HHS Public Access
Author manuscript
Seq Anal. Author manuscript; available in PMC 2019 September 09.

Published in final edited form as:
Seq Anal. 2014 ; 33(1): 39–59. doi:10.1080/07474946.2014.856635.

Flexibly Monitoring Group Sequential Survival Trials When Testing is Based Upon a Weighted Log-Rank Statistic

Sean S. Brummel[1], **Daniel L. Gillen**[2]

[1]Harvard School of Public Health, Center for Biostatistics in AIDS Research, Boston, Massachusetts, USA

[2]Department of Statistics, University of California, Irvine, California, USA

Abstract

We consider the repeated group sequential testing of a survival endpoint with a time-varying treatment effect using a weighted logrank statistic. The emphasis of this paper is on the monitoring of this statistic where information growth is non-linear. We propose using a constrained boundaries approach to maintain the planned operating characteristics of a group sequential design. A simulation study is presented to demonstrate the operating characteristics of the method together with a case study to illustrate the procedure. We show that when monitoring a weighted logrank statistic, the entry and survival distribution needs to be estimated at interim analyses.

Keywords

Constrained boundaries; Group sequential; Information; Monitoring; Nonparametric; Nonproportional hazards; Survival; Weighted logrank

Subject Classifications:

62L05; 62L10; 62N03

1. INTRODUCTION

During the conduct of clinical trials independent Data and Safety Monitoring Committees (DSMCs) may periodically monitor accumulating data to assess the safety, and sometimes, efficacy of the experimental treatment. Interim testing can be formalized using a group sequential framework to attain desired frequentist operating characteristics (Emerson et al. (2007)). To control the type I error rate under repeated tests of significance multiple authors have proposed discrete sequential stopping rules (Armitage et al. (1969); Pocock (1977); O'Brien and Fleming (1979)) and error spending approaches (Demets and Lan (1984); Pampallona (1995)). Most commonly used group sequential stopping rules consider continuation sets of the form $C_j = (a_j, b_j] \cup [c_j, d_j)$ such that $-\infty \le a_j \le b_j \le c_j \le d_j \le \infty$ for $j = 1, \ldots, J$ analyses. Quite often, these boundaries are interpreted as the critical values for a

Address correspondence to Sean S. Brummel, Center for Biostatistics in AIDS Research, Harvard School of Public Health, Boston, MA , 02115; sbrummel@sdac.harvard.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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.

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. For instance, if we calculate a normalized score statistic $Z_j = U_j / \sqrt{\mathrm{Var}\left[U_j\right]}$ at analysis J the proportion of information at analysis J can be calculated as $\Pi_j \equiv \mathrm{Var}\left[U_j\right] / \mathrm{Var}\left[U_J\right]$ where U_j is the score statistic computed at the final analysis of the trial. That is, $\prod_j$ represents the fraction of total statistical information available from all patients at the time of interim analysis j. It then follows that for some specified parametric functions $f_*(\cdot)$, the critical values for a decision rule at analysis j can be given by $a_j = f_a(\prod_j)$, $b_j = f_b(\prod_j)$, $c_j = f_c(\prod_j)$, and $d_j = f_d(\prod_j)$. The functions $f_*(\cdot)$ are generally chosen to maintain the overall type I error rate of the trial as well as other design operating characteristics including power and expected sample size. For critical values on the normalized Z-statistic scale, popular examples of $f_*(\cdot)$ include the two-sided Pocock (Pocock (1977)) stopping rule that takes $f_a(\prod_j) = -G$, $f_b(\prod_j) = f_c(\prod_j)$, and $f_d(\prod_j) = G$ and the two-sided O'Brien-Fleming (O'Brien and Fleming (1979)) stopping rule that takes $f_a\left(\Pi_j\right) = -G$, $f_b\left(\Pi_j\right) = f_c\left(\Pi_j\right)$, $f_a\left(\Pi_j\right) = -G\Pi_j^{-1/2}$, $f_b\left(\Pi_j\right) = f_c\left(\Pi_j\right)$ and $f_d\left(\Pi_j\right) = G\Pi_j^{-1/2}$, where in both cases the value of G is chosen to maintain a pre-specified type I error rate. In the case where U_j is approximately normally distributed and where an independent increments covariance structure holds so that $\mathrm{Cov}[U_{j+1}, U_j] = Var[U_j]$, $j = 1, \ldots, J-1$, the value of G can be computed using the sequential density derived by Armitage, McPherson, and Rowe (Armitage et al. (1969)). As an alternative, the error spending approach of Lan and DeMets (1983) is defined on the cumulative type 1 error scale where the parametric error spending function $f(\prod_j)$ is a monotonically increasing function such that $f(0) = 0$ and $f(1) = \alpha$, the desired overall type I error rate for the trial.

During the design phase of a study it is common to choose the number of analyses and corresponding stopping boundaries, C_j, $j = 1, \ldots, J$, to satisfy desired operating characteristics such as family-wise type I error, unconditional or conditional power at specified alternatives, average sample number (ASN), and maximal sample size (Emerson et al. (2007)). The computation of such statistics requires one to condition upon the total number of analyses to be performed and the proportion of information attained at each analysis. Ideally, the number and timing of analyses actually performed in the trial would not deviate from those assumed at the design stage. However, due to logistical constraints, potential protocol changes, and unforeseen changes in statistical information accrual, the actual timing of interim analyses is likely to vary as the trial proceeds. While stopping rules can be implemented to maintain type I error despite changes to the number and timing of analyses, these changes will likely effect all other operating characteristics of the originally chosen design including power and expected sample size. As such, it can be desirable to implement interim analyses at information fractions as close as possible to those originally

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

envisioned in order to approximately maintain the operating characteristics that had been expected when planning the study. This requires clinical trialists to accurately estimate the variance of the test statistic at each interim analysis relative to the variance of the test statistic observed at the final analysis of the trial. In some cases the information fraction is completely dictated by the number of observations attained at each analysis. This is true when testing population means with uncensored data. However in the case of censored survival data, predicting the information fraction at each interim analysis can be more complex, particularly when time-weighted statistics are used for testing survival differences (Gillen (2009)).

Censored outcomes are commonly investigated in clinical trials. In such settings the hazard function, $\lambda(t)$, is often chosen as a metric for comparison. When the hazards for comparison groups are proportional over time, the logrank statistic is well-known to be locally efficient in a neighborhood about the null hypothesis (Fleming and Harrington (1991)). However, for those cases where trial designers a priori hypothesize a time-varying treatment effect on the hazard, the $G^{\rho,\gamma}$ class of weighted logrank statistics (Harrington and Fleming (1982)) might be considered. These weighted statistics seek to increase power by up-weighting differences in hazards at times where treatment effects are hypothesized to be greatest. Adding to the utility of these statistics, it follows directly from Tsiatis (1982) that members of the $G^{\rho,\gamma}$ class of statistic are asymptotically normally distributed and maintain an independent increments covariance structure for a univariate survival outcome. This implies that the recursive sequential density given by Armitage et al. (1969) may be used for monitoring these statistics in a group sequential fashion. A recent example where a member of the $G^{\rho,\gamma}$ family was used to test the efficacy of a new experimental intervention can be found in the Abbott XINLAY trial that was presented before the FDA’s Oncologic Drugs Advisory Committee (ODAC) in 2005. Briefly, this pivotal study considered the efficacy of atrasentan in men with metastatic hormone-refractory prostate cancer (Abbott (2005)). In this case, data from a phase II study suggested a delayed treatment effect on the hazard for time to disease progression defined as the first occurrence of a radiographic (new bone lesions) or clinical outcome (metastatic pain with the need for additional therapies). Based on these available data, the pivotal trial was conducted using a group sequential stopping rule with testing based upon an *a priori* specified $G^{1,1}$ statistic for comparing the time to disease progression by treatment arm. The Abbot trial points to a need for considering the behavior of weighted logrank survival statistics under a group sequential testing framework.

Despite the resulting independent increments structure of the $G^{\rho,\gamma}$ class of statistics, Lan (1995) and Gillen and Emerson (2005) point out that for time-weighted statistics such as the $G^{\rho,\gamma}$ family, information growth is not only dependent upon the number of events observed in the trial but also the timing of the events as dictated by the underlying survival and censoring distribution in each treatment arm. This implies that in order to maintain operating characteristics originally specified at the design stage, trial monitors must reliably predict the accrual of future events. This is the focus of the current manuscript.

Information growth for the log-rank test in a maximal information and maximal duration trial is described in Lan and Lachin (1990). In Lan and Lachin (1990) they define the proportion of information at a given analysis as the ratio of the number of observed events to

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

the number of maximal events. In this paper we provide an algorithm to maintain pre-specified operating characteristics when there is a non-linear relationship between events and information. Metha (2001) consider monitoring a group sequential trial on the information scale where information is taken to be the squared inverse of the standard error of the test statistic. In their paper, once maximal information is set for a particular power level it is maintained through the duration of the trial. However, the maximal information required for a pre-specified power is conditional on the planned timing of information at interim analyses. Hence if the fraction of information accrued at each analysis is not correctly assumed at the design stage or is estimated incorrectly at previous interim analyses, power will differ from the originally planned design. In this manuscript we consider the projection of information accrual, conditional upon observed estimates of overall survival and accrual patterns in order to guide the timing of future analyses. By correctly predicting information accrual, and planning future analyses accordingly, it is possible to maintain most of those operating characteristics planned for at the design stage of the trial.

As an applied example, we consider data from Trial 002 of the Community Programs for Clinical Research on AIDS (CPCRA) study, a comparative trial of Didanosine (DDI) or Zalcitabine (DDC) after treatment with Zidovudine in patients with human immunodeficiency virus (HIV) infection. The CPCRA study was a multicenter, randomized open-label trial designed to test whether DDC was non-inferior to DDI with respect to the primary endpoint of progression-free survival (PFS). Planned under a proportional hazards framework, the study protocol specified that DDC would be judged non-inferior to DDI if one could rule out that the DDC/DDI hazard ratio was less than 1.25 (Fleming et al. (1995)). Statistical evidence for non-inferiority was based upon the upper limit of a 95% confidence interval for the DDC/DDI hazard ratio for progression events (or death). The statistical analysis plan originally called for five equally spaced (in information time) interim analyses of the accruing data, where information accrual was determined by the number of observed events. A Lan-DeMets error spending implementation of the O'Brien-Fleming guideline was employed by the data and safety monitoring committee (DSMC) to formulate repeated confidence intervals, allowing the DSMC to consider recommendations for early termination of the trial. Figure 1 displays Kaplan-Meier estimates of PFS observed at the second (a) and final (b) analyses of the data. At the second interim analysis (taking place with a maximum of 9 months of follow-up and 116 total events, or 45% of the planned maximal information to be collected on PFS) study results favored DDI with an estimated hazard ratio of 1.24 (95% CI: 0.85, 1.78). However, with extended follow-up to 22 months, a later emerging benefit of DDC was observed resulting in a final estimated hazard ratio of 0.96 (95% CI: 0.76, 1.23). The delay in the beneficial effect of DDC that was observed in the CPCRA trial is not rare in clinical practice. Such delays could be attributed to a minimum time required for the treatment to show an effect in all 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. In these cases, many trial designers may naturally turn to a form of the logrank statistic that upweights later hazard differences, potentially complicating the strategy for monitoring the statistic in a group sequential framework.

For the remainder of the manuscript we consider methods for monitoring the $G^{\rho,\gamma}$ class of weighted logrank statistics in order to maintain desired operating characteristics.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Specifically, we consider forecasting future event accrual rates by estimating the survival and censoring distribution in the trial and using a constrained boundaries approach Burington and Emerson (2003) to account for remaining deviations from the planned timing of analyses. In section 2, we illustrate the effect of incorrectly specified information growth patterns on design operating characteristics when members of the $G^{\rho,\gamma}$ class are sequentially monitored. In section 3, we introduce a constrained boundaries algorithm for implementing a group sequential stopping rule when focus is on a weighted survival statistic. Section 4 presents a simulation study to illustrate the utility of the proposed approach and section 5 applies the proposed methods to data from trial 002 of the CPCRA. Section 6 concludes with a summary of the results and discusses potential extensions of the methodology.

2. INFORMATION GROWTH IN THE $G^{\rho,\gamma}$ FAMILY

For many commonly used statistics, statistical information grows proportional to the number of observations accrued in the study. This makes the translation between relative statistical information growth and sample size a trivial task. In the case of the usual logrank statistic, this relationship is simply determined by the expected number of events. However, for time-weighted survival statistics information growth is not only dependent upon the number of events but also the timing of the events and the risk set size at the event times. This argues for the special consideration of non-linear information growth of weighted log-rank statistics when trying to maintain pre-specified operating characteristics. In this section we describe the $G^{\rho,\gamma}$ family of weighted logrank statistics, briefly review information growth for this family, and illustrate the effect of incorrectly specified information growth patterns on design operating characteristics when members of the $G^{\rho,\gamma}$ family are sequentially monitored.

2.1. The $G^{\rho,\gamma}$ Family of Weighted Logrank Statistics

Let T_{ik}, E_{ik}, and C_{ik} denote the survival, entry and censoring time of individual i, $i = 1, \ldots, m_k$, belonging to group k, $k = 0, 1$, where T_{ik}, E_{ik}, and C_{ik} are assumed to be independent. At analysis time τ define $X_{ik} = \min(T_{ik}, \tau - E_{ik}, C_{ik})$ to be the observed time for individual i in group k, and let $\delta_{ik} = I(X_{ik} = T_{ik})$ denote the indicator that the actual survival time is observed on the i^{th} individual in group k. Finally, let $N_k(t) = \sum_i^{m_k} I\left(X_{ik} \le t, \delta_{ik} = 1\right)$ denote the number of events observed in group k occurring prior to time t and $Y_k(t) = \sum_i^{m_k} I\left(X_{ik} \ge t\right)$ denote the number of patients at risk in group k at time t. The $G^{\rho,\gamma}$ statistic (Fleming and Harrington (1991)) is given by,

$$G^{\rho,\gamma} = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \int_0^\infty w(t) \left\{\frac{Y_1(t) Y_0(t)}{Y_1(t) + Y_0(t)}\right\} \left\{\frac{dN_1(t)}{Y_1(t)} - \frac{dN_0(t)}{Y_0(t)}\right\},$$

where M_i denotes the number of patients initially at risk in group i, $i = 0, 1$, and $w(t) = [\hat{S}(t-)]^{\rho}[1 - \hat{S}(t-)]^{\gamma}$, with $\hat{S}(t-)$ denoting the Kaplan-Meier estimate of the pooled survival distribution of groups 0 and 1 just prior to time t. Under the strong null hypothesis

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$H_0 : S_0(t) = S_1(t)\ \forall\ t > 0$, the variance of the $G^{\rho,\gamma}$ statistic computed at follow-up time τ reduces to

$$\sigma^2 \propto \int_0^{\tau} w^2(t) F_E(\tau - t)\left[1 - F_C(t)\right] dS(t), \qquad (2.1)$$

and a consistent estimate of the variance of can be computed as

$$\hat{\sigma}^2 = \left(\frac{M_1 + M_0}{M_1 M_0}\right)\int_0^{\tau} w^2(t)\left\{\frac{Y_1(t)Y_0(t)}{Y_1(t) + Y_0(t)}\right\}\left\{1 - \frac{\Delta N_1(t) + \Delta N_0(t) - 1}{Y_1(t) + Y_0(t) - 1}\right\}\left\{\frac{\Delta\left[N_1(t) + N_0(t)\right]}{Y_1(t) + Y_0(t)}\right\}.$$

Harrington and Fleming (1982) explore asymptotic efficiencies of members of the $G^{\rho,0}$ class of statistics (taking $\gamma = 0$) under various hazard ratio configurations and present a class of distributions such that members of the $G^{\rho,0}$ family exhibit optimal properties for detecting location alternatives. In particular, Fleming and Harrington (1991) demonstrate that the $G^{\rho,0}$ statistic is efficient against time-transformed shift alternatives corresponding to

$$\lambda_1(t) = \lambda_0(t)e^{\Delta}\left[\left\{S_0(t)\right\}^{\rho} + \left[1 - \left\{S_1(t)\right\}^{\rho}\right]e^{\Delta}\right]^{-1},$$

where $\lambda_i(\cdot)$ denotes the hazard function for group i, $i = 0, 1$. That is, a $G^{\rho,0}$ statistic ($\rho > 0$) is efficient against alternatives in which the hazard ratio decreases monotonically from e^{Δ} (t=0) to unity ($t \rightarrow \infty$). Thus Δ indicates the initial strength of treatment effect while the parameter ρ is related to the rate at which the effect diminishes. For instance when $\rho = 0$ the logrank statistic is efficient against proportional hazards alternatives. In contrast, the $G^{1,0}$ (commonly referred to as the generalized Wilcoxon statistic) applies increased weight to early failure times and is efficient against the alternative given in (2.2) with $\rho = 1$, indicating heavy early treatment effects which wane relatively quickly over time. This type of effect is often due to treatments that are able to postpone the event of interest in individuals for a period of time but are unable to sustain increased survival indefinitely.

The addition of the γ parameter in the $G^{\rho,\gamma}$ family yields a more flexible weighting scheme to address delayed treatment effects. For example, taking $\rho = \gamma = 1$, induces a quadratic weighting scheme, placing increased weight on failures occurring near the pooled median survival time, and less weight on failures occurring early and late during follow-up. This weighting scheme would yield increased efficiency to detect survival differences like that observed in the CPCRA trial (see Figure 1). Taking $\rho = 0$ and $\gamma = 1$ applies increased weight to late failure times, resulting in increased efficiency when little difference in survival is observed early on during follow-up but a large hazard difference is observed late during follow-up. In practice, this delay in the separation of hazards might be considered if it is hypothesized that a minimum time is required for the treatment to show an effect in all patients or that that there may exist a subset of the sickest patients for which the occurrence of an event is inevitable regardless of treatment assignment.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

2.2. Information Growth

Let σ_j^2 equal the variance of the $G^{\rho,\gamma}$ statistic applied at interim analysis j. Then the proportion of information attained at analysis j, relative to that attained at final analysis J, is given by

$$\prod_j \equiv \left(\frac{M_{1,j} + M_{0,j}}{M_{1,j}M_{0,j}}\right)^{-1} \hat{\sigma}_j^2 \Bigg/ \left(\frac{M_{1,J} + M_{0,J}}{M_{1,J}M_{0,J}}\right)^{-1} \hat{\sigma}_J^2,$$

where $M_{k,j}$ is the number of patients accrued in group k at the time of analysis j, $k = 0, 1$, $j = 1, \ldots, J$. From equation (2.1), it is clear that the information the information fraction, $\prod_j$, will depend upon underlying accrual, censoring, and survival distributions observed during the course of a trial.

To illustrate the effects of inappropriately accounting for differential patient accrual patterns in a group sequential design when a weighted logrank statistic is used for testing, consider a cumulative entry distribution given by

$$F_E(t) = \left(\frac{t}{\theta}\right)^r, \theta > 0, r > 0, 0 \le t \le \theta,$$

so that $r = 1$ implies $E_{ik} \sim Unif(0, \theta)$, $r < 1$ yields entry times that tend to be concentrated near time zero (heavy early enrollment), and $r > 1$ yields entry times that tend to be concentrated toward the end of total enrollment time θ (heavy late enrollment). Figure 2 contains an example that highlights the relationship between the information growth curve of a $G^{1,1}$ statistic, the decision boundaries and the operating characteristics of a group sequential trial. For this example, survival is taken to be distributed Exponential(1) in both arms and the group sequential design is taken to be a one-sided test of a lesser alternative with an O'Brien-Fleming (lower) efficacy boundary (O'Brien and Fleming (1979)) and a Pocock (upper) futility boundary (Pocock (1977)). Suppose that the intended group sequential design calls for four analyses equally spaced in information time with a power of 0.90 to detect an alternative hazard ratio of 0.75 and type I error rate of 0.05. The above design requirements would then require 507 maximal events (total) under a proportional hazards framework. For illustrative purposes, we assume that a total of 500 patients are to be enrolled into each arm (1,000 total) and that the duration of the study would be determined by the occurrence of the 507th event.

To demonstrate the impact of a misspecified information growth curve we consider the case where the clinical trialist naively assumes that information for the $G^{1,1}$ statistic will accrue proportional to the number of events (the solid line displayed in Figure 2(a)), and hence performs analyses upon observing 127, 254, 380, and 507 events. However, if one accounts for the accrual of subjects over time, the true information growth can differ substantially. The dashed and dotted lines depicted in Figure 2(a) represent the information accrued as a function of the proportion of maximal events when patients enter the study uniformly over 3 years and over 1 year, then are followed until 507 maximal events, respectively. This plot shows that at 50% of the maximal events the actual proportion of maximal information that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

would be attained is only 40% if accrual were uniform over 3 years, and only 25% if accrual were uniform over 1 year. Generally speaking, we see that information growth is attenuated under a shorter entry distribution. As patients are enrolled more slowly, information growth becomes proportional to the number of events observed.

Figure 2(b) displays the originally intended stopping boundaries on the standardized Z-scale (solid line, with analyses taking place at equal information spacing) and the stopping boundaries that would actually result if analyses were performed at 127, 254, 380, and 507 events when accrual is uniformly distributed over 1 and 3 years. Comparing the three boundaries, we see that when information is assumed to be proportional to the number of events, the proportion of attained information is overestimated at each analysis, leading to boundaries that are shifted forward in true information time.

Naturally, the resulting shift in the timing of analyses leads to changes in the frequentist operating characteristics of the original design. Figure 2(c) and Figure 2(d) show the sample size distributions and power curves for the shifted boundaries resulting from incorrectly specifying the attained statistical information at each analysis. The actual operating characteristics vary depending upon the accrual pattern. Specifically, if patient accrual were uniform over 1 year and the clinical trialist incorrectly assumed that information grew proportional to the number of expected events for the $G^{1,1}$ statistic they would mistakenly believe the ASN and power of the design ($ASN = 312$ and $Power = 0.90$ at an alternative $HR = 0.75$) to be much higher than the true ASN and power resulting from the design with shifted boundaries depicted by the dotted line in 2(b) ($ASN = 222$ and $Power = 0.73$ at an alternative $HR = 0.75$). In addition, the type I error rate would be believed to be correct (Type One Error Rate = 0.05; $ASN = 242$ at an alternative $HR = 1.0$) while in truth the resulting stopping rule is conservative (Type One Error Rate = 0.041; $ASN = 137$ at an alternative $HR = 1.0$).

3. MONITORING VIA CONSTRAINED BOUNDARIES

3.1. Review of the Constrained Boundaries Algorithm

In Section 2, the impact of incorrectly specifying the proportion of information at each analysis was illustrated. However, during the conduct of a trial deviations from the originally planned timing of analyses may occur for a variety of reasons. One approach to account for deviations from a pre-planned schedule of analyses is the constrained boundaries algorithm (Burington and Emerson (2003)). The constrained boundary algorithm is a generalization of the error spending approach of Demets and Lan (1984). Specifically, the use of error spending functions constrained the proportion of type I error spent at previous analyses using a pre-defined function of the proportion of maximal information accrued at those preceding analyses. This approach was later extended by Pampallona (1995) to flexibly implement stopping rules that maintain both type I error and power through the use of type I and type II error spending functions. The constrained boundaries method of Burington and Emerson (2003) generalized these previous monitoring procedures by constraining on additional treatment effect scales including boundary shape functions defined on the estimated of treatment effect scale or the normalized z-statistic scale.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Generally speaking, the constrained boundaries algorithm is implemented as follows: Using the notation from section 1, boundary shape functions for up to four types of decisions are specified as $f_a(\prod_j)$, $f_b(\prod_j)$, $f_c(\prod_j)$, and $f_d(\prod_j)$, where $\prod_j$ denotes the proportion of maximal statistical information attained at interim analysis j. At the first analysis,$\prod_1$ is computed, and stopping boundaries a_1, b_1, c_1, and d_1 are computed. If the power of the trial test to detect a specified design alternative is to be maintained, a schedule of future analyses is assumed and a stopping rule using the design parametric family (constraining the first boundaries to be a_1, b_1, c_1, and d_1) is found which has the desired power. This consists of searching for the maximal sample size 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. At later 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 and all future analyses as well as the new maximal sample size needed to maintain statistical power are computed using the parametric family of designs specified at the design stage and an assumed schedule of future analysis times. When $f_a(\prod_j)$, $f_b(\prod_j)$, $f_c(\prod_j)$, and $f_d(\prod_j)$ are defined on the type I and II error spending scales, this procedure is equivalent to the error spending approach of Pampallona (1995).

As a simple example to illustrate the constrained boundaries approach, consider testing a one sided hypothesis $H_0 : \mu \le 0$ vs. $H_1 : \mu > 0$, based upon independent and identically distributed observations $X_i \sim (\mu, \sigma^2)$, $i = 1, \ldots, N_{\max}$. Further suppose that we wish to test H_0 using a Pocock stopping rule with 4 interim analyses initially specified to take place after data on $N_{\max}/4$, $N_{\max}/2$, $3N_{\max}/4$, and $N_{\max}$ subjects have been observed. Finally, suppose that at the design stage it is assumed that $\sigma^2 = 1$ and we desire to attain 90% power to detect an alternative of $\mu = 0.25$ while maintaining a one-sided type I error rate of 0.025. These design specifications require $N_{\max} = 199$ and the resulting stopping boundaries on the normalized statistic scale (Z-scale) would be $f_d(0.25) = f_d(0.50) = f_d(0.75) = f_d(1.0) = 2.3613$ and $f_*(\Pi) = -\infty$, $* \in \{a, b, c\}$, $\Pi = \{0{:}25, 0{:}50, 0{:}75, 1{:}0\}$. The top portion of Table 1 yields the stopping rule described above. Now, suppose that after $\lceil 199/4 \rceil = 50$ subjects are observed, the data are analyzed and σ^2 is estimated to be $\hat{\sigma}_1^2 = 1.5$. With this current best estimate of σ^2, in order to maintain 90% power the maximal sample size for the study would need to increase from 199 subjects to 302 subjects. With this change in the maximal sample size, the proportion of information at the first analysis is no longer 0.25, but is now 50/302 = 0.166. Projecting that future analyses will take place at 50%, 75%, and 100% of the newly computed maximal sample size, the stopping rule is adjusted to $f_d(0.166) = f_d(0.50) = f_d(0.75) = f_d(1.0) = 2.3774$. Now suppose that the trial proceeds to the second analysis and data are analyzed after 302/2 = 151 subjects are observed. Further suppose that based upon these data, the variance of a single sampling unit is now estimated to be $\hat{\sigma}_2^2 = 1.7$. This would imply that in order to maintain 90% power and a one-sided type I error rate of 0.025, the maximal sample size would need to be increased from 302 subjects to 345 subjects. Because the first decision boundary was already implemented after 50 subjects, we must constrain $f_d(50/345 = 0.145) = 2.3774$ and the current and future stopping boundaries are computed to be $f_d(0.438) = f_d(0.75) = f_d(1.0) = 2.3918$. It is important to note that the current and future boundaries remain constant on the normalized statistic scale (as dictated by the Pocock

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

design), and that the schedule of future analyses must be assumed in order to calculate the new maximal sample size in order to maintain 90% power. Provided that the observed statistic at the second analysis did not cross the stopping boundary, the trial would proceed in an analogous fashion with the variance updated using available data, the preceding boundaries constrained to the actual value used, and the maximal sample size readjusted in order to maintain 90% power and a one-sided type I error rate of 0.025. The lower portion of Table 1 yields the resulting stopping rule if the trial proceeded to the final analysis and $\hat{\sigma}_3^2 = 1.60$ and $\hat{\sigma}_4^2 = 1.65$ were observed. Boxed values in the table highlight those values that are constrained at current and future analyses.

3.2. Implementation of Constrained Boundaries when Testing is Based Upon a Weighted Logrank Statistic

Implementation of a stopping rule via the constrained boundaries approach described above requires one to know, or be able to estimate, $\prod_j$ at each analysis. When information is directly proportional to sample size, $\prod_j = N_j/N$ where N_j is the sample size at interim analysis j and N denotes the maximal planned sample size. However, when monitoring a weighted log rank statistic such as a member of the $G^{\rho,\gamma}$, family $\prod_j$ is dependent upon the unknown censoring and survival distributions and hence must be estimated. To address this, we propose the following algorithm:

<u>Step 1:</u> Specify original design using a parametric design family to satisfy desired operating characteristics with at the desired fraction of information $\prod_j$ for $_j = 1, \ldots, J$.

<u>Step 2:</u> To anticipate the timing of the first analysis, map $\prod_j$, for $j = 1, \ldots, J$, to the proportion of events using a parametric survival model, $S(t; \overrightarrow{\phi})$, and entry distribution, $F_E(t; \overrightarrow{\gamma})$, with assumed values for $\overrightarrow{\theta}$ and $\overrightarrow{\phi}$

<u>Step 3:</u> At the j^{th} analysis when $j < J$ do steps 3.1–3.5 while maintaining the parametric design family.

<u>Step 3.1:</u> Estimate the parameters $\overrightarrow{\phi}$ and $\overrightarrow{\gamma}$ in $S(t; \overrightarrow{\phi})$ and $F_E(t; \overrightarrow{\gamma})$ using pooled data and a parametric model then substituting the parametric estimates for the parameters we obtain $S_j(t; \widehat{\overrightarrow{\phi}})$ and $F_E(t; \widehat{\overrightarrow{\gamma}})$.

<u>Step 3.2:</u> Conditional on $S_j(t; \widehat{\overrightarrow{\phi}})$ and $F_E(t; \widehat{\overrightarrow{\gamma}})$, estimate the information growth curve using equation (2.1) and compute $\widehat{\Pi}_j$ for $j = 1, \ldots, j$, and set $\widehat{\Pi}_j$ for $j = j + 1, \ldots, J$ to desired timing.

<u>Step 3.3:</u> If $j = 1$, using $\widehat{\Pi}_j$ compute J boundaries on the z-scale; If $j > 1$, using $\widehat{\Pi}_j$ constrain the previously used $j-1$ z-scale boundaries and compute the remaining $J-j+1$ z-scale boundaries.

<u>Step 3.4:</u> Test data using a $G^{\rho,\gamma}$ z-statistic and the j^{th} boundary, if the test statistic is in the rejection region go to step 5.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Step 3.5: Map information increments to the proportion of events for timing of future analyses.

Step 4: At the J^{th} analysis use the *observed* variance of the $G^{\rho,\gamma}$ to compute $\widehat{\Pi}_j$ constrain previous $J - 1$ z-scale boundaries, and test data using final boundary.

Step 5: Discontinue trial and compute final inference.

The above algorithm could be trivially modified when the original design is specified via type I and II error spending functions. In this case, one would constrain the exact amount of type I and II error spent at previous interim analyses instead of the exact z-statistics used.

Statistical information of the $G^{\rho,\gamma}$ statistic can be calculated using equation (2.1). However, in order to evaluate equation (2.1) it is necessary to estimate the entry, survival and censoring distributions. Due to unobserved support in these distributions at early analyses, the prediction of information growth will require some form of parametric assumption regarding the entry, survival and censoring distributions. In the current manuscript we consider a Weibull distribution for the survival distribution and estimate $S(t)$ via maximum likelihood, though many other parametric modeling choices are possible. In addition, we consider modeling the entry distribution using the relatively flexibly parametric model given in equation (2.3). At the jth analysis, the maximum likelihood estimate for r is given by $\hat{r} = M_j / \sum_{i=1}^{M_j} \log\left(\theta / e_i\right)$, where M_j is the number of patients accrued at the time of the analysis and e_i is the entry time of patient i centered at study time zero. At a given analysis time, the full support for the entry distribution, θ, will not be observed, so the above estimator can be computed by conditioning on the observed maximal support. This results in an MLE of $\hat{r} = M_j / \sum_{i=1}^{M_j} \log\left(\tau_j / e_i\right)$, where τ_j represents the maximum observed support at the jth analysis. When calculating the proportion of maximal information, the parameter θ (end of enrollment time) for the proposed entry distribution will cancel out, however θ must be assumed since the integration limit in equation (2.1), τ_J, will depend on the time of final enrollment.

Since the proportion of information is estimated from a parametric model, the model or candidate models and model checking procedures should be prespecified. This will keep the choice of the modeling procedure independent of boundaries used at interim analyses. In addition, it is important to note that the observed variance of the $G^{\rho,\gamma}$ statistic is used at the final analysis so that the family-wise type one error rate is robust to model misspecification.

Finally, it should be noted that the above algorithm can be implemented in two different ways depending upon whether the estimated variance of the test statistic is updated at all analyses using the current best estimate at each interim test, or only updating the variance estimate at the current and future analyses. The former approach is analogous to the constrained boundaries example given in the previous section, where the current best estimate of the variance is used at all analyses and the proportion of maximal information for each analysis is updated accordingly. Burington and Emerson (2003) recommend this approach. However, it is shown in the first author's PhD unpublished dissertation (Brummel (2010)) that if the variance estimate changes significantly from one analysis to the next it

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

may be impossible to maintain type I error rates in some cases. This is because updating the variance estimate at previous analyses may lead to poor estimates of the variability of the test statistic used at previous interim analyses, particularly if the patient population is changing from analysis to analysis. In light of this, we recommend that the variance estimate used at previous analyses be held fixed but that the present and future variance estimates be updated with all currently available data, particularly if the patient population will be recruited from diverse backgrounds and the activation of clinical sites is expected to be staggered over time.

4. Simulation study

4.1. Simulation Setup

A simulation study was conducted to investigate the performance of the proposed monitoring procedure when the $G^{1,1}$ statistic is used to periodically test accruing non-proportional hazards data in a group sequential fashion. The simulation study considered two survival scenarios. In the first scenario, data were sampled under the null hypothesis of equal survival in both arms. In this case, the survival experience of the treatment and control groups were assumed to be exponentially distributed with a mean time to event of two years. In the second scenario, the survival experience in the control group was assumed to be exponential with mean 2.0, while the survival distribution of the treatment group was taken to be piecewise exponential with a hazard ratio that changed linearly from 0.95 at time 0 to 0.2 at 2.5 years. Figure 3 yields representative samples from the distributions defining the two simulation scenario. In each case the total number of sampled patients in each arm was taken to be $N = 1,000$ for a total sample size of 2,000.

To examine the proposed methods under multiple accrual patterns, various parameters for the entry distribution were assumed in the simulation study. Using the entry distribution specified in equation (2.3) we set $\theta = 3$ and $r = \{0.50, 0.75, 1.0, 3.0, 5.0\}$. θ represents the year in which the enrollment of 2,000 patients ends. As in Section 2, testing was based upon a group sequential design with four total analyses using O’Brien-Fleming efficacy bound and a Pocock futility bound with the design goal of testing at equally spaced information times, $\prod=\{0.25, 0.50, 0.75, 1.0\}$.

4.2. Simulation Monitoring Strategies

We present four different monitoring strategies: timing analyses based on the actual information growth curve (unknown in practice but used here as a gold standard), the proposed constrained boundaries algorithm with full estimation of equation (2.1) using a Weibull MLE, using a Weibull for the survival distribution to estimate formula (2.1) but always assuming entry is Unif(0,3) ($\theta = 3$, $r = 1$), and a naive strategy that assumes information growth is proportional to the number of events. When the Weibull model is used with the assumed Unif(0,3) entry distribution, the portion of information is misspecified except for the case where the r parameter in the entry distribution is set to one. When the information is assumed to be proportional to the number of events, the portion of information is misspecified for all considered simulations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4.3. Simulation Results

Table 2 contains simulated operating characteristics (type I error, power, and ASN) for the 4 monitoring strategies. From these results, one can see that the type I error deviates from the nominal 0.05 level unless the information and entry distribution is correctly estimated. For example, the estimated type I error under null sampling with $r = 5$ is 0.048 for the proposed constrained boundaries approach, 0.043 when accrual is assumed to be uniform over 3 years (r=1), and 0.045 when the information is assumed to be proportional to the number of events. The power of the design is also affected. Under the alternative, when $r = 5$, the power is maintained at 0.898 for the constrained boundaries strategy, but drops to 0.624 when accrual is assumed to uniform over 3 years, and 0.745 when the information is assumed to be proportional to the number of events. We note that for this example, early information growth was overestimated in the scenario where information was assumed to be proportional to the number of events, leading to earlier analysis times, and hence lower than nominal type I error due to the Pocock futility boundary. However, misspecification of the information growth pattern could also lead to an inflated type one error rate, depending upon the early conservatism of the stopping boundaries, had information been underestimated at early analyses.

ASN also deviates from the original design plan under misspecification of the information growth curve, as one would expect given the above mentioned differences in power. Differences in the ASN occur from over- or underestimation of the information growth curve. When information is overestimated (as is always the case for the scenario when information is assumed to be proportional to the number of events in our simulation study), stopping boundaries are shifted forward in time resulting in a lower ASN but also lower power. When one assumes that accrual of subjects is uniformly distributed, information is underestimated for early accrual patterns ($r < 1$) and overestimated for late accrual patterns ($r > 1$) resulting in higher and lower than planned ASN values, respectively. What is most important is that when using the proposed constrained boundaries algorithm and re-estimating the full information growth curve at each analysis, the observed ASN closely mimics the planned ASN under a known information growth curve. In this case, the biggest observed deviation from the planned ASN is twelve events and occurred under the null hypothesis with a slightly early accrual pattern ($r = 0.75$).

5. APPLICATION TO TRIAL 002 OF THE COMMUNITY PROGRAMS FOR CLINICAL RESEARCH

To illustrate the proposed methods, we return to trial 002 from the CPCRA (Abrams et al. (1994)), a comparative trial of Dianosine (DDI) against Zalcitabine (DDC) after treatment with Zidovudine in patients with human immunodeficiency virus (HIV). The study was originally planned under a proportional hazards framework with DDC being judged non-inferior to DDI if the upper limit of a confidence interval for the hazard ratio comparing DDC to DDI (DDC/DDI) for progression free survival was less than 1.25. In the actual trial the sequential design was using a Lan-DeMets error spending implementation of the O’Brien-Fleming design. As noted in the introduction, the study was originally planned to have 3 interim analyses with one final analysis. At the design stage, the planned number of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

events for the final analysis was 243; however, the study ultimately progressed to 260 observed events before the DSMB recommended that the trial be stopped. One month later, the total number of reported events was 309 due to overrunning data. For this investigation, we will only focus on the observed 260 events. While the timing of interim analyses was originally scheduled to be equally spaced in information time, during the actual monitoring of the trial the number of events at each interim analysis was, 55, 116, 164 and 260 events. We use this same for schedule of analyses to demonstrate the proposed monitoring approach.

For illustration, we consider application of the $G^{1,1}$ statistic to the CPCRA data using an O'Brien-Fleming decision boundary implemented on the normalized Z-statistic scale (Emerson et al. (2007)). Table 3 contains the observed test statistics, the number of observed events, the projected and constrained decision boundaries, and the projected and proportion of information. The decision boundaries are displayed on the normalized statistic scale, and an observed Z-value less than the lower boundary a_j would conclude with a decision for non-inferiority while the upper d_j boundary is a binding futility boundary. The primary comparison in Table 3 is that of the design decision boundaries in the top row to the finally implemented decision boundaries along the diagonal of the upper portion of the table. The boxed decision boundaries along the diagonal are constrained at previous analyses. To compute the future proportion of information, we projected the future information based on the planned number of events. The lower portion of Table 3 depicts the estimated proportion of information at each analysis using formula (2.1) with maximum likelihood parameter estimates for a parametric Weibull model to estimate survival, and the entry distribution is assumed to follow the parametric form as in equation (2.3). Figure 5 shows the design and final decision boundaries as a function of the proportion of maximal information.

In this example we have used the actual number of events that were observed in the CPCRA trial for implementing each analysis. As seen in Table 3, the future proportion of information deviates as much as 15% from the original design specification. However, as can be seen in figure 4, the projected and observed proportion of information are very close. This indicates that if the proposed methodology to project future information growth had been used in the trial, the timing of future analyses could have been corrected early on and the timing of later analyses would have more closely agreed to the originally planned design.

Figure 1 depicts the observed proportion of information attained in the CPCRA trial along with the estimated information growth curve at each of the analysis times. Information growth curves were estimated using the methods described in Section 3. As can be seen, even at the first analysis, the estimated information growth curve approximates the observed information accrual pattern well. As a by-product of using the $G^{1,1}$ statistic, that weights later analysis times heavier than earlier times, the information growth curve is lower for the $G^{1,1}$ than that of an unweighted logrank statistic that would yield information proportional to the number of observed events. The result of this slower information growth can be seen in Table 3 that yields the proportion of maximal information, induced stopping boundaries, and observed test statistics at each of the analysis times. Due to the down weighting of early events, the first and second interim analyses occur at much earlier information levels when monitoring with the $G^{1,1}$ statistic as compared to the logrank statistic. The result is that the induced stopping criteria are more extreme at earlier analyses when monitoring with the $G^{1,1}$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

statistic. This slower information growth can be viewed as protection against late diverging hazards.

6. DISCUSSION

When a weighted logrank statistic is used in a group sequential clinical trial, a necessary component of the trial is to correctly estimate information growth. We have shown that using the incorrect information can greatly affect the resulting operating characteristics of the stopping rule, including possibly inflated type I error rates and/or reduced power. There are two components of information growth that we have investigated: estimation of the survival curve and estimation of the entry distribution. Putting both of these issues together shows, that when a weighted logrank statistic is used, a full estimation of formula (2.1) is necessary. We did not investigate the estimation of $F_C(t)$(random censoring) because it tends to be negligible in many settings, but since it is a component of formula (2.1), it could be estimated in a trial that has random censoring.

In the case where treatment effects vary with time as in the CPCRA trial data, parameter estimates in the predictive survival distribution may be biased when data are analyzed at less than full support. To potentially reduce this bias it may be useful to incorporate early phase data into the estimation of $S(t)$ when it is available. Bayesian methods provide a natural approach to incorporating early phase data into the estimation of the parameters for $S(t)$, and may potentially be useful in helping to maintain the frequentist operating characteristics of the group sequential design through a prior distribution (obtained from pilot data) that incorporates longer support than the observed data at the time of an interim analysis. Of course, one possible disadvantage to using early phase data is that the sampled population of the earlier study may not be representative of the sampled population for the later trial, ultimately yielding poor predictive performance. However, since the true information levels will be used at the final analysis, some operating characteristics (such as type I error), though not all, can be maintained.

The methodology presented here focuses on group sequential procedures where interim analyses are made after a subset of events have been observed. Fully sequential implementations of the logrank statistic have previously been considered by Gombay (2008). In this case, sequential versions of a general family of stopping boundaries including the Pocock (1977) and O'Brien and Fleming (1979) stopping rules were implemented so that testing of survival differences using the logrank statistic could occur at each observed event time. In our experience, the cleaning of clinical trial data for a formal interim analysis can take between one and three months, making a fully sequential test logistically infeasible in many settings. Despite this, extensions of the methods proposed by Gombay (2008) to the $G^{\rho,\gamma}$ class of weighted logrank statistics are possible but would require estimating the variance of the weighted statistic. As demonstrated in the current manuscript, this variance will depend upon the timing of events and hence the survival, censoring, and accrual distributions observed in the trial.

Though not considered here, one could also use additional covariates to predict the survival and information growth curves. In the case where early information was collected on a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

strong surrogate for the survival response, this approach could potentially improve the estimation of information growth patterns. In addition, we note that non-linearity of the information growth curve is not unique to weighted survival statistics but is also present in the case of longitudinal outcomes. For example, in cases where one wished to compare the slope of an outcome over time, information would depend upon the number and timing of repeated measurements as well as the number of accrued subjects. In this setting, projection of future information growth would be required for adequate analysis timing. This remains an area for future research.

ACKNOWLEDGEMENTS

We thank the reviewer and editors for their time and helpful comments on this manuscript. This research was supported in part by grant P30CA062203 from the National Cancer Institute (D.L.G.).

REFERENCES

Abbott (2005) XINLAY, NDA 21–491 Treatment of Patients with Hormone-Refractory Prostate Cancer Metastatic to Bone, Oncologic Drugs Advisory Committee Briefing Document.

Abrams DL, Goldman AL ,Launer C, Korvick JA, Neaton JD, Crane LR, Grodesky M, Wakefield S, Muth K, Kornegay S, Cohn D, Harris A, Luskin-Hawk R, Markowitz N, Sampson JH, Thompson M, Deyton L (1994). A Comparative Trial of Didanosine or Zalcitabine After Treatment with Zidovudine in Patients with Human Immunodeficiency Virus Infection, New England Journal of Medicine 330:657–552. [PubMed: 7906384]

Armitage P, McPherson CK, and Rowe BC (1969). Repeated Significance Tests on Accumulating Data, Journal of the Royal Statistical Society, Series A, General 132:235–244.

Brummel SS(2010). On Monitoring Group Sequential Trials in the Presence of Non-linear Information Growth and Surrogate Information (2010), University of California Irvine (Unpublished Dissertation).

Burington BE and Emerson SS (2003). Flexible Implementations of Group Sequential Stopping Rules Using Constrained Boundaries, Biometrics 59:770–777. [PubMed: 14969454]

Demets DL and Lan KKG (1984). An Overview of Sequential Methods and their Application in Clinical Trials, Communications in Statistics, Part A – Theory and Methods Split from: @J(CommStat) 13:2315–2338.

Emerson SS, Kittelson JM, and Gillen DL (2007). Frequentist Evaluation of Group Sequential Designs, Statistics in Medicine 26:5047–5080. [PubMed: 17573678]

Fleming TR and Harrington DP (1991). Counting Processes and Survival Analysis New York: Wiley.

Fleming TR, Neaton JD, Goldman AL, DeMets DL, Launer C, Korvik J, and Abrams DL (1995). Insights from Monitoring the CPCRA Didanosine/Zalcitabine Trial, Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 10: Suppl 2:S9–18. [PubMed: 7552519]

Gillen DL (2009). A Random Walk Approach for Quantifying Uncertainty in Group Sequential Survival Trials, Computational Statistics & Data Analysis 3:609–620.

Gillen DL and Emerson SS (2005). Information Growth in a Family of Weighted Logrank Statistics Under Repeated Analyses, Sequential Analysis 24:1–22.

Gombay E (2003). Weighted Logrank Statistics in Sequential Tests, Sequential Analysis 27:97–104.

Harrington DP and Fleming TR(1982). A Class of Rank Test Procedures for Censored Survival Data, Biometrika 69:553–566.

Lan KKG and DeMets DL (1983). Discrete Sequential Boundaries for Clinical Trials, Biometrika 70:659–663.

Lan KKG and Lachin JM (1990). Implementation of Group Sequential Logrank Tests in a Maximum Duration Trial, Biometrics 46:759–770. [PubMed: 2242413]

Lan KKG, Rosenberger WF and, Lachin JM (1995). Sequential Monitoring of Survival Data with the Wilcoxon Statistic, Biometrics 51:1175–1183. [PubMed: 7548701]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Metha CR and Tsiatis AA (2001). Flexible Sample Size Considerations Using Information-based Interim Monitoring, Drug Information Journal 35:1095–1112.

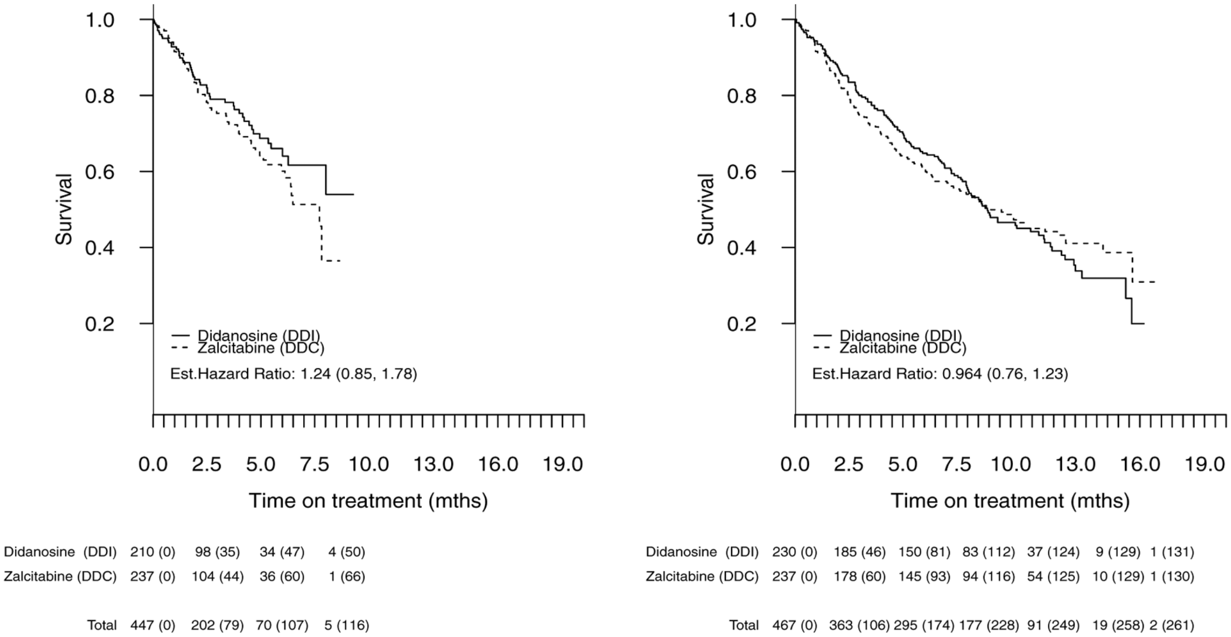
O'Brien PC and Fleming TR (1979). A Multiple Testing Procedure for Clinical Trials, Biometrics 35:549–556. [PubMed: 497341]

Pampallona S, Tsiatis AA and, Kim KM (1995). Spending Functions for the Type I and Type II Error Probabilities of Group Sequential Tests, Technical Report, Harvard University, Dept. of Biostatistics (Unpublished).

Pocock SJ (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials, Biometrika 64:191–200.

Tsiatis AA (1982). Repeated Significance Testing for a General Class of Statistics Used in Censored Survival Analysis, Journal of the American Statistical Association 77:855–861.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



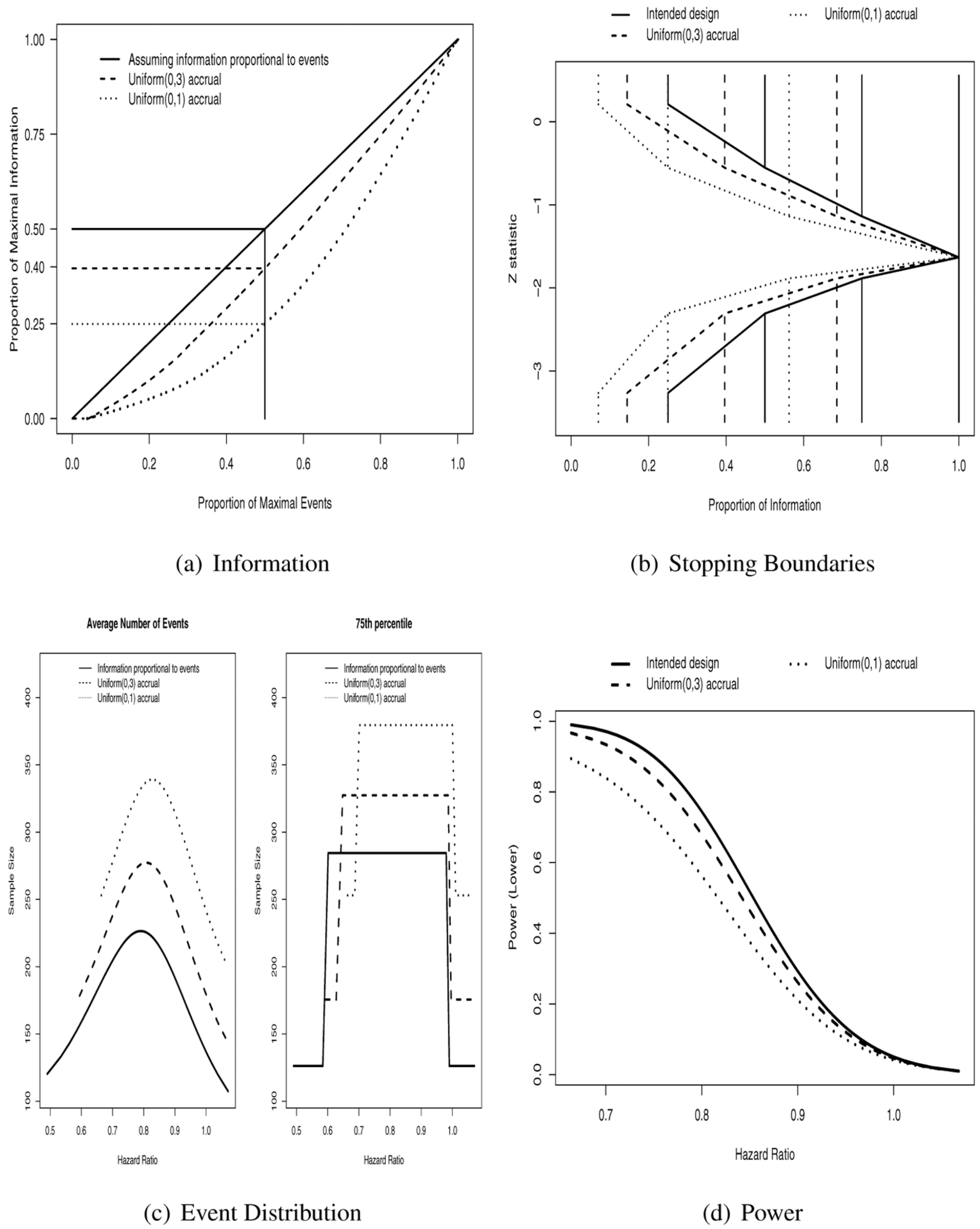
(a) Second Interim Analysis

(b) Final Analysis

Figure 1.

Kaplan-Meier estimates of survival at the second (a) and final (b) analyses of trial 002 of the CPRCA study. Hazard ratios and confidence intervals were estimated from a proportional hazards model. Statistics at the bottom of each figure represent the number of patients as risk (and number of observed events) at 3-month intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



(a) Information

(b) Stopping Boundaries

(c) Event Distribution

(d) Power

Figure 2.

Relationship between the information growth curve of a $G^{1,1}$ statistic, the decision boundaries and the operating characteristics of a group sequential trial under differing accrual patterns. Survival is assumed to be distributed Exponential(1) in both arms and the group sequential design is taken to be a one-sided test of a lesser alternative with an O'Brien-Fleming (lower) efficacy boundary (O'Brien and Fleming (1979)), a Pocock (upper) futility boundary (Pocock (1977)). In addition, the originally intended group sequential design has four analyses, equal spaced in information time, with a power of 0.90 to detect an alternative of 0.75, type I error rate of 0.05, and 507 maximal events. Boundaries are computed when analyses are equally spaced in information time (as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

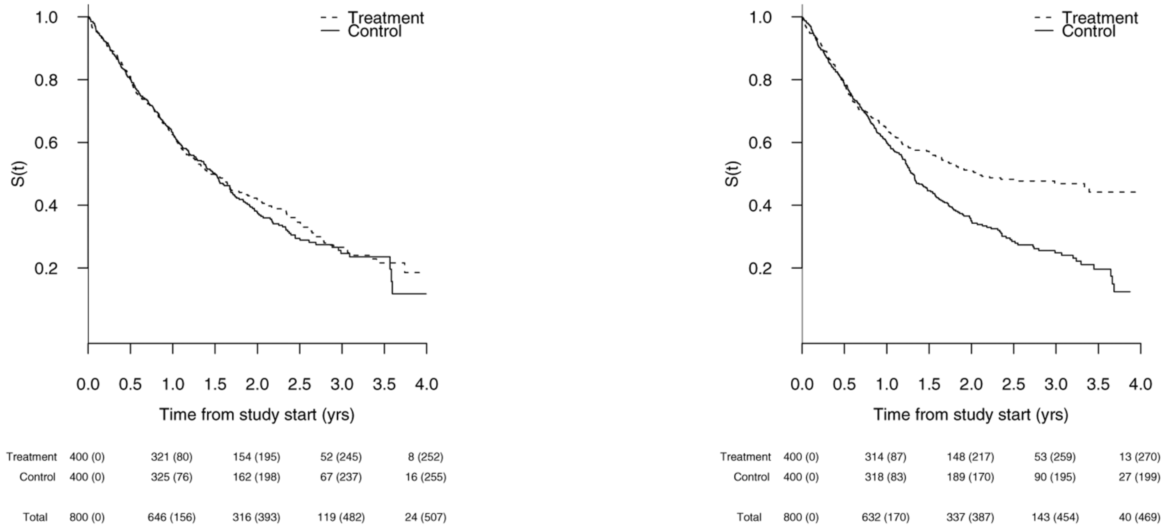
originally intended), and when analyses are performed at equal numbers of accrued events but patients are accrued uniformly over 3 years and 1 year.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



(a) Simulation Scenario 1 : Strong null

(b) Simulation Scenario 2 : Alternative

Figure 3.

Representative survival curves from the two scenarios considered in the simulation study. In the first scenario (a), data were sampled under the strong null hypothesis $H_0 : S_0(t) = S_1(t) \ \forall \ t > 0$ of equal survival in both arms at all followup times. In this case, the survival experience of the treatment and control groups were assumed to be exponentially distributed with a mean time to event of two years. In the second scenario (b), the survival experience in the control group was assumed to be exponential with mean 2, while the survival distribution of the treatment group was taken to be piecewise exponential with a hazard ratio that changed linearly from 0.95 at time 0 to 0.2 at 2.5 years.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

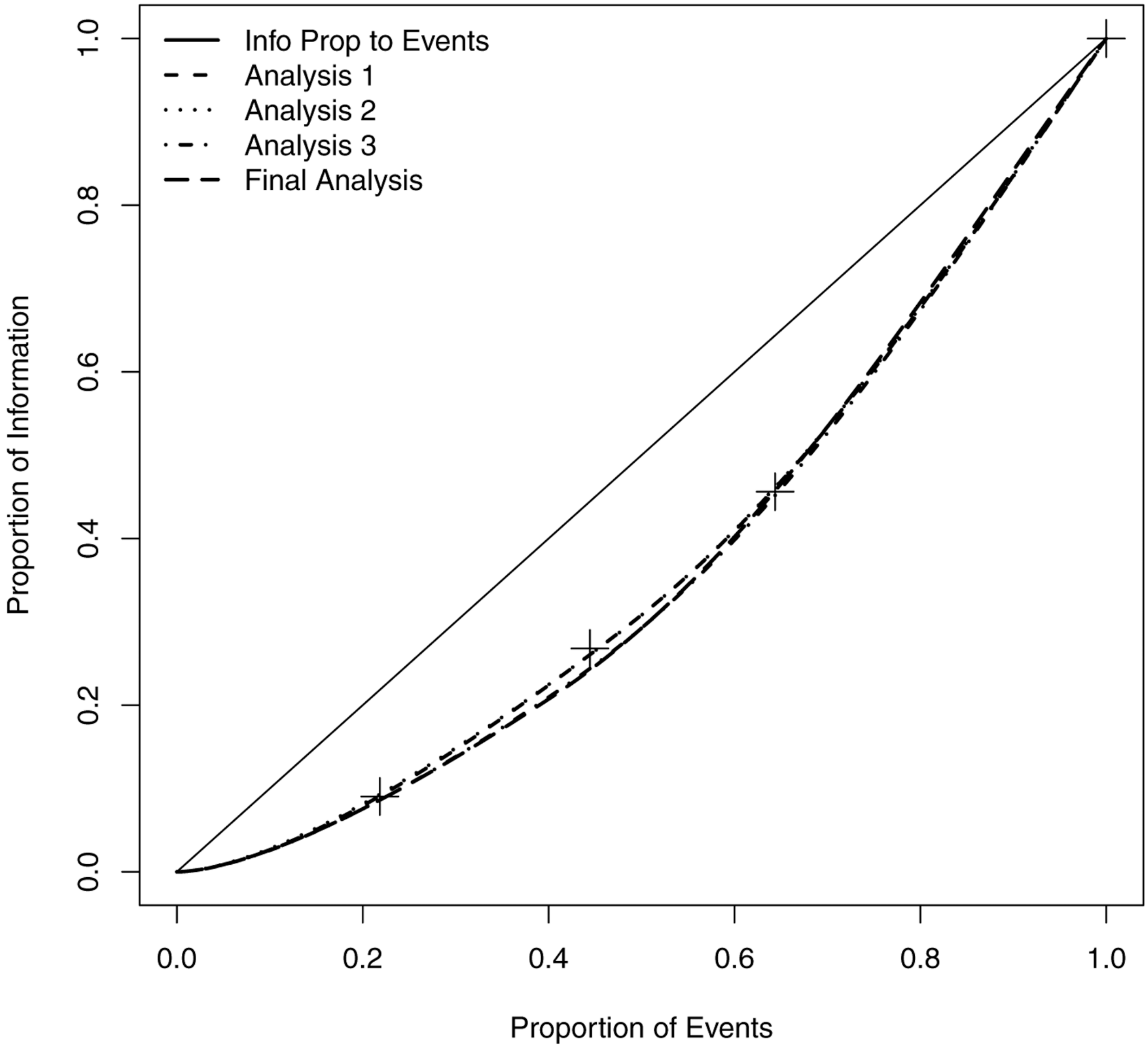


Figure 4.

Information Growth in the CPCRA Data. At each analysis, formula (2.1) was computed with a fixed maximal entry date but estimated time of final analysis. The stars show the observed portion of information for the $G^{1,1}$ statistic. The line $y = x$, the information growth for the logrank is shown for reference.

Author Manuscript

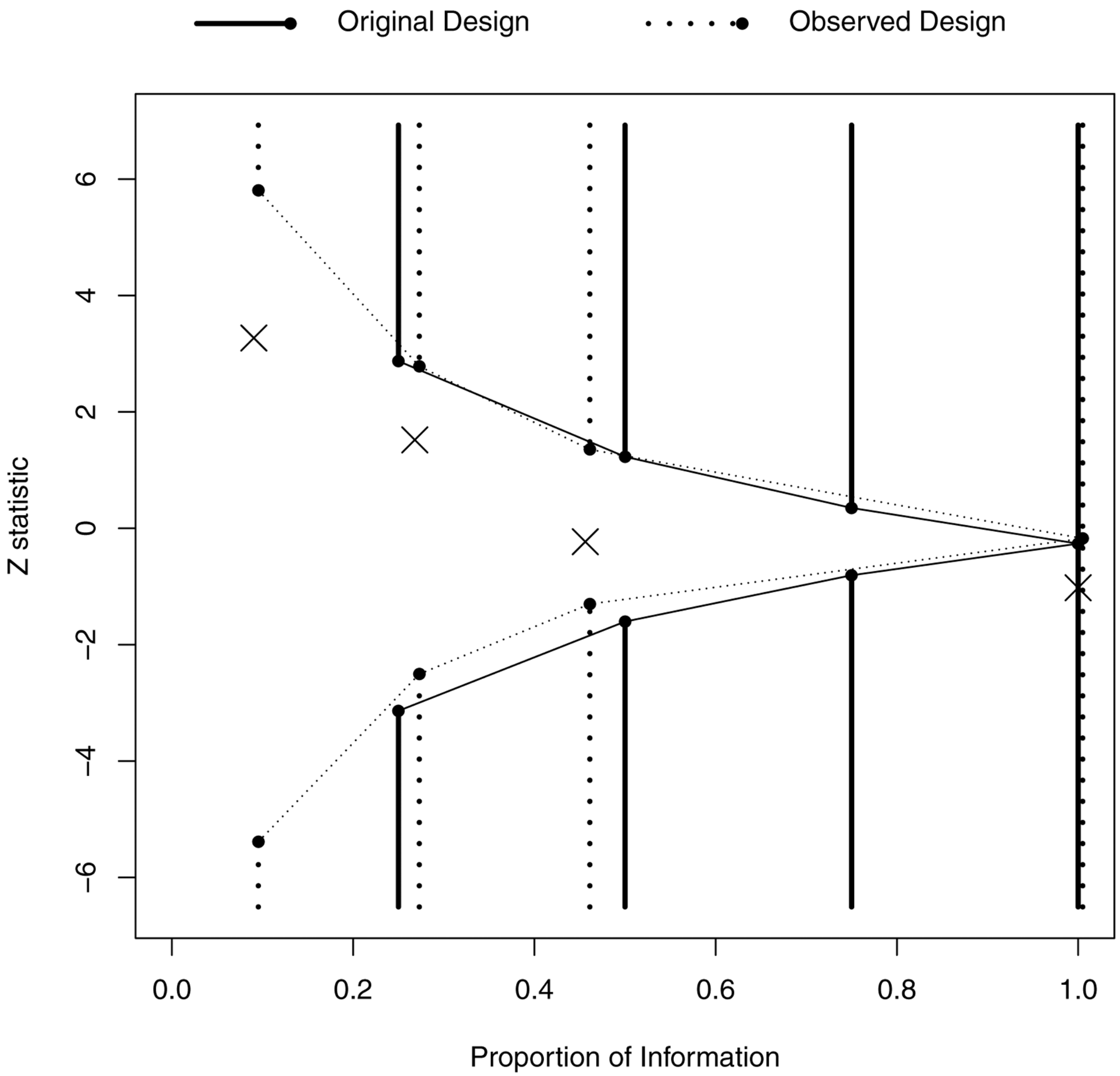


Figure 5.

Comparison of the original decision boundaries (Solid Line) against the final decision boundaries (Dashed Line) plotted on the z-scale as a function of the proportion of information. The crosses are the observed $G^{1,1}$ z-value test statistic.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Illustration of the constrained boundaries algorithm. In this example, the original design is a one-sided test of a greater alternative utilizing a Pocock (1979) stopping boundary for superiority and four equally spaced analyses. Based upon an initial assumption that the sampling variability for single observation is 1.0, the trial is initially designed to collect data on a maximum of 199 subjects to ensure 90% power to detect an alternative of 0.25 with one-sided type I error of 0.025. The implemented design constrains the boundaries on the normalized statistic (Z-scale) and updates the maximal sample size to maintain 90% power for the specified alternative. Boxed values represent implemented decision boundaries that are constrained at future analyses.

Analysis (*j*)	**1**	**2**	**3**	**4**
Planned Design (σ^2=1.0):				
Sample Size	49.75	99.50	149.25	199
Information Fraction (Π_j)	0.25	0.50	0.75	1.00
Decision Boundary (Z-scale)	2.3613	2.3613	2.3613	2.3613
Implemented Design:				
Analysis 1 $\left(\hat{\sigma}_1^2 = 1.5\right)$				
Sample Size	50	151.0	226.5	302
Information Fraction (Π_j)	0.166	0.50	0.75	1.00
Decision Boundary (Z-scale)	2.3774	2.3774	2.3774	2.3774
Analysis 2 $\left(\hat{\sigma}_2^2 = 1.7\right)$				
Sample Size	50	151	258.75	345
Information Fraction (Π_j)	0.145	0.438	0.75	1.00
Decision Boundary (Z-scale)	2.3774	2.3918	2.3918	2.3918
Analysis 3 $\left(\hat{\sigma}_3^2 = 1.6\right)$				
Sample Size	50	151	259	321
Information Fraction (Π_j)	0.156	0.470	0.807	1.00
Decision Boundary (Z-scale)	2.3774	2.3918	2.3735	2.3735
Analysis 4 $\left(\hat{\sigma}_4^2 = 1.6\right)$				
Sample Size	50	151	259	321
Information Fraction (Π_j)	0.156	0.470	0.807	1.00
Decision Boundary (Z-scale)	2.3774	2.3918	2.3735	2.3735

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Simulation results examining the effect of misspecifying information growth on type I error rates, power, and ASN. In all cases, subjects are assumed to accrue over three years with differential accrual patterns as dictated by the *r* parameter. Maximal events are held constant across the different monitoring strategies and chosen to maintain 90% power under the alternative when analyses are equally spaced in information time. Representative survival curves used for the simulation study are depicted in Figure 3. The results are based on 50,000 simulations.

	Portion Rejected					**ASN**				
Alternative/	**Entry Parameter r**					**Entry Parameter *r***				
Monitoring Strategy	**0.5**	**0.75**	**1**	**3**	**5**	**0.5**	**0.75**	**1**	**3**	**5**
Null										
Known information growth	**0.048**	**0.049**	**0.051**	**0.050**	**0.048**	**603**	**543**	**514**	**617**	**728**
Constrained boundaries	**0.049**	**0.049**	**0.051**	**0.049**	**0.048**	**608**	**531**	**519**	**611**	**736**
Assuming entry Unif(0,3) (r = 1)	0.052	0.050	**0.051**	0.044	0.043	691	566	**519**	492	610
Information ∝ events	0.045	0.046	0.049	0.047	0.044	520	494	467	521	613
Alternative										
Known information growth	**0.898**	**0.897**	**0.894**	**0.905**	**0.898**	**681**	**629**	**605**	**751**	**841**
Constrained boundaries	**0.900**	**0.888**	**0.898**	**0.900**	**0.898**	**683**	**618**	**607**	**743**	**842**
Assuming entry Unif(0,3) (*r* = 1)	0.964	0.924	**0.897**	0.702	0.624	746	646	**608**	585	664
Information ∝ events	0.776	0.825	0.831	0.804	0.745	606	586	564	661	732

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3.

Results from monitoring progression free survival in trial 002 of the CPCRA study using constrained boundaries. Results are shown for monitoring based upon the $G^{1,1}$ statistic. The upper d_j boundaries are the futility boundaries and the lower a_j boundaries are the non-inferiority decision boundaries. The boundaries in bold are the boundaries used to decide to stop or continue the trial. Proportion of information is displayed using equation (2.1) at each interim analysis, with a parametric Weibull survival distribution and a parametric entry distribution as given in equation (2.3), and parameters estimated using maximum likelihood.

Characteristic		***j* = 1**	***j* = 2**	***j* = 3**	***j* = 4**
		Decision Boundaries			
	Observed Z-statistic	(a_1, d_1)	(a_2, d_2)	(a_3, d_3)	(a_4, d_4)
Design		(−4.01, 2.00)	(−2.83, 1.11)	(−2.31,−1.16)	(−2.00,−2.00)
Analysis 1	3.27	(−6.23, 4.97)	(−3.57, 1.37)	(−2.56,−0.52)	(−1.98,−1.98)
Analysis 2	1.52	(−6.23, 4.97)	(−3.71, 1.58)	(−2.55,−0.53)	(−1.98,−1.98)
Analysis 3	−0.23	(−6.23, 4.97)	(−3.71, 1.58)	(−2.75,−0.092)	(−1.97,−1.97)
Analysis 4	−1.22	(−6.23, 4.97)	(−3.71, 1.58)	(−2.75,−0.092)	(−1.97,−1.97)
		Proportion of Information			
	Observed Events	Π_1	Π_2	Π_3	Π_4
Design		.25	.50	.75	1.00
Analysis 1	55	0.101	0.309	0.601	1.00
Analysis 2	116	0.102	0.286	0.605	1.00
Analysis 3	164	0.096	0.273	0.517	1.00
Analysis 4	260	0.090	0.268	0.456	1.00