)$ result in boundaries closer to Design A, while larger values of $P^*(A, D)$ result in boundaries closer to Design B. An example with further details is provided in the Appendix. Figure 1 displays one example of the resulting group sequential design boundary, $GSD(A, D)$.

In the next section we explore the behavior of this adaptive design and matching group sequential design in the (unrealistic) setting of no censoring, and find no advantage to this adaptive design. With no censoring, the statistical information is proportional to the number of subjects accrued. As such, the same results will hold for comparisons of means or differences of proportions. Then in the following section we add in censoring and explore the tradeoffs of number of subjects versus calendar time, where we do find instances in which the adaptive design exhibits some advantages over a traditional group sequential design.

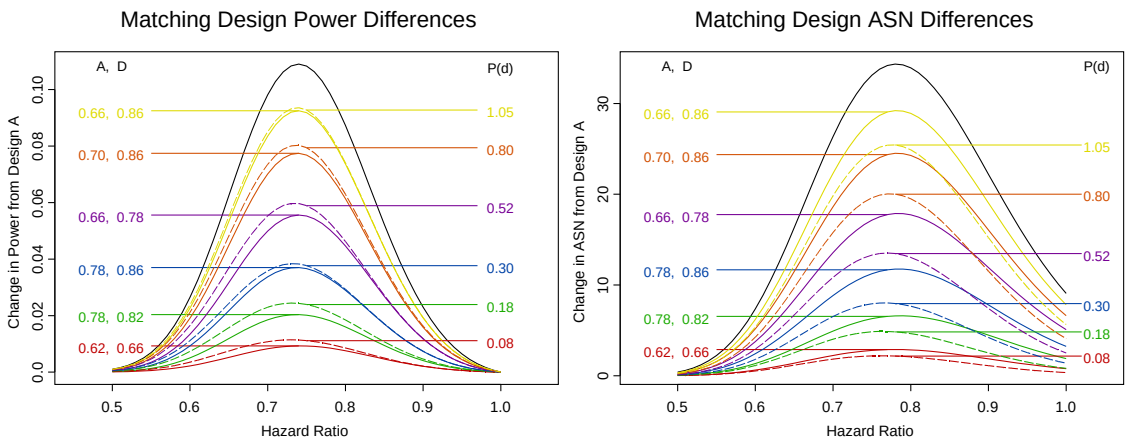
3 No censoring

In Figure 3 we show power and ASN comparisons for a selection of six adaptive designs and the corresponding matched group sequential designs. As these figures show, the group sequential designs $GSD(A, D)$ are more efficient than the adaptive designs $ASD(A, D)$ in terms of ASN, with equal or slightly superior power across the range of true effect size considered. Similar results were obtained as we explored the behavior of the adaptive designs and the group sequential designs over the complete range of possible values of A and D. Thus in settings where information is measured by number of subjects, these group sequential designs are observed to be uniformly superior to the corresponding adaptive designs over the range of alternatives that would typically be considered during design of the study. In explorations not shown here, these results were found to generalize to clinical trial settings using means or binomial proportions.

4 Censoring

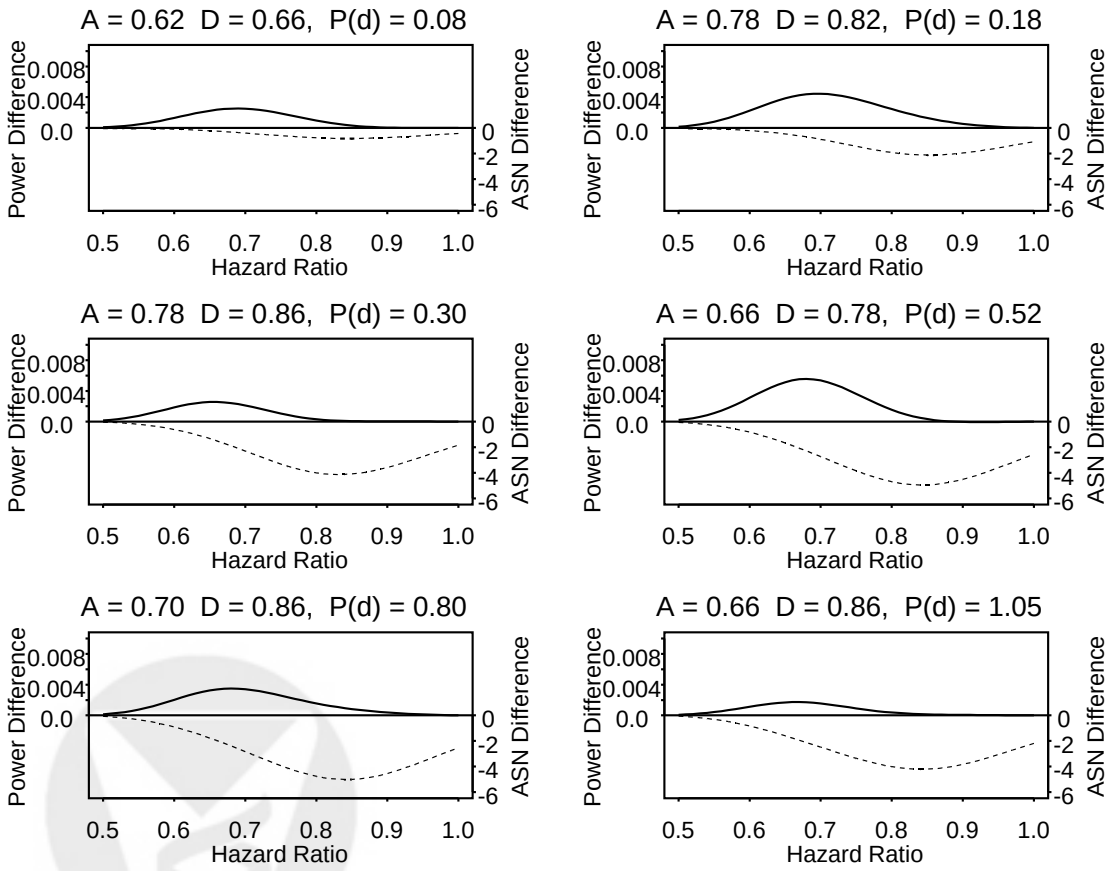
Censoring occurs when we accrue subjects and only follow them for a certain amount of time, as opposed to following them indefinitely until an event occurs. Generally the amount of follow-up time is determined by a certain date at which the study ends, at which time subjects who have not yet had an event are censored. Note that in this administrative censoring scenario the follow-up time for individual subjects may differ, depending on when they were accrued to the study.

In a setting with censoring, we now have to consider the costs associated with accrual of subjects and follow-up time. Follow-up time can be reduced by accruing more subjects, and conversely, the number of subjects required may be reduced by extending the follow-up time. Here the adaptive design has the advantage of allowing accrual to stop earlier when it is determined that the maximum number of events needed is only 200 instead of 300. We explore the behavior of study duration versus number of subjects required under a variety of accrual patterns and event rates. Following Schoenfeld and Richter [15], as



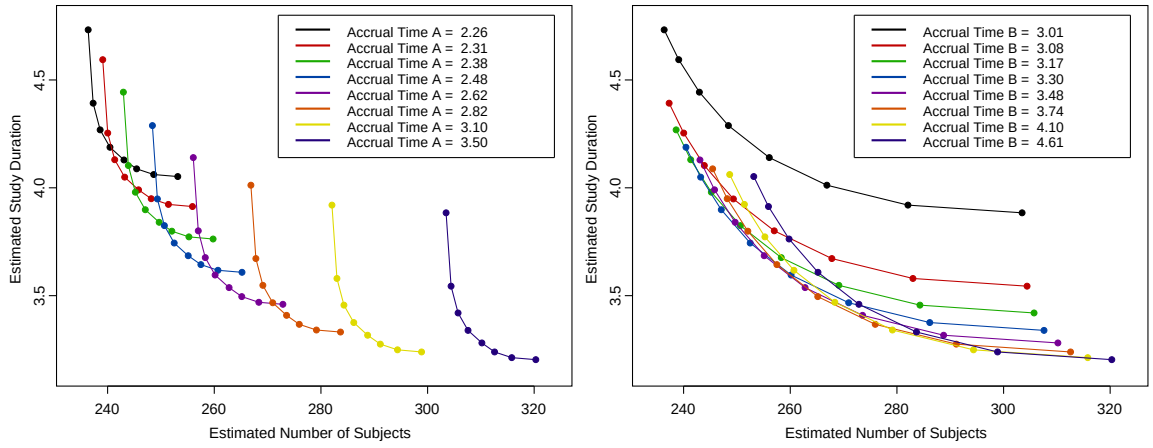
(a)

(b)



(c)

Figure 3: Matched designs $ASD(A, D)$ and $GSD(A, D)$ operating characteristics comparison: (a) Change in power from Design A, matched designs shown in same color with a solid line for $ASD(A, D)$, and a dashed line for $GSD(A, D)$; (b) Change in ASN from Design A, matched designs shown in same color with a solid line for $ASD(A, D)$, and a dashed line for $GSD(A, D)$; (c) Difference in power and ASN ($GSD(A, D) - ASD(A, D)$) for matched pairs. Note that we were able to find group sequential designs that have higher power and smaller ASN over the range of alternatives that would typically be considered in the design of a clinical trial.



(a) (b)

Figure 4: Example number of subject versus study duration plots for adaptive designs, with the same points represented in both panels. Note the different effects of changing the accrual times for Design A and Design B, respectively: (a) Connected points have constant accrual time for Design A, t_A, with t_A increasing from left to right; (b) Connected points have constant accrual time for Design B$^*(A, D)$, t_B, with t_B increasing from top to bottom.

implemented in S+SeqTrial [16], the accrual patterns we consider have a constant number of subjects accrued per time unit, and event rates are modeled as exponential with $h_T = \theta h_C$ where h_T is the hazard rate for the treatment arm, h_C is the hazard for the control arm, and θ is the hazard ratio.

To reduce the dimension of the space of parameters that we consider, we fix the control group event rate to have a median of 1. Note that this reduction still allows us to explore the complete space of accrual patterns and event rates, as it is equivalent to changing the unit of time used. For instance, an accrual rate of 20 patients per month with a median control event time of 6 months is equivalent to an accrual rate of 120 patients per six-month period with a median control event time of 1 six-month period. Similarly, any combination of accrual rate r and median control event time m_C expressed in time units u may also be expressed as an accrual rate of $r^* = rm_C$ and a median control event time of 1 in time units $u^* = um_C$. We choose to explore accrual rates in $\{40, 60, 100, 150, 200, 250\}$ as a reasonably comprehensive representation of possible accrual scenarios.

To compare the behavior of the adaptive designs $ASD(A, D)$ to the matching group sequential designs $GSD(A, D)$ under a particular accrual rate r and effect size θ, we considered the range of possible accrual times and the resulting estimated number of subjects and trial duration. We will assume that if accrual ends before the first analysis then the accrual time t must have been sufficient to obtain at least 300 subjects to ensure that analyses at 300 events will be possible (otherwise, if fewer than 300 subjects were accrued and if the Design B$^*(A, D)$ boundary were selected at the first analysis, it could prove necessary to restart accrual–a practice that is generally avoided). In this case, the adaptive design clearly offers no benefit of curtailed accrual, and is therefore slightly less efficient than the matching group sequential design. In the setting we consider here, accrual rates higher than 250 subjects per unit time were not explored, as they tend to result in accrual ending before the first analysis. For the adaptive designs, if accrual continues beyond the first analysis, we must consider two independent accrual times t_A and t_B depending on which boundary is adaptively chosen for the later analyses. Each combination of t_A and t_B produces an estimated number of subjects and trial duration. The factors involved in choosing accrual times t_A and t_B for the adaptive design are:

- Each of t_A and t_B are constrained (as discussed above) to be larger than the time of the first analysis.

We consider the case of accrual finishing before the first analysis separately.

- t_A and t_B are required to be large enough to obtain at least 200 and 300 subjects respectively, in order to ensure that we will be able to observe the necessary number of events.
- The maximum accrual time considered for each design was chosen to be the accrual time that resulted in zero follow-up time after the end of accrual for that design. Accrual times greater than this maximum would be pointless, as this would mean accruing subjects after the study finished.
- Since accrual is fixed such that it never ends before the first analysis, we can freely decide which combination of accrual times we will use to achieve the adaptive design. Therefore we consider all possible combinations of accrual times for Design A and accrual times for Design B.

We explored a range of 20 values for each of t_A and t_B subject to the above constraints. For a given combination of accrual times t_A and t_B we calculate the estimated number of subjects and estimated study duration as follows. Define the following quantities:

$$
\begin{aligned}
r &= \text{ rate of accrual} \\
\\
p_1 &= \text{ probability of stopping at the first analysis} \\
p_{2_A} &= \text{ probability of stopping at the second analysis, using Design A} \\
p_{2_B} &= \text{ probability of stopping at the second analysis, using Design B} \\
p_3 &= \text{ probability of stopping at the third analysis} \\
\\
\tau_1(t_A) = \tau_1 &= \text{ estimated time of first analysis} \\
\tau_{2_A}(t_A) = \tau_{2_A} &= \text{ estimated time of second analysis for Design A} \\
\tau_{2_B}(t_B) = \tau_{2_B} &= \text{ estimated time of second analysis for Design B} \\
\tau_3(t_B) = \tau_3 &= \text{ estimated time at third analysis} \\
\\
S &= \text{ number of subjects accrued} \\
T &= \text{ total study duration}
\end{aligned}
$$

Then the expected number of subjects S, and the expected trial duration T, for the adaptive design with accrual times t_A and t_B are given by:

$$
\begin{aligned}
\mathbf{E}[S] &= [p_1 \times \tau_1 \times r] + [p_{2_A} \times t_A \times r] + [p_{2_B} \times \min(\tau_{2_B}, t_B) \times r] + [p_3 \times t_B \times r] \\
\mathbf{E}[T] &= [p_1 \times \tau_1] + [p_{2_A} \times \tau_{2_A}] + [p_{2_B} \times \tau_{2_B}] + [p_3 \times \tau_3]
\end{aligned}
$$

Figure 4 illustrates an example of a plot resulting from these calculations. Each dot in the figures represents the expected number of subjects and expected trial duration resulting from one combination of t_A and t_B. In the left panel, contours connect points corresponding to a constant choice of t_A; in the right panel, contours connect points corresponding to a constant choice of t_B.

The number of subjects versus study duration curves for the matching group sequential designs $GSD(A, D)$ were similarly obtained by considering a range of accrual times t_G:

- The minimum accrual time for t_G is required to be large enough to achieve the maximum possible number of events (300).

- The maximum accrual time considered was chosen to be the accrual time that resulted in zero follow-up time after the end of accrual.

A range of 30 accrual times were considered between the minimum and maximum accrual times for t_G. For a given accrual time t_G we calculate the expected number of subjects and expected duration of study as follows. Define the following quantities:

$$
\begin{aligned}
r &= \text{ rate of accrual} \\
\\
p_1 &= \text{ probability of stopping at the first analysis} \\
p_2 &= \text{ probability of stopping at the second analysis} \\
p_3 &= \text{ probability of stopping at the third analysis} \\
\\
\tau_1(t_G) = \tau_1 &= \text{ estimated time of first analysis} \\
\tau_2(t_G) = \tau_2 &= \text{ estimated time of second analysis} \\
\tau_3(t_G) = \tau_3 &= \text{ estimated time of third analysis} \\
\\
S &= \text{ number of subjects accrued} \\
T &= \text{ total study duration}
\end{aligned}
$$

Then the expected number of subjects S, and the expected trial duration T, for the group sequential design with accrual time t_G are given by:

$$
\begin{aligned}
\mathbf{E}[S] &= [p_1 \times \min(\tau_1, t_G) \times r] + [p_2 \times \min(\tau_2, t_G) \times r] + [p_3 \times t_G \times r] \\
\mathbf{E}[T] &= [p_1 \times \tau_1] + [p_2 \times \tau_2] + [p_3 \times \tau_3]
\end{aligned}
$$

The estimated analysis times are computed using the S+SeqTrial function `seqPHSubjects`, which computes the expected number of events observed by a time τ, and then solves for τ_j such that the expected number of events observed by time τ_j is N_j. Note that this is very slightly different from the expected time at which N_j events are observed, but the differences are of an insignificant order of magnitude in the examples we are considering, so we will interchangeably use the phrases "estimated study duration" and "expected study duration", and similarly "estimated number of subjects" and "expected number of subjects".

Figures 5 and 6 demonstrate some of the possible relationships between the number of subjects needed and trial duration for the adaptive and matching group sequential designs. Figure 5 presents results for the adaptive design with $A = 0.62, D = 0.66$, for accrual rates of 60 and 200 in panels 5(a) and 5(b) respectively. Figure 6 shows results for the adaptive design with $A = 0.70, D = 0.86$, for accrual rates of 40 and 250 in panels 6(a) and 6(b) respectively. In each figure, the black dots correspond to results for the adaptive design, and the solid green line corresponds to the matching group sequential design.

For the adaptive design and accrual rate of 60 in Figure 5(a), there is a clear potential for benefit using the adaptive design over the group sequential design. With this slow accrual rate, there is a limited range for t_B (and also for t_G), and thus the dots corresponding to the same value of t_B are very close together. The lines corresponding to the group sequential design are at the far right end of the plots, demonstrating the benefit of accrual modification in Design A of the adaptive design, for this scenario. The adaptive design allows the possiblity of reducing the number of subjects required by 20 to 50 depending on the effect size. Of course, there is a tradeoff of increasing the trial duration, but a moderate reduction in number of subjects does not produce a dramatic increase in study time. For instance, at a hazard ratio of $\theta = 0.7$, the number of subjects can be reduced from a minimum of 263 for the group sequential design to 230 for the adaptive

design, while only increasing the expected study duration from 4.5 to 4.7 time units. As the accrual rate increases, the benefit of the adaptive design becomes less pronounced, and eventually disappears.

At an accrual rate of 200 subjects per time unit (median survival time on the control arm), the same adaptive design shows much less potential for subject reduction. Depending on the treatment effect and the relative costs of time versus number of subjects, either the adaptive or the group sequential design may be preferable. For strong treatment effect, e.g., a hazard ratio of $\theta = 0.5$, the group sequential design is very slightly superior. Also, when study duration is the primary concern, the group sequential design dominates across the range of treatment effect. However, there are points in Figure 5(b) that still demonstrate a marginal benefit to using the adaptive design. When the hazard ratio is $\theta = 0.8$, the group sequential design requires an average of 2.285 time units when the expected number of subjects is 297.31. In comparison, the adaptive design requires an average of only 2.185 time units when the expected number of subjects is 297.27. Clearly these are not dramatic differences, but for this pair of designs, the adaptive design seems to provide more flexibility in number of subjects required at some treatment effect sizes.

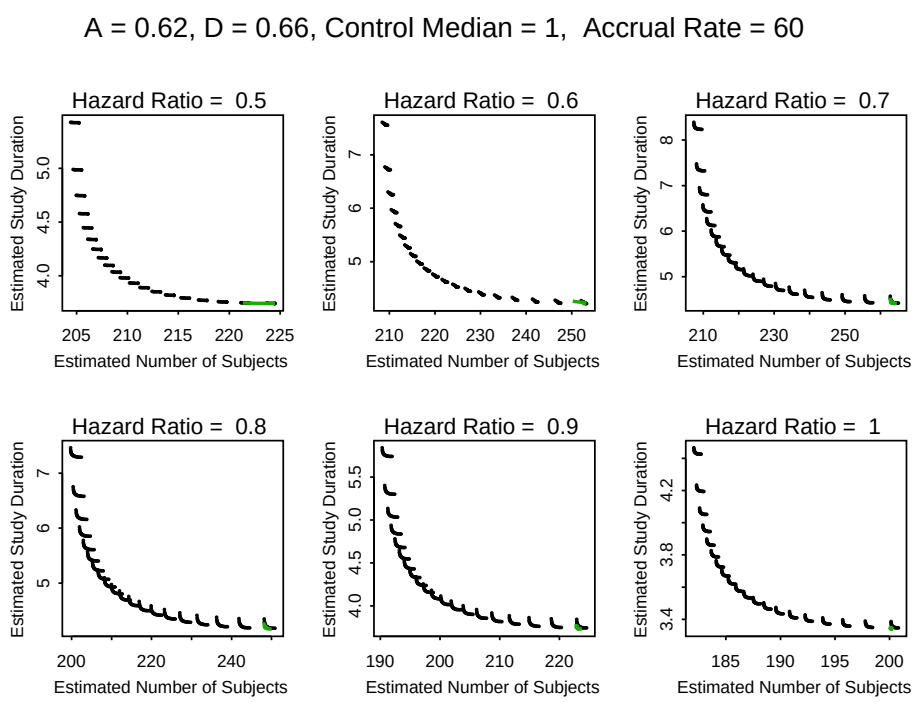
The adaptive design and accrual rates presented in Figure 6 illustrate a rather different scenario. For this pair of designs there is no clear benefit to using the adaptive design over the group sequential design. For example, in Figure 6(a) when the accrual rate is 40, we can see that for hazard rates of $\theta = 0.8$ and higher, the group sequential design may allow a shorter trial duration for the same number of subjects. As the accrual rate increases to 250 in Figure 6(b) for this design, the curves for the group sequential design and adaptive design become quite close, with the group sequential design dominating.

Clinical trialists would choose from among the spectrum of adaptive designs considered in Figure 2 based on the efficiency and power curves desired. We present the two examples in Figures 5 and 6 to demonstrate the patterns of behavior resulting from the range of adaptive designs and accrual scenarios, with similar trends observed for other choices of A, D, and accrual rates. The following general trends were observed: As accrual rate increases, the difference between an adaptive design and the corresponding group sequential design tends to disappear, with the group sequential design tending to be slightly more efficient. For lower accrual rates, the ability of either the adaptive design or the group sequential design to improve upon the other, in terms of expected trial duration at a given expected number of subjects, will depend on the choices of A, D, and the effect size. More generally, there are tradeoffs between the adaptive designs and the matching group sequential designs that depend on the relative importance of minimizing the number of subjects versus minimizing trial duration. In the next section, we attempt to explore these tradeoffs.

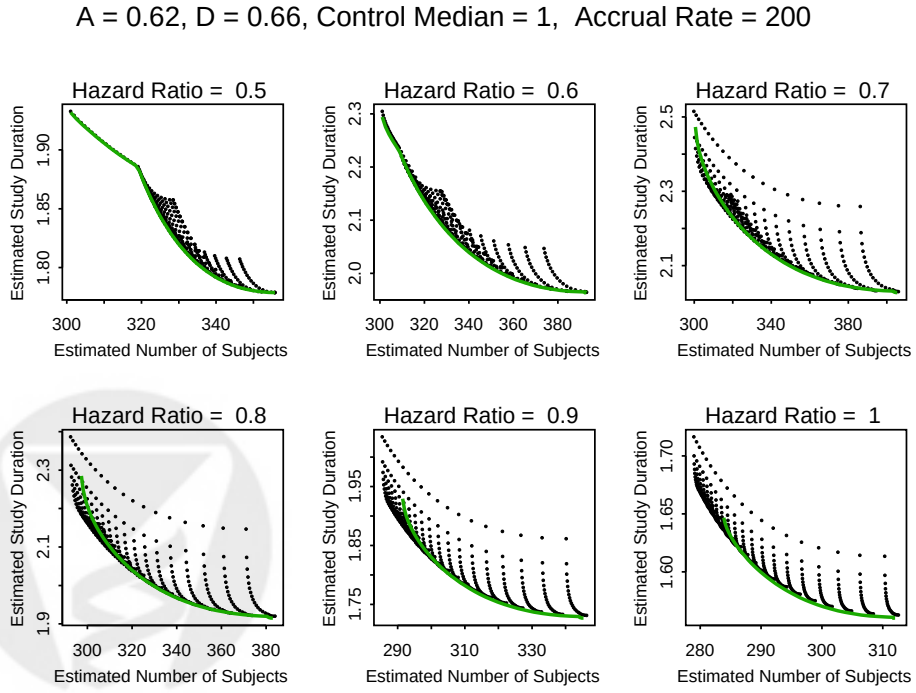
5 Cost Estimation

In order to explore the tradeoffs between increased sample size and decreased study duration, we consider the cost to the sponsor using a simple discrete time economic model. We presume the setting of the design of a Phase III clinical trial. At the start of the trial, the sponsor will have incurred costs related to treatment development and early phase clinical trials, and there are costs to the sponsor associated with the money invested in that development program. For instance, the cost of prior development might total \$10 million. Then, the Phase III trial might engender costs on the order of, say, \$10,000 per patient. In our simple model, we consider the cost of that prior investment, as well as the cost of the patients accrued. We then further consider the cost of study duration by allowing for interest to be paid by the sponsor on its investment. Letting n_t represent the number of subjects accrued between time $t-1$ and time t and letting p be the per patient costs, we can then calculate the total cost $C(t)$ up to calendar time t as $C(t) = n_t \times p + (1+\omega) \times C(t-1)$, where ω is the cost of money per unit time. Without loss of generality, it is sufficient for us to consider merely the ratio $C(0)/p$. The interest rate is used to represent the cost of time, i.e., study duration, and may serve as a surrogate for all time-related costs such as the expenses related to maintaining databases, personnel, and borrowing money. The ratio of the prior costs to the per-patient costs determines the direction of the tradeoff between trial duration and number of subjects. For relatively higher per-patient costs, the total trial cost is minimized when fewer patients are used and a longer study duration is permitted.

With this cost model, for each design and each accrual scenario, we can calculate the ratio of the optimal trial cost for $GSD(A, D)$ to the optimal trial cost for $ASD(A, D)$ for a range of θ values. As an example, we

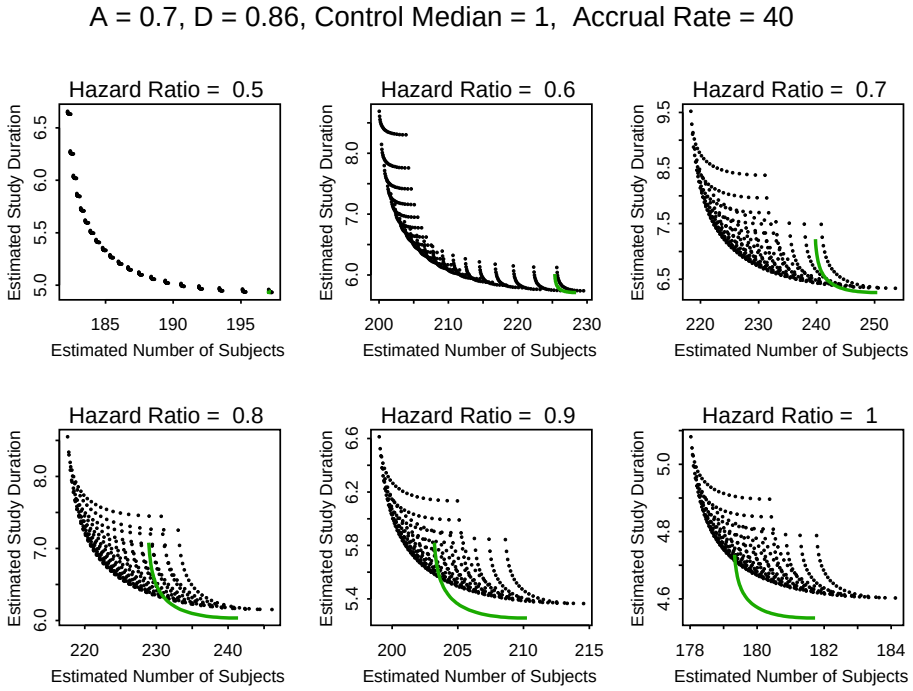


(a)

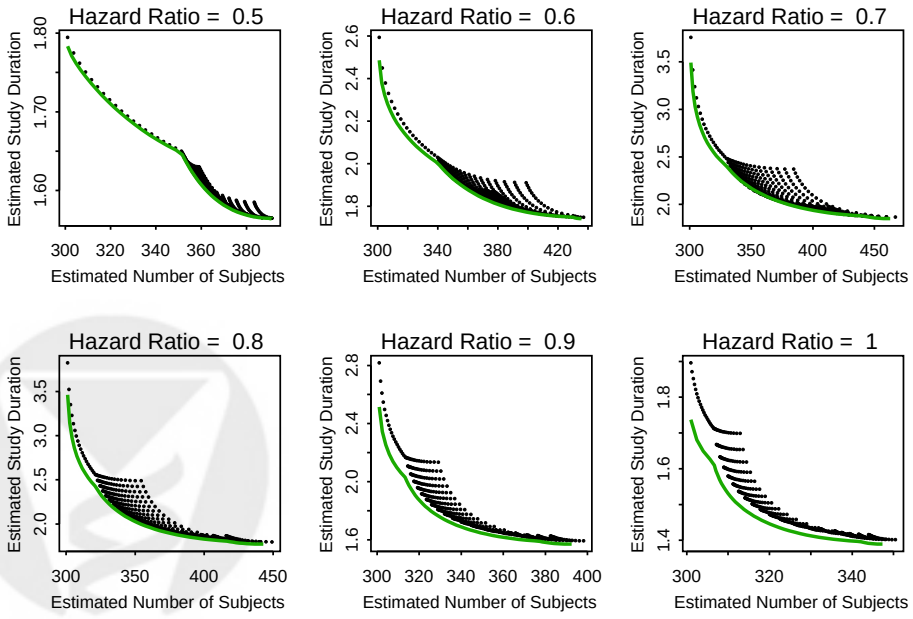


(b)

Figure 5: Number of subjects versus study duration for the adaptive design with $A = 0.62, D = 0.66$ and the matching group sequential design, under two different accrual rates. The black points represent adaptive design results, and the green line represents the group sequential design results: (a) Accrual rate = 60 subjects per unit time; (b) Accrual rate = 200 subjects per unit time.



(a)



(b)

Figure 6: Number of subjects versus study duration for the adaptive design with $A = 0.70, D = 0.86$ and the matching group sequential design, under two different accrual rates. The black points represent adaptive design results, and the green line represents the group sequential design results: (a) Accrual rate = 40 subjects per unit time; (b) Accrual rate = 250 subjects per unit time.

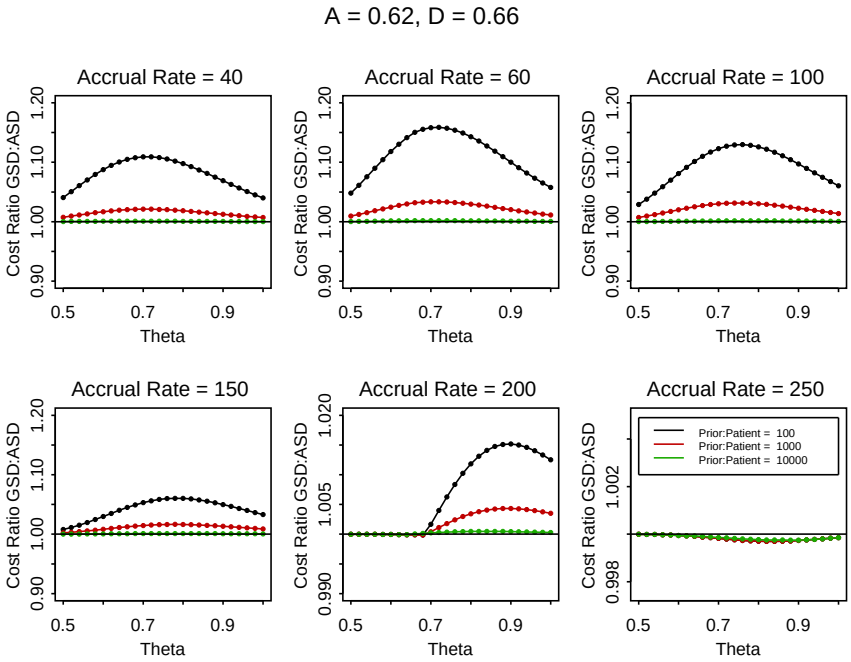
consider a time cost of money based on $\omega = 0.005$, which if the unit of time is one month would correspond to a 6% yearly interest rate. Higher interest rates tend to favor shorter trial duration, which would cause the group sequential designs to be more advantageous, as indicated by Figures 5 and 6. The resulting cost ratio plots for the adaptive designs considered in Section 4 are shown in Figure 7. As indicated by these plots, the adaptive design typically offers some cost improvements over the group sequential design for slow accrual rates, but as the accrual rate increases the cost improvement disappears. Figure 7(a) shows the cost ratio plots corresponding to the adaptive design of Figure 5 for the full range of accrual rates considered, and for prior to patient cost ratios of 100, 1000, and 10000. The most significant benefit of the adaptive design is seen for mid-range values of the true effect size ($\theta \in (0.65, 0.8)$), when the patient costs are high (at a prior cost to per-patient cost ratio of 100), and for accrual rates near the low end of the spectrum (40 - 100 patients per unit time). In this range, the group sequential design may be 10–15% more expensive than the matched adaptive design. However, when accrual rates are high, the group sequential design is actually very slightly more cost effective than the adaptive design, saving a small fraction of a percent over the adaptive design. Figure 7(b) shows the cost ratio plots summarizing the adaptive design of Figure 6. The results for this design are similar to those of Figure 7(a), though the cost benefit of the adaptive design is somewhat attenuated. Again, for the highest accrual rate of 250 subjects per unit time, the group sequential design offers a very slight improvement over the adaptive design. We note that an anonymous referee commented that the prior to patient cost ratio may well exceed 10,000, which we acknowledge. The general trend presented here indicates that as the prior to patient cost ratios increase, the ratio of the adaptive to group sequential trial costs will tend toward 1. As the ratio of prior costs to per patient costs increases with the simple economic model considered here, the cost difference, which may be of greater importance than the cost ratio, tends to favor designs which minimize calendar time regardless of the sample size.

6 Discussion

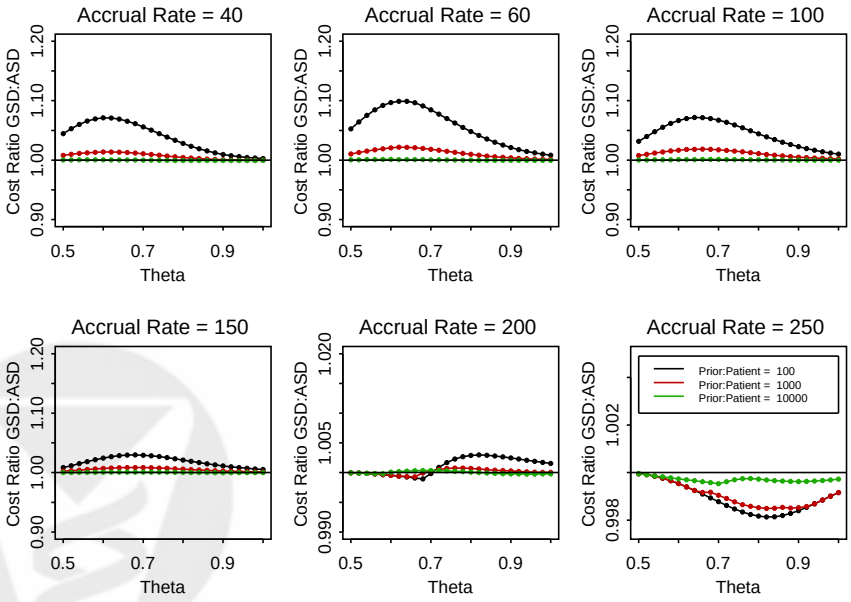
In our comparisons considered here, we compared adaptive designs of a form similar to those initially proposed for an industry sponsored clinical trial to traditional group sequential designs that might have had the same operating characteristics. There are many parameters that can be considered in a group sequential stopping rule including the type I error, the power under some design alternative, the number of interim analyses, the relative timing of the interim analyses, and boundary shape functions for each of the decisions that might be reached. The boundary shape functions can in turn be defined for any one of several different statistics: partial sum of (potentially transformed) observations, the maximum likelihood estimate, the standardized Z statistic, the fixed sample P value, the error spending function, the conditional power under some hypothesized treatment effect, the Bayesian predictive power under some prior distribution for the true treatment effect, or the Bayesian posterior probability of some hypothesis. For each of these statistics, the boundary shape function relates the early conservatism with which the boundary would allow termination of the study at the earliest analyses. In the unified family of Kittelson and Emerson [5], a user may choose from a broad spectrum of boundary shape functions through the choice of three parameters that can be chosen separately for each stopping boundary.

This wide flexibility of group sequential stopping rules means that there are likely many different group sequential designs with the same power curves and ASN curves, for instance. Hence, when evaluating the ability of adaptive designs to improve on standard group sequential methods, we must ensure that we understand the design constraints that are to be held constant, and those that are allowed to vary. In the investigations presented in this paper, we presumed that we needed to maintain the criteria for stopping at the earliest analyses, as well as the schedule and timing of interim analyses. As part of our interest was to see how easily we could match the adaptive designs, we considered only a single method of modifying the group sequential boundaries at the second analysis in order to achieve a comparable group sequential design.

While it is clear that the adaptive designs offer no advantages in uncensored settings such as when evaluating the difference of means or proportions, there are some distinct advantages to the adaptive design in certain survival analysis settings, because we may gain efficiency in the number of subjects needed and/or the calendar time required for the study to complete. Such considerations will also be relevant in trials



(a)



(b)

Figure 7: Cost ratio plots for the adaptive designs considered in Figures 5 and 6, for a range of accrual rates. Cost ratio is calculated as the ratio of the minimal cost for the group sequential design to the minimal cost for the adaptive design, where costs are estimated as described in Section 5. Three different prior to patient cost scenarios are considered, representing high patient costs (black line), mid-range patient costs (red line), and low patient costs (green line). Note that the y-axis scale changes for the two highest accrual rates. (a) Cost ratio plots comparing $ASD(0.62, 0.66)$ to $GSD(0.62, 0.66)$; (b) Cost ratio plots comparing $ASD(0.70, 0.86)$ to $GSD(0.70, 0.86)$.

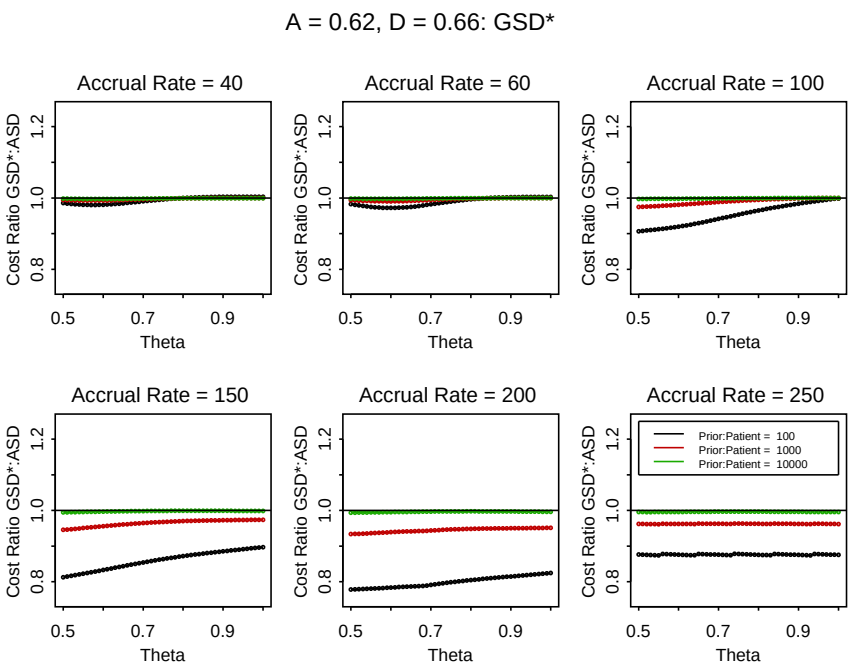


Figure 8: Cost ratio plots comparing a less-constrained, optimized group sequential design $GSD^*(0.62, 0.66)$ to the adaptive design $ASD(0.62, 0.66)$, for a range of accrual rates. Cost ratio is calculated as the ratio of the minimal cost for the group sequential design to the minimal cost for the adaptive design, where costs are estimated as described in Section 5. Three different prior to patient cost scenarios are considered, representing high patient costs (black line), mid-range patient costs (red line), and low patient costs (green line).

with a delayed response or in longitudinal studies, as observed by an anonymous referee. We found that the degree of benefit depends on the distributions of event times and accrual rate as well as on the particular adaptive design under consideration. It is therefore worth considering the cost-effectiveness of using such an adaptive design in time-to-event endpoints. Even the simple cost model considered here, when used with trial specific values of prior development costs, per patient costs, accrual rates, and the current interest rates reflecting the time cost of money, could provide useful insight into the tradeoffs between potentially higher patient accrual or longer calendar time.

It is worth noting however, that our decision to hold the number and schedule of interim analyses constant may represent an unreasonable restriction. To briefly explore the effects of relaxing these constraints, we consider the example with $A = 0.62, D = 0.66$, where the adaptive design appeared to offer the most potential for benefit. We relaxed the constraint on the timing of the second and third analyses to explore a broader class of group sequential designs, but we continued to enforce the stopping boundary for the first analysis at 100 events. Having searched across a range of possible maximal sample sizes and P parameters to find an improved group sequential design within these relaxed constraints, we found that a design with maximum sample size of 210 events (analyses at 100, 155, and 210 events) and $P = (1.3, 1.3)$ matched the power curve of $ASD(0.62, 0.66)$ while dramatically improving ASN. The cost ratio plots resulting from comparing this design to $ASD(0.62, 0.66)$ are shown in Figure 8, from which we can see that significant reductions in total trial cost are possible for certain accrual and cost scenarios.

We do acknowledge that the above exercise is not totally fair. We presumed that the adaptive person tipped his/her hand first. Thus we only had to show we could improve over their choice. In the context of this example, which was based loosely on the type of design proposed for an industry sponsored study, we found

that we could easily find a group sequential design that met the same general operating characteristics. Given the small number of operating characteristics that were actually considered relative to the large number of group sequential test parameters at our disposal, this is not surprising. It would be similarly unsurprising to find that a proponent of adaptive designs could match the specified constraints of any particular group sequential design, and perhaps improve on some others. However, we believe the advantages of the well-developed group sequential trial theory makes it advantageous to use the group sequential design whenever the two approaches are roughly comparable.

As noted in Emerson [6], there remain problems with inference following the use of such an adaptive design, so in cases where there is questionable or insignificant gain from the adaptive design it may be wiser to continue to use a standard group sequential design where inferential methods are readily available in commercially available statistical software. Thus, we would argue that the time of clinical trialists is probably better spent exploring the wide range of group sequential trials already described and implemented, rather than trying to find ad hoc adaptive designs. Our belief is that the careful evaluation of candidate group sequential designs can largely address the issues that have motivated the development of adaptive designs.

One such area of evaluation that we have not explored here, but one that should receive careful attention in a time-to-event setting, is that of the ability to assess time varying treatment effects: In the setting of treatment effects that might be of greater magnitude either soon after randomization or after some delay, the tradeoffs between sample size and calendar time take on great importance. A study that terminates early with most events corresponding to short periods of treatment may not detect a clinically important difference in treatment behavior with additional follow-up.

Appendix

We consider the adaptive switching from a pre-specified group sequential design A to a pre-specified design group sequential design B$^*(A, D)$ according to whether the test statistic $\hat{\theta}(N_1)$ computed at the first analysis is between the values of A and D.

Notationally we define group sequential design A as a level α one-sided test of a lesser alternative having continuation sets $\mathcal{C}_1 = (a_1, d_1)$ and $\mathcal{C}_2 = \emptyset$ for $\hat{\theta}(N_1)$ and $\hat{\theta}(N_2)$, respectively, computed at analyses performed when the accrued sample sizes are $N_1 = n_1$ and $N_2 = N_1 + n_2$, respectively. The threshold a_2 for statistical significance at the second analysis is defined to guarantee an experimentwise error of α. Hence

$$P\left(\hat{\theta}_1 \leq a_1 \;\middle|\; \theta = 1\right) + P\left(a_1 \leq \hat{\theta}_1 \leq d_1,\ \hat{\theta}_2 \leq a_2 \;\middle|\; \theta = 1\right) = \alpha.$$

Now, suppose that if we do not terminate the clinical trial at the first analysis, we want to switch to an alternative stopping rule whenever $\hat{\theta}_1$ is observed between pre-specified values of A and D satisfying

$$a_1 \leq A \leq D \leq d_1.$$

If $a_1 < \hat{\theta}(N_1) < A$ or $D < \hat{\theta}(N_1) < d_1$, we will continue to use the sampling plan that specified a maximal sample size of N_2, with a threshold for statistical significance of a_2 at that last analysis.

Based on the pre-specified values of A and D, we further prospectively identify a group sequential design B$^*(A, D)$ having continuation sets $\mathcal{C}_2^* = (a_2^*, d_2^*)$ and $\mathcal{C}_3^* = \emptyset$ for $\hat{\theta}(N_2^*)$ and $\hat{\theta}(N_3^*)$, respectively, computed at analyses performed when the accrued sample sizes are $N_2^* = N_1 + n_2^*$ and $N_3^* = N_2^* + n_3^*$, respectively. Values of $\hat{\theta}(N_2^*) \leq a_2^*$ or $\hat{\theta}(N_3^*) \leq a_3^*$ will be judged cause to reject the null hypothesis. Hence, we need to pre-specify values of a_2^*, d_2^*, and a_3^* that, when used in conjunction with the group sequential design A and

the adaptation pre-specified through the choice of A and D, will preserve the experimentwise error of α:

$$\begin{aligned} P\left(\text{Reject } H_0 \mid \theta = 1\right) &= P\left(\hat{\theta}(N_1) \le a_1 \;\middle|\; \theta = 1\right) \\ &\quad + P\left(D \le \hat{\theta}(N_1) < d_1,\ \hat{\theta}(N_2) \le a_2 \;\middle|\; \theta = 1\right) \\ &\quad + P\left(a_1 < \hat{\theta}(N_1) \le A,\ \hat{\theta}(N_2) \le a_2 \;\middle|\; \theta = 1\right) \\ &\quad + P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2^*) \le a_2^* \;\middle|\; \theta = 1\right) \\ &\quad + P\left(A < \hat{\theta}(N_1) < D,\ a_2^* < \hat{\theta}(N_2^*) < d_2^*,\ \hat{\theta}(N_3^*) \le a_3^* \;\middle|\; \theta = 1\right) \\ &= \alpha. \end{aligned}$$

Now using the fact that the original specification of group sequential design A was a level α test, and the fact that $(a_1, A]$, (A, D), and $[D, d_1)$ form a partition of (a_1, d_1), we have that we only need

$$\begin{aligned} P\left(\text{Reject } H_0 \mid \theta = 1\right) &= \alpha - P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2) \le a_2 \;\middle|\; \theta = 1\right) \\ &\quad + P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2^*) \le a_2^* \;\middle|\; \theta = 1\right) \\ &\quad + P\left(A < \hat{\theta}(N_1) < D,\ a_2^* < \hat{\theta}(N_2^*) < d_2^*,\ \hat{\theta}(N_3^*) \le a_3^* \;\middle|\; \theta = 1\right) \\ &= \alpha \end{aligned}$$

which in turn yields that we only need find a_2^*, d_2^*, and a_3^* to satisfy

$$\begin{aligned} P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2) \ge d_2 \;\middle|\; \theta = 1\right) &= P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2^*) \le a_2^* \;\middle|\; \theta = 1\right) \\ &\quad + P\left(A < \hat{\theta}(N_1) < D,\ a_2^* < \hat{\theta}(N_2^*) < d_2^*,\ \hat{\theta}(N_3^*) \le a_3^* \;\middle|\; \theta = 1\right) \end{aligned}$$

In particular, we can define a group sequential design using the constrained boundary approach of Burington and Emerson (2003) in which analyses are performed at sample sizes N_1, N_2^*, and N_3^*, the continuation set (a_1^*, d_1^*) at the first analysis is constrained to be $a_1^* = A$ and $d_1^* = D$, and a_2^*, d_2^*, and a_3^* can be chosen as any resulting group sequential design that has type I error of

$$\alpha^* = P\left(\hat{\theta}_1 \le A \;\middle|\; \theta = 1\right) + P\left(A < \hat{\theta}(N_1) < D,\ \hat{\theta}(N_2) \le a_2 \;\middle|\; \theta = 1\right).$$

Any such choice will thus preserve an experimentwise error of α for the adaptive procedure.

It should be noted that it is immaterial the group sequential design family that is used to parameterize the specification of group sequential designs A and B$^*(A, D)$. Hence, subject to the specification representing a valid design, the group sequential designs could be specified in the unified family of Kittelson and Emerson (1999), a family of error spending functions, a specification of Bayesian posterior probabilities, or conditional or predictive power families.

The operating characteristics of the adaptive design considered here can be calculated using standard group sequential software, so long as the software allows the specification of arbitrary boundaries and the calculation of stopping probabilities at each analysis. The following approach will work in the program S+SeqTrial.

Design A can be computed according to the specified design parameters, which might include specifying the desired type I error, the number of analyses, the boundary shape parameters for both efficacy and futility, and any two of the design alternative, the desired statistical power, and the maximal number of events. For example, using the unified family of [5], the following code specifies Design A to be a one-sided level 0.025 test of a lesser hazard ratio having a maximum of two analyses after 100 and 200 events have been observed and using an O'Brien-Fleming efficacy (lower) boundary (so boundary shape parameters of

$P_a = 1$, $R_a = 0$, $A_a = 0$) and a Pocock futility (upper) boundary (with boundary shape parameters of $P_d = 0.5$, $R_d = 0$, $A_d = 0$), and also specifies Design B with three analyses at 100, 200, and 300 events using the same parameters, constrained to match Design A at the first analysis:

```
> designA <- seqDesign(prob.model="hazard", test.type="less", size=0.025,
      sample.size=c(100, 200), power=0.975, nbr.analyses=2, P=c(1, 0.5))

> bou <- seqBoundary(designA)
> bou <- matrix(NA, 3, 4)
> bou[1,] <- seqBoundary(designA)[1,]

> designB <- update(designA, sample.size=c(100, 200, 300), nbr.analyses=3,
      exact.constraint=bou)
```

The actual stopping boundaries are printed with the command:

```
> designA

PROBABILITY MODEL and HYPOTHESES:
   Two arm study of censored time to event response variable
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
          Null hypothesis : Theta >= 1         (size  = 0.025)
    Alternative hypothesis : Theta <= 0.5596    (power = 0.975)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N= 100) 0.5792 0.8645
    Time 2 (N= 200) 0.7611 0.7611
```

Then, a modification of Design A, Design $\mathrm{A}_{comp}(A, D)$, is specified in order to assist in computations reflecting the adaptive switching from Design A to a modification of Design B. For specified A and D, we modify the boundary of Design A to allow stopping and switching to the modified Design B if the observed hazard ratio is between A and D at the first analysis. We make use of the facility for constrained boundaries (Burington and Emerson, 2003). For instance, if we choose $A = 0.62$ and $D = 0.66$,

```
> bouA <- seqBoundary(designA)
> bouA[1,2] <- 0.62
> bouA[1,3] <- 0.66
> designAcomp.AD <- update(designA, test.type="two.sided", exact.constraint=bouA)
```

The boundaries of the Design $\mathrm{A}_{comp}(A, D)$ are obtained as:

```
> seqBoundary(designAcomp.AD)

STOPPING BOUNDARIES: Sample Mean scale
                         a    b    c      d
    Time 1 (N= 100) 0.5792 0.62 0.66 0.8645
    Time 2 (N= 200) 0.7611   NA   NA 0.7611
```

Stopping probabilities computed under Design $\mathrm{A}_{comp}(A, D)$ then reflect the probabilities of decisions made using Design A. For instance, under the null hypothesis of a hazard ratio of 1.0:

```
> seqOC(designAcomp.AD, theta=1)
```

```
Operating characteristics at theta= 1
ASN= 121.9633
Lower Power= 0.0218
Upper Power= 0.9677

Stopping Probabilities:
                 Lower   Null  Upper  Total
Analysis time 1 0.0032 0.0105 0.7668 0.7804
Analysis time 2 0.0187 0.0000 0.2010 0.2196
```

From the above, we see that for these values of A and D under the null hypothesis there is a probability of 0.0032 of stopping at the first analysis (when 100 events have been observed) with a decision for efficacy, a probability of 0.7668 of stopping at the first analysis with a decision for futility, a probability of 0.0187 of staying with Design A and then deciding for efficacy at the second analysis (when 200 events have been observed), and a probability of 0.2010 of staying with Design A and then deciding for futility at the second analysis. The remaining probability of 0.0105 corresponds to deciding to switch to Design $\mathrm{B}^*(A, D)$ based on the observation of an estimated hazard ratio between 0.62 and 0.66.

Design $\mathrm{B}^*(A, D)$ is a modification of Design B found in such a way as to ensure the experimentwise type I error of 0.025. There are an infinite number of ways to proceed. For the purposes of this paper, we considered maintaining the parameterization of the boundary shapes within the unified family, but constraining the stopping boundaries at the first analysis to agree with the boundaries of Design A, as discussed previously. In order to perform the necessary computations, we must specify a design, Design $\mathrm{B}_{comp}(A, D)$ that will have the same boundary as Design $\mathrm{B}^*(A, D)$, except that the continuation region at the first analysis will be defined by A and D rather than matching the Design A boundary. The operating characteristics of Design $\mathrm{B}_{comp}(A, D)$ will then correctly reflect the probabilities resulting from adaptively switching to Design $\mathrm{B}^*(A, D)$. The specified type I error for Design $\mathrm{B}_{comp}(A, D)$ was found to guarantee the experimentwise error, which is computed as the probability of 0.0218 of deciding for efficacy using Design A plus the probability of deciding for efficacy at either the second or third analysis using Design $\mathrm{B}_{comp}(A, D)$. This is easily computed from a single iteration: The specified size for Design $\mathrm{B}_{comp}(A, D)$ should equal the desired experimentwise type I error of 0.025 minus the probability of 0.0218 for declaring efficacy when using Design $\mathrm{A}(A, D)$ plus the probability of observing an estimated hazard ratio less than $A = 0.62$ at the first analysis in Design $\mathrm{B}_{comp}(A, D)$. In the example presented in this Appendix, a type I error of 0.0116 was found to satisfy the constraint. Hence, we used code:

```
> bouB <- matrix(NA, 3, 4)
> bouB[1,1] <- 0.62
> bouB[1,4] <- 0.66
> bouB <- seqBoundary(bouB)
> designBcomp.AD <- update(designA, size=0.0116, nbr.analyses=3, sample.size=c(100, 200, 300),
      exact.constraint=bouB)
```

The stopping boundaries when using Design $\mathrm{B}_{comp}(A, D)$ are found to be:

```
> seqBoundary(designBcomp.AD)

STOPPING BOUNDARIES: Sample Mean scale
                         a      d
    Time 1 (N= 100) 0.6200 0.6600
    Time 2 (N= 200) 0.7283 0.9386
    Time 3 (N= 300) 0.8095 0.8095
```

We can verify the experimentwise error for the design resulting from adaptively switching from Design A to Design $\mathrm{B}^*(0.62, 0.66)$ by using the operating characteristics of Design $\mathrm{A}_{comp}(A, D)$ and Design $\mathrm{B}_{comp}(A, D)$:

```
> seqOC(designBcomp.AD, theta=1)

Operating characteristics at theta= 1
ASN= 101.8519
Lower Power= 0.0116

Stopping Probabilities:
                 Lower Null  Upper  Total
Analysis time 1 0.0084    0 0.9811 0.9895
Analysis time 2 0.0018    0 0.0006 0.0024
Analysis time 3 0.0014    0 0.0067 0.0081
```

In the adaptive design based on Design A and Design B$^*(A, D)$ with parameters $A = 0.62, D = 0.66$, we thus find an experimentwise error of 0.025: A probability of deciding for efficacy of 0.0032 at the first analysis and 0.0187 at the second analysis when using Design A and a probability of 0.0018 at the second analysis and 0.0014 at the third analysis when using Design B$^*(A, D)$. Using similar computations of stopping probabilities under the design alternative of a hazard ratio of 0.5596, we find an experimentwise power of 0.9757: A probability of deciding for efficacy of 0.5684 at the first analysis and 0.3079 at the second analysis when using Design A and a probability of 0.0970 at the second analysis and 0.0024 at the third analysis when using Design B$^*(A, D)$. The average sample size (ASN) for the adaptive design can be found by multiplying the number of events at study termination by the probability of stopping with a decision for futility or efficacy at the first or second analyses using Design A and the probability of stopping at the second or third analyses for Design B$^*(A, D)$.

In order to find a comparable pre-specified group sequential design, we can again use the constrained boundary approach in order to match the decision boundaries at the first analysis. There are then an infinite number of ways that the design parameters can be modified at future analyses in order to closely match the unconditional power curve or the ASN curve to that of the adaptive approach. For instance, the following code could be used if the boundary shape parameters were to be modified:

```
> bouGS <- seqBoundary(designB)
> bouGS[2,] <- NA

> designGS <- update(designB, exact.constraint=bouGS, P=c(0, 0.08))

> designGS

PROBABILITY MODEL and HYPOTHESES:
   Two arm study of censored time to event response variable
   Theta is hazard ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
         Null hypothesis : Theta >= 1         (size  = 0.025)
    Alternative hypothesis : Theta <= 0.5617    (power = 0.975)

STOPPING BOUNDARIES: Sample Mean scale
                          a      d
    Time 1 (N= 100) 0.5792 0.8645
    Time 2 (N= 200) 0.7589 0.7665
    Time 3 (N= 300) 0.8025 0.8025
```

In this example, the group sequential design has marginally higher power and slightly lower ASN than the adaptive design given above for all alternatives corresponding to hazard ratios between 0.3 and 1.2.

References

[1] Anastasios A. Tsiatis and Cyrus R. Mehta. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90:367–378, 2003.

[2] Christopher Jennison and Bruce W. Turnbull. Adaptive and nonadaptive group sequential tests. *Biometrika*, 93(1):1–21, 2006.

[3] Scott S. Emerson, John M. Kittelson, and Daniel L. Gillen. Frequentist evaluation of group sequential designs. *Statistics in Medicine*, 26(28):5047–5080, 2007.

[4] Scott S. Emerson, John M. Kittelson, and Daniel L. Gillen. Bayesian evaluation of group sequential designs. *Statistics in Medicine*, 26(7):1431–1449, 2007.

[5] John M. Kittelson and Scott S. Emerson. A unifying family of group sequential test designs. *Biometrics*, 55:874–882, 1999.

[6] Scott S. Emerson. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine*, 25(19): 3270–3296, 2006.

[7] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[8] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–200, 1977.

[9] Samuel K. Wang and Anastasios A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–199, 1987.

[10] Scott S. Emerson and Thomas R. Fleming. Symmetric group sequential test designs. *Biometrics*, 45: 905–923, 1989.

[11] *S+SeqTrial*. Insightful Corporation, Seattle, Washington, 2002.

[12] Michael A. Proschan and Sally A. Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, 51:1315–1324, 1995.

[13] Lu Cui, H. M. James Hung, and Sue-Jane Wang. Modification of sample size in group sequential clinical trials. *Biometrics*, 55:853–857, 1999.

[14] Bart E. Burington and Scott S. Emerson. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics*, 59:770–777, 2003.

[15] David A. Schoenfeld and Jane R. Richter. Nomograms for calculating the number fo patients needed for a clinical trial with survival as an endpoint. *Biometrics*, 38:163–170, 1982.

[16] Scott S. Emerson. *S+SeqTrial Technical Overview*. Insightful Corporation, 2003.