Published in final edited form as:

Stat Biopharm Res. 2014 ; 6(4): 289–301. doi:10.1080/19466315.2014.923325.

Single-Arm Phase II Group Sequential Trial Design with Survival Endpoint at a Fixed Time Point

Jianrong Wu and Xiaoping Xiong

Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

Abstract

In this paper, three non-parametric test statistics are proposed to design single-arm phase II group sequential trials for monitoring survival probability. The small-sample properties of these test statistics are studied through simulations. Sample size formulas are derived for the fixed sample test. The Brownian motion property of the test statistics allowed us to develop a flexible group sequential design using a sequential conditional probability ratio test procedure (Xiong, 1995). An example is given to illustrate the trial design by using the proposed method.

Keywords

Brownian motion; Interim analysis; Nelson-Aalen estimate; Single-arm phase II trial; Sample size; Survival probability

1 Introduction

Phase II trial evaluation is an important step in the development of new anti-cancer drugs and treatments. The goal of phase II cancer clinical trials is assessing the efficacy and toxicity of a new agent or treatment. Evidence of anti-cancer efficacy demonstrated in a phase II trial often determines whether the new agent or treatment warrants further study in a large and confirmatory randomized phase III trial. Phase II trials are often designed as single-arm studies with a limited number of subjects. Due to ethical concerns, interim monitoring rules are frequently used to terminate the trial early for futility, thereby reducing the number of subjects exposed to an ineffective therapy. The proportion of subjects achieving complete or partial response is often chosen as a surrogate endpoint of the study. Such trials with binary endpoints are often designed and monitored using Simon's optimal or minimax two-stage design. However, for many phase II cancer trials, the probability of a subject surviving after some clinically meaningful landmark point x is the primary endpoint. Designing such trials with interim monitoring rules is challenging because of incomplete follow-up for some subjects at the time of the interim analyses. A naive approach is to estimate the survival probability as a simple proportion of all subjects surviving the required time at the interim analyses. There are several disadvantages to this approach. First, subjects who have not been followed for the required time will not contribute to the estimate of the survival probability, resulting in an inefficient estimate. Second, simply treating subjects who are lost to follow-up as treatment failures or excluding them from the analysis will result in a biased estimate. Finally, suspending accrual to wait for all the subjects to be

followed up for the required length of time for an interim analysis is often impractical. Long trial suspensions can ruin a trial's momentum, increase study length, and increase costs (Case and Morgan, 2003).

To design a trial with incomplete follow-up data, survival analysis can be used. There is much literature available to assist in the design of randomized phase III survival trials with multi-stage interim analyses, such as Jennison and Turnbull (2000) and references therein. However there are few such methods for the design of single-arm phase II survival trials with interim monitoring rules. Recently, several methods have been proposed for monitoring the survival probabilities at a fixed time point for single-arm phase II trials. For example, Case and Morgan (2003) proposed an optimal two-stage phase II design to minimize the expected sample size or expected study length under the null hypothesis. Huang et al. (2010) modified Case and Morgan's approach to protect the type I error rate and improve the robustness of the design. Zhao et al. (2011) proposed a simulation based Bayesian decision theoretic two-stage design for a phase II clinical trial with survival endpoint. All these designs are restricted to a two-stage design only. Lin et al. (1996) developed a multi-stage sequential procedure for monitoring the survival probabilities, but it was developed for the purpose of testing a specific umbrella hypothesis of the National Wilms Tumor Study.

In phase II drug testing trials, it is also important to know as early as possible if the drug is highly promising. For this purpose and to allow more flexible monitoring of the trial, a sequential conditional probability ratio test (SCPRT) procedure (Xiong, 1995; Xiong et al., 2003) can be used to design a study with multi-stage interim analyses, and the trial can be monitored or stopped early for either futility or efficacy. The maximum sample size required for the SCPRT test is the same as for a fixed sample test, and the SCPRT boundaries have analytical solutions. Therefore, the proposed method can be easily applied to design a group sequential trial for monitoring the survival probability for a single-arm phase II study.

The rest of the paper is organized as follows: Section 2 discusses three non-parametric test statistics based on Nelson-Aalen estimate. Sample size formulas for a fixed sample test are derived in Section 3. A group sequential design using the SCPRT procedure is discussed in Section 4. The empirical type I error and power of the three test statistics are studied in Section 5 through simulations. An example is given in Section 6 to illustrate trial design using the proposed method. Section 7 is the concluding discussion.

2 Test Statistics

In a single-arm phase II trial with survival outcome, such as the time-to-failure (disease progression, relapse, or death) at some clinically meaningful landmark point x is often the primary interest. Let $S(x)$ be the survival probability at x and $S_0(x)$ be the level of survival probability at x at which investigators are no longer interested in the treatment. Then the design of the single-arm phase II trial can be based on testing the following one-sided hypotheses:

$$H_0: S(x) \leq S_0(x) \quad \text{vs} \quad H_a: S(x) > S_0(x),$$

for a given x ; it is equivalent to test the hypotheses based on the cumulative hazard function $\Lambda(x) = -\log S(x)$

$$H_0^\dagger: \Lambda(x) \geq \Lambda_0(x) \quad \text{vs} \quad H_a^\dagger: \Lambda(x) < \Lambda_0(x), \quad (1)$$

where $\Lambda_0(x) = -\log S_0(x)$

Suppose during the accrual phase of the trial, n subjects are enrolled in the study at calendar times $Y_1, \dots, Y_n$ which are measured from the start of the study. Let T_i and C_i denote, respectively, the failure time and censoring time of the i^{th} subject, with both being measured from the time of study entry of this subject. We assume that the failure time T_i is independent of the censoring time C_i and entry time Y_i and $\{(Y_i, T_i, C_i); i = 1, \dots, n\}$ are independent and identically distributed.

When the data are examined at calendar time $t \leq \tau$, where t is measured from the start of study and τ is the study duration, including the periods of enrollment and follow-up, we observe the time to failure $X_i(t) = T_i \wedge C_i \wedge (t - Y_i)^+$ and failure indicator $\delta_i(t) = I(T_i \leq C_i \wedge (t - Y_i)^+)$, $i = 1, \dots, n$. Based on the observed data $\{X_i(t), \delta_i(t), i = 1, \dots, n\}$, the Nelson-Aalen estimate for the cumulative hazard function $\Lambda(x) = \int_0^x \lambda(u) du$ is

$$\hat{\Lambda}(x; t) = \sum_{\{i: X_i(t) \leq x\}} \frac{\Delta_i(t)}{R_i(t)}, \quad (2)$$

where $R_i(t) = \sum_{j=1}^n I\{X_j(t) \geq X_i(t)\}$. It has been shown (Lin et al., 1996) that the process

$$n^{1/2} \{\Lambda_0(x) - \hat{\Lambda}(x; t)\}; \quad x < t$$

converges weakly to a Gaussian process with time variable t and independent increments and variance

$$\sigma^2(x; t) = \int_0^x \frac{\lambda(u)}{\pi(u; t)} du, \quad (3)$$

as a function of t , where $\pi(u, t) = P(X_1(t) \geq u)$, and $\sigma^2(x, t)$ can be estimated by

$$\hat{\sigma}^2(x; t) = \sum_{\{i: X_i(t) \leq x\}} \frac{\Delta_i(t)}{R_i^2(t)/n} \quad (4)$$

(Lin et al., 1996; Breslow and Crowley, 1974; Fleming and Harrington, 1991).

However, the distribution of $\hat{\Lambda}(x; t)$ is often skewed, and it requires large sample sizes to maintain its type I error rate. This is partly because $\hat{\Lambda}(x; t)$ is restricted to nonnegative values. However, since a logarithmic transformation of the Nelson-Aalen estimate $\log \hat{\Lambda}(x; t)$ takes the value over the entire real line, the asymptotic normality is expected to be more accurate. Therefore, Lin et al. (1996) suggested testing the hypothesis (1) using the following test statistic:

$$Z_1(x; t) = \frac{n^{1/2} \{\log \Lambda_0(x) - \log \hat{\Lambda}(x; t)\} \hat{\Lambda}(x; t)}{\hat{\sigma}(x; t)} \quad (5)$$

and showed that $Z_1(x; t)$ as a process in t can be approximated by a Gaussian process. Several other transformations have also been discussed in the literature (Klein et al., 2007), and the corresponding test statistics will be introduced as follows. The Nelson-Aalen estimate for the survival function is given as

$$\hat{S}(x; t) = e^{-\hat{\Lambda}(x; t)}, \quad (6)$$

which is asymptotically equivalent to the Kaplan-Meier estimator. The next test statistic is constructed based on an arcsine-square root transformation on $\hat{S}(x; t)$, which is given by

$$Z_2(x; t) = \frac{n^{1/2} \{\arcsin \sqrt{\hat{S}(x; t)} - \arcsin \sqrt{S_0(x)}\}}{\hat{\nu}(x; t)}, \quad (7)$$

where

$$\hat{\nu}^2(x; t) = \frac{\hat{S}(x; t) \hat{\sigma}^2(x; t)}{4(1 - \hat{S}(x; t))}.$$

The third test statistic discussed in the paper is based on the logit transformation, which is given by

$$Z_3(x; t) = \frac{n^{1/2} \{\text{logit} \hat{S}(x; t) - \text{logit} S_0(x)\} (1 - \hat{S}(x; t))}{\hat{\sigma}(x; t)}. \quad (8)$$

The small-sample properties of these three test statistics $Z_1 - Z_3$ will be studied and compared through simulations in Section 5.

3 Sample Size for Fixed Sample Test

A key step in designing a clinical trial is to calculate the required minimal sample size with constrained type I error and power. If the study is interested in testing hypotheses (1), then

given type I error rate α and power $1 - \beta$ at the alternative $\Lambda_1(x)(< \Lambda_0(x))$ for a given x , the sample sizes for fixed sample tests can be derived as follows. The power of a test is the probability of rejecting the null hypothesis when the alternative is true. Hence, for a fixed sample test, the power of the test statistic $Z_1 = Z_1(x; \tau)$ under the alternative $\Lambda(x) = \Lambda_1(x)$ is determined by

$$\begin{aligned} 1 - \beta &= P(n^{1/2} \hat{\sigma}_1^{-1} \{\log \Lambda_0(x) - \log \hat{\Lambda}(x; \tau)\} \hat{\Lambda}(x; \tau) > z_{1-\alpha} | \Lambda(x) = \Lambda_1(x)) \\ &\simeq \Phi\{n^{1/2} \sigma_1^{-1} (\log \Lambda_0(x) - \log \Lambda_1(x)) \Lambda_1(x) - z_{1-\alpha}\}, \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function, and $\sigma_1^2 = \sigma_1^2(x; \tau)$ is the asymptotic variance of $\hat{\Lambda}(x; \tau)$ for the alternative at the end of study duration $t = \tau$. Thus, the sample size for a fixed sample test for test statistic Z_1 is given by

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma_1^2(x; \tau)}{(\log \Lambda_0(x) - \log \Lambda_1(x))^2 \Lambda_1^2(x)}. \quad (9)$$

Similarly, the sample size for Z_2 is given by

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \nu_1^2(x; \tau)}{\{\arcsin \sqrt{S_1(x)} - \arcsin \sqrt{S_0(x)}\}^2}, \quad (10)$$

where

$$\nu_1^2(x; \tau) = \frac{S_1(x) \sigma_1^2(x; \tau)}{4(1 - S_1(x))},$$

and for Z_3 is given by

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma_1^2(x; \tau)}{\{\logit S_1(x) - \logit S_0(x)\}^2 (1 - S_1(x))^2}. \quad (11)$$

To calculate the sample size n given in above formulas, we have to calculate the asymptotic variance $\sigma_1^2(x; \tau)$ under the alternative hypothesis, which is given by (3) with $t = \tau$.

Assuming subjects were recruited with a uniform distribution over the accrual period t_a and followed for an additional period of length t_f , then the study duration $\tau = t_a + t_f$. We further assume that the survival function of censoring time is $G_1(u)$. Then under the alternative, $\pi(u; t) = P(T_1 \wedge C_1 \wedge (t - Y_1)^+ | u) = P(T_1 > u)P(C_1 > u)(P(Y_1 < t - u) = S_1(u)G_1(u)\min\{(t - u)/t_a, 1\}, t > u$. Thus from equation (3), we have

$$\sigma_1^2(x; t) = \int_0^x \frac{\lambda_1(u)}{S_1(u)G_1(u)\min\{(t-u)/t_a, 1\}} du, \quad x < t \leq \tau, \quad (12)$$

where $\lambda_1(\cdot)$ and $S_1(\cdot)$ are the hazard and survival distribution function at the alternative, respectively.

As a special case, if there is no loss to follow-up, then $G_1(u) = P(C_1 = u) = 1$ and $\min\{(t_a + t_f - u)/t_a, 1\} = 1$ for any $0 < u < x < t_f$; hence equation (12) can be simplified as

$$\sigma_1^2(x; \tau) = \int_0^x \frac{\lambda_1(u)}{S_1(u)} du = S_1^{-1}(x) - 1. \quad (13)$$

This implies that if there is no loss to follow-up and $x < t_f$, then the sample size for a fixed sample test depends only on the single value $S_1(x)$, but not on the specification of the underlying survival distribution $S_1(\cdot)$.

In traditional randomized oncology trials with survival as the endpoint, the study power is determined by the number of events (see, e.g., Collett, 2003). Similarly here, equation (12) showed that the variance $\sigma_1^2(x; \tau)$ depends on the survival probability and the censoring distribution. The more events in $[0, x]$, the smaller is $S(u)$ for $u < x$; the heavier is censoring, the smaller is $G(u)$ for $u < x$; and then the larger is variance $\sigma_1^2(x, \tau)$, which leads to larger sample size n . Therefore, even though the power of the test statistic is expressed explicitly in terms of sample size n , it is actually determined by the number of events. Unfortunately, there is no explicit formula available here to express the power in terms of number of events; however, the number of events is not as convenient as the sample size to be used for the design and operation of a planned study with survival data, in general.

4 Group Sequential Procedure

In this section, we will apply an SCPRT procedure (Xiong, 1995) to the test statistic $Z_1(x; t)$ as a sequential test statistic with time variable t based on its Brownian motion property (Lin et al., 1996; Breslow and Crowley, 1974). The SCPRT has some unique features: (1) the maximum sample size of the sequential test is the same as the sample size of the fixed sample test; (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; (3) the power function of the SCPRT is virtually the same as that of the fixed sample test (Xiong et al., 2007; Xiong, 1995). All these features make the SCPRT more practical and attractive to be used for designing a sequential trial.

Letting $\{B_t : 0 < t \leq 1\}$ be a 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 normal distribution with mean $\mu = (\theta t, \theta)$ and variance matrix $\Sigma = (\sigma_{ij})_{2 \times 2}$ with $\sigma_{11} = \sigma_{12} = \sigma_{21} = t$

and $\sigma_{22} = 1$. Therefore, according to multivariate normal conditional distribution theory (see, e.g., Anderson, 1958), the conditional density $f(B_t|B_1)$ is 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

$$a_k = z_{1-\alpha}t_k - \{2at_k(1-t_k)\}^{1/2}; \quad b_k = z_{1-\alpha}t_k + \{2at_k(1-t_k)\}^{1/2}, \quad (14)$$

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 interim looks and the final look. The a in (14) 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 is a , the smaller is the discordance probability, and the wider apart are the upper and lower boundaries, making it harder for the sample path to reach boundaries and stop early and resulting in larger expected sample sizes. An appropriate a can be determined by choosing an appropriate discordance probability (Xiong, 1995; Xiong et al., 2003). As an illustration, the calculations of the operating characteristics of a multi-stage group sequential design are given in the Appendix.

We now apply the SCPRT to the test statistic $Z_1(x; t)$ in (5). Let

$$U(x; t) = n^{1/2} I(x; t) \{ \log \Lambda_0(x) - \log \hat{\Lambda}(x; t) \} \hat{\Lambda}(x; t), \quad (15)$$

where $I(x; t) = 1/\sigma_1^2(x; t)$. For a fixed x , $U(x; t)$ is asymptotically a Brownian motion in the time scale of the information $I(x; t)$, with mean $E\{U(x; t)\} = \delta I(x; t)$ and covariance $\text{cov}\{U(x; t), U(x; s)\} = I(x; t)$ for any t and s such that $x < t < s < \tau$, where $\delta = n^{1/2}(\log \Lambda_0(x) - \log \Lambda_1(x))\Lambda_1(x)$ (Lin et al., 1996; Breslow and Crowley, 1974). Then we can transform this process into a process with time variable t^* on $[0, 1]$ by letting

$$t^* = I(x; t)/I(x; \tau) = \sigma_1^2(x; \tau)/\sigma_1^2(x; t) \quad (16)$$

be the information time, where $\sigma_1^2(x; \tau)$ is calculated by (3) or (12) or (13), depending on given conditions in the design, and letting

$$B(x; t^*) = U(x; t) / I^{1/2}(x; \tau). \quad (17)$$

Then the sequential test statistic $B(x; t^*) \sim \mathcal{N}(\theta t^*, t^*)$ is approximately a Brownian motion in information time t^* , where the drift parameter $\theta = \delta I^{1/2}(x; \tau) = \delta / \sigma_1^2(x; \tau)$. Based on the SCPRT procedure presented above, the lower and upper boundaries for $B_{t_k^*} = B(x; t_k^*)$ at the k^{th} look are given by

$$a_k = z_{1-\alpha} t_k^* - \{2\alpha t_k^* (1 - t_k^*)\}^{1/2}; \quad b_k = z_{1-\alpha} t_k^* + \{2\alpha t_k^* (1 - t_k^*)\}^{1/2}, \quad (18)$$

for $k = 1, \dots, K$, where $t_k^* = I(x; t_k) / I(x; \tau)$ is the information time at the k^{th} look with calendar time t_k . The nominal critical p -values for testing hypotheses (1) are

$$P_{a_k} = 1 - \Phi(a_k / \sqrt{t_k^*}); \quad P_{b_k} = 1 - \Phi(b_k / \sqrt{t_k^*}). \quad (19)$$

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

$$P_{B_{t_k^*}} = 1 - \Phi(B_{t_k^*} / \sqrt{t_k^*}).$$

The stopping rule for monitoring the trial can be executed by stopping the trial when, for the first time, $P_{B_{t_k^*}} \geq P_{a_k}$ (accept H_0 and stop for futility) or $P_{B_{t_k^*}} \leq P_{b_k}$ (reject H_0 and stop for efficacy). Since $Z_1(x; t_k)$ and $B_{t_k^*} / \sqrt{t_k^*}$ are asymptotically the same standard normal distribution under the null hypothesis, the observed p -value at k -stage can be calculated from the test statistic $Z_1(x; t_k)$ by applying all observations up to stage k . The derived sequential boundaries (18) are also applied to the test statistics Z_2 in (7) and Z_3 in (8) because the two sequential test statistics obviously have the same Brownian process property based on arguments by Lin et al. (1996). The maximum sample sizes (n) for the three sequential test statistics $Z_1(x; t)$, $Z_2(x; t)$, and $Z_3(x; t)$ are same as the sample sizes for corresponding fixed test statistics and can be obtained by equations (9), (10), and (11), respectively. The operating characteristics of a multi-stage group sequential design can be calculated using computation methods illustrated in the Appendix.

For convenience, an SCPRT procedure can be summarized as follows: a) Compute the sample size n for a fixed sample test using the formulas given in Section 2. b) Given the number of looks, K , and the maximum conditional discordance probability ρ , determine the boundary coefficient a (Xiong et al., 2003). c) Given calendar times $t_1, \dots, t_K$, calculate information times using formulas (16) and (3) or (12) or (13), depending on the given

conditions of the design. d) Calculate SCPRT boundaries using equation (18). e) Calculate the nominal critical p -values using formula (19).

5 Simulation Study

To show the small-sample properties of the three test statistics discussed in Section 2, we conducted simulation studies to compare the type I errors and powers for the three test statistics Z_1 , Z_2 , and Z_3 under various scenarios. In simulation studies, the survival distribution was taken as $S(x) = e^{-(\omega x)^\kappa}$, which is the Weibull distribution $W(\omega, \kappa)$ with scale parameter ω and shape parameter κ , where the shape parameter is set to be $\kappa = 0.5, 1$, and 2. The physical meaning of κ for the shape of the survival function by the Weibull distribution can be understood by observing plots of survival functions in Figure 1. The scale parameter ω is determined by the given values of $S_0(2)$ and $S_1(2)$ under null and alternative hypotheses, where $x = 2$ is the clinically meaningful landmark point (e.g., 2 months or 2 years) selected for the simulations. The simulations were performed for a variety of study designs of the fixed sample test and are shown in Table 1. The survival functions under null $S_0(x)$ and alternative $S_1(x)$ of the Weibull distribution with the various settings of simulation parameters in Table 1 are plotted in Figure 1 for illustration and for helping one to choose an appropriate shape parameter κ for sample size calculations in real applications. The null and alternative hypothesis value of survival probabilities $S_0(2)$ and $S_1(2)$ are set up to be in a normal hypothesis range 0.2–0.8 to detect a difference of 0.1–0.2 between $S_0(2)$ and $S_1(2)$.

We assumed that subjects were recruited with a uniform distribution over the accrual period $t_a = 5$ and followed for $t_f = 3$, with the study duration $\tau = t_a + t_f = 8$. By the definition of $X_{\lambda}(t)$ and equation (2), at the interim look time t , a subject was censored if his/her failure time was longer than either $x = 2$, or censoring time C_i or $t - u_i$, where C_i follows an exponential distribution with hazard rate of 0.1 and u_i is an uniform distributed entry time on interval $[0, t_a]$ for i^{th} subject. The empirical type I error and power of the test statistics by simulation were evaluated at the end of the study $\tau = 8$ for the fixed sample test and at $t_1 = 4$ and $t_2 = 8$ for the two-stage group sequential test. The fixed sample sizes for each design were calculated using formula (9), (10), and (11), respectively, for $Z_1(x; t)$, $Z_2(x; t)$, and $Z_3(x; t)$; the $\sigma_1^2(x; \tau)$ in these formulas was calculated by (12) with $t = \tau$. For each parameter configuration of the corresponding design, 100,000 observed samples of censored failure times were generated from the Weibull distribution $W(\omega, \kappa)$ to calculate the test statistics under the null or alternative hypothesis. The nominal type I error and power were set to be 5% and 80%, respectively. The proportions rejecting the null under the true null hypothesis represent the estimated empirical type I error. The proportions rejecting the null under the alternative hypothesis represent the estimated empirical power. The simulated empirical type I errors and powers based on samples (excluding those with $\hat{S}(2; \tau) = 0$ and $\hat{S}(2; \tau) = 1$) are summarized in Tables 1–4 for Z_1 – Z_3 in various scenarios.

From Table 1, we may have following observations. *First*, under each hypothetical setting of the Weibull distribution, being the same for Z_1 – Z_3 , the calculated sample sizes of the three statistics could be quite different, with the largest for Z_1 ; the empirical type I errors by simulation for the three statistics are slightly different but all close to the targeted goal (0.05), with the smallest for Z_1 ; the empirical powers for the three statistics could be quite

different, but all are above the targeted goal (0.80), with the largest for Z_1 . *Second*, for each test statistic (e.g., Z_1) under the same hypothetical setting (e.g., 0.2 vs 0.35), the sample sizes under different values of κ are slightly different, increasing with the largest for $\kappa = 2$; for the same test statistic, the empirical type I errors and powers of this statistic under three values of κ are very close to one another. *Third*, for each test statistic (e.g., Z_1) under the same value of κ setting (e.g., $\kappa = 1$), the sample sizes under different hypothetical settings (e.g., (0.2 vs 0.3) and (0.5 vs 0.65)) could be quite different; the sample size increases as the difference between $S_0(2)$ and $S_1(2)$ gets smaller.

In summary, the simulation results for empirical type I errors and powers showed that the test statistic Z_1 is slightly liberal (e.g., $\alpha = 0.055$) when the null survival probability $S_0(2)$ is set low (e.g., $S_0(2) = 0.2$), and slightly conservative (e.g., $\alpha = 0.045$) when $S_0(2)$ is set high (e.g., $S_0(2) = 0.7$); whereas the empirical type I errors for Z_2 and Z_3 are liberal in most scenarios. The Z_1 requires largest sample size and it is overpowered in all scenarios. Therefore, the sample size calculated based on Z_1 is overestimated. Z_2 and Z_3 require relatively small sample size, but both have inflated type I error.

Sample size calculations and simulations were also done under the log-normal distribution $LN(\mu, \sigma^2)$ using the same design scenarios as discussed above for the Weibull distribution. The sample sizes and simulation results under the log-normal are almost identical to those under the Weibull distribution when $\sigma = 2, 1$, and 0.5 for the log-normal distribution are set corresponding to $\kappa = 0.5, 1$, and 2 for the Weibull distribution. To help us understand this phenomenon, the Weibull distribution $W(\omega, \kappa)$ and log-normal distribution $LN(\mu, \sigma^2)$ are plotted in Figure 2 with scale parameters (ω and μ) equal to survival probability under the null hypothesis (e.g., $S_0(2) = 0.4$) and shape parameters (κ and σ) ranging from 0.5 to 2 .

To study the properties of the sequential tests with Z_1 – Z_3 , simulations were done for a two-stage SCPRT design under the Weibull distribution using the same fixed design scenarios as discussed above (Tables 3–4). The simulated empirical type I error of test Z_1 were slightly liberal when the null hypothesis survival probability was set low, and slightly conservative when the null hypothesis survival probability was set high, in particular at the first stage with a small sample size, whereas the empirical type I error for Z_2 and Z_3 was liberal for most scenarios.

Overall, the test statistic Z_1 has a satisfactory type I error rate and hence is recommended for use in phase II trial design. However, the sample size calculated based on Z_1 is overestimated. A further simulation (Table 2) showed that by reducing sample size by about 10%, the power of the test Z_1 will stay at the nominal 80% level without inflating the type I error. The sample size calculation results also showed that even one misspecified κ value in design stage for the Weibull distribution or a misspecified underlying distribution from the Weibull to the log-normal had little impact on the study power or sample size.

6 An Example

The example is to design a single-arm phase II trial for patients with refractory metastatic colorectal cancer. The primary objective of the study was to evaluate overall survival at 6

months from enrollment (Huang, et al., 2010). The null overall survival probability at 6 months was 0.45 based on historical data. An alternative survival probability of 0.60 at 6 months was selected as a success rate warranting further development of the new treatment. Assume that overall survival time is exponentially distributed (i.e., Weibull with $\kappa = 1$) and during the trial patients could be censored exponentially with a hazard rate of 0.1 in addition to the administrative censoring. Under these assumptions, the sample size for the fixed sample test calculated based on the test Z_1 is 112 patients with a nominal level $\alpha = 0.05$ and power of 80% at the alternative. However, a simulation of 100,000 runs based on sample size 112 showed that the empirical type I error and power are 0.05 and 84.6%, respectively. Therefore, the study is overpowered. A further simulation showed that sample size of 100 for the fixed test would have 81% empirical power and type I error of 0.051. Hence the study with a sample size of 100 maintains the study power and does not inflate the type I error. If an average of 3.7 patients can be accrued per month, then the accrual period is $t_a = 27$ months for 100 patients. Assuming a follow-up period $t_f = 6$ months, then the study duration is $\tau = 33$ months. The investigators are planning a three-stage group sequential design with two interim looks and a final look planned at calendar times $t_1 = 18$, $t_2 = 24$, and $t_3 = 33$ months. Then the corresponding variances at interim look t_k can be calculated by

$$\sigma_1^2(6; t_k) = \int_0^6 \frac{\kappa \omega^\kappa u^{\kappa-1}}{e^{-(\omega u)^\kappa} e^{-0.1u} \min\{(t_k - u)/t_a, 1\}} du, \quad 6 < t_k, \quad k=1, 2, 3,$$

where $\kappa = 1$ and ω is determined by the alternative $S_1(6) = 0.6$. By numerical integration, $[\sigma_1^2(6; 18), \sigma_1^2(6; 24), \sigma_1^2(6; 33)] = (2.105, 1.404, 0.937)$ and the information time $t^* = (0.445, 0.667, 1)$. Assuming a maximum conditional probability of discordance $\rho = 0.02$ (recommended by Xiong et al., 2003), we have $a = 2.604$ for $K = 3$ (Xiong et al., 2003). The lower boundary and upper boundary calculated by (6) are $(a_1, a_2, a_3) = (-0.402, 0.0216, 1.645)$ and $(b_1, b_2, b_3) = (1.866, 2.173, 1.645)$, respectively. The acceptance and rejection nominal critical significance levels are $(0.727, 0.489, 0.05)$ and $(0.0026, 0.0039, 0.05)$, respectively. Therefore, at the first interim look, if the observed p -value < 0.0026 , then reject the null (declaim or stop for efficacy), or if the observed p -value < 0.727 , then accept the null (stop for futility); otherwise, the trial continues to the second stage. At the second interim look, if the observed p -value < 0.0039 , then reject the null (declaim or stop for efficacy), or if the observed p -value < 0.489 , then accept the null (stop for futility); otherwise, the trial continues to the final stage. We performed 100,000 simulation runs under the exponential model to evaluate the operating characteristics of the proposed group sequential design. The empirical (nominal) type I error and power of the three-stage sequential test were 0.0516 (0.0507) and 0.8079 (0.799), respectively. The empirical (nominal) probabilities of stopping under null and alternative hypothesis were 0.247 (0.276) and 0.124 (0.139) at the first look; 0.251 (0.259) and 0.169 (0.171) at the second look, respectively. The details of the operating characteristics for the proposed group sequential design are shown in Table 4.

7 Conclusion

A single-arm phase II group sequential design is proposed for monitoring the survival probability at a fixed time point. A practical issue to design such a study is to select the time point at which the survival probability to be tested. First, because the trial is monitored to stop the futility and/or efficacy, therefore, the fixed time point can be chosen such that the survival probability at fixed time point should not be too high or too low so that there is a room to detect a specified effect size of the study design. Second, there is often a landmark time point for the survival probability of the historical data in a particular disease research area. Therefore, the survival probability at such a landmark time point can be chosen for the trial design. Third, the landmark time point could be obtained after careful deliberation as a compromising of accrual rate, urgency of investigation, expectation of improvement, and prevention of worst situation.

Three non-parametric test statistics for testing survival probability at a fixed time point are discussed in this paper. The sample size formulas for fixed sample tests are also derived. These test statistics are non-parametric and applicable to test any survival function; however, the sample size calculation may depend on the underlying failure time distribution, censoring distribution, and accrual distribution, as shown by equation (12). Simulation results showed that the underlying failure time distribution has little impact on the study power or sample size. However, the censoring distribution and accrual distribution could have large impacts on the study power. In this paper, we have studied the property of three test statistics under the assumption of an exponential censoring distribution and an uniform accrual distribution. It would be worth exploring the property of the test statistics under other censoring distribution and accrual distribution assumptions.

The common basis of the three test statistics is the cumulative hazard function $\Lambda(x; t)$, and its estimate $\hat{\Lambda}(x; t)$ in equation (2) shows that any of the test statistics does include subjects who have not been followed for the required time x . This implies that these test statistics should be more efficient than the naive approach of estimating the survival probability as a simple proportion of all subjects surviving the required time at the interim analyses.

Among the three test statistics, test statistic Z_1 is recommended. The sample size calculated based on Z_1 is robust against the underlying distribution assumption; however, it is overestimated. Roughly a 10% sample size reduction could be appropriate to keep the study power at the design level but we recommend that a simulation study should be done to reduce the sample size while keeping the type I error and power at the nominal levels. The trial can be monitored by a multi-stage group sequential design based on the SCPRT procedure. A practically useful and unique feature of SCPRT is that the investigators may have freedom to ignore any predetermined interim stopping during the study without harming the integrity of study because the skipping does not change the overall significance level and power of the testing. This feature is useful when a study has reached significance or futility at an interim look but investigators wish to continue the study to the planned end for some reasons, e.g., either wanting to collect more data for later analysis or wanting to convince others (or even themselves) that the interim result can be trusted. For this situation, one may continue the study and take the result at the final end; they should ignore the

interim result if there is contradiction between the two results. For an SCPRT design, the probability of this contradiction is very small, e.g., its maximum is $\rho_{max} = 0.0054$ if the conditional discordant probability $\rho = 0.02$ is used for the design. Therefore, the proposed method can be easily used to design a single-arm phase II group sequential trial for monitoring survival probability.

Acknowledgments

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

References

- Anderson, TW. An introduction to multivariate statistical analysis. New York: Wiley; 1958.
- Breslow N, Crowley J. A large sample study of the life table and product limit estimates under random censorship. The Annals of Statistics. 1974; 2:437–453.
- Case LD, Morgan TM. Design of phase II cancer trials evaluating survival probabilities. BMC Medical Research Methodology. 2003; 3:1–12. [PubMed: 12515580]
- Collett, D. Modeling survival data in medical research. 2. London: Chapman and Hall; 2003.
- Fleming, TR., Harrington, DP. Counting processes and survival analysis. New York: John Wiley and Sons; 1991.
- Huang B, Talukder E, Thomas N. Optimal two-stage phase II designs with long-term endpoints. Statistics in Biopharmaceutical Research. 2010; 2:51–61.
- Jennison, C., Turnbull, WD. Group sequential methods with applications to clinical trials. New York: Chapman & Hall; 2000.
- Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. Statistics in Medicine. 2007; 26:4505–4519. [PubMed: 17348080]
- Lin DY, Shen L, Ying Z, Breslow NE. Group sequential designs for monitoring survival probabilities. Biometrics. 1996; 52:1033–1042. [PubMed: 8805766]
- Tan M, Xiong X. A flexible multi-stage design for phase II oncology trials. Pharmaceutical Statistics. 2010; 10:369–373. [PubMed: 22328328]
- Xiong X. A class of sequential conditional probability ratio tests. Journal of American Statistical Association. 1995; 15:1463–1473.
- Xiong X, Tan M. Sufficiency-Based methods for evaluating sequential tests and estimations. ASA Proceedings, Biometrics Section. 1999:268–273.
- Xiong X, Tan M. Evaluating sequential tests for a class of stochastic processes. Computing Science and Statistics. 2001; 33:30–34.
- 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]
- Xiong X, Tan M, Boyett J. SCPRT: a sequential procedure that gives another reason to stop clinical trials early. Advances in Statistical Methods for the Health Sciences Statistics for Industry and Technology. 2007:419–434.
- Xiong X, Tan M, Kutner MH. Computation methods for evaluating sequential tests and post-estimations via sufficiency principle. Statistica Sinica. 2002; 12:1027–41.

Appendix

Computation for the sequential test normalized with information time

Assumption

Let $B(t) \sim N(\theta t, t)$ 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 $l_{t_k}^*(\cdot)$

Define $l_{t_k}^*(s)$ 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) ,

$$l_{t_k}^*(s) = \frac{1}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}} \int_{a_{k-1}}^{b_{k-1}} l_{t_{k-1}}^*(x) \phi\left(\frac{x - st_{k-1}/t_k}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}}\right) dx, \quad (20)$$

where $\phi(\cdot)$ in (20) 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 (20) 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

$$l_{t_k}(s) = \frac{1}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}} \int_{a_{k-1}}^{b_{k-1}} l_{t_{k-1}}^*(x) \phi\left(\frac{x - st_{k-1}/t_k}{\sqrt{t_{k-1}(1 - t_{k-1}/t_k)}}\right) dx, \quad (21)$$

where $\phi(\cdot)$ in (21) 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 (21), the power function $P(\theta)$ or the probability to reject H_0 under true mean θ is

$$P(\theta) = \sum_{k=1}^K (1/\sqrt{t_k}) \int_{b_k}^{\infty} l_{t_k}(x) \phi((x - \theta t_k)/\sqrt{t_k}) dx. \quad (22)$$

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

$$P_{t_k}(\theta) = (1/\sqrt{t_k}) \left(\int_{-\infty}^{a_k} + \int_{b_k}^{\infty} \right) l_{t_k}(x) \phi((x - \theta t_k)/\sqrt{t_k}) dx, \quad (23)$$

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 $P_{t_k}(\theta)$ by (23), the expected stopping $ET(\theta)$ is a function of θ as

$$ET(\theta) = \sum_{k=1}^K t_k P_{t_k}(\theta), \quad (24)$$

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 as

$$EN(\theta) = n \times ET(\theta), \quad (25)$$

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

For SCPRT design

To test $H_0: \theta = 0$ vs. $H_a: \theta > 0$ with significance level α and power $1 - \beta$ by an SCPRT design, the cutoff value at 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 (22), (24), (23), and (25), we can compute type I error, power, probability of stopping at given t_k , and expected sample sizes under the null and alternative hypotheses for the SCPRT design.

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

Figure 1. Survival functions for the Weibull distribution under the null and alternative hypotheses for different design scenarios.

Figure 2. Survival functions for the Weibull and log-normal distributions under the null hypothesis for different shape parameters.

Sample size and simulated empirical type I error (α) and power ($1 - \beta$) based on 100,000 simulation runs for Weibull and log-normal distributions with nominal level 0.05 and power of 80% for fixed test statistics $Z_1 - Z_3$

Table 1

Dist.	Design	0.2 vs 0.35			0.2 vs 0.4			0.3 vs 0.45		
$W(\omega, \kappa)$	κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α
0.5		Z_1	63	.055	.837	38	.056	.849	77	.051
		Z_2	59	.058	.825	35	.060	.834	70	.056
		Z_3	51	.069	.803	30	.078	.817	65	.062
1		Z_1	65	.054	.837	40	.055	.854	79	.051
		Z_2	61	.058	.825	36	.061	.833	72	.057
		Z_3	52	.071	.802	31	.077	.818	68	.063
2		Z_1	67	.056	.838	41	.055	.854	82	.051
		Z_2	63	.057	.823	37	.059	.831	74	.057
		Z_3	54	.068	.802	31	.077	.811	70	.062
$LN(\mu, \sigma^2)$	σ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α
2		Z_1	63	.056	.837	39	.056	.855	77	.051
		Z_2	59	.056	.821	35	.059	.832	70	.058
		Z_3	51	.071	.804	30	.078	.816	66	.063
1		Z_1	65	.053	.834	40	.055	.854	80	.051
		Z_2	62	.056	.827	36	.060	.831	73	.057
		Z_3	53	.071	.806	31	.076	.815	68	.062
0.5		Z_1	67	.054	.836	41	.054	.852	82	.052
		Z_2	63	.057	.823	37	.059	.828	75	.056
		Z_3	54	.071	.798	32	.076	.817	70	.062
Dist.	Design	0.5 vs 0.65			0.6 vs 0.75			0.7 vs 0.8		

β	$1 - \beta$	α	u	$\beta - 1$	α	u	$\beta - 1$	α	u	$\beta - 1$	α	u	Test	κ	$W(\omega, \kappa)$
.857	.150	.045	.152	.869	.048	.82	.048	.856	.86	.048	.856	.82	Z_1	0.5	
.816	.181	.058	.124	.825	.060	.65	.060	.817	.72	.058	.817	.65	Z_2		
.847	.150	.043	.143	.852	.055	.74	.055	.833	.77	.055	.833	.74	Z_3		
.857	.146	.046	.157	.869	.049	.85	.049	.838	.89	.049	.838	.85	Z_1	1	
.815	.181	.059	.128	.825	.058	.67	.058	.820	.75	.058	.820	.67	Z_2		
.847	.150	.050	.148	.852	.055	.77	.055	.835	.80	.055	.835	.77	Z_3		
.856	.146	.046	.162	.866	.048	.87	.048	.854	.92	.048	.854	.87	Z_1	2	
.815	.151	.058	.132	.823	.058	.69	.058	.819	.77	.057	.819	.69	Z_2		
.847	.151	.051	.153	.851	.055	.79	.055	.833	.82	.055	.833	.79	Z_3		
.858	.147	.047	.156	.872	.047	.84	.047	.856	.88	.048	.856	.84	Z_1	2	
.815	.158	.058	.127	.822	.059	.66	.059	.822	.74	.058	.822	.66	Z_2		
.847	.151	.051	.147	.850	.051	.76	.051	.830	.78	.055	.830	.76	Z_3		
.855	.147	.047	.162	.869	.047	.87	.047	.855	.93	.048	.855	.87	Z_1	1	
.815	.158	.058	.132	.825	.061	.68	.061	.818	.76	.057	.818	.68	Z_2		
.847	.150	.050	.152	.852	.052	.79	.052	.834	.81	.055	.834	.79	Z_3		
.857	.147	.047	.166	.869	.046	.89	.046	.852	.93	.048	.852	.89	Z_1	0.5	
.813	.157	.057	.135	.825	.059	.70	.059	.818	.78	.057	.818	.70	Z_2		
.847	.150	.050	.157	.852	.053	.81	.053	.832	.83	.055	.832	.81	Z_3		

Simulated empirical type I error (α) and power ($1 - \beta$) based on 100,000 simulation runs for the Weibull and log-normal distributions with the reduced sample size, adjusting the empirical power of Z_1 to nearly 80%, for fixed tests with $Z_1 - Z_3$

Table 2

Dist.	Design	0.2 vs 0.35			0.2 vs 0.4			0.3 vs 0.45		
$W(\omega, \kappa)$	κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α
0.5		Z_1	56	.055	.800	33	.057	.806	70	.051
		Z_2	56	.057	.807	33	.061	.815	70	.056
		Z_3	56	.070	.830	33	.073	.841	70	.062
1		Z_1	58	.054	.803	34	.056	.802	72	.052
		Z_2	58	.059	.810	34	.060	.814	72	.057
		Z_3	58	.072	.831	34	.075	.843	72	.063
2		Z_1	60	.055	.802	35	.055	.806	74	.052
		Z_2	60	.058	.810	35	.059	.814	74	.057
		Z_3	60	.069	.833	35	.074	.842	74	.062
$LN(\mu, \sigma^2)$	κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α
2		Z_1	56	.055	.801	33	.055	.804	70	.052
		Z_2	56	.057	.807	33	.060	.815	70	.058
		Z_3	56	.070	.830	33	.074	.838	70	.063
1		Z_1	58	.054	.799	34	.056	.802	72	.051
		Z_2	58	.058	.810	34	.059	.812	72	.057
		Z_3	58	.070	.831	34	.075	.841	72	.062
0.5		Z_1	60	.055	.802	35	.057	.804	74	.051
		Z_2	60	.056	.809	35	.058	.811	74	.057
		Z_3	60	.070	.831	35	.073	.839	74	.062
Dist.	Design	0.5 vs 0.65			0.6 vs 0.75			0.7 vs 0.8		

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$W(\omega, \kappa)$	κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
0.5		Z_1	75	.048	.805	68	.046	.806	132	.047	.807	132	.047	.807
		Z_2	75	.058	.830	68	.059	.839	132	.058	.837	132	.058	.837
		Z_3	75	.054	.826	68	.052	.821	132	.050	.819	132	.050	.819
1		Z_1	77	.048	.805	70	.047	.804	136	.046	.801	136	.046	.801
		Z_2	77	.057	.830	70	.059	.837	136	.058	.833	136	.058	.833
		Z_3	77	.055	.822	70	.052	.821	136	.050	.811	136	.050	.811
2		Z_1	80	.048	.808	72	.046	.801	139	.046	.801	139	.046	.801
		Z_2	80	.058	.830	72	.058	.835	139	.058	.833	139	.058	.833
		Z_3	80	.055	.826	72	.053	.820	139	.049	.813	139	.049	.813
$LN(\mu, \sigma^2)$	κ	Test	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
2		Z_1	75	.048	.801	68	.046	.796	132	.047	.797	132	.047	.797
		Z_2	75	.058	.826	68	.062	.832	132	.058	.826	132	.058	.826
		Z_3	75	.056	.821	68	.053	.814	132	.050	.808	132	.050	.808
1		Z_1	77	.048	.799	70	.045	.792	136	.047	.795	136	.047	.795
		Z_2	77	.058	.823	70	.059	.830	136	.058	.826	136	.058	.826
		Z_3	77	.055	.814	70	.052	.813	136	.050	.805	136	.050	.805
0.5		Z_1	80	.048	.803	72	.046	.792	139	.046	.792	139	.046	.792
		Z_2	80	.058	.827	72	.060	.828	139	.057	.825	139	.057	.825
		Z_3	80	.055	.819	72	.053	.812	139	.051	.807	139	.051	.807

Table 3

Simulated empirical type I error and power of the two-stage SCPRT design based on 100,000 simulation runs for Weibull distribution $W(\omega, \kappa)$ with nominal level 0.05 and power of 80% for fixed test statistics $Z_1 - Z_3$

Design	κ	Test	At k^{th} interim look	Type I error			Power		
				$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total
0.2 vs 0.35	0.5	Z_1	Empirical	.0058	.0511	.0569	.2883	.5507	.8384
		Z_2	Empirical	.0082	.0512	.0594	.3118	.5112	.8230
		Z_3	Empirical	.0149	.0603	.0752	.3423	.4659	.8082
			Nominal	.0046	.0462	.0507	.2091	.5901	.7992
	1	Z_1	Empirical	.0050	.0506	.0557	.2434	.5957	.8391
		Z_2	Empirical	.0082	.0518	.0599	.2720	.5542	.8262
		Z_3	Empirical	.0137	.0625	.0762	.2988	.5066	.8054
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987
	2	Z_1	Empirical	.0046	.0520	.0566	.2060	.6307	.8367
		Z_2	Empirical	.0067	.0532	.0599	.2402	.5840	.8242
		Z_3	Empirical	.0133	.0622	.0755	.2706	.5376	.8082
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987
0.2 vs 0.4	0.5	Z_1	Empirical	.0060	.0519	.0579	.2935	.5556	.8495
		Z_2	Empirical	.0103	.0541	.0643	.3396	.4955	.8352
		Z_3	Empirical	.0181	.0653	.0834	.3675	.4549	.8224
			Nominal	.0046	.0461	.0507	.2120	.5871	.7991
	1	Z_1	Empirical	.0049	.0524	.0573	.2473	.6084	.8557
		Z_2	Empirical	.0105	.0542	.0647	.2960	.5375	.8334
		Z_3	Empirical	.0171	.0668	.0839	.3263	.4961	.8224
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987

		Type I error					Power		
Design	κ	Test	At k^{th} interim look	$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total
	2	Z_1	Empirical	.0042	.0530	.0573	.2037	.6507	.8543
		Z_2	Empirical	.0092	.0542	.0634	.2621	.5704	.8325
		Z_3	Empirical	.0149	.0678	.0827	.2838	.5286	.8124
		Nominal		.0043	.0469	.0512	.1454	.6534	.7987
0.3 vs 0.45	0.5	Z_1	Empirical	.0052	.0483	.0535	.2712	.5698	.8411
		Z_2	Empirical	.0080	.0502	.0582	.3073	.5173	.8246
		Z_3	Empirical	.0100	.0547	.0647	.3112	.5046	.8158
		Nominal		.0046	.0461	.0507	.2145	.5848	.7992
	1	Z_1	Empirical	.0044	.0490	.0534	.2257	.6137	.8394
		Z_2	Empirical	.0078	.0508	.0586	.2674	.5565	.8239
		Z_3	Empirical	.0096	.0563	.0659	.2744	.5448	.8192
		Nominal		.0043	.0469	.0512	.1454	.6534	.7987
	2	Z_1	Empirical	.0034	.0492	.0525	.1904	.6504	.8408
		Z_2	Empirical	.0073	.0523	.0596	.2347	.5880	.8227
		Z_3	Empirical	.0083	.0574	.0656	.2412	.5757	.8169
		Nominal		.0043	.0469	.0512	.1454	.6534	.7987
0.5 vs 0.65	0.5	Z_1	Empirical	.0035	.0446	.0481	.2538	.6013	.8551
		Z_2	Empirical	.0085	.0512	.0605	.3110	.5102	.8211
		Z_3	Empirical	.0060	.0504	.0564	.2780	.5566	.8366
		Nominal		.0047	.0460	.0507	.2217	.5756	.7993
	1	Z_1	Empirical	.0032	.0464	.0495	.2008	.6536	.8544
		Z_2	Empirical	.0084	.0530	.0614	.2690	.5571	.8261
		Z_3	Empirical	.0053	.0523	.0576	.2334	.6031	.8365

		Type I error				Power			
Design	κ	Test	At k^{th} interim look	$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total
0.6 vs 0.75	2		Nominal	.0043	.0469	.0512	.1454	.6534	.7987
		Z_1	Empirical	.0025	.0451	.0477	.1633	.6910	.8543
		Z_2	Empirical	.0080	.0529	.0608	.2359	.5868	.8227
		Z_3	Empirical	.0052	.0516	.0568	.1980	.6349	.8329
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987
	0.5	Z_1	Empirical	.0030	.0451	.0482	.2403	.6283	.8686
		Z_2	Empirical	.0100	.0524	.0623	.3239	.5040	.8280
		Z_3	Empirical	.0049	.0488	.0536	.2645	.5879	.8524
	1		Nominal	.0047	.0459	.0507	.2275	.5718	.7993
		Z_1	Empirical	.0026	.0444	.0470	.1818	.6883	.8701
		Z_2	Empirical	.0100	.0532	.0632	.2789	.5477	.8266
		Z_3	Empirical	.0038	.0495	.0533	.2069	.6452	.8521
0.7 vs 0.8	2		Nominal	.0043	.0469	.0512	.1454	.6534	.7987
		Z_1	Empirical	.0017	.0461	.0478	.1397	.7285	.8682
		Z_2	Empirical	.0097	.0530	.0627	.2477	.5783	.8260
		Z_3	Empirical	.0037	.0498	.0535	.1671	.6833	.8504
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987
	0.5	Z_1	Empirical	.0033	.0443	.0477	.2396	.6174	.8570
		Z_2	Empirical	.0087	.0509	.0596	.3054	.5128	.8181
		Z_3	Empirical	.0042	.0474	.0516	.2531	.5931	.8462
	1		Nominal	.0047	.0459	.0506	.2288	.5705	.7993
		Z_1	Empirical	.0024	.0457	.0480	.1857	.6702	.8559
		Z_2	Empirical	.0079	.0520	.0600	.2609	.5571	.8180

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

		Type I error				Power				
Design	κ	Test	At k^{th} interim look	$k = 1$	$k = 2$	total	$k = 1$	$k = 2$	total	
		Z_3	Empirical	.0032	.0470	.0502	.2006	.6461	.8466	
			Nominal	.0043	.0469	.0512	.1454	.6534	.7987	
		2	Z_1	Empirical	.0021	.0449	.0470	.1456	.7083	.8539
			Z_2	Empirical	.0085	.0534	.0619	.2311	.5881	.8191
		Z_3	Empirical	.0032	.0480	.0511	.1676	.6782	.8458	
		Nominal	.0043	.0469	.0512	.1454	.6534	.7987		

Table 4

Operating characteristics of the three-stage SCPRT design based on 100,000 simulation runs under Weibull distribution $W(\omega, \kappa)$ with nominal level 0.05 and power of 80% of the test Z_1 for the example in Section 6

At k^{th} interim look	$k = 1$	$k = 2$	$k = 3$	total
Type I error				
Nominal	0.0026	0.0028	0.0452	0.0507
Empirical	0.0019	0.0025	0.0472	0.0516
Power				
Nominal	0.1274	0.1554	0.5165	0.7993
Empirical	0.1174	0.1608	0.5297	0.8079
Probability of stopping under null				
Nominal	0.2759	0.2585	0.4658	1.0000
Empirical	0.2472	0.2510	0.5017	1.0000
Probability of stopping under alternative				
Nominal	0.1393	0.1712	0.6897	1.0000
Empirical	0.1244	0.1689	0.7067	1.0000
Expected stopping time and sample size	$ET(0)$	$ET(\theta_a)^*$	$EN(0)$	$EN(\theta_a)$
Nominal	0.7608	0.8659	77	87
Empirical	0.7791	0.8747	78	88

* : $\theta_a = z_1 - \alpha + z_1 - \beta$ is the drift parameter at the alternative hypothesis.