Published in final edited form as:

J Biopharm Stat. 2015 ; 25(6): 1190–1205. doi:10.1080/10543406.2014.971165.

Group Sequential Design for Randomized Phase III Trials under the Weibull Model

Jianrong Wu and Xiaoping Xiong

Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA, Phone: 901-595-2850

Xiaoping Xiong: jianrong.wu@stjude.org

Abstract

In this paper, a parametric sequential test is proposed under the Weibull model. The proposed test is asymptotically normal with an independent increments structure. The sample size for fixed sample test is derived for the purpose of group sequential trial design. In addition, a multi-stage group sequential procedure is given under the Weibull model by applying the Brownian motion property of the test statistic and sequential conditional probability ratio test methodology.

Keywords

Brownian motion; Group sequential trial; Randomized clinical trial; Sample size; Time-to-event; Weibull distribution

1 Introduction

For ethical reasons, clinical trials are often monitored for early stopping if a sufficiently large treatment difference is observed during an interim analysis. Various group sequential monitoring methods have been developed in the past few decades, such as the procedures of Haybittle (1971), Pocock (1977), and O'Brien and Fleming (1979); the type I error spending function approach of Lan and DeMets (1983); the triangular test of Whitehead and Stratton (1983); the sequential conditional probability ratio test (SCPRT) of Xiong (1995); and many others. Comprehensive reviews of these methods are provided by Jennison and Turnbull (2000, references therein).

In cancer clinical trials, time-to-event is often a primary endpoint for the study design, such as overall survival and event-free survival, where the event could be disease progression, relapse, or death. The primary interest is to compare the survival distributions between treatment groups. The non-parametric log-rank test is the most popular test statistic used to design such a study (Collett, 2003). Its Brownian motion property makes it easy to monitor such trials by using the group sequential procedure (Tsiatis, 1982; Sellke and Siegmund, 1983; Slud, 1984; Kim and Tsiatis, 1990).

For survival data, the exponential and Weibull distributions are the two most frequently used parametric models. Of the two distribution forms, the Weibull distribution is usually more appropriate to describe time-to-event data than the exponential distribution because it includes the shape parameter in addition to the scale parameter, with a decreasing or

increasing hazard. In advanced-stage cancer studies, the survival rate usually drops dramatically toward the end of the study. Such characteristics of the survival distribution can be better approximated by a Weibull distribution. In general, a cancer survival trial under the Weibull model can also be designed under the proportional hazards model using the log-rank test. However, a parametric test derived under the Weibull model has better small sample properties than the non-parametric log-rank test because the latter has to be general and thus information from continuous quantities derived from specific parametric model cannot be included for inference (Wu, 2013). The maximum sample size is often large for a phase III group sequential trial. However, the available data could be small in the early stages of interim monitoring; therefore, a study with group sequential design under the Weibull model may perform better in the early stages than that a general proportional hazards model. Tsiatis et al. (1995) derived asymptotic sequential distributions for score and Wald tests in general parametric survival models, but the method has not been applied to group sequential trial design under the Weibull model. Recently, Jiang et al. (2012) proposed a simulation method for group sequential trial design under the Weibull model, but it is a computationally intensive method with restrictive assumptions. Heo et al. (1998) and Wu (2013) proposed a sample size formula for a fixed sample test under the Weibull model. Lu et al. (2012) derived a sample size formula for designing a two-stage seamless adaptive design under the Weibull model. However, a general multi-stage group sequential design under the Weibull model is not available in the literature.

The rest of this paper is organized as follows. In Section 2, a parametric sequential test statistic is proposed under the Weibull model. The sample size for a fixed sample test is given in Section 3. A general multi-stage group sequential procedure is discussed in Section 4. In Section 5, the empirical type I error and power of the proposed parametric sequential test are compared with those of the well-known non-parametric log-rank test. An example is given in Section 6 to illustrate the proposed method. The final conclusion is presented in Section 7.

2 Sequential Test Statistics

A parametric sequential test statistic is discussed in this section to provide group sequential design for randomized two-arm survival trials under the Weibull model. Assume that time-to-event variable T_j of a subject from the j^{th} group follows the Weibull distribution with a common shape parameter κ and scale parameter ρ_j , $j = 1, 2$. That is, T_j has survival distribution function

$$S_j(t) = e^{-(\rho_j t)^\kappa}$$

and hazard function

$$h_j(t) = \kappa \rho_j^\kappa t^{\kappa-1}.$$

The shape parameter κ indicates the degree of acceleration ($\kappa > 1$) or deceleration ($\kappa < 1$) of the hazard over time. In a cancer trial, the median survival time is an intuitive endpoint for clinicians. The median survival time of the j^{th} group for the Weibull distribution can be calculated as $m_j = \rho_j^{-1} \{\log(2)\}^{1/\kappa}$. Therefore, the Weibull survival distribution can be expressed as

$$S_j(t) = e^{-\log(2) \left(\frac{t}{m_j}\right)^\kappa}, j=1, 2.$$

The one-sided hypotheses of a randomized two-arm trial defined by median survival times can be expressed as

$$H_0: m_1 \geq m_2 \text{ vs. } H_1: m_1 < m_2.$$

For notation convenience, we convert the scale parameter ρ_j to a hazard parameter $\lambda_j = \rho_j^\kappa = \log(2)/m_j^\kappa$. Then the above hypotheses on median survival times are equivalent to the following:

$$H_0^*: \delta \leq 1 \text{ vs. } H_1^*: \delta > 1,$$

where the hazards ratio $\delta = \lambda_1/\lambda_2 = R^\kappa$, with $R = m_2/m_1$. Then the survival distribution is $S_j(t) = e^{-\lambda_j t^\kappa}$ with hazard function $h_j(t) = \kappa \lambda_j t^{\kappa-1}$, in which κ is taken as a known constant. This indicates that the testing problem can also fit into a proportional hazard model and the log-rank test is applicable for the intended testing.

Now, suppose during the accrual phase of the trial, n_j subjects of the j^{th} group are enrolled in the study, and let T_{ij} and C_{ij} denote, respectively, the event time and censoring time of the i^{th} subject of the j^{th} group, with both being measured from the time of study entry, Y_{ij} . We assume that the event time T_{ij} is independent of the censoring time C_{ij} and entry time Y_{ij} , and $\{(Y_{ij}, T_{ij}, C_{ij}); i = 1, \dots, n_j\}$ are independent and identically distributed. When the data are examined at calendar time $t = \tau$, where τ is the study duration, we observe the time-to-event $X_{ij}(t) = T_{ij} \wedge C_{ij} \wedge (t - Y_{ij})^+$ and failure indicator $\delta_{ij}(t) = I(T_{ij} \leq C_{ij} \wedge (t - Y_{ij})^+)$, $i = 1, \dots, n_j$. Based on the observed data $\{X_{ij}(t), \delta_{ij}(t), i = 1, \dots, n_j, j = 1, 2\}$, the observed likelihood function at time t is proportional to (see, e.g., Cox & Oakes, 1984, Chapter 3)

$$L(\lambda_1, \lambda_2; t) = \lambda_1^{d_1(t)} \lambda_2^{d_2(t)} e^{-\lambda_1 U_1(t) - \lambda_2 U_2(t)},$$

where $d_j(t) = \sum_{i=1}^{n_j} \delta_{ij}(t)$ is the total number of events observed in the j^{th} group by time t , and $U_j(t) = \sum_{i=1}^{n_j} X_{ij}^\kappa(t)$ is the cumulative follow-up time by time t penalized by the Weibull shape parameter κ . The maximum likelihood estimate of $\lambda_j(t)$ can be derived as

$$\hat{\lambda}_j(t) = d_j(t)/U_j(t),$$

and its variance is approximately $\hat{\lambda}_j^2(t)/d_j(t) (= - \left[\frac{\partial^2 \log L}{\partial \lambda_j^2} \right]_{\hat{\lambda}_j}^{-1})$. Therefore, under the null hypothesis, the Wald statistic of the log-hazard ratio $\gamma = \log(\delta)$ at calendar time t is given by (see Appendix 1)

$$Z(t) = \log\{U_2(t)d_1(t)/U_1(t)d_2(t)\} (d_1^{-1}(t) + d_2^{-1}(t))^{-1/2} \quad (1)$$

and has approximately a standard normal distribution. To derive the group sequential design, let

$$U(t) = \log\{U_2(t)d_1(t)/U_1(t)d_2(t)\} (d_1^{-1}(t) + d_2^{-1}(t))^{-1}; \quad (2)$$

then under the alternative $\gamma = \log(\delta) > 0$, the statistic $U(t)$ is approximately normal with mean $\gamma V(t)$ and variance $V(t)$ and has an independent increment structure, where $V(t) = (d_1^{-1}(t) + d_2^{-1}(t))^{-1}$. The above results can be derived from Tsiatis et al. (1995), who proved similar results for general survival parametric models. Since

$$n_1^{-1}V(t) \rightarrow D(t) = (p_1^{-1}(t) + \pi^{-1}p_2^{-1}(t))^{-1}, \quad (3)$$

where $p_j(t) = P(I_j(t) = 1)$ and $\pi = n_2/n_1$ is the treatment allocation ratio. Thus, $B(I) = n_1^{-1/2}U(t)/D^{1/2}(\tau) \sim N(\theta I, I)$ is approximately a Brownian motion with drift parameter $\theta = n_1^{1/2} \log(\delta) D^{1/2}(\tau)$ and information time $I = D(t)/D(\tau)$, where $D(\tau)$ is the value of $D(t)$ at $t = \tau$.

3 Sample Size for Fixed Sample Test

The sample size for a fixed sample test is calculated for the situation at the end of the study. Based on test statistic $Z(t)$ at $t = \tau$, under the null hypothesis,

$$Z(\tau) = \log\{U_2(\tau)d_1(\tau)/U_1(\tau)d_2(\tau)\} (d_1^{-1}(\tau) + d_2^{-1}(\tau))^{-1/2}$$

has an approximate standard normal distribution. To calculate the power, let $p_j(\tau)$ be the probability of a subject from the j^{th} group having an event during the study. Then under the alternative $\delta = \lambda_1/\lambda_2 > 1$, $Z(\tau)$ is an approximately normal distribution with mean $n_1^{1/2} \log(\delta) D^{1/2}(\tau)$ and unit variance. Therefore, given a significance level α , the power $(1 - \beta)$ of the $Z(\tau)$ test under the alternative is given by

$$power \simeq \Phi\{n_1^{1/2} \log(\delta)(p_1^{-1}(\tau) + \pi^{-1} p_2^{-1}(\tau))^{-1/2} - z_{1-\alpha}\},$$

where $\Phi(\cdot)$ is the standard normal distribution function and $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$. Thus, the sample size of the first group based on the $Z(\tau)$ test can be calculated as

$$n_1 = \frac{(z_{1-\alpha} + z_{1-\beta})^2 (p_1^{-1}(\tau) + \pi^{-1} p_2^{-1}(\tau))}{[\log(\delta)]^2}, \quad (4)$$

where $\delta = R^\kappa$ (Wu, 2013). Therefore, the total sample size for the two groups is given by

$$n = \frac{(\pi + 1)(z_{1-\alpha} + z_{1-\beta})^2 (p_1^{-1}(\tau) + \pi^{-1} p_2^{-1}(\tau))}{[\log(\delta)]^2}. \quad (5)$$

To calculate the number of subjects required for the study, we need to calculate $p_j(\tau)$, the probability of a subject in the j^{th} group having an event during the study. Typically, we assume that subjects are accrued over an accrual period of length t_a with an additional follow-up period of length t_f . A subject enters the study at time u , the entry time is uniformly distributed on $[0, t_a]$, and no subject is lost to follow-up during the study. Then the probability of a subject having an event during the study under the Weibull model can be calculated by (Collett, 2003):

$$p_j(\tau) = 1 - \frac{1}{t_a} \int_{t_f}^{t_a+t_f} e^{-\log(2)(\frac{u}{m_j})^\kappa} du. \quad (6)$$

Therefore, given the design parameters δ (or κ), m_1 , m_2 , α , β , π , t_f and t_a , the number of subjects n required for the study can be calculated using formula (5).

In designing an actual trial, given the accrual time t_a , calculating the sample size is often impractical because we may not be able to enroll the total number of subjects as planned in the given accrual duration. It is more practical to design the study starting with the accrual rate r and then calculate the required accrual time t_a . This can be accomplished under the Weibull model assumption. First, the integration in the probability formula (4) can be simplified by approximation using Simpson's rule,

$$p_j(\tau) = 1 - \frac{1}{6} \{S_j(t_f) + 4S_j(0.5t_a + t_f) + S_j(t_a + t_f)\}. \quad (7)$$

Then, combining the sample size formula based on (5) with equation (7), we can define a root function of the accrual time t_a

$$\text{root}(t_a) = rt_a - \frac{(\pi+1)(z_{1-\alpha}+z_{1-\beta})^2(p_1^{-1}(t_a+t_f)+\pi^{-1}p_2^{-1}(t_a+t_f))}{[\log(\delta)]^2}. \quad (8)$$

Now the accrual time t_a can be obtained by solving the root equation $\text{root}(t_a) = 0$ numerically in Splus using the **uniroot** function. The total sample size required for the study is approximately $n = [rt_a]^+$, where $[x]^+$ denotes the smallest integer greater than x .

The total sample sizes for two groups under each scenario for fixed sample tests recorded in Table 1 were calculated from formula (5) for parametric test $Z(\tau)$ and from a formula given by Collett (2003) for the log-rank test $L(\tau)$,

$$n = \frac{(z_{1-\alpha}+z_{1-\beta})^2}{\pi_1\pi_2[\log(\delta)]^2P(\tau)}, \quad (9)$$

where $\pi_1 = 1/(1+\pi)$ and $\pi_2 = \pi/(1+\pi)$ are proportions of subjects assigned to treatment 1 and treatment 2, respectively, and $P(\tau) = \pi_1 p_1(\tau) + \pi_2 p_2(\tau)$ is the combined probability of failure in $[0, \tau]$ for subjects from the two groups. As shown in Table 1, the sample sizes for the parametric test and log-rank test are very similar. In an unpublished manuscript on the log-rank test, Xiong (2014) obtained a precise analytical formula of $E(\tau)$, where $E(\tau) = \mu(\tau)^2/v(\tau)$ and $\mu(\tau) = \lim_{d \rightarrow \infty} \mu(\tau)/d$ and $v(\tau) = \lim_{d \rightarrow \infty} V(\tau)/d$, where $\mu(\tau)$ and $V(\tau)$ are the mean and variance of log-rank score statistic, and d is the total number of events in $[0, \tau]$. The precise number of failures in calendar time interval $[0, \tau]$ for the log-rank test should be $d = (z_{1-\alpha}+z_{1-\beta})^2/E(\tau)$, and the precise sample size should be $n = d/P(\tau)$, where $P(\tau)$ is defined above as the combined probability of failure on $[0, \tau]$ for the two groups. $E(\tau)$ is a function of the survival distributions, the hazards ratio, the entry time distribution, the censoring distribution, and allocation proportions of subjects in the two groups. By numerical computation using $E(\tau)$ as a criterion, we evaluated the existing formula of the sample size for the log-rank test and proposed a new formula for it. The computation indicates that $[\log(\delta)]^2 \pi_1 \pi_2 \approx E(\tau)$ when the allocation is balanced (i.e., $\pi_1(\pi_2)$ is close to 0.5) and $|\log(\delta)| \approx 1$; this verifies the accuracy of the formula (9) for this range of parameters in the application. The computation also indicates that $[\log(\delta)]^2 \pi_1 \pi_2 p_1(\tau)p_2(\tau)/[P(\tau)]^2 \approx E(\tau)$ for any $0 < \pi_1 < 1$ and $|\log(\delta)| \approx 2$, which leads to

$$n = \frac{(z_{1-\alpha}+z_{1-\beta})^2 P(\tau)}{\pi_1 \pi_2 [\log(\delta)]^2 p_1(\tau) p_2(\tau)}, \quad (10)$$

as a formula for the sample size calculation of the log-rank test that is more accurate than the formula in (9). We will give a numerical example in Section 5 to illustrate this feature. It is straightforward to check that equations (5) and (10) are mathematically equivalent, which implies that the formula in (5) not only works for the proposed parametric test for the Weibull distribution, but also works well for the log-rank test, especially when the assignment of subjects is unbalanced for the two groups.

By the way, in practice the attrition from clinical trials should be considered in the design stage. Patients who are lost to follow-up for various causes during the study are censored in the survival analysis. Assume that losses to follow-up are random and independent of the survival distribution. Thus it is part of the censoring distribution and can be incorporated in the trial design. Sample sizes can also be adjusted for the attrition due to patients' dropout or noncompliance to the study (see e.g. Lachin and Foulkes, 1986).

4 Group Sequential Procedure

In this section, we will apply an SCPRT procedure (Xiong, 1995) to the test statistic $Z(t)$. The SCPRT has two unique features: (1) the maximum sample size of the sequential test is not greater than the size of the reference fixed sample test; and (2) the probability of discordance, or the probability that the conclusion of the sequential test would be reversed if the experiment were not stopped according to the stopping rule but continued to the planned end, can be controlled to an arbitrarily small level (Xiong et al., 2007). Furthermore, the power function of the SCPRT is virtually the same as that of the fixed sample test (Xiong, 1995). The SCPRT boundaries derived in this paper have analytical solutions. All these features make the SCPRT attractive and simple to use.

Let $\{B_t; 0 < t \leq 1\}$ be the Brownian motion $B_t \sim N(\theta t, t)$ and B_1 be the B_t at the final stage with full information $t = 1$. Then the joint distribution of (B_t, B_1) has a bivariate distribution with mean $\mu = (\theta t, \theta)$ and a variance matrix $\Sigma = (\sigma_{ij})_{2 \times 2}$ with $\sigma_{11} = \sigma_{12} = \sigma_{21} = t$ and $\sigma_{22} = 1$. Therefore, according to multivariate conditional distribution theory (e.g., Anderson, 1958), the conditional density $f(B_t|B_1)$ is the normal density of $N(B_1 t, (1-t)t)$. Let $s_0 = z_{1-\alpha}$ be the critical value of B_1 to reject the null for the fixed sample test. Then the conditional maximum likelihood ratio for the stochastic process on information time t is (Xiong, 1995; Xiong, et al., 2003)

$$L(t, B_t|z_{1-\alpha}) = \frac{\max_{\{s > s_0\}} f(B_t|B_1=s)}{\max_{\{s \leq s_0\}} f(B_t|B_1=s)}.$$

Taking the logarithm, the log-likelihood ratio can be simplified as

$$\log(L(t, B_t|z_{1-\alpha})) = \pm \frac{(B_t - z_{1-\alpha}t)^2}{2(1-t)t},$$

which has a positive sign if $B_t > z_{1-\alpha}t$ and a negative sign if $B_t < z_{1-\alpha}t$. This equation leads to lower and upper boundaries for B_{t_k} as

$$a_k = z_{1-\alpha}t_k - \{2at_k(1-t_k)\}^{1/2}; b_k = z_{1-\alpha}t_k + \{2at_k(1-t_k)\}^{1/2}, \quad (11)$$

for $k = 1, \dots, K$, where K is the total number of looks, and $t_1, t_2, \dots, t_K (= 1)$ are the information times of the interim looks and the final look. The a in (11) is the boundary

coefficient, and it is crucial to choose an appropriate a for the design such that the probability of conclusion by the sequential test being reversed by the test at the planned end is small but not unnecessarily small. The larger a is, the smaller is the discordance probability, and the wider apart the upper and lower boundaries are, making it harder for the sample path to reach the boundaries and stop early and resulting in larger expected sample sizes. Thus, an appropriate a can be determined by choosing an appropriate discordance probability (Xiong, 1995; Xiong et al., 2003).

Now we apply the SCPRT to the test statistic $B(I) = n_1^{-1/2} U(t) / D^{1/2}(\tau) \sim N(\theta I, I)$ which is a Brownian motion in information time $I = D(t)/D(\tau)$ on $[0, 1]$, and the drift parameter

$\theta = n_1^{1/2} \log(\delta) D^{1/2}(\tau)$ where $D(t)$ is defined by (3) in Section 2. Suppose k^{th} interim looks are planned at calendar time $t_k, k = 1, \dots, K$. Then based on the SCPRT procedure presented above, the lower and upper boundaries for $B_{I_k} = B(I_k)$ at the k^{th} look are given by

$$a_k = z_{1-\alpha} I_k - \{2a I_k (1 - I_k)\}^{1/2}; b_k = z_{1-\alpha} I_k + \{2a I_k (1 - I_k)\}^{1/2}, \quad (12)$$

for $k = 1, \dots, K$, where $I_k = D(t_k)/D(\tau)$ is the information time at the k^{th} look at calendar time t_k . The nominal critical p -values for testing H_0 are

$$P_{a_k} = 1 - \Phi(a_k / \sqrt{I_k}); P_{b_k} = 1 - \Phi(b_k / \sqrt{I_k}). \quad (13)$$

The observed p -value at the k^{th} look is

$$P_{S_{I_k}} = 1 - \Phi(B_{I_k} / \sqrt{I_k}). \quad (14)$$

The stopping rule for monitoring the trial can be executed by stopping the trial when, for the first time, $P_{B_{I_k}} \leq P_{a_k}$ (accept H_0 and stop for futility) or $P_{B_{I_k}} \leq P_{b_k}$ (reject H_0 and stop for efficacy). Since $Z(t_k)$ and $B_{I_k} / \sqrt{I_k}$ have the same asymptotic distribution under the null hypothesis, the observed p -value at the k^{th} stage can be calculated from the test statistic $Z(t_k)$ by applying all observations up to stage k . As an illustration, the calculations of the operating characteristics of a multi-stage group sequential design are given in the Appendix 2.

5 Simulation Studies

In this section, we conducted simulation studies to compare the power and type I error of the proposed parametric test statistic $Z(t)$ and the non-parametric log-rank test $L(t)$ under various scenarios. In the simulations, the survival distribution of the j^{th} group was taken as $S_j(t) = e^{-\log(2)(t/m_j)^\kappa}$, which is the Weibull distribution with shape parameter κ and median survival time $m_j, j = 1, 2$, where the shape parameter κ was taken as 0.5, 1, and 2 to reflect cases of decreasing, constant, and increasing hazard functions.

The null hypothesis was set to $H_0 : m_1 = m_2 (= 1)$, and the ratio of medians $R = m_2/m_1$ under the alternative was taken as 1.5 – 2.0. Furthermore, we assumed that subjects were recruited with a uniform distribution over the accrual period $t_a = 5$ (years) and followed for $t_f = 2$ (years), and no subject was lost to follow-up during the study period $\tau = t_a + t_f = 7$. Therefore, a subject was censored at calendar time t if his/her event time was longer than $t - u$, where u is the time when the subject entered the study. The sample sizes of the two groups were balanced, and thus $\pi = n_2/n_1 = 1$; this setting was based on the facts that the total sample size $n = n_1 + n_2$ is minimized with π close to 1 and a study with groups of equal sample sizes is easier to plan and manage.

In each design parameter configuration, 100,000 observed samples of censored event times were generated from the Weibull distribution to calculate the test statistics under the null or alternative hypothesis. The nominal significance level was set to 0.05 and power was set to 80% and 90%. Sample sizes, empirical type I error, and power for the fixed sample test were calculated at the end of the study, $\tau = 7$. The empirical type I error and power for the two-stage SCPRT design were calculated at calendar time $t_1 = 4$ and $t_2 = 7$. The simulated empirical powers and type I errors in various scenarios for the fixed sample test and two-stage SCPRT test are summarized in Table 1 and Table 2, respectively.

Sample sizes under each scenario for fixed sample tests recorded in Table 1 were calculated by formula (5) for the parametric test $Z(\tau)$ and by formula (9) for the log-rank test $L(\tau)$; the latter formula was given by Collett (2003). As shown in Table 1, the sample sizes for the two tests were very similar, both assuming $\pi = 1$, which is a favorable condition as discussed in the last paragraph of Section 3. For π not close to 1, the sample sizes for the two tests could be different. For example, in Table 1, assume the conditions are same as before except letting $\pi = 7/13$ (or $\pi_1 = 0.65$ and $\pi_2 = 0.35$); then for $\kappa = 0.5$ and $R = 1.5$, the total sample size n from formula (9) is 1249, whereas the total sample size n from formula (5) or (10) is 1289. The latter is close to the precise sample size, 1287, from the log-rank test using $E(\tau)$ (see the last paragraph of Section 3). The simulation results in Table 1 for the fixed sample tests showed that both the log-rank test $L(\tau)$ and parametric test $Z(\tau)$ had adequate empirical type I error and power for moderate to large sample sizes. However, the log-rank test $L(\tau)$ was liberal and underpowered when the sample size was small. The type I error and power of the parametric test $Z(\tau)$ were satisfactorily close to the nominal type I error of 0.05 and power of 90% or 80%, respectively, even when the sample size was small. The difference in performance for the two tests with a small sample size may be explained by the fact that $L(\tau)$ is non-parametric and includes only counting data, whereas $Z(\tau)$ is parametric and includes counting data $d_1(\tau)$ and $d_2(\tau)$ as well as continuous data $U_1(\tau)$ and $U_2(\tau)$. The results for the two-stage SCPRT design (Table 2) showed again that the empirical type I error and power of both tests were close to the nominal level at each stage for moderate to large sample sizes. However, the log-rank test $L(t)$ was liberal and underpowered when the sample size was small. The parametric test $Z(t)$ performed better with adequate empirical type I error and power at each stage.

To study the Brownian motion property of the statistic $U(t)$ in equation (2), the empirical correlation matrix of the increments of the $U(t)$ at times $t = 3, 4, 5, 6, 7$ with sample size $n =$

100 per group were also computed through 100,000 simulations. All of the correlations were close to theoretical value zero (see Table 3).

6 An Example

Rhabdoid tumors are aggressive pediatric malignancies with a poor prognosis. Over the past 5 years, St. Jude Children's Research Hospital accrued 14 pediatric patients with recurrent or refractory non-CNS rhabdoid tumors treated with conventional chemotherapy. The median event-free survival is only about 1 year, where the event is defined as disease relapse or death. All 14 patients had events within about 3 years. The Weibull model was fitted in Splus to the data, resulting an estimate (standard error) of the shape parameter $\kappa = 1.37(0.28)$ and median event-free survival time of $m_1 = 0.936$ years, which provides a more satisfactory model than the exponential model (Wu, 2013). Now, suppose that we would like to design a multi-center randomized two-arm trial to assess the effectiveness of the small molecule inhibitor alisertib (treatment 2) versus conventional chemotherapy (treatment 1) for this group of patients. Patients will be randomized with equal allocation to each treatment group and hence $\pi = n_2/n_1 = 1$. The hypotheses of the planned study are $H_0 : m_2 \leq m_1$ vs. $H_1 : m_2 > m_1$. The investigators would like to detect a half-year increase in the median event-free survival of the alisertib treatment group over that of the conventional chemotherapy group, with 90% power, 5% type I error, and 2 years of follow-up ($t_f = 2$) after the last patient is enrolled in the study. Then for the alternative hypothesis, $m_2 = m_1 + 0.5 = 1.436$ and $\delta = \lambda_1/\lambda_2 = (m_2/m_1)^\kappa = 1.797$ and $\gamma = \log(\delta) = 0.586$. Assume this multi-center trial has the capacity to enroll and treat 20 patients per year. Then under the assumption of the Weibull model, with uniform entry and no loss to follow-up, the required total accrual time is $t_a = 5.3$ years, calculated by equation (8), in which $p_1(\tau)$ and $p_2(\tau)$ are found by (7) with $\tau = t_a + t_f$. Then the study duration is $\tau = t_a + t_f = 7.3$ years, and the total sample size is 106 patients (53 per group). Now assuming interim and final looks are planned at calendar times $t_1 = 4$, $t_2 = 5$, and $t_3 = \tau = 7.3$ years, the corresponding information time at each planned interim look t_k can be calculated by $I_k = D(t_k)/D(\tau)$, where

$$D(t) = (p_1^{-1}(t) + \pi^{-1} p_2^{-1}(t))^{-1} \text{ with}$$

$$p_j(t) = \frac{1}{t_a} \int_0^{t \wedge t_a} \{1 - S_j(t-u)\} du,$$

$S_j(t) = e^{-\log(2)(\frac{t}{m_j})^\kappa}$ and $\pi = 1$. By calculation, the corresponding information times are $I_1 = 0.511$, $I_2 = 0.706$, and $I_3 = 1$. Assuming the maximum conditional probability of discordance $\rho = 0.02$, then the boundary coefficient $a = 2.593$ for $K = 3$ (Xiong et al., 2003), and the maximum probability of discordance is $\rho_{max} = 0.0043$. That is, under the most unfavorable setting of mean parameter, i.e., the underlying true drift θ of Brownian motion $B(I)$ is pointing to the cutoff point $z_{1-\alpha}$ of the test at the final stage, on average, for every 1,000 such sequential tests, there would be only 4.3 cases of later reversal of conclusion (significant or futility) at early stopping if the sequential test were not stopped as it should be but continued to the planned end. The lower boundary and upper boundary calculated from (12) are $(a_1, a_2, a_3) = (-0.298, 0.124, 1.645)$ and $(b_1, b_2, b_3) = (1.979, 2.199, 1.645)$,

respectively. The acceptance and rejection nominal critical significance levels are (0.6615, 0.4415, 0.050) and (0.0028, 0.0044, 0.050).

We performed 100,000 simulation runs under the Weibull distribution to evaluate the operating characteristics of the proposed group sequential design. The empirical (nominal) type I error and power of the sequential test are 0.0514 (0.0505) and 0.9034 (0.8995), respectively. The empirical (nominal) probabilities of stopping under the null and alternative hypotheses are 0.3446 (0.3412) and 0.2456 (0.2406) at the first look and 0.5902 (0.5817) and 0.4739 (0.4690) at the second look. The details of the operating characteristics for the proposed group sequential design are shown in Table 4.

7 Conclusion

A parametric sequential test statistic under the Weibull model is proposed. Simulation results showed that the proposed parametric sequential test $Z(t)$ has better small-sample properties than the log-rank test $L(t)$ under the Weibull model. A multi-stage group sequential procedure is given based on the SCPRT test proposed by Xiong (1995). The maximum sample size of the sequential test is the same as the sample size of the fixed sample test, and group sequential boundaries have analytical solutions. Therefore, the proposed group sequential procedure is attractive and simple to use. The study can be monitored by pre-planned multi-stage interim analyses to stop for either efficacy or futility of the new treatment. By the way, if the hypothesis of a randomized phase III trial is a two-sided hypothesis, one can replace $z_{1-\alpha}$ by $z_{1-\alpha/2}$ in equations (5) or (8) to obtain a sample size formula, and in equation (12) to obtain SCPRT boundaries for the two-sided hypotheses for a group sequential trial design.

Acknowledgments

This work was supported in part by National Cancer Institute (NCI) support grant CA21765 and the American Lebanese Syrian Associated Charities (ALSAC).

Appendix 1: Derivation of the Sequential Test Statistic $Z(t)$

First, for notation convenience, we convert (λ_1, λ_2) to (γ, λ) , where $\gamma = \log(\lambda_1/\lambda_2)$ is the log hazards ratio and $\lambda = \lambda_2$. Then the log-likelihood at calendar time t for (γ, λ) is given by

$$\ell_\gamma(\gamma, \lambda; t) = d_1(t) \log(\lambda) + \gamma d_1(t) + d_2(t) \log(\lambda) - \lambda e^\gamma U_1(t) - \lambda U_2(t).$$

By solving the following score equations:

$$\begin{aligned} \ell_\gamma(\gamma, \lambda; t) &= d_1(t) - \lambda e^\gamma U_1(t), \\ \ell_\lambda(\gamma, \lambda; t) &= \frac{d_1(t)}{\lambda} + \frac{d_2(t)}{\lambda} - e^\gamma U_1(t) - U_2(t), \end{aligned}$$

the maximum likelihood estimates of γ and λ are

$$\hat{\gamma}(t) = \log \left\{ \frac{U_2(t)d_1(t)}{U_1(t)d_2(t)} \right\} \text{ and } \hat{\lambda}(t) = \frac{d_2(t)}{U_2(t)}.$$

The observed Fisher information matrix is given by

$$j(\hat{\gamma}, \hat{\lambda}; t) = \begin{pmatrix} d_1(t) & U_2(t)d_1(t)d_2^{-1}(t) \\ U_2(t)d_1(t)d_2^{-1}(t) & U_2^2(t)d_2^{-2}(t)(d_1(t)+d_2(t)) \end{pmatrix},$$

and then the variance of γ can be estimated by $\hat{j}^{\gamma\gamma}(t) = (d_1^{-1}(t) + d_2^{-1}(t))$, which is the (1,1) entry in the inverse of the Fisher information matrix $j^{-1}(\gamma, \lambda, t)$. Therefore, the Wald test statistic of $\gamma(t)$ is given by

$$Z(t) = (\hat{\gamma}(t) - \gamma) \{ \hat{j}^{\gamma\gamma}(t) \}^{-1/2}.$$

Under the null hypothesis $H_0: \gamma = 0$,

$$Z(t) = \log \{ U_2(t)d_1(t)/U_1(t)d_2(t) \} (d_1^{-1}(t) + d_2^{-1}(t))^{-1/2}$$

has an approximate standard normal distribution.

Appendix 2: Computation for the sequential test normalized with information time

Assumption

Let $B(t) \sim N(\theta, t)$ be a Gaussian process with the time variable t on interval $[0, 1]$ and drift θ . Let $0 < t_1 < \dots < t_K = 1$ be the information times of the looks for a sequential test with K looks. Let $a_k < b_k$ be lower and upper boundaries for $B(t)$ at time t_k for $k = 1, \dots, K-1$, and $a_K = b_K$.

Function $l_{t_\kappa}^*(\cdot)$

Define $l_{t_\kappa}^*(s)$ as a function of s on interval (a_k, b_k) for $k = 1, \dots, K-1$; this series of functions can be determined recursively as follows. Let $l_{t_1}^*(s) = 1$ for s on (a_1, b_1) ; for $k = 2, \dots, K$, for s in (a_k, b_k) ,

$$l_{t_\kappa}^*(s) = \frac{1}{\sqrt{t_{\kappa-1}(1 - t_{\kappa-1}/t_\kappa)}} \int_{a_{\kappa-1}}^{b_{\kappa-1}} l_{t_{\kappa-1}}^*(x) \phi \left(\frac{x - st_{\kappa-1}/t_\kappa}{\sqrt{t_{\kappa-1}(1 - t_{\kappa-1}/t_\kappa)}} \right) dx, \quad (15)$$

where $\phi(\cdot)$ in (15) is the density function of the standard normal distribution.

Function $l_{t_k}(\cdot)$

Define $l_{t_k}(s)$ as a function of s on interval $(-\infty, a_k) \cup (b_k, \infty)$, for $k = 1, \dots, K$; this series of functions can be determined using functions $l_{t_k}^*(\cdot)$ defined in (15) as follows. Let $l_{t_1}(s) = 1$ for s on $(-\infty, a_1) \cup (b_1, \infty)$; for $k = 2, \dots, K$, for s in $(-\infty, a_k) \cup (b_k, \infty)$, let

$$l_{t_k}(s) = \frac{1}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}} \int_{a_{k-1}}^{b_{k-1}} l_{t_{k-1}}^*(x) \phi\left(\frac{x - st_{k-1}/t_k}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}}\right) dx, \quad (16)$$

where $\phi(\cdot)$ in (16) is the density function of the standard normal distribution.

Power Function

For testing $H_0: \theta = 0$ vs. $H_a: \theta > 0$, with functions $l_{t_k}(\cdot)$ by (16), the power function $P(\theta)$ or the probability of rejecting H_0 under the true mean θ is

$$P(\theta) = \sum_{k=1}^K (1/\sqrt{t_k}) \int_{b_k}^{\infty} l_{t_k}(x) \phi((x - \theta t_k)/\sqrt{t_k}) dx. \quad (17)$$

For the sequential test design, the significance level is $\alpha = P(0)$ and the power is $1 - \beta = P(\theta_a)$, where θ_a is the value of θ under H_a .

Probability of Stopping

The probability of stopping at time t_k is a function of θ as

$$P_{t_k}(\theta) = (1/\sqrt{t_k}) \left(\int_{-\infty}^{a_k} + \int_{b_k}^{\infty} \right) l_{t_k}(x) \phi((x - \theta t_k)/\sqrt{t_k}) dx, \quad (18)$$

with which the probability of stopping at t_k is $P_{t_k}(0)$ for the null hypothesis and $P_{t_k}(\theta_a)$ for the alternative hypothesis.

Expected Stopping Time

With the probability of stopping $P_{t_k}(\theta)$ from (18), the expected stopping $ET(\theta)$ is a function of θ as

$$ET(\theta) = \sum_{k=1}^K t_k P_{t_k}(\theta), \quad (19)$$

with which the expected stopping time is $ET(0)$ for the null hypothesis and $ET(\theta_a)$ for the alternative hypothesis.

Expected Sample Size

Suppose the maximum sample size for the sequential test is n . The expected sample size for the sequential test is a function of θ and can be obtained by

$$EN(\theta) = n \times ET(\theta), \quad (20)$$

with an expected sample size of $EN(0)$ for the null hypothesis and $EN(\theta_a)$ for the alternative hypothesis.

For SCPR design

To test $H_0: \theta = 0$ vs. $H_a: \theta > 0$ with significance level α and power $1 - \beta$ by an SCPR design, the cutoff value at the final stage $t_K = 1$ is $a_K = b_K = z_{1-\alpha}$, the drift at the null hypothesis is $\theta_0 = 0$, and the drift at the alternative hypothesis is $\theta_a = z_{1-\alpha} + z_{1-\beta}$ which are the same as those for the fixed test at the final stage with information time $t = 1$. Including the cutoff value θ_0 and θ_a into equations (17), (18), (19), and (20), we can compute the type I error, power, probability of stopping at a given t_k , and expected sample sizes under the null and alternative hypotheses for the SCPR design.

For details of the derivation of these computational formulas, please refer to Xiong and Tan (1999, 2001) and Xiong et al. (2002).

References

- Anderson, TW. An introduction to multivariate statistical analysis. New York: Wiley; 1958.
- Collett, D. Modeling survival data in medical research. 2nd. London: Chapman and Hall; 2003.
- Cox, DR.; Oakes, DV. Analysis of Survival Data. London: Chapman and Hall; 1984.
- Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. British Journal of Radiology. 1971; 44:793–797. [PubMed: 4940475]
- Heo M, Faith MS, Allison DB. Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest. Mechanisms of Ageing and Development. 1998; 102:45–53. [PubMed: 9663791]
- Jennison, C.; Turnbull, BW. Group sequential methods with applications to clinical trials. New York: Chapman and Hall; 2000.
- Jiang Z, Wang L, Li C, Xia J, Jia H. A practical simulation method to calculate sample size of group sequential trials for time-to-event under exponential and Weibull distribution. PLOS ONE. 2012; 7:1–12.
- Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983; 70:659–663.
- Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics. 1986; 42:507–519.
- Lu Q, Tse SK, Chow SC, Lin M. Analysis of time-to-event data nonuniform patient entry and loss to follow-up under a two-stage seamless adaptive design with Weibull distribution. Journal of Biopharmaceutical Statistics. 2012; 22:773–784.
- Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64:191–199.
- Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983; 39:499–503. [PubMed: 6354290]

- Schoenfeld DA, Ritche JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics*. 1982; 38:163–170. [PubMed: 7082758]
- Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika*. 1983; 79:315–326.
- Slud EV. Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*. 1984; 12:551–571.
- Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*. 1982; 77:855–861.
- Tsiatis AA, Boucher H, Kim K. Sequential methods for parametric survival models. *Biometrika*. 1995; 70:165–173.
- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
- Whitehead J, Stratton I. Group sequential clinical trial with triangular continuation regions. *Biometrics*. 1983; 39:227–236. [PubMed: 6871351]
- Wu J. Power and sample size for randomized phase III survival trials under the Weibull model. *Journal of Biopharmaceutical Statistics*. 2013 inprint.
- Xiong X. A class of sequential conditional probability ratio tests. *Journal of American Statistical Association*. 1995; 15:1463–1473.
- Xiong X. A precise approach for sequential test design on comparing survival distributions by log-rank test. Un-published Manuscript. 2014
- Xiong X, Tan M, Boyett J. Sequential conditional probability ratio tests for normalized test statistic on information time. *Biometrics*. 2003; 59:624–631. [PubMed: 14601763]

Sample size and simulated empirical type I error (α) and power ($1 - \beta$) based on 100,000 simulation runs for the Weibull distribution for fixed sample tests $Z(\tau)$ and $L(\tau)$ with a nominal type I error of 0.05 and powers of 80% and 90% (one-sided test).

Table 1

κ	Design	(90% Power)	$R=1.5$			$R=1.6$			$R=1.7$		
			n^*	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
0.5	$Z(\tau)$		578	0.049	0.902	434	0.051	0.901	344	0.050	0.903
	$L(\tau)$		577	0.049	0.900	433	0.050	0.901	342	0.050	0.899
1	$Z(\tau)$		118	0.052	0.903	89	0.050	0.902	71	0.051	0.907
	$L(\tau)$		118	0.052	0.900	89	0.051	0.899	70	0.051	0.897
2	$Z(\tau)$		27	0.051	0.903	20	0.052	0.902	16	0.053	0.903
	$L(\tau)$		27	0.054	0.887	20	0.055	0.881	16	0.057	0.876
<hr/>											
κ	Design	(90% Power)	$R=1.8$			$R=1.9$			$R=2.0$		
			n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
0.5	$Z(\tau)$		283	0.049	0.902	239	0.050	0.903	207	0.050	0.903
	$L(\tau)$		281	0.050	0.899	237	0.050	0.900	205	0.049	0.900
1	$Z(\tau)$		58	0.051	0.904	50	0.051	0.909	43	0.052	0.910
	$L(\tau)$		58	0.052	0.898	49	0.052	0.898	43	0.053	0.902
2	$Z(\tau)$		13	0.054	0.903	11	0.055	0.906	10	0.054	0.916
	$L(\tau)$		13	0.058	0.871	11	0.059	0.867	10	0.061	0.876
<hr/>											
κ	Design	(80% Power)	$R=1.5$			$R=1.6$			$R=1.7$		
			n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
0.5	$Z(\tau)$		418	0.051	0.802	314	0.050	0.801	248	0.050	0.802
	$L(\tau)$		417	0.049	0.798	313	0.049	0.800	247	0.049	0.800
1	$Z(\tau)$		85	0.052	0.802	64	0.051	0.805	51	0.052	0.804
	$L(\tau)$		85	0.052	0.796	64	0.053	0.798	51	0.053	0.796

Design		R=1.5			R=1.6			R=1.7		
κ	Test	n^*	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
2	$Z(\tau)$	20	0.052	0.816	15	0.053	0.815	12	0.054	0.819
	$L(\tau)$	20	0.055	0.789	15	0.056	0.781	12	0.059	0.778
Design		R=1.8			R=1.9			R=2.0		
κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
0.5	$Z(\tau)$	204	0.050	0.805	173	0.051	0.806	149	0.049	0.804
	$L(\tau)$	203	0.051	0.800	172	0.051	0.803	148	0.051	0.799
1	$Z(\tau)$	42	0.053	0.807	36	0.052	0.811	31	0.052	0.811
	$L(\tau)$	42	0.053	0.798	36	0.052	0.801	31	0.054	0.798
2	$Z(\tau)$	10	0.054	0.827	8	0.055	0.810	7	0.056	0.816
	$L(\tau)$	10	0.061	0.778	8	0.062	0.748	7	0.063	0.746

* n is the sample size per group (with equal allocation), and the sample size for the log-rank test was calculated from formula (9) given by Collett (2003).

Table 2

Simulated empirical type I error and power of the two-stage SCPRT designs based on 100,000 simulation runs for sequential test statistics $Z(t)$ and $L(t)$ with a nominal type I error of 0.05 and power of 90% (one-sided test).

<i>R</i>	κ	Test	At k^{th} interim look	Type I error			Power		
				$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total
1.5	0.5	$Z(t)$	Empirical	0.005	0.045	0.050	0.382	0.519	0.901
		$L(t)$	Empirical	0.005	0.044	0.049	0.385	0.514	0.899
			Nominal	0.005	0.046	0.051	0.382	0.518	0.899
	1	$Z(t)$	Empirical	0.005	0.047	0.052	0.325	0.577	0.902
		$L(t)$	Empirical	0.005	0.047	0.053	0.324	0.575	0.899
			Nominal	0.005	0.046	0.051	0.325	0.574	0.899
	2	$Z(t)$	Empirical	0.005	0.047	0.052	0.323	0.579	0.902
		$L(t)$	Empirical	0.006	0.049	0.055	0.293	0.593	0.886
			Nominal	0.004	0.046	0.051	0.326	0.573	0.899
	0.5	$Z(t)$	Empirical	0.005	0.045	0.050	0.378	0.524	0.902
		$L(t)$	Empirical	0.005	0.046	0.051	0.377	0.522	0.899
			Nominal	0.005	0.046	0.051	0.382	0.518	0.899
	1	$Z(t)$	Empirical	0.005	0.047	0.052	0.315	0.588	0.903
		$L(t)$	Empirical	0.005	0.048	0.053	0.315	0.582	0.897
			Nominal	0.005	0.046	0.051	0.314	0.585	0.899
	2	$Z(t)$	Empirical	0.005	0.049	0.054	0.286	0.614	0.900
		$L(t)$	Empirical	0.006	0.053	0.060	0.236	0.633	0.869
			Nominal	0.004	0.046	0.051	0.285	0.615	0.899
	0.5	$Z(t)$	Empirical	0.005	0.046	0.051	0.378	0.525	0.903
		$L(t)$	Empirical	0.005	0.045	0.050	0.375	0.524	0.899
			Nominal	0.005	0.046	0.051	0.377	0.523	0.899
	1	$Z(t)$	Empirical	0.005	0.048	0.052	0.309	0.599	0.908

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Design	Test	At k^{th} look	Type I error			Power		
			$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total
R	$L(t)$	Empirical	0.006	0.048	0.054	0.312	0.589	0.901
		Nominal	0.005	0.046	0.051	0.307	0.593	0.899
	$Z(t)$	Empirical	0.006	0.050	0.055	0.275	0.637	0.913
		Empirical	0.007	0.056	0.063	0.214	0.660	0.874
	$L(t)$	Nominal	0.004	0.046	0.051	0.285	0.615	0.899

Table 3

Simulated empirical correlation matrix of the statistic $U(t)$ based on 100,000 simulation runs with sample size $n = 100$.

κ	Calendar time t_k						
	3	4	5	6	7		
0.5	1	0.0092	0.0061	0.0059	0.0056		
		1	0.0036	0.0066	0.0070		
			1	0.0088	0.0087		
				1	0.0034		
						1	
	1	0.0089	0.0068	0.0060	0.0057		
		1	0.0060	0.0079	0.0075		
			1	0.0056	0.0062		
				1	0.0049		
						1	
2	1	0.0024	-0.0002	-0.0012	-0.0003		
		1	-0.0004	-0.0004	0.0010		
			1	0.0018	0.0027		
				1	0.0019		
						1	

Table 4

The empirical operating characteristics of the three-stage SCPRT design for the example were estimated based on 100,000 simulation runs under the Weibull distribution.

At k^{th} interim look	$k = 1$	$k = 2$	$k = 3$	Total
Type I error				
Nominal	0.0028	0.0031	0.0446	0.0505
Empirical	0.0031	0.0029	0.0453	0.0514
Power				
Nominal	0.2494	0.2066	0.4435	0.8995
Empirical	0.2499	0.2106	0.4429	0.9034
Probability of stopping under null				
Nominal	0.3412	0.2405	0.4184	1.0000
Empirical	0.3446	0.2456	0.4098	1.0000
Probability of stopping under alternative				
Nominal	0.2554	0.2136	0.5311	1.0000
Empirical	0.2557	0.2182	0.5261	1.0000
Expected stopping time and sample size	$ET(0)$	$ET(\theta_a)^*$	$EN(0)$	$EN(\theta_a)$
Nominal	0.7625	0.8124	81	87
Empirical	0.7593	0.8108	81	86

* $\theta_a = z_{1-\alpha} + z_{1-\beta}$ is the draft parameter at the alternative hypothesis.