UW Biostatistics Working Paper Series
University of Washington

Year 2007 *Paper* 307

Evaluating a Group Sequential Design in the Setting of Nonproportional Hazards

Daniel L. Gillen
University of California, Irvine

Scott S. Emerson
University of Washington

Evaluating a Group Sequential Design in the Setting of Nonproportional Hazards

Abstract

Group sequential methods have been widely described and implemented in a clinical trial setting where parametric and semiparametric models are deemed suitable. In these situations, the evaluation of the operating characteristics of a group sequential stopping rule remains relatively straightforward. However, in the presence of nonproportional hazards survival data nonparametric methods 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. We describe an approach for constructing alternatives under nonproportional hazards using pre-existing pilot data, allowing one to evaluate various operating characteristics of candidate group sequential stopping rules. The method is illustrated via a case study in which testing is based upon a weighted logrank statistic.

1. Introduction

Prior to conducting large scale confirmatory, or phase III, clinical trials it is common for researchers to first conduct phase I and II exploratory trials concerned with toxicology and pharmacology. Although these preliminary trials are generally not adequately powered to prove efficacy or non-inferiority, important information regarding potential treatment benefits can often be gleamed before proceeding to a larger trial. Figure 1 displays survival curves resembling those observed during a recent phase II trial considering the potential efficacy of a new cancer chemotherapy treatment on all-cause mortality. In this phase II trial, researchers observed that the experimental treatment being considered yielded a nonproportional hazards effect on survival, generally characterized by a delayed effect on the hazard. Study sponsors hypothesized that this delay in the separation of hazards may have been attributed to several factors, including the need for a minimum time required for the treatment to show an effect within patients or because there may exist a subset of the sickest patients for which the occurrence of an event is inevitable regardless of treatment assignment. Given the observed treatment effect over mid to late followup times, it was agreed by the sponsors that it would be appropriate to proceed to a larger confirmatory trial investigating the efficacy of the experimental treatment. The statistical process involved in bringing this study to fruition involved three main steps: (1) choosing a test statistic that would be likely to capture alternatives from the null hypothesis of equal survival that were felt most scientifically relevant and plausible; (2) constructing alternatives that might reasonably arise in a future study in order to evaluate the performance of the proposed test statistic; and (3) choosing a group sequential stopping rule that would allow for early stopping in the event that sufficient confidence in favor of a decision for efficacy or futility of the treatment were observed. It was recognized by the study designers that, at least some, focus on the previously observed phase II data would be necessary to suggest appropriate probability models for data which might be obtained during the confirmatory trial and to evaluate the adequacy of potential stopping rules used to monitor accruing data. In this manuscript we report our approach to the statistical design of the proposed trial and provide one framework for designing and evaluating group sequential survival studies when pilot data suggest alternatives which deviate from the usual proportional hazards assumption.

Early in the design process, it was agreed by the study sponsors and trial designers that some form of a weighted logrank statistic, highlighting treatment effects occurring during mid to late followup, would be a

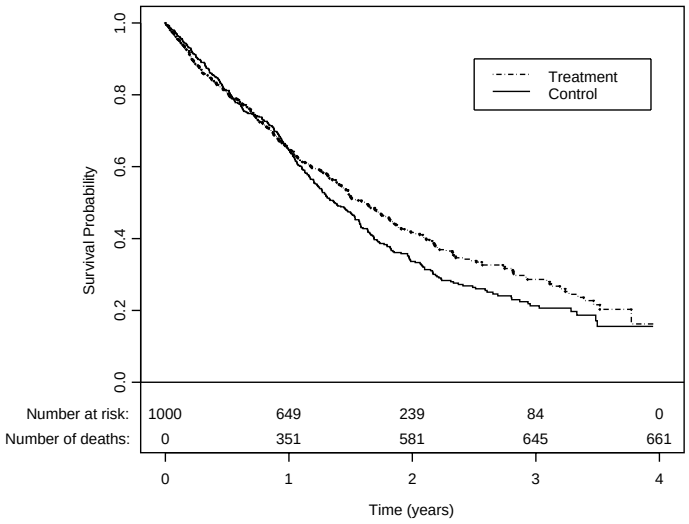


Figure 1: Kaplan-Meier estimates of survival curves resembling those observed in a phase II trial considering an experimental treatment for cancer on all cause mortality.

reasonable choice for use in the confirmatory trial. The $G^{\rho,\gamma}$ class of weighted logrank statistics as defined by Fleming and Harrington (1991) is given by

$$G^{\rho,\gamma} = \left(\frac{N_1 + N_0}{N_1 N_0}\right)^{1/2} \sum_{j \in \mathcal{F}} w_j \left[\hat{\lambda}_{1j} - \hat{\lambda}_{0j}\right], \tag{1}$$

where N_i denotes the initial sample size of group i, $i = 0, 1$, $\mathcal{F}$ denotes the set of distinct observed failure times in the pooled sample, $w(t)$ is defined as $\{(n_{1t}n_{0t})/(n_{1t} + n_{0t})\}\,\hat{S}(t-)^{\rho}[1 - \hat{S}(t-)]^{\gamma}$, n_{it} denotes the number of persons at risk in group i at time t, $\hat{S}(t-)$ denotes the Kaplan-Meier estimate of the pooled sample just prior to time t, $\hat{\lambda}_i(t) \equiv d_{it}/n_{it}$ denotes the estimated hazard for group i at time t, and d_{it} represents the number of deaths observed in group i at time t. Thus, the $G^{\rho,\gamma}$ statistic represents a sum, over all failure times, of weighted differences in the estimated hazards. Based upon the available pilot data, trial designers concluded that the $G^{1,1}$ statistic, which places increased weight to hazard differences estimated near the median survival time of the pooled groups, would be provide a functional of the survival distributions that would efficiently detect delayed treatment effect alternatives taking the general form of those observed in the phase II data.

In designing the phase III trial, it was deemed ethically necessary that a group sequential analysis plan be implemented to monitor data and stop the trial as soon as sufficient confidence in favor of efficacy or futility was obtained. 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 Emerson, 2003), and analysis (Whitehead, 1986; Emerson and Fleming, 1990) of group sequential trials. In particular, the body of group sequential methodology is well-defined for situations in which the within-individual treatment effect is constant with respect to time. Common statistical techniques that have been developed in this setting include the comparison of means of continuous data, proportions or odds resulting from binomial data, and proportional hazards from censored time to event data. In comparison, little attention has been devoted to the development of group sequential methods for monitoring a treatment effect which may vary with time, as the pilot data in Figure 1 suggest.

In order to ensure desirable operating characteristics of the group sequential design ultimately selected for use in the phase III trial, a comprehensive evaluation of multiple stopping rules was necessary. Emerson et al. (2004b) 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 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.

Although these operating characteristics are relatively straightforward to evaluate in the setting of a time-invariant treatment effect, as described by Emerson et al. (2004b), they have not been fully addressed in longitudinal situations where the observed treatment effect within individuals may vary over the course of follow-up.

In order to evaluate power curves, sample size distributions, and futility of continuing the trial, it is necessary to construct alternatives at which these operating characteristics are to be estimated. Because our choice of test statistic, $G^{1,1}$, does not necessarily correspond to a parameter of clinical interest, characterizing meaningful alternatives from the null hypothesis can be difficult. Further, in order to fully address the scientific question of interest it is generally preferred to present stopping boundaries based upon a statistic which represents some clinically meaningful measure of treatment efficacy. This logic no longer holds when using the nonparametric $G^{1,1}$ statistic since this statistic does not correspond to a specific parameter of interest. Hence care must be taken when considering potential stopping rules in order to examine what point estimates for clinically meaningful measures arise upon study termination.

In the remainder of the manuscript we describe the procedure used for evaluating and selecting a group sequential stopping rule to be applied to the phase III confirmatory study. In order to evaluate the operating characteristics of these designs, it was first necessary to construct alternatives which were deemed plausible in the confirmatory trial given the prior information available regarding a nonproportional hazards treatment effect. Section 2 introduces a bootstrapping procedure used to simulate potential hypothesis testing alternatives using the observed pilot data displayed in Figure 1. In Section 3, we introduce the four candidate stopping boundaries that were originally considered in the design evaluation process. Section 4 presents simulation results illustrative of an approach used to choose between the candidate rules defined in Section 3. Finally, Section 5 concludes with a discussion of the differences and challenges that arise when evaluating group sequential stopping rules where early evidence indicates that the treatment effect of interest may vary with time.

2. Construction of hypothesis alternatives using observed pilot data

Frequentist operating characteristics are based on the sampling distribution of test statistics under various alternative hypotheses. Hence, the definition of alternatives which should be considered when investigating the operating characteristics of potential stopping rules is an essential component to the evaluation of group sequential procedures. When parametric and semiparametric models are to be evaluated, the specification of alternatives is trivial since they are generally defined by a particular parameter of interest (eg. the hazard ratio in the case of the proportional hazards model). However, in a nonparametric setting, no such parameterization exists and alternatives from the null hypothesis are no longer clearly defined. In this section we consider the use of the pilot data presented in Figure 1 to simulate potential alternatives at which candidate design operating characteristics are to be evaluated.

In general, the definition of alternatives can be based on such necessary information as patient enrollment rates, treatment effects over time, and the survival experience and censoring distribution for each comparison group as derived from pilot data. We propose the simulation of alternatives by resampling repeatedly from the single set of observed Kaplan-Meier estimates of survival obtained from the phase II trial, first considering the null survival distribution from which alternatives should deviate. In deciding upon a null survival distribution, one can draw on the statistical questions raised by the available pilot data. That is, the question posed by the pilot data is whether any observed differences might reasonably be obtained by drawing two samples from a single population. Such spurious differences might arise, for example, when the combined samples were representative of a true null distribution, but randomization into two treatment groups produced large separation between survival curves. Alternatively, it might be the case that the control group is representative of the true null distribution, but that random sampling led to a treatment group with better survival than expected. Under the first scenario, we might choose the null survival distribution to be a 50-50 mixture of the estimated survival experience of the control and treatment samples from the pilot study, while under the second scenario we might use estimates of the survival experience from the control sample alone. Other scientifically reasonable options for the null survival distribution exist, including oversampling healthier or sicker patients which may account for possible changes in eligibility criteria.

Given the existence of pilot data, one natural alternative to the chosen null distribution is the observed

survival experience of the comparison group. However, one must also consider a variety of plausible alternatives when evaluating the operating characteristics of a statistical test. To construct a range of alternatives we consider mixtures of the control and comparison Kaplan-Meier estimates of survival from the pilot data. For example if the null survival distribution is taken to be that of the control sample then we may consider 'mixing in' the survival experience of the comparison sample in order to obtain various alternatives. Thus 0% mixing would be indicative of no treatment effect on survival, 50% mixing would indicate a treatment effect in which the survival experience of the treated group represents a 50-50 mixture of the control and comparison survival experience from the available pilot data, and 100% mixing would correspond to a treatment effect that results in a survival experience that is equivalent to that of the comparison sample in the pilot study. Letting $(t_{1i}, t_{2i}, ..., t_{n_i i})$ denote the n_i observed failure times in treatment group i, $i = 0, 1$, the construction of a single sample of size N from an alternative defined by mixing parameter m would proceed as follows:

1. Compute the Kaplan-Meier estimate of the survival distribution for the control and treatment groups in the pilot study, $\hat{S}_0$ and $\hat{S}_1$, respectively.

2. Define the alternative via the percentage that the control and treatment groups are to be mixed, $0 \leq m \leq 1$.

3. For $i = 0, 1$ do

   (a) Let $N_i = \text{round}(N * |(1 - i) - m|)$.

   (b) Sample N_i survival times $\vec{t_i} = (t_1^*, t_2^*, ..., t_{N_i}^*)$ with replacement from $(t_{1i}, t_{2i}, ..., t_{n_i i}, \infty)$ with probability $(1 - \hat{S}_i(t_{1i}), \hat{S}_i(t_{1i}) - \hat{S}_i(t_{2i}), ..., \hat{S}_i(t_{n_i i}) - 0)$.

   (c) For $j = 1, ..., N_i$, if $t_j^* = \infty$ set $\delta_j = 0$, otherwise set $\delta_j = 1$.

4. Combine the sampled survival times $\vec{t} = (\vec{t_0}, \vec{t_1})$ and event indicators $\vec{\delta} = (\vec{\delta_0}, \vec{\delta_1})$.

Of course, the creation of alternatives need not be restricted simply to mixing survival experiences from the pilot study. Another option is to construct alternatives by oversampling the healthiest patients from the pilot study in order to represent various treatment effects. In this setting, the alternative might be

quantified by the weighting assigned to the healthiest patients. One advantage to this approach is that it does allow for alternatives which correspond to greater effects than seen in the pilot data. To define similar alternatives when using the mixing approach described above, one could potentially use a combination of both mixing and oversampling healthy patients (with or without possible semiparametric extensions based on proportional hazards or accelerated failure time models), to consider alternatives which correspond to greater differences than were observed in the pilot data.

In this case study, we define the null survival distribution to be the Kaplan-Meier estimate of the control survival distribution depicted in Figure 1 and consider alternatives which are defined in terms of mixtures of the survival experience of the previously observed comparison sample. Specifically, when constructing alternatives we will consider 0%, 20%, 40%, 60%, 80%, and 100% mixtures, where 0% mixing corresponds to the control population from the original pilot data, and 100% mixing corresponds to the treated population from the original pilot data.

3. Introduction of candidate stopping rules

In defining stopping boundaries we use the unified design family as proposed by Kittelson and Emerson (1999). This particular design family encompasses all previously reported classes of group sequential designs, including the Wang and Tsiatis (1987) family of boundary shape functions and the class of triangular tests as proposed by Whitehead and Stratton (1983). Briefly, the unified family utilizes three parameters: the P-parameter which controls the curvature of the stopping boundary (larger values of P make early stopping more difficult), the R-parameter which allows for even greater flexibility of the curvature of boundaries (larger values of R make early stopping easier), and the A-parameter for which choices of A with small absolute value make stopping at early analyses more difficult.

In this and the following section, we will consider the operating characteristics of four group sequential designs. We assume that the maximal number of accrued patients is 1000 (N=500 per treatment arm) uniformly accrued to the study over a period of 3 years and each design is constructed to allow for 4 interim analyses taking place at 12 months, 18 months, when 51% of subjects have experienced an event, and when 650 subjects have experienced an event. In selecting candidate designs, one-sided symmetric designs with

early stopping under the null hypothesis were considered, however we note that equivalent stopping rules for any one-sided test could have been created using a two-sided framework in the full parameterization of the unified family. Also considered in the design selection process was the degree of early conservatism displayed by candidate stopping rules. Due to our focus on late differences in survival, the majority of the proposed group sequential designs evaluated in the current manuscript are highly conservative with respect to early stopping in favor of futility (since the particular trial of interest would be required to carry on in order to witness any treatment effect), but not so conservative that they fail to protect against harm.

Ultimately, in addition to a reference fixed sample design with one test occurring at 650 events, four one-sided designs were chosen, with varying degrees of conservatism early on. The definitions of chosen group sequential designs are as follows:

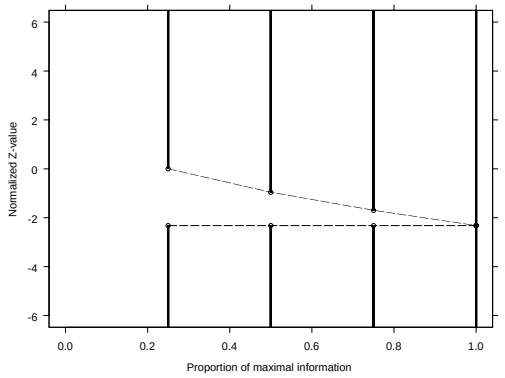
- $\underline{DSN1}$: A one-sided level .025 Pocock (1977) stopping rule (corresponding to $P = .5$, $R = 0$, and $A = 0$) on both the lower (efficacy) and upper (futility) boundaries. This design is constant on the Z-statistic scale (See Figure 3(a)) and is generally regarded as being quite anti-conservative at early analyses.

- $\underline{DSN2}$: A one-sided level .025 test utilizing the O'Brien and Fleming (1979) stopping rule (corresponding to $P = 1$, $R = 0$, and $A = 0$) on both the lower (efficacy) and upper (futility) boundaries; See Figure 3(b). Although this particular design is generally regarded as being highly conservative in early analyses, in the setting of late diverging hazards it still may not yield the amount of conservatism required early on.

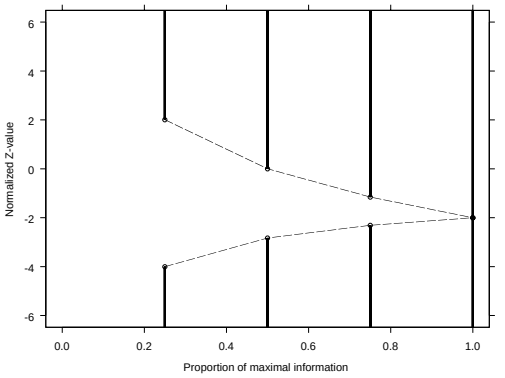
- $\underline{DSN3}$: A one-sided level .025 test parameterized using the Wang and Tsiatis (1987) family of shape functions. The stopping rule for this particular design has an O'Brien-Fleming lower (efficacy) boundary corresponding to $P = 1.0$, $R = 0$, and $A = 0$, and an upper (futility) boundary corresponding to $P = 1.5$, $R = 0$, and $A = 0$, which is extremely conservative at early analyses; See Figure 3(c).

- $\underline{DSN4}$: A one-sided level .025 test parameterized using the full flexibility of the unified design family. The lower (efficacy) boundary takes $P = 1.2, R = 0$, and $A = 0$ (more conservative than the O'Brien-Fleming stopping rule), while the upper (futility) boundary takes $P = 0, R = 0.5$, and $A = 0.3$; See

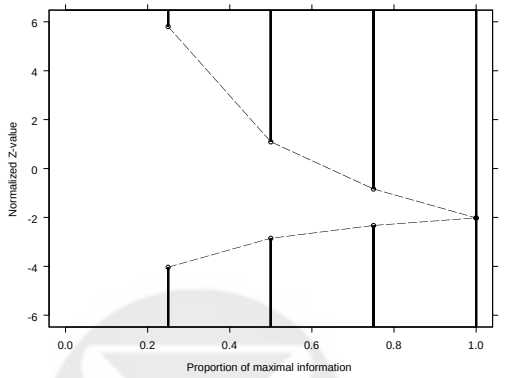
Figure 3(d). *DSN*4 was chosen to increase the degree of conservatism of the previous designs on the efficacy boundary, while shifting the futility boundary to increase the overall power of the design.



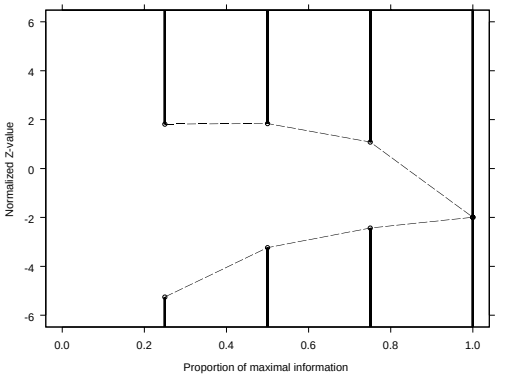
(a) DSN1 (Pocock)



(b) DSN2 (O'Brien-Fleming)



(c) DSN3



(d) DSN4

Figure 2: Proposed group sequential stopping rules. Boundaries are plotted on the Z-statistic scale, assuming equally spaced analysis times.

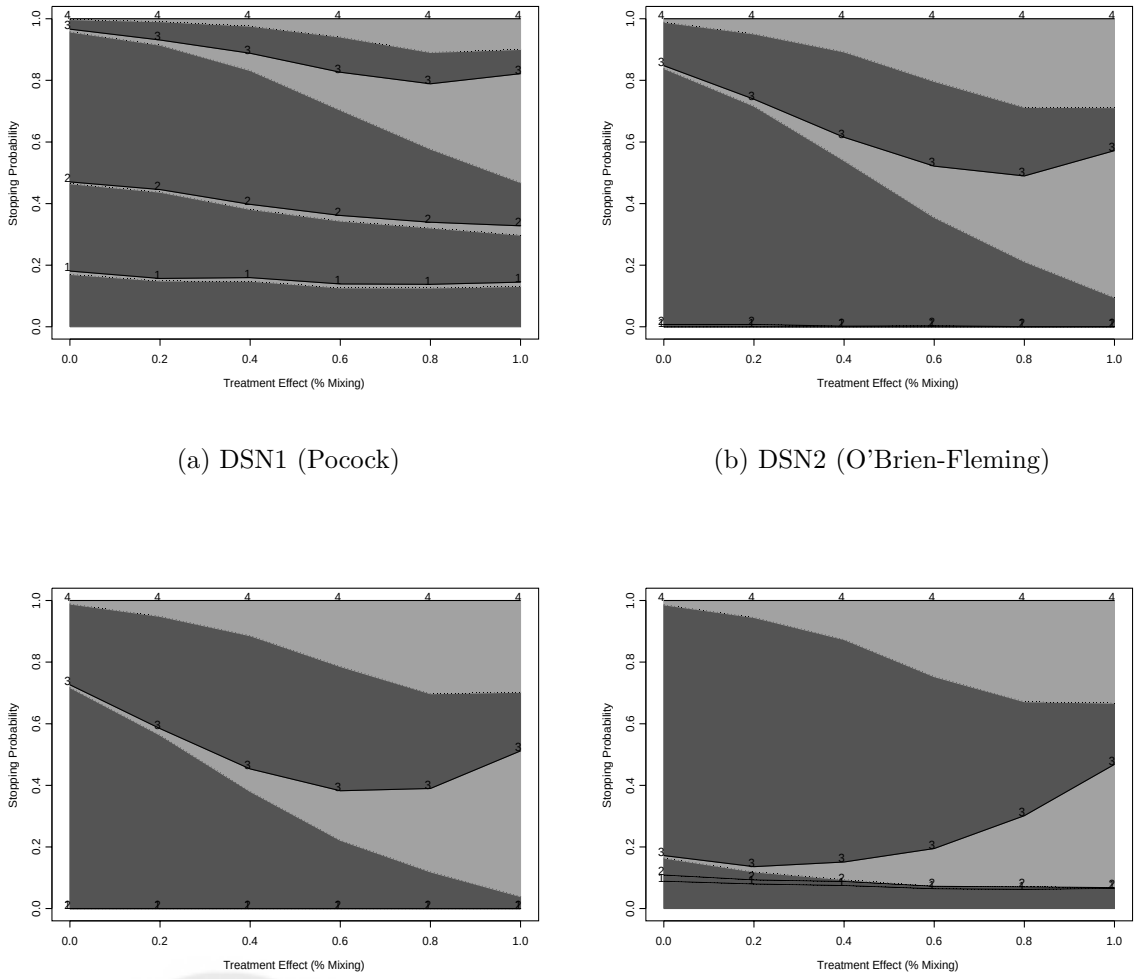
4. Evaluation of candidate stopping rules

In this section, we consider the evaluation of those group sequential designs defined in Section 3 when alternatives are constructed from the nonproportional hazards pilot data depicted in Figure 1(b). As noted earlier, all evaluations are based upon the $G^{1,1}$ weighted logrank test statistic, the test statistic chosen by the study sponsor to best capture the general hypothesized alternative. In a manner analogous to Emerson et al. (2004b), the presented evaluation focuses on a variety operating characteristics, including stopping probabilities under various alternatives, power curves, sample size distributions, and measures of inference on the decision boundaries.

4.1 *Stopping probabilities*

Due to ethical and economic concerns, clinical trial researchers are often concerned with the likelihood of stopping a trial early due to high confidence in favor of efficacy, futility, or harm. To address this, simulated stopping probabilities by treatment effect, defined in terms of the percent mixing as described in Section 2, for each of the previously described group sequential stopping rules are presented in Figure 4. Stopping probabilities were estimated by repeatedly sampling from the pilot data depicted in Figure 1(b) under various levels of mixing, then counting the number of simulations that were stopped at each analysis time. Presented estimates are based upon 5000 simulations performed at each alternative. In each plot, the numbered contours represent the cumulative probability of stopping at the analysis time given by the number on the contour. In addition, the vertical length encompassed by light colored regions between two numbered contours reflects the probability of stopping in favor of efficacy at the latter analysis time, while the vertical length encompassed by dark regions between two contours indicates the estimated probability of stopping in favor of futility at the latter analysis time.

Figure 4(a) displays stopping probabilities for the Pocock design. Under the null hypothesis of 0% mixing, the probability of stopping in favor of efficacy was estimated to be 1.10%, .60%, .90%, and .025% at the first, second, third, and fourth analyses, respectively, resulting in an estimated type I error rate of 2.85%. Under the full alternative of 100% mixing, the probability of stopping in favor of efficacy was estimated to be 1.40%, 3.05%, 35.40%, and 9.95% at the first, second, third, and fourth analyses, respectively, resulting



(a) DSN1 (Pocock)

(b) DSN2 (O'Brien-Fleming)

(c) DSN3

(d) DSN4

Figure 3: Simulated stopping probabilities when selected group sequential designs are applied to alternatives constructed from the nonproportional hazards pilot data depicted in Figure 1(b).

in an overall power of 49.80% at the 100% mixing alternative. Of particular interest is the high stopping probabilities in favor of futility at early analyses under the full alternative when the Pocock boundary was applied in this setting of late diverging hazards.

Estimated stopping probabilities for the O'Brien-Fleming design are displayed in Figure 4(b). We can

see that the contour lines representing the first two analysis times are nearly horizontal at zero for any treatment effect, indicating that there is little to no chance of early stopping in either direction at either of these analyses when the O'Brien-Fleming design is invoked. Under the null hypothesis, the probability of stopping in favor of treatment was estimated to be 0 at the first two analyses, 0.95% at the third analysis and 1.10% at the fourth and final analysis, resulting in an overall type I error of 2.05%. Under an alternative of 100% mixing, the probability of stopping in favor of efficacy was estimated to be 47.70% at the third analysis and 28.85% at the final analysis, resulting in an overall power of 76.55%.

Due to the extreme early conservatism of $DSN3$, no trials were stopped at the first two analyses (see Figure 4(c)). Under 100% mixing, the probability of stopping in favor of efficacy at the third analysis when using the $DSN3$ stopping rule was estimated to be 47.30%, while the probability of stopping for efficacy at the final analysis was estimated to be 29.80%.

Finally, in Figure 4(d) we can see that although no simulations stopped in favor of efficacy at the first or second analysis, there does exist a reasonable chance of stopping early in favor futility at these early analyses when $DSN4$ was invoked. The probability of stopping early in favor of futility under the null hypothesis was estimated to be 8.85% at the first analysis, 2.00% at the second analysis, 5.55% at the third analysis, and 81.40% at the final analysis, implying that the estimated type I error for the design was 2.20%. Under a treatment effect of 100% mixing, the probability of stopping early in favor of efficacy was estimated to be zero at the first two analyses, 40.00% at the third analysis and 33.30% at the final analysis, resulting in an estimated power of 73.30%.

4.2 *Statistical power*

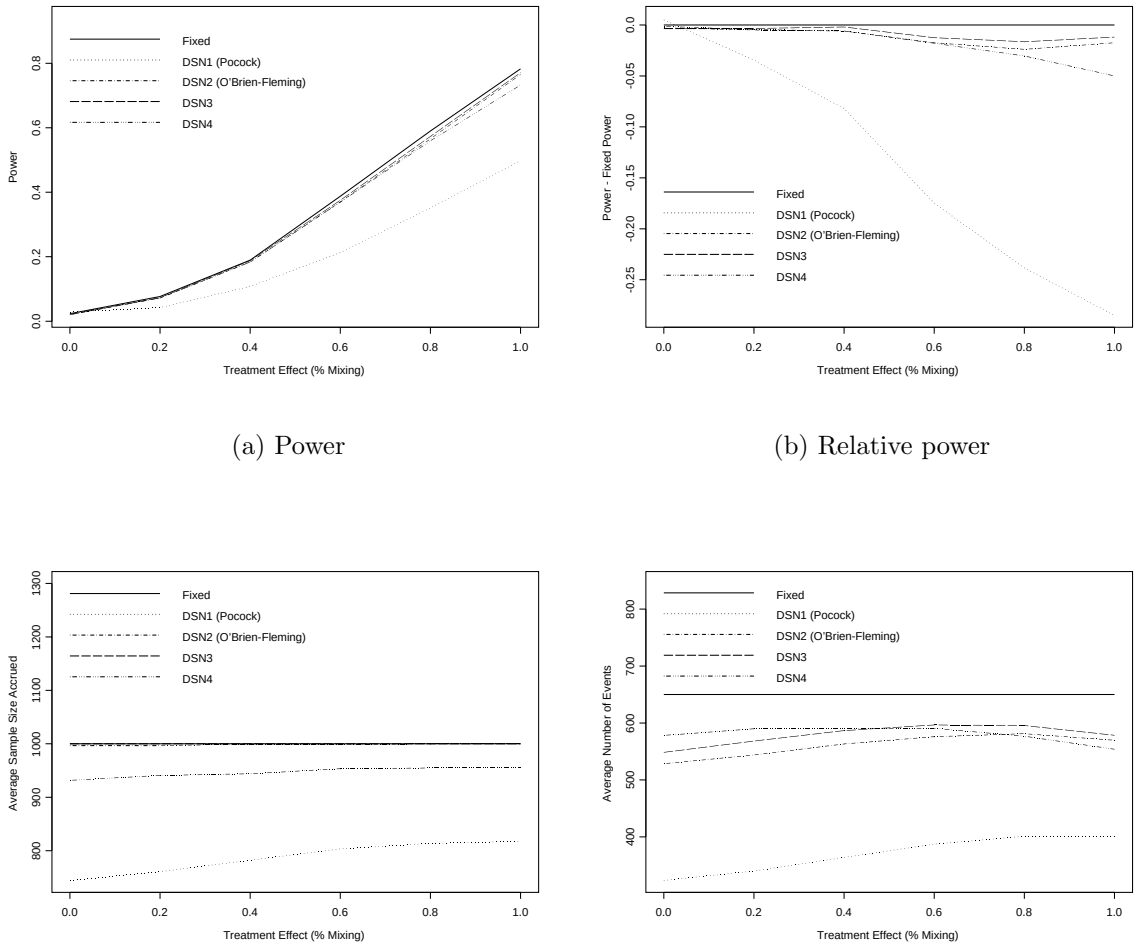
From a statistical perspective, it is clearly of great interest to examine the power curve associated with a given testing procedure. To complement the stopping probabilities displayed in Figure 5, plots of power and relative power as a function of treatment for each of the considered designs are provided in Figure 6. From Figure 6(a) we can see that great differences in the power curves were not found when comparing the fixed sample design, $DSN2$, $DSN3$, and $DSN4$, however the $DSN1$ (Pocock) design obtained substantially lower power than any of the other four designs considered. We note that this loss in power is attributed

to the rather anti-conservative Pocock futility boundary, allowing many trials to stop early before the late occurring treatment effect had been observed. Figure 6(b) better reflects decreases in power due to using a group sequential design over a fixed sample design with the same maximal sample size. The general trend of the plot suggests that as treatment effect increases, larger disparities in power are witnessed between the fixed sample test and the considered group sequential designs. Under the full alternative, $DSN2$ obtained 1.75% lower power when compared to the fixed sample test, $DSN3$ suffered a decrease of 1.20% power relative to the fixed sample design, and $DSN4$ revealed a 5.00% drop in power compared to the fixed sample design. As mentioned above, under an alternative of 100% mixing $DSN1$ yielded the largest disparity in power relative to the fixed sample test, with an estimated drop of 28.5%.

4.3 *Sample size distribution*

Noting that in the group sequential setting the sample size is random, one should also consider the sample size distribution before deciding upon a stopping rule. Figures 5(c) and 5(d) yield plots of the average number of patients and the average number of events required by treatment effect for each of the considered designs. The average number of patients at each alternative was estimated by repeatedly simulating 5000 survival curves under the respective alternative and applying each stopping rule to the simulated data. This process yielded an estimate of the sample size for each simulation, allowing the number of patients required for each sample to be averaged over the total number of simulations. In Figure 5(c), we can see that little to no difference in the average number of required patients is found when comparing the 1000 accrued patients required by the fixed sample design to $DSN2$ and $DSN3$, regardless of the alternative considered. However, there is a decrease in the average number of patients required when either $DSN1$ or $DSN4$ are invoked. Under the null hypothesis, the average number of accrued patients was estimated to be 744 for $DSN1$, and 932 for $DSN4$. Hence we see efficiency gains, relative to the fixed sample test, for the loss of power noted in Figures 5(a) and 5(b).

The average number of events required by each design was also evaluated. Figure 5(d) displays the average number of required events by treatment effect. Here we can see that all of the group sequential designs considered require (on average) fewer than the 650 events planned for the fixed sample test, regardless of treatment effect. Under the null hypothesis of 0% mixing, the average required number of events was



(a) Power

(b) Relative power

(c) Average number of patients

(d) Average number of events

Figure 4: Estimated power curves (Figures (a) and (b)) and sample size distributions (Figures (c) and (d) when selected group sequential designs are applied to alternatives constructed from the nonproportional hazards pilot data depicted in Figure 1(b). Figure 5(b) represents power relative to the fixed sample design. Figure 5(c) and 5(d) display the average number of patients and average number of events by treatment effect, respectively.

estimated to be 323 for the Pocock design ($DSN1$), 529 for the O'Brien-Fleming design ($DSN2$), 549 for $DSN3$, and 579 for $DSN4$. This ordering of the average number of events by design is nearly the same throughout all treatment effects, with exception to the number of events required by $DSN4$ relative to $DSN2$

and $DSN3$. In this case, as the treatment effect began to grow, the average number of events required by $DSN4$ fell at a quicker rate than either $DSN2$ or $DSN3$. At the full alternative of 100% mixing, the expected number of events was estimated to be 401 for $DSN1$, 570 for $DSN2$, 579 for $DSN3$, and 554 for $DSN4$.

4.4 *Inference on the boundary*

Finally, to examine clinical measures of treatment effect which correspond to boundary decisions, the right-hand column of Table 1 yields minimal (for efficacy) and maximal (for futility) estimates of treatment effect encountered when the O'Brien-Fleming ($DSN2$) and $DSN4$ stopping rules were applied to the range of alternatives from 0% to 100% mixing. For this particular case study we present estimates corresponding to the Cox estimate of the hazard ratio and a trimmed hazard ratio considering only the inner 50% of the observed data. Again, estimates are based upon a total of 5000 simulated trials at each alternative. At the first analysis, we can see that no trials were stopped under the one-sided O'Brien-Fleming boundary, while stoppage only occurred in favor of futility under $DSN4$. In this case, the hazard ratio of 0.963 implies that it is plausible, under the $DSN4$ stopping rule, that a study could be stopped early in favor of futility when testing is based upon the $G^{1,1}$ statistic, despite the data showing a 4.7% overall decrease in the hazard.

At the second analysis time, none of the simulated trials were stopped in favor of efficacy using either the O'Brien-Fleming stopping rule or the $DSN4$ stopping rule, while early stoppage in favor of futility occurred at the second analysis under both designs. With respect to the O'Brien-Fleming design, the hazard ratio of 1.278 implies that of all simulated trials stopped early for futility, the maximal treatment effect observed suggested a 27.8% increase in the hazard associated with treatment. The hazard ratio of 1.079 observed under $DSN4$ suggests that a 7.9% increase in the hazard associated with treatment was required for any of the simulated studies to be stopped early in favor of futility at the second analysis time.

At the third interim analysis, trials were stopped early in favor of both efficacy and futility for each of the considered designs. The hazard ratio of 0.946 observed under $DSN2$ suggests that it is plausible that a study could be stopped early in favor of efficacy despite the data showing only a 5.4% overall decrease in the hazard. With respect to futility, the hazard ratio of 0.842 implies that at least one trial stopped early for

Table 1: Minimal (for efficacy) and maximal (for futility) estimates of treatment effect by analysis time when testing is based upon the $G^{1,1}$ statistic. Alternatives were constructed using the pilot data depicted in Figure 1. Hazard ratios represent the hazard of the treatment group relative to the control group.

Summary Statistic	DSN1 (Pocock)		DSN2 (OBF)		DSN3		DSN4	
	Efficacy	Futility	Efficacy	Futility	Efficacy	Futility	Efficacy	Futility
				Time 1				
Z statistic	-7.362	6.285	-12.306	1.293	-6.708	5.524	-10.824	1.344
Hazard Ratio	–	–	–	0.643	–	–	–	0.963
Trimmed HR	–	–	–	$-\infty$	–	–	–	$-\infty$
				Time 2				
Z statistic	-4.825	3.180	-6.810	1.823	-4.287	2.434	-5.783	1.883
Hazard Ratio	–	1.770	–	0.961	–	1.278	–	1.079
Trimmed HR	–	1.826	–	0.877	–	1.367	–	1.134
				Time 3				
Z statistic	-2.388	-0.933	-2.545	1.583	-2.355	-1.019	-2.500	1.394
Hazard Ratio	0.908	0.803	0.888	1.058	0.946	0.842	0.924	1.013
Trimmed HR	0.914	0.754	0.877	1.115	0.926	0.759	0.902	1.032
				Time 4				
Z statistic	-1.991	-1.991	-1.973	-1.973	-1.993	-1.993	-1.979	-1.979
Haz Ratio	0.924	0.759	0.924	0.759	0.958	0.808	0.970	0.808
Trimmed HR	0.911	0.722	0.922	0.722	0.944	0.724	0.944	0.724

futility at the third analysis even though the data revealed a 15.7% overall decrease in the hazard associated with treatment. Similar results for the efficacy boundary were also observed under $DSN4$ at the third and final analyses, though $DSN4$ was found to be much more conservative with respect to stopping in favor of futility at the third analysis when compared to the O'Brien-Fleming stopping rule.

5. Discussion

The use of group sequential methodology has become widespread in the conduct of clinic trials. Because each clinical trial presents unique scientific and statistical issues it is important to carefully evaluate candidate group sequential designs to ensure desirable operating characteristics. Although this methodology is well-defined for situations in which the within-individual treatment effect is constant with respect to time (see for example Emerson et al. (2004b)), when prior evidence for time-varying treatment effects is present the evaluation of potential designs is not a trivial task and this problem is made more complicated when testing is based upon a nonparametric statistic.

In order to evaluate power curves, sample size distributions, and measures of futility, one must specify

the alternatives at which test statistics are to be computed. When parametric and semiparametric models are to be evaluated, the specification of alternatives is trivial since they are generally defined by a particular parameter of interest (eg. the hazard ratio in the case of the proportional hazards model). However, under nonproportional hazards when no parametric model is to be assumed, alternatives from the null hypothesis are no longer clearly defined. We have proposed a procedure for the simulation of hypothesis testing alternatives that does not require the assumption of a parametric model by using observed pilot data. Specifically, we propose that potential alternatives could be constructed by considering various mixtures of the estimated survival experience observed in the pilot data or by oversampling healthier or sicker patients. Ultimately, upon simulating alternatives one is able to estimate commonly examined operating characteristics such as power curves, sample size distributions, stopping probabilities, and estimates of treatment effect that occur on the boundaries, allowing for the comparison of potential group sequential stopping rules.

In order to fully address the scientific question posed by a trial it is generally preferred to present stopping boundaries based upon a statistic which represents some clinically meaningful measure of treatment efficacy. This logic no longer holds when using a nonparametric test statistic such as a weighted logrank statistic since one is no longer testing a specific parameter of interest. Hence care must be taken when considering potential stopping rules in order to examine what point estimates for clinically meaningful measures arise upon study termination. Although we have demonstrated our methods in a hypothetical setting, the approach described here was one used in the planning of a Phase III study designed to investigate the efficacy of an experimental treatment for lung cancer. Of great interest to the investigators were the tradeoffs between efficiency (as measured by average sample sizes) and loss of power (in the absence of increasing sample size to accommodate interim analyses), as well as the potential magnitude of the treatment effect corresponding to statistically significant results. Through the presented case studies we demonstrated that contradictions between decisions based upon particular weighted logrank statistics and clinically meaningful measures of treatment effect can frequently arise. In particular, it was demonstrated that in cases where our test statistic rejected in favor of efficacy, commonly used measures of treatment effect were sometimes found to indicate harm. Similar contradictions were found when decisions in favor of futility were made. This potential for contradiction can have bearing on the functioning of the data safety monitoring committee, regulatory agencies, and the eventual marketing of a new treatment. Careful evaluation of the design is therefore crucial to ensure that everyone understand and agree upon the appropriateness of a stopping rule selected for a

particular study.

When evaluating designs in the setting of nonproportional hazards, we have not extended the methods of Emerson et al. (2004a), Emerson et al. (2004b), and Emerson et al. (2004c) for describing inference and futility measures. The presentation of frequentist confidence intervals and Bayesian posterior probabilities and credible intervals is made quite difficult due to the lack of a single parameter measuring treatment effect. We note that futility measures such as conditional power could be estimated in this setting, but are computationally quite complicated when using simulations because conditioning on interim results requires a prohibitively large number of simulations. Further, Gillen and Emerson (2005) compare the choice of orderings of the sample space with respect to the calculation of corrected P-values under nonproportional hazards. They show that the Z-statistic ordering (Chang, 1989) consistently results in lower P-values relative to the analysis time ordering (Tsiatis et al., 1984) under late occurring treatment effects.

Although the evaluation techniques proposed here are based on simulation and can be time intensive, they are relatively straightforward in nature and do provide reasonable estimates of the operating characteristics generally considered when evaluating potential designs. Thus, given the importance of a priori planning for large scale clinical trials, this investment in time should be deemed negligible.

The research was supported in part by grant # HL69719 from the National Heart, Lung, and Blood Institute.

References

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

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247–254.

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. (2004a). Bayesian evaluation of group sequential designs. *In Press : Statistics in Medicine* .

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2004b). Frequentist evaluation of group sequential designs. *In Revision : Statistics in Medicine* .

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2004c). On the use of stochastic curtailment in group sequential clinical trials. *In Revision : Statistics in Medicine* .

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

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

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.

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

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

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

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