Designing, Monitoring, and Analyzing Group Sequential Clinical Trials Using the RCTdesign Package for R

Daniel L. Gillen[1]
Department of Statistics, University of California, Irvine, USA

and

Scott S. Emerson
Department of Biostatistics, University of Washington, Seattle, WA, USA

June 4, 2012

Abstract

The use of group sequential methodology has become widespread in the conduct of clinic trials. As each clinical trial presents unique scientific, statistical and logistical constraints, it is important to carefully evaluate candidate group sequential designs to ensure desirable operating characteristics. At the implementation stage of a clinical trial design it is also essential to account for deviations from original design specifications in order to control operating characteristics such as type I and II error rates. These changes might include the number and/or timing of analyses as well as deviations from the originally assumed variability of outcome measures . Due to the computational complexity involved in evaluating, monitoring, and analyzing a group sequential procedure, specialized software is required. In the current manuscript we demonstrate how the RCTdesign package (www.rctdesign.org) in R can be used to select, implement, and analyze a group sequential stopping rule. Throughout, we illustrate trial design and monitoring in the context of a group sequential survival trial of an experimental monoclonal antibody in patients with relapsed chronic lymphocytic leukemia.

1. Introduction

The use of group sequential methodology has become widespread in the conduct of clinic trials. Many authors have addressed the design (Pocock, 1977; O'Brien and Fleming, 1979; Whitehead and Stratton, 1983; Wang and Tsiatis, 1987; Emerson and Fleming, 1989), implementation (Lan and DeMets, 1983, Burington and

[1]Corresponding author:
Daniel L. Gillen
Department of Statistics
2226 Donald Bren Hall
University of California, Irvine
Irvine, CA 92697-1250

e-Mail: dgillen@uci.edu
Tel: +1-949-824-9862
Fax: +1-949-824-9863

Emerson, 2003), and analysis (Whitehead, 1986b; Emerson and Fleming, 1990) of group sequential trials.

In the general case, a stopping rule is defined for a schedule of analyses occurring at times t_1, t_2, ..., t_J, which may be random. Often, the analysis times are in turn defined according to the statistical information available at each analysis. In the case of a statistical model that has statistical information proportional to the sample size accrued to the study, such an approach is equivalent to defining the sample sizes N_1, N_2, ..., N_J at which the analyses will be performed. For $j = 1, \ldots, J$, we calculate a specified test statistic T_j based on observations available at time t_j. The outcome space for T_j is then partitioned into stopping set $\mathcal{S}_j$ and continuation set C_j. Starting with $j = 1$, the clinical trial proceeds by computing T_j, and if $T_j \in \mathcal{S}_j$, the trial is stopped. Otherwise, T_j is in the continuation set C_j, and the trial gathers additional observations until time t_{j+1}. By choosing $C_J = \emptyset$, the empty set, the trial must stop at or before the J-th analysis.

All of the most commonly used group sequential stopping rules are included if we consider continuation sets of the form $\mathcal{C}_j = (a_j, b_j] \cup [c_j, d_j)$ such that $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$. Quite often, these boundaries are interpreted as the critical values for a decision rule. For instance, in a clinical trial comparing two active treatments A and B, test statistics less than a_j might correspond to decisions for the superiority of treatment A, test statistics exceeding d_j might correspond to decisions for the inferiority of treatment A, and test statistics between b_j and c_j might correspond to decisions for approximate equivalence between the two treatments.

As each clinical trial presents unique scientific, statistical and logistical constraints, it is important to carefully evaluate candidate group sequential designs to ensure desirable operating characteristics. Emerson et al. (2007b) describe a variety of frequentist design characteristics which might be examined in the most commonly encountered statistical problems. Among them are

1. The scientific measures of treatment effect which will correspond to early termination for futility and/or efficacy.

2. The sample size requirements as described by the maximal sample size and summary measures of the sample size distribution (e.g., mean, 75th percentile) as a function of the hypothesized treatment effect.

3. The probability that the trial would continue to each analysis as a function of the hypothesized treat-

ment effect.

4. The frequentist power to reject the null hypothesis as a function of the hypothesized treatment effect, with the type I error corresponding to the power under the null hypothesis.

5. The frequentist inference (adjusted point estimates, confidence intervals, and P values) which would be reported were the trial to stop with results corresponding exactly to a boundary.

At the implementation stage of a clinical trial design it is also essential to account for deviations from original design specifications in order to control operating characteristics such as type I and II error rates (Lan and DeMets, 1983, Burington and Emerson, 2003). These changes might include the number and/or timing of analyses as well as deviations from the originally assumed variability of outcome measures. Finally, at the completion of a group sequential test it is important that point and interval estimates be adjusted to account for bias that arises through repeated testing, particularly when the implemented stopping boundaries allow for early stopping under more modest effect sizes (Whitehead, 1986b; Emerson and Fleming, 1990).

Due to the computational complexity involved in evaluating, monitoring, and analyzing a group sequential procedure, specialized software is required. Multiple software packages can be used for the design and/or analysis of group sequential trials (PEST, 2000; EaSt, 2000; SAS/SEQDESIGN, 2011). The `RCTdesign` package (`www.rctdesign.org`) for `R` statistical software is an extension of the `SeqTrial` module for `SPlus` (S+SeqTrial, 2002). `RCTdesign` is a comprehensive package that allows users to choose from a full array of previously proposed group sequential stopping rules, monitor an ongoing trial using standard constrained boundaries techniques, and report bias adjusted results at the conclusion of a clinical trial.

In the current manuscript we demonstrate how the `RCTdesign` package can be used to select, implement, and analyze a group sequential stopping rule. Throughout, we illustrate trial design and monitoring in the context of a clinical trial of an experimental monoclonal antibody in patients with relapsed chronic lymphocytic leukemia. Section 2 provides an evaluation of candidate clinical trial designs based upon commonly considered frequentist operating characteristics. Section 3 describes previously proposed methods for flexibly monitoring a group sequential test. An example implementing the constrained boundaries algorithm (Burington and Emerson, 2003) is presented and adjusted inference is discussed. In Section 4, we present additional issues that should be considered when designing and monitoring a clinical trial to investigate an

intervention for which the effect may be hypothesized to vary with the duration of time since initiation. Section 5 concludes with a discussion of the importance of thorough evaluation in the selection of a group sequential stopping rule along with areas of current and future research.

2. Evaluation of a Group Sequential Trial for a Censored Time-To-Event Endpoint

In this section we illustrate the evaluation of statistical operating statistics in the context of a randomized, double-blind, placebo-controlled clinical trial of an experimental monoclonal antibody in patients with relapsed chronic lymphocytic leukemia (CLL). Treatment of CLL tends to focus on controlling disease symptoms through the use of chemotherapy, radiation therapy, biological therapy, or bone marrow transplantation. Recently there have been multiple trials to assess the efficacy of treating CLL via monoclonal antibodies that target markers which are heavily expressed by CLL cells. In one of these trials, patients with relapsed CLL were randomly assigned to receive an experimental antibody or placebo, in addition to a standard chemotherapeutic regime. The intervention was administered intravenously once a week for four weeks and patients were followed for the primary endpoint of overall survival. It was anticipated that the median survival time among patients treated with placebo would be approximately 16 months and that the distribution of survival times among this group would be approximately exponentially distributed. It was hoped that patients receiving the antibody would experience a 33% reduction in the hazard for death and that the effect of treatment on the hazard would remain roughly constant over time.

In the discussion of the operating characteristics which follows, we will use comparisons similar to (but not exactly the same as) those explored by the collaborators in the CLL study. In all cases, we consider level 0.025 one-sided hypothesis tests appropriate for testing a null hypothesis $H_0 : \theta \geq 1$ versus the lesser alternative $H_1 : \theta \leq 0.67$, where θ represents the hazard ratio comparing treatment to control. Throughout, a 1-to-1 randomization scheme is assumed.

To illustrate the evaluation process, we consider candidate designs as derived from the unified family of group sequential stopping rules (Kittelson and Emerson, 1999). As noted in Section 1, all of the most commonly used group sequential stopping rules are included if we consider continuation sets of the form $\mathcal{C}_j = (a_j, b_j] \cup [c_j, d_j)$ such that $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$. Particular families of group sequential

designs correspond to parameterized boundary functions which relate the stopping boundaries at successive analyses according to the proportion of statistical information accrued and the hypothesis rejected by the boundary. For instance, letting Π_j represent the proportion of the maximal statistical information available at the j-th analysis (e.g., $\Pi_j = N_j/N_J$ for the most commonly used analytic models), then for some specified parametric function $f_d()$, the boundary function for the upper boundary might be given by $d_j = f_d(\theta_d, \Pi_j)$, where θ_d is the hypothesis rejected when $T_j > d_j$. Furthermore, many of the group sequential design families previously described can be expressed in a parameterization which has $d_j = f(\theta_d, g(\Pi_j; A_d, P_d, R_d, G_d))$ with boundary shape function

$$g(\Pi; A, P, R, G) = (A + \Pi^{-P}(1 - \Pi)^R)G$$

where parameters A, P, and R are typically specified by the user to attain some desired level of conservative behavior at the earliest analyses, and critical value G might be found in an iterative search to attain some specified operating characteristics (e.g., frequentist type I error and power) when the stopping rule is to be used as the basis of a decision rule. In this parameterization, taking $A = R = 0$ yields a 1-parameter family of stopping boundaries where larger values of P result in increased conservatism of the stopping rule meaning that it is more difficult to stop at early analyses for a given treatment effect. In the unified family (Kittelson and Emerson, 1999), the boundaries are expressed on the treatment effect scale and the boundary hypothesis is merely a shift of the boundary shape function so that

$$\begin{aligned} d_j &= \theta_d + g(\Pi_j; A_d, P_d, R_d, G_d) \\ a_j &= \theta_a + g(\Pi_j; A_a, P_a, R_a, G_a). \end{aligned}$$

For the remainder of the manuscript we will focus on the following candidate designs. RCTdesign code to compute each of the above stopping rules is provided in Appendix A.

1. *Fixed.Sample*: a fixed sample study with 263 events providing 90.1% power to detect the alternative H_1.

2. *SymmOBF.2*, *SymmOBF.3*, *SymmOBF.4*: one-sided symmetric stopping rules that treat the null and alternative hypotheses symmetrically (Emerson and Fleming (1989)) and utilize O'Brien-Fleming

boundary relationships having a total of 2, 3, and 4 equally spaced analyses, respectively, and a maximal sample size of 263 events.

3. *SymmOBF.Power*: one-sided symmetric stopping rule with O'Brien-Fleming boundary relationships, a total of 4 equally spaced analyses and the total sample size selected to provide 90.1% power to detect the alternative H_1.

4. *Futility.5*, *Futility.8*, *Futility.9*: one-sided stopping rules from the unified family Kittelson and Emerson (1999) with a total of 4 equally spaced analyses, with a maximal sample size of 263 events, and having O'Brien-Fleming lower (efficacy) boundary relationships and upper (futility) boundary relationships corresponding to boundary shape parameters $P = 0.5$, 0.8, and 0.9, respectively. In this parameterization of the boundary shape function, parameter P is a measure of conservatism at the earliest analyses. $P = 0.5$ corresponds to Pocock boundary shape functions, and $P = 1.0$ corresponds to the more conservative O'Brien-Fleming boundary relationships.

5. *Eff11.Fut8*, *Eff11.Fut9*: one-sided stopping rules from the unified family Kittelson and Emerson, 1999 with a total of 4 equally spaced analyses, with a maximal sample size of 263 events, and having lower (efficacy) boundary relationships corresponding to boundary shape parameter $P = 1.1$ and upper (futility) boundary relationships corresponding to boundary shape parameters $P = 0.8$ and 0.9, respectively.

6. *Fixed.Power*: a fixed sample study which provides the same power to detect H_1 as the *Eff11.Fut8* trial design.

2.1 *Evaluation of Stopping Boundaries*

It is important that clinical trialists not be surprised by the conditions under which a particular stopping rule suggests that a trial might continue or stop early. As such, we believe that it is of paramount importance that the stopping boundary at each analysis be considered as the stopping rule is selected. Emerson et al. (2007b) note that there are a number of scales on which the boundaries can be examined. While there exists a one-to-one relationship between these scales, the statistical and scientific utility of the scales varies depending upon one's background. In the context of the CLL trial, we may consider any of the following test statistics as the basis for the definition of the stopping rule at interim analysis j:

1. *Partial sum statistic*: S_j, the partial likelihood based score function for $\log(\theta)$ in a proportional hazards regression model.

2. *Crude estimate of treatment effect*: $\hat{\theta}_j$, the estimated hazard ratio from a proportional hazards model.

3. *Normalized Z statistic*: Z_j, the score statistic for testing H_0.

4. *Fixed sample P value statistic*: $P_j = \Phi(Z_j)$, where $\Phi(\cdot)$ represents the cumulative distribution function corresponding to the standard normal distribution.

5. *Error spending statistic*: An error spending statistic can be defined for any of the four boundaries based on an arbitrary hypothesized value for the true treatment effect. For instance, if a group sequential stopping rule were defined for the partial sum statistic and the observed value of the test statistic at the j-th analysis were $S_j = s_j$, a lower type I error spending statistic defined for the null hypothesis $H_0 : \theta = \theta_0$ would have

$$E_{aj} = \frac{1}{\alpha_L}\left(Pr\left[S_j \leq s_j,\ \bigcap_{k=1}^{j-1} S_k \in C_k \mid \theta = \theta_0 \right] + \sum_{\ell=1}^{j-1} Pr\left[S_\ell \leq a_\ell,\ \bigcap_{k=1}^{\ell-1} S_k \in C_k \mid \theta = \theta_0 \right] \right),$$

where α_L is the lower type I error of the stopping rule defined by

$$\alpha_L = \sum_{\ell=1}^{J} Pr\left[S_\ell \leq a_\ell, \bigcap_{k=1}^{\ell-1} S_k \in C_k | \theta = \theta_0 \right].$$

6. *Bayesian posterior probabilities*: $B_j(\theta_0) = Pr(\theta \leq \theta_0 \,|\, S_j = s_j)$, the posterior probability that the null hypothesis $H_0 : \theta \geq \theta_0$ is false under a specified prior distribution

7. *Conditional power statistics*: The conditional probability that the test statistic at the final (J-th) analysis would exceed the threshold for declaring statistical significance, where we condition on the observed statistic $S_j = s_j$ at the j-th analysis and assume some particular value for the true treatment effect θ. For instance, we might define a conditional power statistic using a threshold a_J defined for the partial sum statistic. Such a threshold would represent the critical value for declaring statistical significance at the J-th analysis. Using an alternative hypothesis $H_1 : \theta = \theta_1$ conditional power would

be computed as

$$C_j(a_J, \theta_1) \quad = \quad Pr(S_J < a_J \,|\, S_j = s_j; \theta = \theta_1)$$

Alternatively, a conditional power statistic might use the current best estimate of the treatment effect $\hat{\theta}_j$ in place of θ_1.

8. *Predictive probability statistics*: The Bayesian predictive probability that the test statistic would exceed some specified threshold at the final analysis. In the case of a threshold a_J defined for the partial sum statistic, a predictive probability statistic may be computed as

$$H_j(a_J, \zeta, \tau^2) \quad = \quad \int Pr(S_J < a_J \,|\, S_j = s_j, \theta)\, p(\theta \,|\, S_j = s_j)\, d\theta$$

Returning to the CLL trial, we find it most useful to consider stopping boundaries on the scientifically relevant scale of the estimated treatment effect. By graphing the stopping boundaries versus the number of events (or statistical information) available at each analysis, we can see both the degree of conservatism employed at the earliest analyses and the worst case sample size requirements for the study. In Figure 1 we display the stopping boundaries for the *SymmOBF.4*, *Eff11.Fut9*, and *Eff11.Fut8* stopping rules, all of which use the same level of significance. These three designs all have a maximal sample size of 263 events but differ in the boundary shape function used for the efficacy (lower) and futility (upper) boundary, ranging in conservatism (higher values of P yield a more conservative stopping rule). By comparing *Eff11.Fut9* and *Eff11.Fut8*, it can be seen that altering the futility boundary has only minimal effects on the efficacy boundary. We can also see from Figure 1 that at the planned first analysis (N=66 events) the O'Brien-Fleming boundary shape function would suggest early termination for futility only if the estimated hazard ratio were 1.639 or larger – a difference that may be deemed too large. The futility boundary shape function for the *Eff11.Fut8* stopping rule, on the other hand, would allow early termination for futility when the observed hazard ratio is 1.319. Similar comparisons may be made with respect to the efficacy bound. It is worth noting that in many cases the extreme conservatism of the *Eff11.Fut9* and *Eff11.Fut8* efficacy bounds may be desired at early analyses because stopping a trial early for efficacy would preclude the collection of longer term safety data in a controlled setting. `RCTdesign` code to generate the resulting boundary plot and

to create a table of the stopping boundaries on different scales is provided in Appendix A.

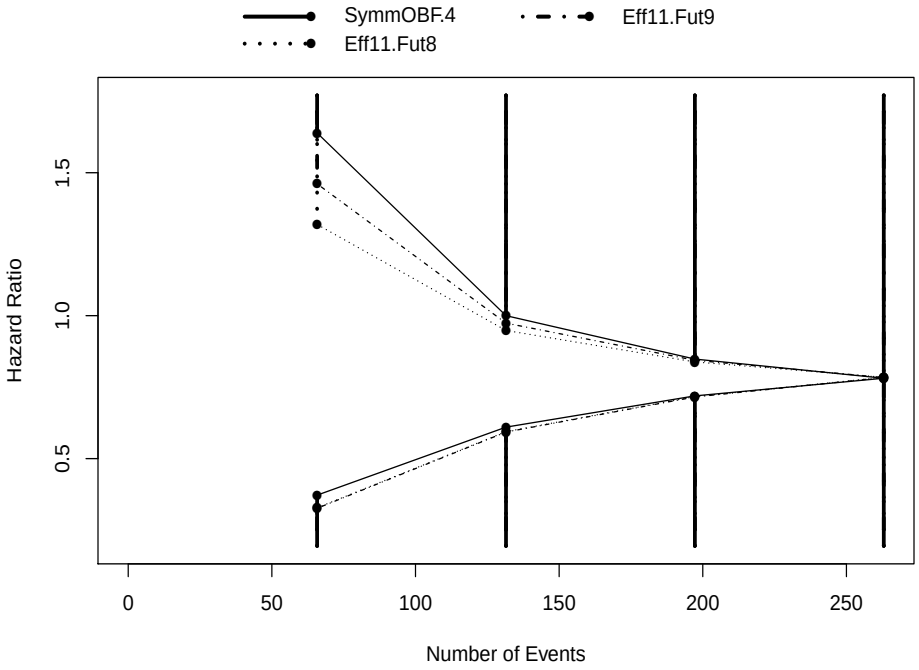


Figure 1: Stopping boundaries on the scale of the crude estimate of treatment effect (estimated hazard ratio). In the case of the CLL trial, stopping boundaries for level 0.025 one-sided stopping rules for a maximum of 263 events and various levels of conservatism for the efficacy (lower) and futility (upper) boundary relationships.

2.2 *Frequentist Type I Error and Power*

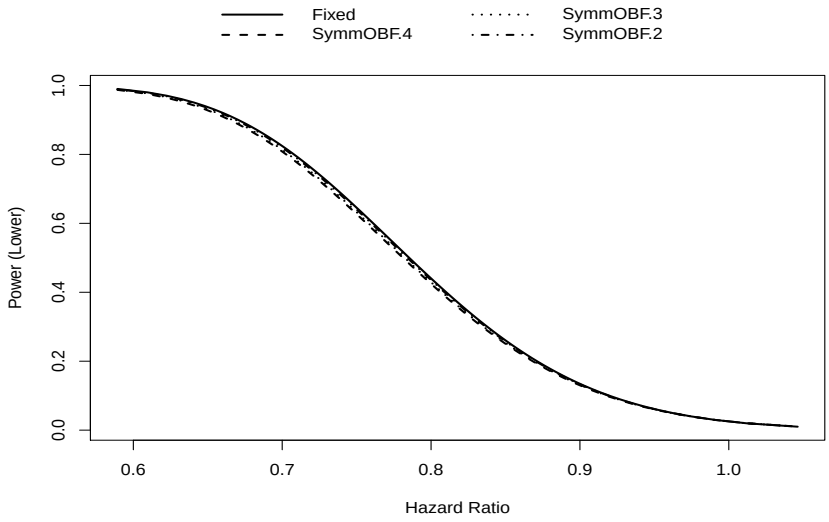
The most commonly used definition for statistical evidence against a null hypothesis is to consider the probability of falsely rejecting the null hypothesis. In fact, regulatory agencies often use this criterion as a *de facto* standard for strength of evidence that will be attained in a clinical trial design. Thus, when specifying a group sequential stopping rule, clinical trialists most often constrain the type I error associated with a decision boundary to some prescribed level, typically 0.05 for a two-sided test and 0.025 for a one-sided test.

Similarly, it is often the case that the sample size to be used in a clinical trial is determined by computing the sample size that will allow estimation of the treatment effect with specified precision (often according to

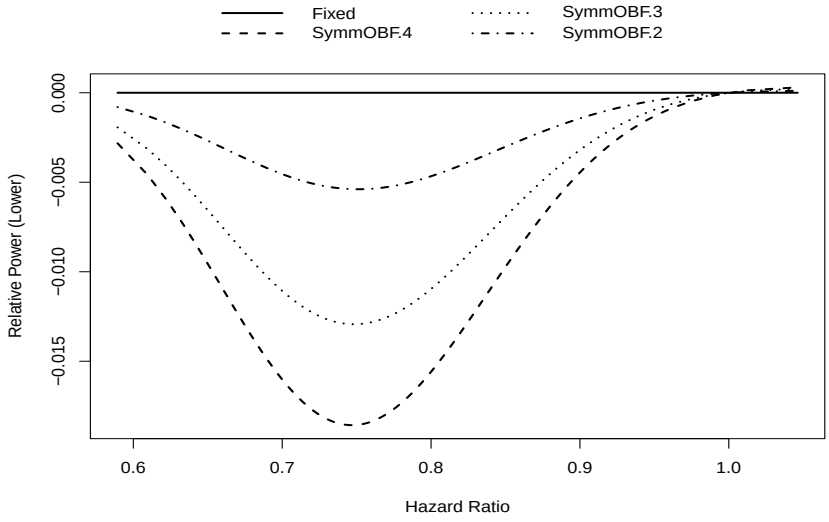
the width of a 95% confidence interval) or that will allow a decision to reject the null hypothesis to be made with high probability (e.g., 80%, 90%, 95%, or 97.5% statistical power) when a specific alternative hypothesis is true. This criterion of statistical power is of particular interest from a scientific standpoint: It describes the probability that the clinical trial will discriminate between the two viable scientific hypotheses represented by the null and alternative hypotheses. Hence, basic scientists, clinical researchers, epidemiologists, and biostatisticians often focus on the statistical power of the study to detect a hypothesis representing the minimal treatment effect which is of clinical importance.

Figure 2 displays power curves for some stopping rules considered in the design of the CLL trial. In this figure we compare the effect of increasing the number of interim analyses on the statistical power when the maximal sample size is maintained at 263 events. Rather than display the absolute power curve as in Figure 2a, we often find it most convenient to display the power relative to some reference design. In Figure 2b, we examine the loss of power relative to a fixed sample clinical trial for several stopping rules which vary in the number of interim analyses. With the O'Brien-Fleming boundary relationships considered in this figure, we see relatively little loss of power: A one-sided symmetric design with O'Brien-Fleming relationships and a total of 4 equally spaced analyses Emerson and Fleming (1989) loses at most 0.019 power (from 68.5% to 66.6%) relative to a fixed sample analysis with the same maximal sample size.

Table 1 compares the power of the *Fixed.Sample*, *Eff11.Fut9*, and *Eff11.Fut8* stopping rules under specific hypotheses and provides the alternative hypotheses for which the various designs have prescribed statistical power. From this table it is apparent that the introduction of either of these stopping rules has relatively minimal impact on the statistical power of the study. This in turn means that the introduction of either of these stopping rules has relatively little effect on the scientific interpretation of a failure to reject the null hypothesis. More specifically, using a confidence level of 95% as the statistical criterion for evidence, a failure to reject the null can be interpreted as a rejection of a hazard ratio corresponding to the alternative for which the design attains a power of 0.975: equal to 0.617 using the *Fixed.Sample* design, 0.610 using the *Eff11.Fut9* design, and 0.607 using the *Eff11.Fut8* design. We note that this difference in rejected alternatives is due to the fact that the maximal sample size was not increased when a stopping rule was introduced. With an increase in the maximal sample size, we can maintain the magnitude of the alternative rejected by a failure to reject the null hypothesis.



(a)



(b)

Figure 2: Power curves and difference in power relative to a fixed sample design for a fixed sample design and one-sided symmetric tests with O'Brien-Fleming ($SymmOBF.J$) boundary relationships with J = 2, 3, or 4 analyses. All designs have type I error of 0.025 under the null hypothesis $H_0 : \theta \geq 1$ and a maximal sample size of 263 events.

Table 1: Comparison of operating characteristics of three candidate stopping rules: (upper panel) detectable alternatives and mean sample size for fixed power; (lower panel) power and mean sample size for fixed alternative.

	Fixed.Sample Stopping Rule		*Eff11.Fut9* Stopping Rule		*Eff11.Fut8* Stopping Rule	
Power	Hazard Ratio	Average Samp Size	Hazard Ratio	Average Samp Size	Hazard Ratio	Average Samp Size
0.800	0.708	263	0.703	207	0.702	204
0.900	0.670	263	0.665	196	0.663	194
0.950	0.641	263	0.635	185	0.633	184
0.975	0.617	263	0.610	176	0.607	174
Hazard Ratio	Power	Average Samp Size	Power	Average Samp Size	Power	Average Samp Size
1.00	0.025	263	0.025	163	0.025	154
0.75	0.645	263	0.628	214	0.624	211
0.67	0.901	263	0.889	198	0.885	196
0.60	0.985	263	0.981	172	0.980	172

2.3 *Sample Size Distribution*

In a fixed sample clinical trial, if the sample size is chosen to attain some prespecified statistical power, one of the first operating characteristics considered is whether obtaining that sample size is feasible logistically and financially. Clinical trial collaborators also have to consider whether the sample size would provide credible scientific evidence. In the presence of data collected using a stopping rule, the actual sample size obtained during the conduct of a clinical trial is a random variable with a distribution that depends on the magnitude of the true treatment effect– a dependence that is, of course, behind the ethical motivation for interim analyses: We want to use fewer patients when one treatment is markedly inferior to another or not sufficiently superior to warrant further investigation. Thus, when examining the sample size requirements of a particular clinical trial design, we will be interested in summary measures of the probability distribution for the sample size. The maximal sample size will be of interest for the feasibility of accrual, just as it is in a fixed sample trial. Examination of the curves for the average sample size (ASN = average sample number) and various quantiles of the sample size distribution provides some indication of the values that might reasonably be attained under various hypotheses. In the case of a survival endpoint, statistical information is (at least partially) dictated by the number of observed events. In this case it is natural to consider summary measures of the distribution of the required number of events.

In Figure 3 we compare the average and 75th percentile of the distributions of required events for the group sequential stopping rules considered for the futility boundary in the CLL trial. From this figure it can be seen that substantially smaller numbers of events would be accrued on average as the futility (upper) boundary becomes successively less conservative. Of course, because the maximal number of events does not differ among these stopping rules, the power curves will vary. Therefore, the ultimate selection of a stopping rule involved simultaneous graphical comparisons of the average event curves and the respective power curves (not shown here, but analogous to those shown in Figure 2) in order to judge the acceptability of tradeoffs between the loss of power and gains in average number of events.

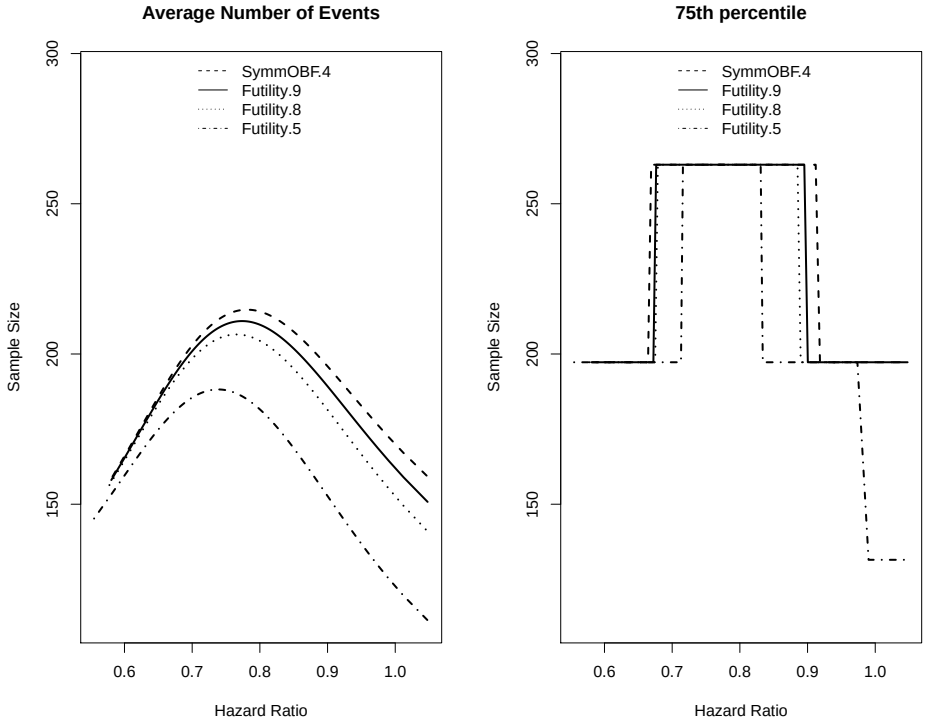


Figure 3: Average and 75th percentile of the distribution of required events as a function of the hypothesized treatment effect. In the case of the CLL trial, stopping boundaries for level 0.025 one-sided stopping rules for a maximum of 263 events and having O'Brien-Fleming efficacy (lower) boundary relationships and various levels of conservatism for the futility (upper) boundary relationships.

2.4 *Stopping Probabilities*

When more detail about the stopping behavior of the group sequential trial design is desired, the probability of stopping at each analysis time can be examined as a function of the hypothesized true treatment effect. Figure 4 displays the cumulative stopping probability at each analysis versus the true treatment effect. For any given hypothesized hazard ratio (x-axis) the vertical distance up to each contour line represents the cumulative probability that the trial will stop at or before the corresponding analysis time (as depicted by the number of the contour). In this figure, the shading indicates the probability with which a decision at stopping will be made for the alternative hypothesis (i.e., when the test statistic is less than the lower boundary) or the null hypothesis (i.e., when the test statistic is greater than the upper boundary). Thus from this figure we can see that under the *Eff11.Fut8* stopping rule, when the true treatment effect corresponds to a hazard ratio of $\theta = 0.70$, the probability of stopping at or before the third analysis is approximately 0.67, and as the shading below that curve is generally the darker color, the predominant decision will be one to reject the null hypothesis. Furthermore, by examining the stopping boundaries on the scale of the crude estimate of treatment effect (Figure 1), it can be seen that the stopping rule would only recommend continuing past the third analysis if the observed hazard ratio comparing treatment to placebo were between 0.717 and 0.838, a situation that may look promising enough to invest in the larger sample size.

2.5 *Frequentist Inference at the Stopping Boundaries*

In order to ensure the scientific and statistical credibility of the study results, it is important to examine the statistical inference that would be reported if the study were to be terminated early. Of particular interest is whether estimates of treatment effect would indeed be extreme enough to convince the scientific community that action should be taken with less precision in the estimates. When using frequentist inference, we typically consider point estimates of treatment effect with small bias and mean squared error, and we consider the precision of such estimates using 95% confidence intervals. Strength of evidence against a null hypothesis is often quantified by the P value– the probability that results as or more extreme than those actually obtained would be observed when the null hypothesis is true. These same frequentist measures are possible in the setting of group sequential stopping rules, though the calculation of the estimates, confidence intervals, and P values must use the correct sampling distribution. Further discussion of inferential procedures that account

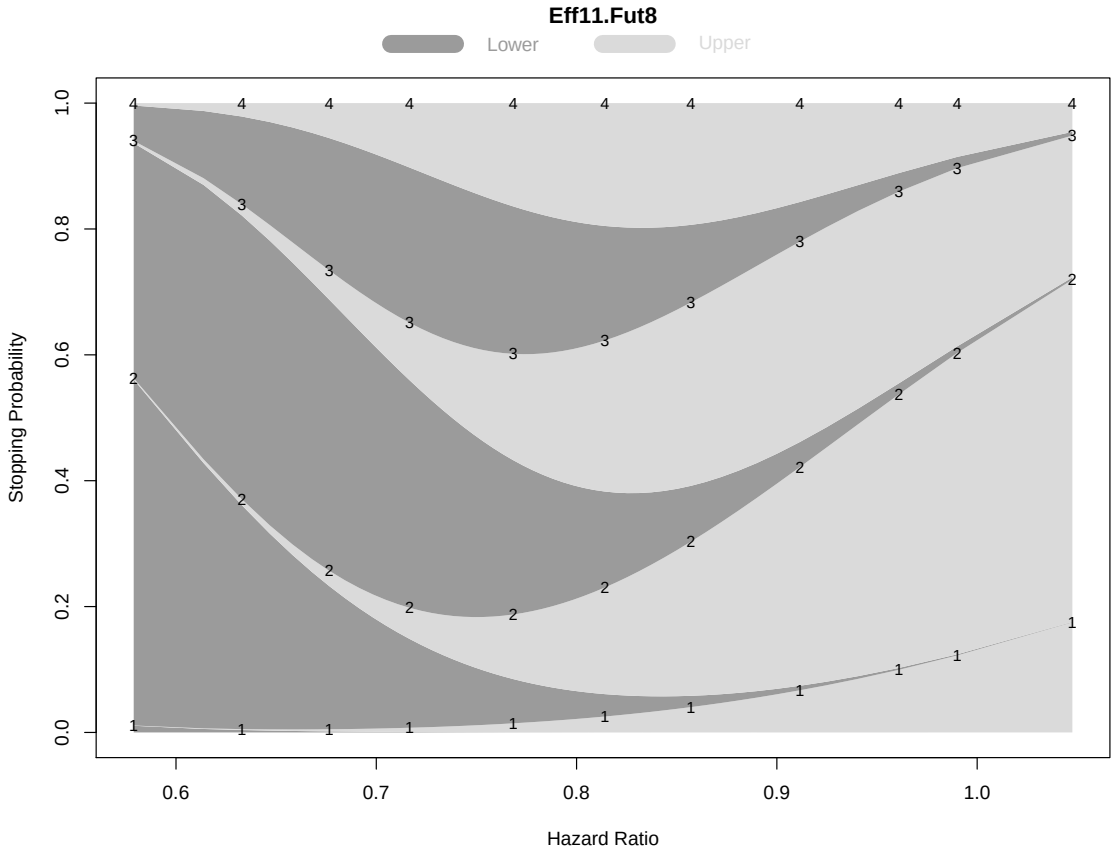


Figure 4: Cumulative stopping probabilities at each analysis as a function of hypothesized treatment effect. The one-sided level 0.025 stopping rule to test the null hypothesis $H_0 : \theta \geq 0$ has a maximum of 4 equally spaced analyses with an efficacy (lower) boundary shape function corresponding to $P = 1.1$ and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs.

for group sequential testing is presented in Section 3.2.

In the process of evaluating group sequential designs, it is useful to consider the inference associated with outcomes which correspond exactly to the stopping boundaries. Clearly, if such outcomes are scientifically and statistically convincing, more extreme results would also be acceptable. Figure 7 displays such hypothetical inference for the stopping boundaries of the *Eff11.Fut8* stopping rule. The top and bottom panels display the adjusted point estimates (bias adjusted mean as described by Whitehead (Whitehead (1986a)) and the sample mean ordering based 95% confidence intervals and P values for hypothetical results which

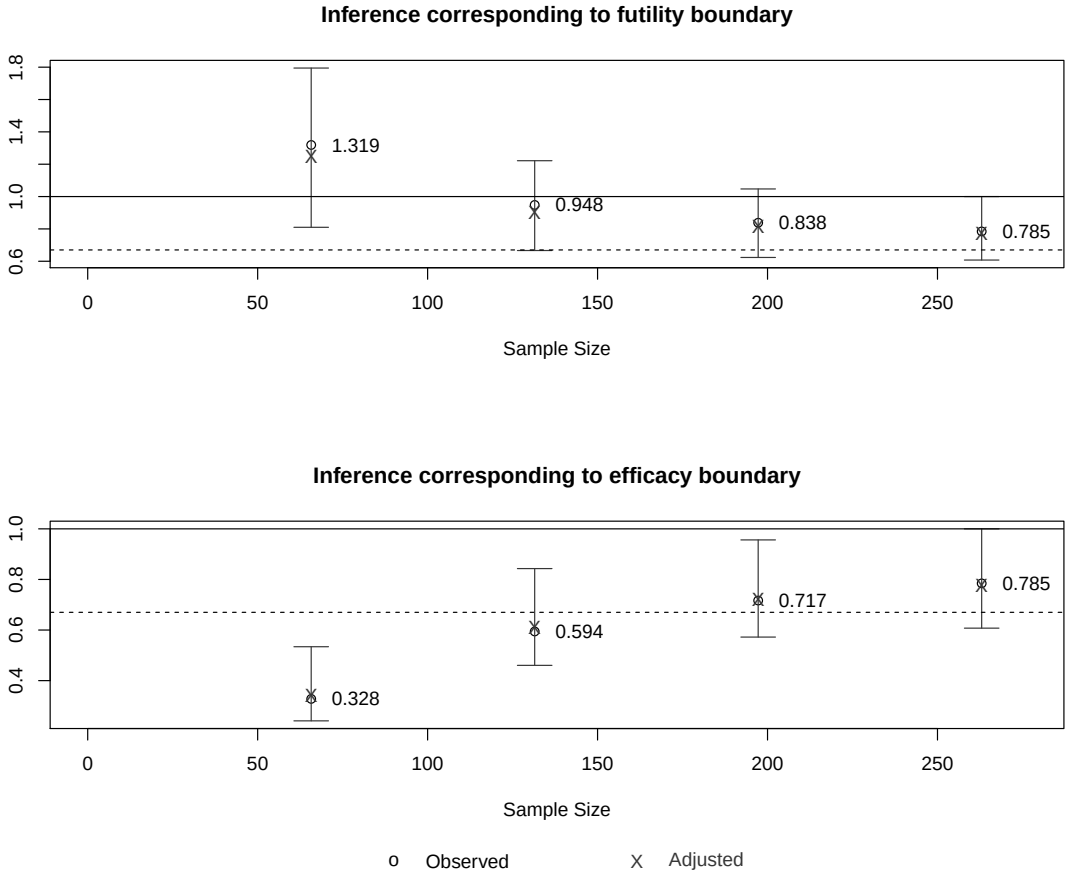


Figure 5: Display of estimates and confidence intervals for observed trial results which correspond exactly to the stopping boundaries of a one-sided level 0.025 stopping rule to test the null hypothesis $H_0 : \theta \geq 1$ and having a maximum of 4 equally spaced analyses, an efficacy (lower) boundary shape function corresponding to $P = 1.1$ and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs. Inference for the futility (upper) boundary is displayed in the upper panel, and inference for the efficacy (lower) boundary is displayed in the lower panel. All estimates, confidence intervals, and P values are adjusted for the stopping rule. Horizontal lines correspond to the null hypothesis $\theta = 1$ and the alternative hypothesis $\theta = 0.67$.

correspond to the futility (upper) and efficacy (lower) boundaries, respectively. Also displayed for reference are horizontal lines corresponding to the null hypothesis $\theta = 1$ and the alternative hypothesis $\theta = 0.67$. From this plot we see the extreme conservatism of the efficacy boundary. At the first analysis, we would stop the study early with a decision for efficacy only if we could with high confidence rule out that the treatment effect was less extreme than an alternative far beyond that which we considered in the design of the trial (i.e., the

95% confidence interval not only excludes the null hypothesis, but also excludes an alternative corresponding to a hazard ratio of 0.54). On the other hand, the futility boundary is less conservative as evidenced by the fact that although results which would cause termination have ruled out a markedly beneficial effect of treatment, they have not established with high confidence that the treatment might have some small beneficial effect (i.e., the 95% confidence interval corresponding to results at the futility stopping boundary includes the null hypothesis of $\theta = 1$.)

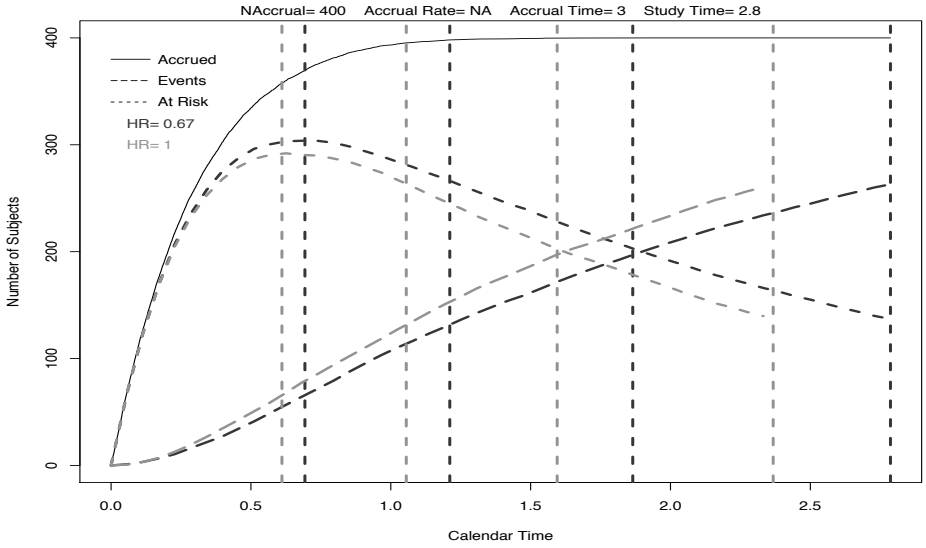
2.6 *Assessing the implications of varying patient accrual patterns*

The rate at which patients accrue will directly impact the observed censoring distribution in the trial when testing a time-to-event endpoint. For example, if accrual to the study were heavy at early times and slowed as the study progressed then a majority of patients would tend to have high censoring times relative to the maximal followup of the trial. On the hand, if the rate of accrual were low at the initial stages of the trial but increased towards the end of trial then a majority of patients would tend to have low censoring times relative to the maximal followup. More rigorously, if T_L denotes the total followup for the trial and T_A denotes the duration of accrual, the probability that a subject is observed for an event over the course of the study is given by

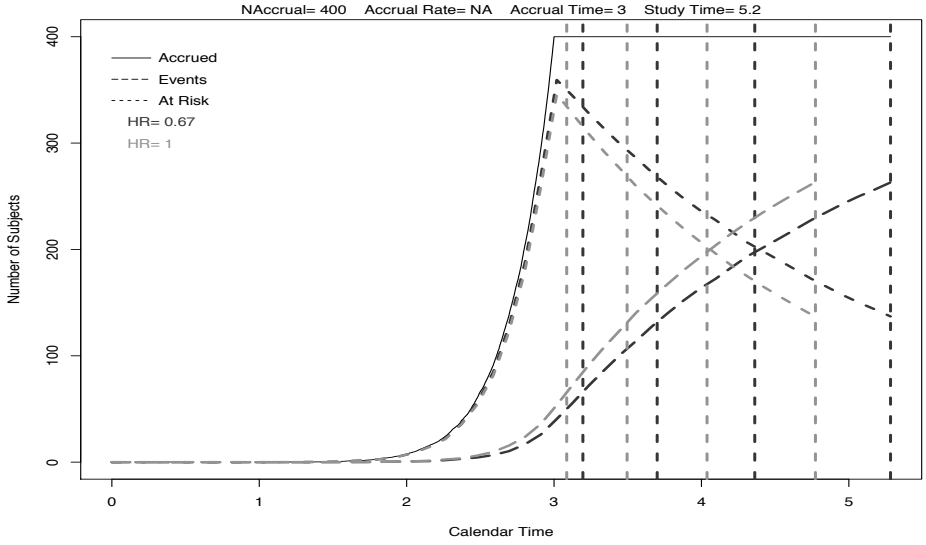
$$1 - \int_0^{T_A} S_T(T_L - u) f_A(u) du,$$

where S_T denotes the survival function of the subject and f_A denotes the probability density function of the accrual distribution of the subject. Because of this, changes to the accrual distribution can have economic (the duration of the study), statistical (the rate of statistical information growth) and scientific (the length of time treatments are to be compared) implications on a clinical trial.

Under a proportional hazards treatment effect, the statistical information as derived from the partial likelihood is directly proportional to the number of events observed on the trial. This reduces the complexity of planning interim analyses but one must still translate between the number of observed events and when those events are to be expected in calendar time so that a Data Monitoring Committee can be convened for interim analyses. The translation from events to calendar time requires specification of the accrual rate of patients, the duration of accrual, the duration of continued followup after accrual is closed, the baseline survival distribution, and the treatment effect. Most, if not all, of these parameters are unknown to



(a)



(b)

Figure 6: Expected event accrual rates and analysis times under fast (a) and slow (b) patient accrual patterns. Analysis times are for one-sided level 0.025 stopping rule to test the null hypothesis $H_0 : \theta \geq 1$ and having a maximum of 4 equally spaced analyses, an efficacy (lower) boundary shape function corresponding to $P = 1.1$ and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs. In each figure, the solid line represents the cumulative number of subjects accrued to the trial as a function of calendar time, the small dashed line represents the number of subjects still at risk in the trial, and the large dashed line represents the cumulative number of events observed in the trial.

investigators at the design stage of a trial and must be assumed. Therefore we find it useful to explore the potential impact of varying assumptions on the timing of analyses and the overall duration of the trial. In RCTdesign patient accrual patterns can be explored at the time of design specification. To provide flexibility in the exploration process, accrual rates may be parameterized via a Beta(a, b) distribution or simulated from existing pilot data. Similarly, baseline survival may be parameterized via a Weibull distribution, a piece-wise constant hazard function, or simulated from existing pilot data.

In the context of the CLL trial, Figure 6 depicts the event accrual rates and expected analysis times under under fast (a) and slow (b) patient accrual patterns. In both cases, a total of 400 subjects were assumed to enroll over a period of three years and baseline survival in the placebo group was assumed to follow an exponential distribution with a median survival of 16 months. In Figure 6a, a Beta(1,10) accrual distribution was assumed while in Figure 6b, a Beta(10,1) distribution was assumed. In each figure, the solid line represents the cumulative number of subjects accrued to the trial as a function of calendar time, the small dashed line represents the number of subjects still at risk in the trial, and the large dashed line represents the cumulative number of events observed in the trial. Lighter dashed lines depict estimates under the null hypothesis ($\theta = 1$) and darker dashed lines depict estimates under the full design alternative ($\theta = 0.67$). Entry distributions have been chosen to be extreme to highlight the impact of the patient accrual patterns. Specifically, under fast accrual the first interim analysis is estimated to take place between 7 and 8 months after study start. From a clinical perspective, this may be too soon to begin assessing efficacy because longterm survival effects are generally of primary interest. Conversely, under slow early accrual the first interim analysis is expected to take place more than three years after recruitment into the trial began. In the context of the trial, this may be too long of a wait to assess futility, particularly in light of the fact that by this time all 400 patients will have been recruited to the trial and treated. Also of note is the total expected duration of the study under each each scenario. Under the full alternative, the total study time could be as little as 2.8 years if early accrual is fast and as long as 5.2 years if early accrual is slow. Beyond the obvious clinical implications of estimating the treatment effect in a controlled setting for these different periods of time, the cost of running a trial for longer periods of time may not be feasible for a sponsor.

3. Implementing a group sequential design

3.1 *Flexible Implementation of Stopping Rules Based on Constrained Boundaries*

The stopping rule chosen in the design of a clinical trial serves as a guideline to a Data Monitoring Committee as it makes the decision to recommend continuing or stopping a clinical trial. If all aspects of the conduct of the clinical trial adhered exactly to the conditions stipulated during the design, the stopping rule obtained during the design phase could be used directly. However there are usually at least two complicating factors that must be dealt with during the conduct of the clinical trial. First, the schedule of interim analyses does not follow that used in the design of the trial. Often, meetings of the Data Monitoring Committee are scheduled according to calendar time, and thus the sample sizes available for analysis at any given meeting is a random variable. Similarly, accrual may be slower or faster than planned, thereby resulting in a different number of interim analyses than was originally planned. Either of these eventualities will necessitate modifications of the stopping rule, because the exact stopping boundaries are dependent upon the number and timing of analyses. Second, the estimate for response variability that was used at the design phase is typically incorrect. Often very crude estimates of response variability or baseline event rates are used at the design phase. As the trial progresses, more accurate estimates are to be used. Clearly the operating characteristics of particular stopping rules are heavily dependent on the variability of response measurement. In order to address these issues, flexible methods of implementing stopping rules have been developed which allow the clinical trialist to maintain at least some of the operating characteristics of the stopping rule. Typically such flexible methods always maintain the size (type I error) at the prescribed level. A choice must then be made as to whether the maximal sample size or the power to detect the design alternative should be maintained.

The flexible methods of implementing stopping rules in `RCTdesign` are based on the idea of computing a stopping boundary for the current interim analysis in such a way that the desired operating characteristics are satisfied and that the stopping rule is constrained to agree with the stopping boundaries used at all previously conducted interim analyses. Algorithmically, the monitoring strategy proceeds as follows:

1. At the first analysis, the stopping boundaries are derived by using the parametric boundary shape

family specified in the design. The exact stopping boundary is computed by considering the proportion Π_1 of statistical information available at that first analysis. The value of Π_1 depends on which operating characteristics of the stopping rule the trial designer chooses to preserve:

(a) If the maximal sample size (or number of events), N, is to be maintained, $\Pi_1 = N_1/N$. Here N_1 represents the number of subjects (or events) accrued at the first analysis.

(b) If the power of the test to detect the design alternative is to be maintained, a schedule of future analyses is assumed and a stopping rule using the design parametric family (possibly constrained) is found which has the desired power. This consists of searching for the value of N which has the correct type I error and power to detect the alternative for the parametric design family for the assumed schedule of interim analyses. In either case, interpolation of the exact, minimum, or maximum constraints specified at the design stage is used to derive any constraints for the interim analyses specified by the assumed schedule of future analyses (which may differ from the schedule specified at the design stage). In cases where statistical information is dependent upon a variance parameter, the current best estimate of the statistical information contributed by a single sampling unit is used instead of the estimate supplied at the design stage.

2. At later interim analyses, the exact stopping boundaries used at previously conducted interim analyses are used as exact constraints at those analysis times, and the stopping boundaries at the current analysis and all future analyses specified by an assumed schedule of future analyses are computed using the parametric family of designs specified at the design stage. The basic approach is that described for the first analysis, in which the proportion of statistical information at the j-th analysis is computed based either on the planned maximal sample size N if that operating characteristic is to be maintained, or it is computed based on a recomputation of a sample size which takes into account the new schedule of interim analyses and the current best estimate of the statistical information contributed by a single sampling unit. In either case, $\Pi_j = Nj/N$ is used as the proportion of statistical information available at the j-th analysis.

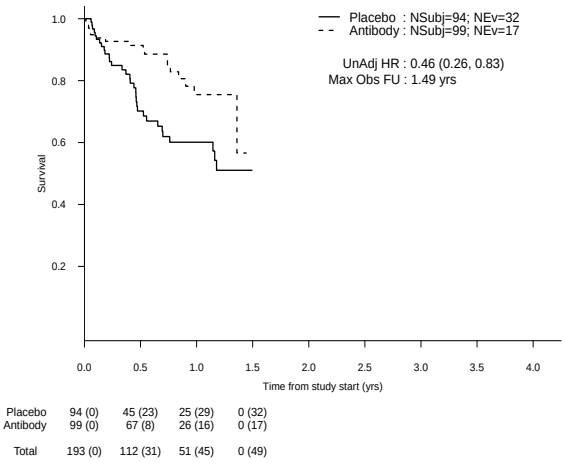
It should be noted that when a variance parameter is re-estimated at each analysis, the stopping boundaries at previously conducted interim analyses depend upon which boundary scale is used when constraining the stopping rules at those analyses. That is, if the value of the variance parameter used in computing

the stopping rule is constant over the course of the study, it is irrelevant which boundary scale is used for the constraints at previously conducted analyses. If, as is usually the case, the estimate of the variance parameter varies over the study, there will be some difference between the boundaries obtained. There is no clear advantage for one such scale over another.

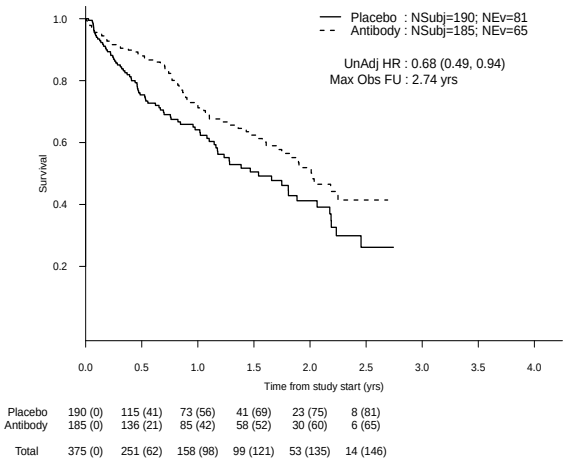
This approach based on constrained boundaries is a generalization of the error spending approach of Lan and DeMets (1983) and Pampallona et al. (1995): That approach corresponds to boundary constraints specified on the error spending scale. More recently, Burington and Emerson (2003) suggested the above constrained boundaries algorithm to allow a clinical trialist to constrain the stopping rule on any scale (eg. the sample mean scale) and for any parametric family of designs (eg. the unified family of group sequential designs).

To illustrate the use of the constrained boundaries approach in RCTdesign, Figure 7 depicts data simulated in the context of the CLL trial. To demonstrate the method, suppose that the *Eff11.Fut8* design (boundaries depicted in Figure 1) was the chosen stopping rule for the trial. Data were simulated under uniform patient accrual over three years assuming exponential survival times, with a median survival of 16 months in the placebo arm and a median survival of 22.9 months in the antibody arm (corresponding to a hazard ratio of 0.70). Figures (a)-(c) depict the observed survival curves and estimates of treatment effect at three interim analyses taking place at 1.5, 2.75, and 3.5 years after the start of trial enrollment. For reference, subfigure (d) depicts the data that would have been observed if a fixed sample design were performed after a total of 263 events were observed (occurring at 4.11 years after the start of trial enrollment).

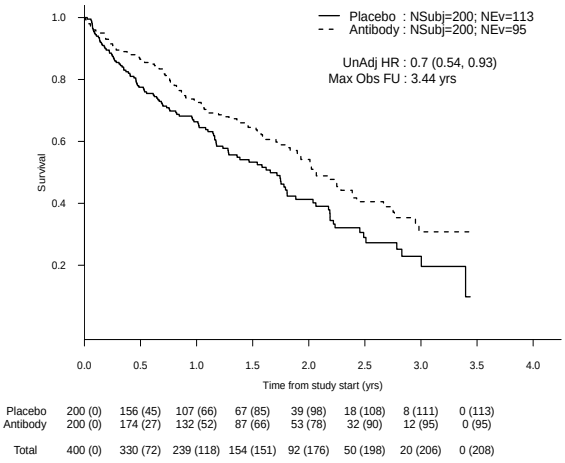
Table 2 depicts the observed statistics at each of the three interim analyses, including the total number of observed events, the estimated hazard ratio (not adjusted for the stopping rule), and the normalized Z statistic (also not adjusted or the stopping rule). In addition to the observed statistics at each analysis, Table 2 yields the modified stopping rule obtained from the constrained boundaries algorithm. Of note, the first analysis took place after 49 events were observed in the study and not the originally planned 66 events. This earlier analysis time results in a much wider continuation interval at the first analysis when compared to the original *Eff11.Fut8* stopping boundaries depicted in Figure 1. Given the observed hazard ratio of 0.46, the stopping rule suggested continuation of the trial and new stopping thresholds were derived using the



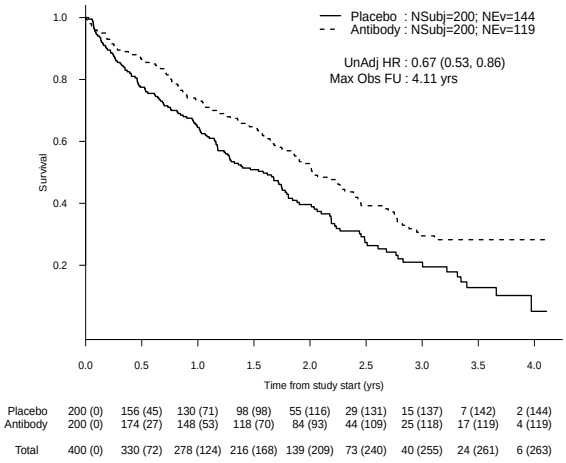
(a) Interim Analysis 1



(b) Interim Analysis 2



(c) Interim Analysis 3



(d) Fixed Sample Analysis

Figure 7: Data simulated in the context of the CLL trial. Figures (a)-(c) depict the observed survival curves and estimates of treatment effect at the three interim analysis times taking place at 1.5, 2.75, and 3.5 years after the start of trial enrollment. Figure (d) depicts the data that would have been observed if a fixed sample design were performed after a total of 263 events were observed (occurring at 4.11 years after the start of trial enrollment). In each plot, numbers under the x-axis indicate the number of patients at risk and (the cumulative number of events) observed at 6-month intervals.

original parametric family under a specified timing for the future analyses while maintaining an overall type I error rate of .025. The algorithm was again implemented at the second analysis occurring after 146 events

Table 2: Implementation of the original *Eff11.Fut8* using constrained boundaries to account for variation in the originally planned analysis times. Original analysis times were planned at 66, 132, 198, and 263 events. Data were simulated in the context of the CLL trial and are depicted in Figure 7. Due to deviations from patient and event accrual rates, the first three interim analysis actually took place at 49, 146, and 208 events.

	Observed Statistics			Modified Stopping Boundaries[1,2]			
Analysis	No. Events	Crude HR	Normalized Z statistic	Time	Efficacy	Futility	Stopping Rule Recommendation
1	49	0.462	-2.628	NEv= 49	0.214	1.619	Continue
	–	–	–	NEv=132	0.595	0.947	
	–	–	–	NEv=198	0.718	0.837	
	–	–	–	NEv=263	0.784	0.784	
2	49	0.462	-2.628	NEv= 49	0.214	1.619	Continue
	146	0.678	-2.342	NEv=146	0.629	0.915	Continue
	–	–	–	NEv=198	0.717	0.837	
	–	–	–	NEv=263	0.784	0.784	
3	49	0.462	-2.628	NEv= 49	0.214	1.619	Continue
	146	0.678	-2.342	NEv=146	0.629	0.915	Continue
	208	0.704	-2.522	NEv=208	0.730	0.826	Stop (Efficacy)
	–	–	–	NEv=263	0.784	0.784	

1: NEv denotes the observed number of events
2: Stopping boundaries are displayed on the scale of the estimated hazard ratio

were observed (differing from the previously assumed 132 events). We note that the stopping thresholds at the first analysis remained unchanged, while future stopping boundaries are recomputed. Again, the stopping rule suggested continuation of the study. The process was repeated at the third interim analysis, where a hazard ratio of 0.70 was computed (not adjusted for the stopping rule) after observing 208 events. At this analysis, the stopping rule suggested stopping the trial in favor of efficacy. Figure 8 provides a visual comparison of the original *Eff11.Fut8* (solid lines), having a maximum of 4 equally spaced analyses, an efficacy (lower) boundary shape function corresponding to $P = 1.1$ and a futility (upper) boundary shape function corresponding to P = 0.8 in the unified family of group sequential designs, and the final implemented design (dashed lines) using constrained boundaries to account for variation in the originally planned timing of analyses. In the figure, 'X' denotes the observed point estimate (hazard ratio) at the three interim analyses.

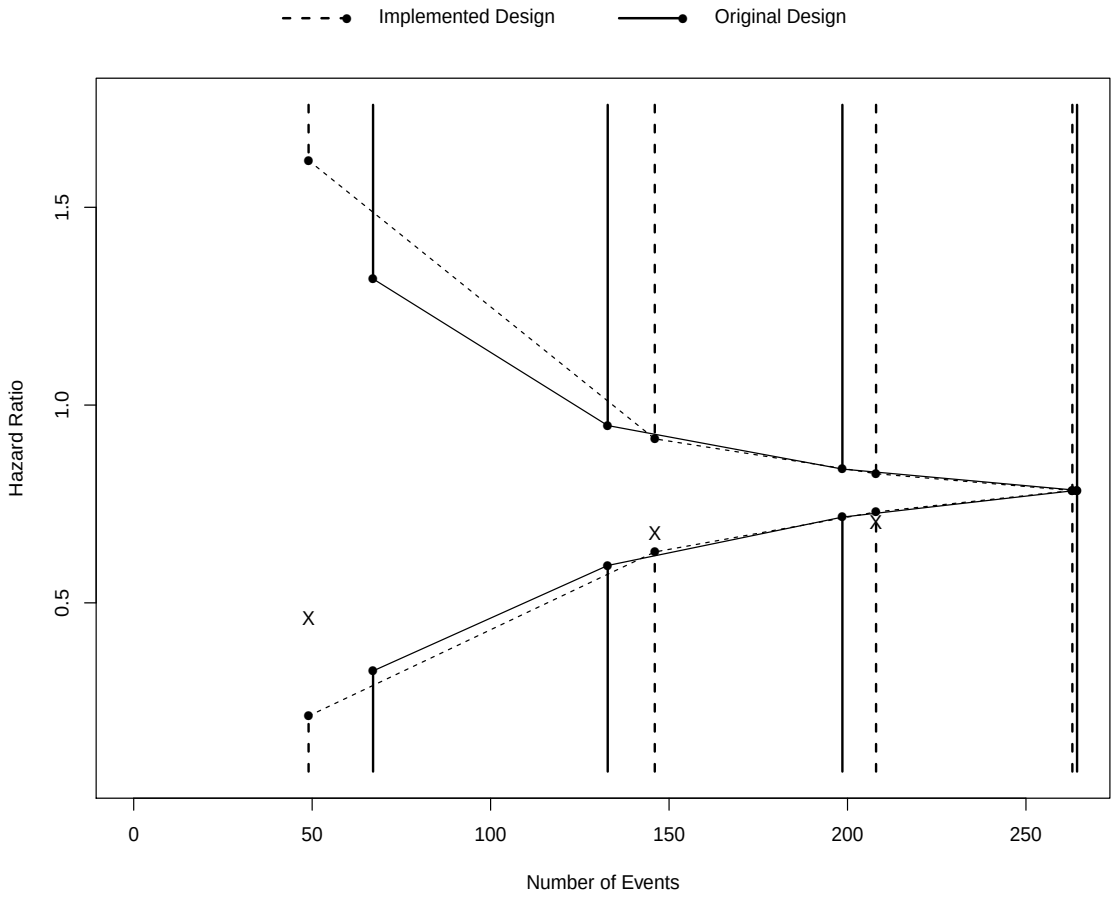


Figure 8: Comparison of the original *Eff11.Fut8* (solid lines), having a maximum of 4 equally spaced analyses, an efficacy (lower) boundary shape function corresponding to $P = 1.1$ and a futility (upper) boundary shape function corresponding to $P = 0.8$ in the unified family of group sequential designs, and the final implemented design (dashed lines) using constrained boundaries to account for variation in the originally planned timing of analyses. 'X' denotes the observed point estimate (hazard ratio) at the first three analyses. As indicated in the plot, the stopping rule recommended stopping at the third analysis so that no observed estimate is provided at the final analysis.

3.2 *Adjusted inference*

As previously stated, the use of a group sequential stopping rule generally alters the sampling distribution of usual fixed sample statistics. Therefore special techniques must be used to compute point estimates, interval estimates and P values. Commonly reported point estimates include the usual maximum likelihood

estimate (MLE), the median unbiased estimator (MUE; see Whitehead, 1997), the bias adjusted mean (BAM; Whitehead, 1986a), and the Rao-Blackwell adjusted unbiased estimate (RBUE; Liu and Hall, 1999).

In order to compute a MUE, P value, or confidence interval which adjusts for the stopping rule used in a group sequential trial, an ordering of possible clinical trial outcomes must be chosen. There is no uniformly optimal choice for such an ordering. In group sequential testing, the issue is how to treat outcomes observed at different analyses (see Emerson and Fleming, 1990). RCTdesign offers two approaches: the sample mean ordering (Emerson and Fleming, 1990) and the analysis time ordering (Tsiatis et al., 1984).

The sample mean ordering judges one result more extreme than another according to whether the estimate of the treatment effect is more extreme. Thus, a treatment effect measured by a hazard ratio of 0.6 is lower than a treatment effect measured by a hazard ratio of 0.7, regardless of the analysis time. In the analysis time ordering , results that led to earlier termination of the study are judged to be more extreme than those observed at later analyses. Results that exceed an upper boundary for the treatment effect at a specific analysis are higher than all results exceeding the upper boundary at later analyses, and also higher than all results less than the lower boundary at any analysis. Thus, a treatment effect measured by a hazard ratio of 0.6, which was judged so high as to warrant early termination of the study, is less extreme than a hazard ratio of 0.7 which was similarly judged high enough to warrant termination of the study at an earlier analysis.

Emerson and Fleming (1990) investigated the relative behavior of the sample mean and analysis time orderings with respect to the average width of confidence intervals. The sample mean ordering tends to average shorter confidence interval lengths for the same coverage probabilities. Gillen and Emerson (2005b) more recently showed that under a time-varying treatment effect, the sample mean ordering tends to attain higher power relative to the analysis time ordering in the sense that the probability of attaining a small P value is higher with the sample mean ordering when compared to the analysis time ordering. Finally, the analysis time ordering is not defined for two-sided group sequential tests that allow early stopping under both the null and alternative analyses. For these reasons, the sample mean ordering is the recommended method of computing ordering-dependent inference in RCTdesign. However, the sample mean ordering does depend on the number and timing of future analyses, but such dependence was found to be fairly slight by

Table 3: Inference adjusted for the *Eff11.Fut8* stopping rule. Data were simulated in the context of the CLL trial and are depicted in Figure 7. Due to variability in patient and event accrual rates, the first three interim analysis actually took place at 49, 146, and 208 events. Based upon the sequential boundaries the trial was stopping at the third interim analysis.

Result	Analysis Time Ordering	Sample Mean Ordering
Unadjusted Estimate	0.7044	
Adjusted Estimates		
BAM	0.7127	
RBadj	0.7167	
MUE	0.7074	0.7153
Adjusted Inference		
95% CI	(0.5382, 0.9313)	(0.5469, 0.9347)
P value	0.006906	0.007238

Emerson and Fleming (1990).

Table 3 depicts the resulting inference at the conclusion of the simulated CLL trial. Based upon the implemented *Eff11.Fut8* design the trial was stopped at the third analysis. The observed hazard ratio (unadjusted for the stopping rule) was 0.7044. As can be seen from Table 3, each of the adjusted estimates are attenuated towards the null hypothesis. This adjustment for bias is slight in the example due to the conservativeness of the *Eff11.Fut8* stopping rule. Had a less conservative design been chosen, a larger difference between the unadjusted and adjusted estimates would have been observed. Also reported in Table 3 are the corrected 95% confidence intervals and P values based upon the analysis time and sample mean orderings. Again the two orderings produce similar results. This is because the trial continued to the penultimate analysis before stopping. Had the trial stopped earlier, at the first analysis for example, the difference in the inference obtained from the two orderings would have been more extreme.

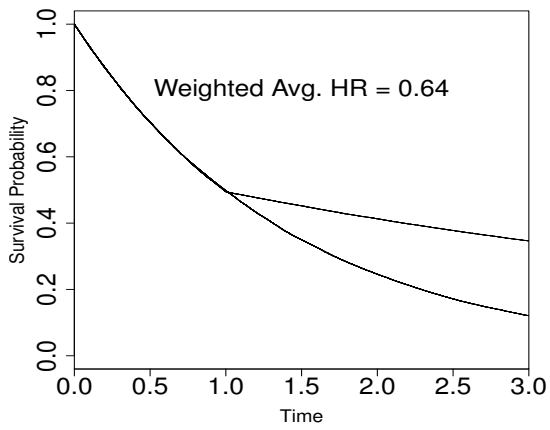
4. Consideration of potential time-varying treatment effects

The methods discussed in this manuscript have focused on settings in which the measure of treatment effect does not vary with time. However, it is often the case that a given treatment might have a delayed effect within individuals or that the effect of treatment might dissipate over time. Special issues arise in such settings. For instance, when using nonparametric statistics to analyze survival data exhibiting nonproportional hazards one must consider (among other things):

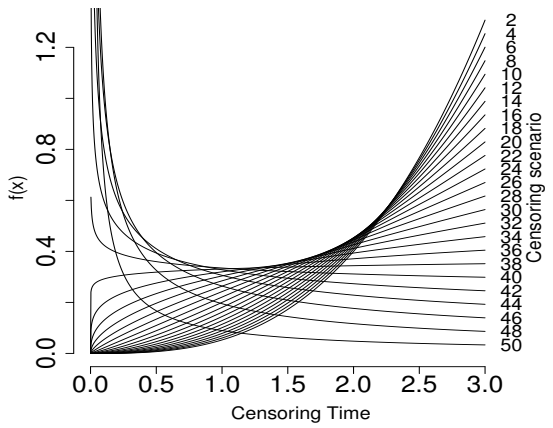
1. The formulation of alternatives at which operating characteristics are to be evaluated.
2. The rate of information growth of the test statistic for appropriately timing interim analyses.
3. The changing censoring distribution across interim analyses and its impact on the asymptotic distribution of the test statistic under alternatives.

In a further extension to the evaluation paradigm demonstrated here, Gillen and Emerson (2011) describe one general approach to the evaluation of clinical trial designs in the setting of nonproportional hazards. Gillen and Emerson (2011) note that in the presence of nonproportional hazards survival data, nonparametric methods such as the $G^{\rho,\gamma}$ family of weighted logrank statistics (Fleming and Harrington, 1991) are often used and the evaluation of stopping rules is no longer a trivial task. Specifically, nonparametric test statistics do not necessarily correspond to a parameter of clinical interest, thus making it difficult to characterize alternatives at which operating characteristics are to be computed. It is shown that this sometimes leads to contradictions when reporting clinically meaningful measures of treatment effect in the event that they do not correspond to the nonparametric statistic on which testing is based. Gillen and Emerson (2011) go on to describe re-sampling approaches which might be used to construct alternatives under nonproportional hazards when pre-existing pilot data are available. Those methods can be implemented using the `RCTdesign` package as a foundation for generating stopping boundaries.

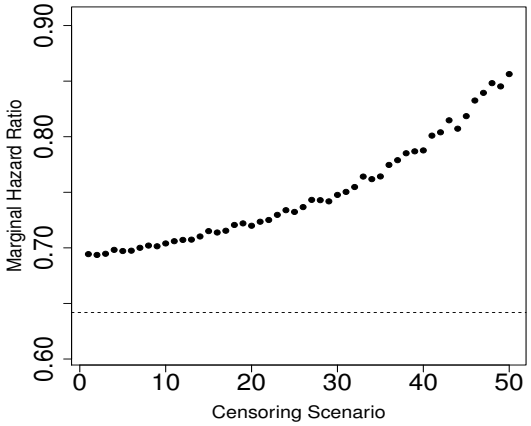
It was noted in Section 2.6 that under a proportional hazards treatment effect, statistical information is directly proportional to the number of events observed on the trial. Because of this, the timing of interim analyses need only map the the number of accrued events to calendar time under a proportional hazards assumption. However, when testing is based upon a weighted statistic, such as the $G^{\rho,\gamma}$ family of weighted logrank statistics (perhaps to emphasize particular time intervals where treatment effects are of greatest clinical importance), the growth of statistical information is non-linear with respect to the cumulative number of observed events (see Gillen and Emerson, 2005a). Specifically, the amount of information contributed by each event is dependent upon when the event occurred as well as the accrual distribution. Building on the work of Gillen and Emerson (2005a) and Burington and Emerson (2003), Brummel and Gillen (2011) describe a general constrained boundaries algorithm that can be used to flexibly monitor a group sequential survival trial under non-linear information growth patterns. This procedure modifies the usual constrained boundaries



(a) True Survival Configurations



(b) Censoring pdfs



(c) Estimand from the Cox Model

Figure 9: Example illustrating the effect of the censoring distribution on the parameter consistently estimated by the Cox proportional hazards model under 50 different censoring distributions (shown in subfigure (b)). Data were generated under the survival configurations presented in subfigure (a) and and the resulting estimand from each censoring scenario is plotted in subfigure (c).

algorithm described in Section 3.1 by using observed survival and accrual data at each interim analysis to predict the information growth curve, then mapping information accrual to calendar time. Because the method is an extension of the constrained boundaries approach implemented in `RCTdesign` it can easily be implemented in the current software.

Finally, multiple authors have noted that the parameter consistently estimated by the Cox proportional hazards model and the logrank statistic are dependent upon the observed censoring distribution when the

proportional hazards assumption does not hold (cf. Struthers and Kalbfleisch, 1986; Gillen and Emerson, 2007). This dependence is not only on the support of the underlying censoring distribution but also the shape of the distribution. As an illustration, Figure 9 depicts the effect of the censoring distribution on the parameter consistently estimated by the Cox proportional hazards model under 50 different censoring distributions (shown in Figure 9b). In this simulation study, survival curves were generated under the piecewise constant hazard ratio alternative depicted in Figure 9a. With no censoring over three years the Cox model consistently estimates a 'marginal' or 'weighted' hazard ratio of 0.64. However, as censoring is introduced Figure 9c illustrates that the parameter consistently estimated by the Cox model can vary from 0.67 (no censoring) to 0.87 (heavy early censoring as depicted in scenario 50). To remove this dependence, Xu and O'Quigley (2000) suggested an inverse probability of censoring estimator that assumes a common censoring pattern across all comparison groups. Boyd et al. (2011) later extended the weighted estimator to allow for group dependent censoring in the case of a two sample comparison and derived a consistent variance estimator in this case. Most recently Nguyen and Gillen developed a censoring robust reweighted estimator for discrete survival outcomes in the two sample setting (Nguyen and Gillen, 2011b) and proposed a method to provide robust estimation of survival effects under covariate-dependent censoring in observational studies (Nguyen and Gillen, 2011a) . Because it is necessary to *a priori* specify the estimation and testing procedure to be used in a clinical trial, the above estimators are attractive in that they limit the influence of study accrual/dropout patterns on trial results under a misspecified model. In addition, although these estimators are not directly implemented in `RCTdesign` at the present time, the ability to monitor a normalized Z statistic using `RCTdesign` provides clinical trialists with a software tool that can be adapted to monitor any statistic that can be suitably normalized.

5. Discussion

In this manuscript we have demonstrated how the `RCTdesign` package can be used to select, implement, and analyze a group sequential stopping rule. The `RCTdesign` package aids in the complete evaluation of a clinical trial design by easily allowing clinical trialists to compare a broad range of candidate designs with respect to:

1. The scientific measures of treatment effect which will correspond to early termination for futility and/or

efficacy.

2. The sample size requirements as described by the maximal sample size and summary measures of the sample size distribution (e.g., mean, 75th percentile) as a function of the hypothesized treatment effect.

3. The probability that the trial would continue to each analysis as a function of the hypothesized treatment effect.

4. The frequentist power to reject the null hypothesis as a function of the hypothesized treatment effect, with the type I error corresponding to the power under the null hypothesis.

5. The frequentist inference (adjusted point estimates, confidence intervals, and P values) which would be reported were the trial to stop with results corresponding exactly to a boundary.

6. The frequentist power to obtain a point estimate above some relevant threshold.

7. The expected timing of interim analyses as a function of patient accrual patterns.

After the selection of a group sequential stopping rule, flexible implementation of the sequential boundaries using a constrained boundaries approach was demonstrated. This method easily accounts for deviations in planned variance, timing, and number of analyses in order to maintain some of those operating characteristics specified at the design stage. With careful evaluation of stopping rules and methods for flexibly implementing those rules under changing circumstances, there seems little reason to resort to less efficient adaptive designs such as those based on using conditional power to re-design a study (Proschan and Hunsberger, 1995) or Fisher's "self-designing clinical trial" (Fisher, 1998). The most frequently cited motivation for using such adaptive designs include the possibility that at an interim analysis a clinical trialist might observe treatment effects that were promising, but not statistically significant, and thus want to continue the clinical trial to obtain a larger sample size. Of course, as noted in this manuscript, by examining the stopping boundary on the scale of the estimated treatment effect, all such possibilities can truly be considered at the design stage, and there is no real need to accommodate adaptive designs based solely on the estimate of the primary measure of treatment effect. Additionally, if conditions external to the trial suggest a change in the clinical or economic importance of particular alternative hypotheses or estimates of treatment effect, redesign of the clinical trial can proceed without materially affecting the type I error, because in that setting

the factors affecting the redesign of the trial are not based on the trial results. This then argues that there is no real need for using adaptive designs. Furthermore, there are distinct disadvantages to the adaptive methods, most notably those related to the loss of statistical efficiency (Tsiatis and Mehta, 2003; Emerson, 2006).

The CLL case study used throughout this manuscript considered a survival endpoint and the design approach relied on a semi-parametric (proportional hazards) model. Robust methods for the analysis of survival data under a time-varying treatment effect remains an active area of research. Section 4 discusses multiple approaches to modify common survival statistics in order to limit the impact the censoring distribution in these settings. While these methods are effective for removing the dependence of the resulting estimand on the censoring distribution under fixed support, they do not address the dependence of these estimators on the underlying length of trial followup. This is a particular problem in the case of sequential testing where interim analyses inherently truncate the observed support of the survival distribution. Gillen (2009) proposes one method for quantifying uncertainty in future treatment effects by utilizing a random walk approach to generate future alternatives which might reasonably be observed conditional upon data collected up to the time of an interim analysis. Similar methods could also be used in the design, evaluation, and monitoring of longitudinal studies, since the potential for time-varying treatment effects in these settings forces one to consider future alternative which might arise following an interim analysis.

Finally, the current manuscript has focused on the evaluation of frequentist operating characteristics. Increasingly, however, there has been much interest in the design and analysis of clinical trials under a Bayesian paradigm. While not demonstrated here, the RCTdesign package also allows for the Bayesian evaluation of group sequential designs. For further reading on this topic, the reader should see Emerson et al. (2007a).

References

Boyd, A., Kittelson, J. and Gillen, D. (2011). Estimation of treatment effect under nonproportional hazards and covariate dependent censoring. *Statistics in Medicine* (In Revision).

Brummel, S. and Gillen, D. (2011). Flexibly monitoring group sequential survival trials using constrained boundaries. *Under Review* .

Burington, B. E. and Emerson, S. S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* **59**, 770–777.

EaSt (2000). The Cytel Software Corp. Cambridge, Massachusetts.

Emerson, S. S. (2006). Issues in the use of adaptive clinical trial designs. *In Press : Statistics in Medicine* .

Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905–923.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875–892.

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2007a). Bayesian evaluation of group sequential designs. *Statistics in Medicine* **26**, 1431–1449.

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2007b). Frequentist evaluation of group sequential designs. *Statistics in Medicine* **26**, 5047–5080.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–1562.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley.

Gillen, D. L. (2009). A random walk approach for quantifying uncertainty in group sequential survival trials. *Computational Statistics and Data Analysis* **53**, 603–620.

Gillen, D. L. and Emerson, S. S. (2005a). Information growth in a family of weighted logrank statistics under repeated analyses. *Sequential Analysis* **24**, 1–22.

Gillen, D. L. and Emerson, S. S. (2005b). A note on P-values under group sequential testing and nonproportional hazards. *Biometrics* **61**, 546–551.

Gillen, D. L. and Emerson, S. S. (2007). Non-transitivity in a class of weighted logrank statistics under non-proportional hazards. *Statistics and Probability Letters* **77**, 123–130.

Gillen, D. L. and Emerson, S. S. (2011). Evaluating a group sequential design in the setting of non-

proportional hazards. *Submitted* .

Kittelson, J. M. and Emerson, S. S. (1999). A unifying family of group sequential test designs. *Biometrics* **55**, 874–882.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Liu, A. and Hall, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika* **86**, 71–78.

Nguyen, V. and Gillen, D. (2011a). Robust inference in semiparametric discrete hazard models for observational studiesSubmitted for publication.

Nguyen, V. and Gillen, D. (2011b). Robust inference in semiparametric discrete hazard models for randomized clinical trialsSubmitted for publication.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pampallona, S., Tsiatis, A. and Kim, K. (1995). Spending functions for the type i and type ii error probabilities of group sequential tests.

PEST (2000). *Planning and Evaluation of Sequential Trials*. The MPS Research Unit, The University of Reading, Reading, U.K.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–200.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

SAS/SEQDESIGN (2011). SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

S+SeqTrial (2002). Insightful Corporation, Seattle, Washington.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

Tsiatis, A. A. and Mehta, C. R. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–378.

Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential

trials. *Biometrics* **43**, 193–199.

Whitehead, J. (1986a). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**, 573–581.

Whitehead, J. (1986b). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42**, 461–471.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. John Wiley & Sons.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions (corr: V39 p1137). *Biometrics* **39**, 227–236.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics (Oxford)* **1**, 423–439.

Appendix A : RCTdesign code to recreate the CLL examples in Sections 2 and 3 using R

```
##
##### Definition of candidate designs for the CLL trial
##
Fixed.Sample <- seqDesign( prob.model = "hazard", arms = 2, null.hypothesis = 1.,
      alt.hypothesis = 0.67, ratio = c(1., 1.), nbr.analyses = 1,
      test.type = "less", sample.size=263, power = "calculate", alpha = 0.025 )
SymmOBF.2 <- update( Fixed.Sample, nbr.analyses=2, P=c(1,1), sample.size=263, power="calculate" )
SymmOBF.3 <- update( SymmOBF.2, nbr.analyses = 3 )
SymmOBF.4 <- update( SymmOBF.2, nbr.analyses = 4 )
SymmOBF.Power <- update( SymmOBF.4, power = 0.901 )
Futility.5 <- update( SymmOBF.4, P=c(1,.5) )
Futility.8 <- update( SymmOBF.4, P=c(1,.8) )
Futility.9 <- update( SymmOBF.4, P=c(1,.9) )
Eff11.Fut8 <- update( SymmOBF.4, P=c(1.1,.8) )
Eff11.Fut9 <- update( SymmOBF.4, P=c(1.1,.9) )
Fixed.Power <- update( Fixed.Sample, nbr.analyses=1, power=0.8853 )

##
##### Figure 1 : Comparison of stopping boundaries on crude estimate of treatment effect scale
##
seqPlotBoundary(  SymmOBF.4, Eff11.Fut8, Eff11.Fut9, lty=c(1,3,4), col=1, stagger=0, fixed=FALSE )
seqBoundary( Eff11.Fut8, scale="X" )
seqBoundary( Eff11.Fut8, scale="Z" )
1-seqBoundary( Eff11.Fut8, scale="P" )

##
##### Figure 2 : Comparison of statistical power curves
##
seqPlotPower(SymmOBF.4,SymmOBF.3,SymmOBF.2, lty=1:4, col=1, lwd=2 )
seqPlotPower(SymmOBF.4,SymmOBF.3,SymmOBF.2, reference=TRUE, lty=1:4, col=1, lwd=2 )
```

```
##
#####  Table 1 : Computation of power and alternative tables for the Eff11.Fut8 design
##
seqOC( Eff11.Fut8, power=c(.8,.9,.95,.975) )
seqOC( Eff11.Fut8, theta=c(1,.75,.67,.60) )

##
##### Figure 3 : Comparison of sample size distributions
##
seqPlotASN(SymmOBF.4,Futility.9,Futility.8,Futility.5, fixed=FALSE, lty=c(2,1,3,4), col=1, lwd=2)

##
##### Figure 4 : Depiction of stopping probabilities
##
seqPlotStopProb(Eff11.Fut8)

##
##### Figure 5 : Statistical inference on the boundaries
##
plot(seqInference(Eff11.Fut8))

##
##### Figure 6a : Patient accrual patterns (early accrual)
##
Eff11.Fut8Extd.early <- seqDesignExtd(prob.model = "hazard", arms = 2, null.hypothesis = 1.,
      alt.hypothesis = 0.67, ratio = c(1., 1.), nbr.analyses = 4,
      test.type = "less", alpha = 0.025, sample.size=263, power="calculate", P=c(1.1,.8),
      accrualSize=400, accrualTime=3, bShapeAccr=10, eventQuantiles=16/12, nPtsSim=10000, seed=0)
seqPlotPHNSubjects(Eff11.Fut8Extd.early)
```

```
##
##### Figure 6b : Patient accrual patterns (late accrual)
##
Eff11.Fut8Extd.late <- seqDesignExtd(prob.model = "hazard", arms = 2, null.hypothesis = 1.,
      alt.hypothesis = 0.67, ratio = c(1., 1.), nbr.analyses = 4,
      test.type = "less", alpha = 0.025, sample.size=263, power="calculate", P=c(1.1,.8),
      accrualSize=400, accrualTime=3, aShapeAccr=10, eventQuantiles=16/12, nPtsSim=10000, seed=0)
seqPlotPHNSubjects(Eff11.Fut8Extd.late)

##
##### Simulation of CLL data
##
set.seed( 123456 )
n <- 200
grp1 <- rexp( n, rate=.75*log(2) )
grp2 <- rexp( n, rate=(.75*log(2))*.70 )
trueSurv <- c( grp1, grp2 )
entry <- runif( 2*n, 0, 3 )
grp <- rep( 0:1, each=n )

## First analysis at 1.5 years after study start
analysisTime <- 1.5
obsSurv <- ifelse( trueSurv + entry <= analysisTime, trueSurv, analysisTime-entry )
event <- ifelse( obsSurv == trueSurv, 1, 0 )
cllData <- as.data.frame( cbind( grp, entry, obsSurv, event ) )
cllData <- cllData[ cllData$obsSurv > 0, ]
resp <- Surv( cllData$obsSurv, cllData$event )
interim1 <- seqMonitor( Eff11.Fut8, response=resp, treatment=cllData$grp, future.analyses=c(132,198,263) )

## Second analysis at 2.75 years after study start
analysisTime <- 2.75
obsSurv <- ifelse( trueSurv + entry <= analysisTime, trueSurv, analysisTime-entry )
```

```
event <- ifelse( obsSurv == trueSurv, 1, 0 )
cllData <- as.data.frame( cbind( grp, entry, obsSurv, event ) )
cllData <- cllData[ cllData$obsSurv > 0, ]
resp <- Surv( cllData$obsSurv, cllData$event )
interim2 <- seqMonitor( interim1, response=resp, treatment=cllData$grp, future.analyses=c(198,263) )

## Third analysis at 3.5 years after study start
analysisTime <- 3.5
obsSurv <- ifelse( trueSurv + entry <= analysisTime, trueSurv, analysisTime-entry )
event <- ifelse( obsSurv == trueSurv, 1, 0 )
cllData <- as.data.frame( cbind( grp, entry, obsSurv, event ) )
cllData <- cllData[ cllData$obsSurv > 0, ]
resp <- Surv( cllData$obsSurv, cllData$event )
interim3 <- seqMonitor( interim2, response=resp, treatment=cllData$grp, future.analyses=c(263) )

##
##### Figure 8 : Comparison of implemented and original design
##
plot( interim3, dsnLbls=c("Implemented Design", "Original Design") )

##
##### Table 3 : Inference adjusted for the stopping rule
##
print( interim3 )
```