Published in final edited form as:

J Biopharm Stat. 2016 ; 26(4): 644–656. doi:10.1080/10543406.2015.1052494.

Single-Arm Phase II Cancer Survival Trial Designs

Jianrong Wu

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

Abstract

The current practice for designing single-arm phase II trials with time-to-event endpoints is limited to using either a maximum likelihood estimate test under the exponential model or a naive approach based on dichotomizing the event time at a landmark time point. A trial designed under the exponential model may not be reliable, and the naive approach is inefficient. The modified one-sample log-rank test statistic proposed in this paper fills the void. In general, the proposed test can be used to design single-arm phase II survival trials under any parametric survival distribution. Simulation results showed that it preserves type I error well and provides adequate power for phase II cancer trial designs with time-to-event endpoints.

Keywords

Clinical trial design; One-sample log-rank test; Time-to-event; Sample size; Single-arm phase II; Survival analysis

Introduction

The single-arm phase II clinical trial has been frequently used in oncology to determine whether new treatment agents have sufficient antitumor activity to warrant further investigation in randomized phase III trials. Antitumor activity is often quantified by tumor response to an upfront therapy; however, tumor response may not be associated with improved survival. Furthermore, the emergence of molecularly targeted therapy offers promising novel agents; the activity of which may not be appropriately evaluated by tumor response. Therefore, a time-to-event endpoint such as progression-free survival (PFS) or event-free survival (EFS) is also often evaluated in phase II oncology trials. There are many methods available for designing single-arm phase II trials with tumor response as a primary endpoint. The scope of this paper is limited to single-arm phase II trial designs with time-to-event endpoints only.

Various statistical methods have been developed for designing randomized phase III trials with time-to-event endpoints (George and Desu, 1977; Lachin, 1981; Rubenstein, 1981; Schoenfeld, 1983; and Lakatos, 1986). However, the literature on designing single-arm phase II trials with time-to-event endpoints is relatively scarce. The current practice for designing such trials is limited to using either a parametric maximum likelihood estimate (MLE) test under the exponential model or a naive approach based on dichotomizing the event time at a landmark time point. A trial design under the exponential model may not be reliable, and the naive approach is inefficient. In a discussion paper by Owzar and Jung

(2008), the naive method was compared to a parametric MLE test under the exponential model and a nonparametric test based on the Nelson-Aalen estimator of the cumulative hazard function. The authors concluded that the naive method should be considered for the statistical design and decision rule for phase II trials with time-to-event endpoints. However, the naive approach has several drawbacks. First, subjects who experience events after the landmark time point will not be included, thereby resulting in an inefficient estimator and study design. Second, simply treating subjects who are lost to follow-up as treatment failures or excluding them from the analysis will result in a biased estimator. Finally, suspending accrual until all of the subjects have been followed for the required time for an interim analysis is often impractical. Long trial suspensions can ruin a trial's momentum and increase the duration and costs of the study (Case and Morgan, 2003).

Recently, several methods have been proposed for designing and monitoring single-arm phase II trials with survival endpoints at fixed time points (Case and Morgan, 2003; Lin et al., 1999; and Huang et al., 2010). There are two major disadvantages to this approach. First, the nonparametric test does not preserve the type I error well when sample size is small, which is typically the case in phase II trials. Second, events that occurred after the fixed landmark time point are not used in the test. Thus, the test may result in an inefficient study design. The STPlan software version 4.5 (Brown et al., 2010) and the web-based software offered by SWOG provide sample size and power calculation for single-arm phase II trial designs, but both are limited to the exponential model only. Kwak and Jung (2013) proposed a phase II survival trial design using the one-sample log-rank test (OSLRT) (Breslow, 1975; and Woolson, 1981). However, the OSLRT is conservative and does not preserve the type I error well, and the proposed two-stage procedure is restricted to the exponential distribution only. The Weibull, gamma, log-normal, log-logistic, Gompertz, and other survival distributions are also often seen in oncology studies, but no method is available for designing single-arm phase II trials based on these survival distributions. Therefore, it is necessary to develop methods that can incorporate these survival distributions. In this article, a modified one-sample log-rank test (MOSLRT) is proposed, and a sample size formula is derived for designing single-arm phase II trials under these distributions. The efficacy of the proposed test is compared to a naive method, a nonparametric test proposed by Lin et al. (1999), a parametric MLE test, and the OSLRT under the exponential distribution.

The rest of the paper is organized as follows. The OSLRT and MOSLRT are introduced in Section 2. In Section 3, sample size formulae of the two tests are derived. Comparisons of the proposed test to the other methods under the exponential model are carried out in Section 4. Simulations are conducted to study the performance of the proposed test in Section 5. An actual example is given in Section 6 to illustrate trial design using the proposed methods. Concluding remarks are made in Section 7.

2 Test Statistics

Let $S_0(t)$ denote the survival distribution of a historical control that is chosen for a single-arm phase II trial design, and let $S(t)$ denote the survival distribution of an experimental therapy. Then improvement in the survival distribution of an experimental therapy compared to that of the historical control can be tested by the following one-sided hypothesis:

$$H_0: S(t) \leq S_0(t) \text{ vs. } H_a: S(t) > S_0(t), \text{ for all } t > 0. \quad (1)$$

Suppose that during the accrual phase of the trial n subjects are enrolled. For the i^{th} subject, let T_i denote the failure time and let C_i denote censoring time. For convenience, the words "failure" and "event" are used interchangeably in this paper. We assume that the failure time T_i and censoring time C_i are independent and that $\{T_i, C_i, i = 1, \dots, n\}$ are independent and identically distributed. Then the observed failure time and failure indicator are $X_i = T_i \wedge C_i$ and $\Delta_i = I(T_i \leq C_i)$, respectively, for the i^{th} subject. On the basis of the observed data $\{X_i, \Delta_i, i = 1, \dots, n\}$, we define $O = \sum_{i=1}^n \Delta_i$ as the observed number of failures, and $E = \sum_{i=1}^n \Lambda_0(X_i)$ as the expected number of failures (asymptotically), where $\Lambda_0(t) = -\log S_0(t)$ is the cumulative hazard function under the null hypothesis. The one-sample log-rank test is then defined by

$$L_1 = \frac{O - E}{\sqrt{E}}.$$

To study the asymptotic distribution of the OSLRT, we formulate it using counting-process notations (Fleming and Harrington, 1991). Specifically, let $N_i(t) = I\{X_i \leq t\}$ and $Y_i(t) = I\{X_i > t\}$ be the failure and at-risk processes, respectively, then

$$O = \sum_{i=1}^n \int_0^\infty dN_i(t), \quad \text{and} \quad E = \sum_{i=1}^n \int_0^\infty Y_i(t) d\Lambda_0(t).$$

Thus, the counting-process formulation of the OSLRT is given by

$$L_1 = W / \hat{v},$$

where

$$W = n^{-1/2} \sum_{i=1}^n \int_0^\infty \{dN_i(t) - Y_i(t) d\Lambda_0(t)\},$$

and

$$\hat{v}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty Y_i(t) d\Lambda_0(t).$$

Under the null hypothesis H_0 , $n^{-1} \sum_{i=1}^n Y_i(t) \rightarrow G(t) S_0(t)$, where $G(t)$ is the survival distribution of censoring time C . Thus, $\hat{v}^2$ converges to $\sigma_0^2 = \int_0^\infty G(t) S_0(t) d\Lambda_0(t)$, which is the exact variance of W under the null hypothesis. As shown in the Appendix, the exact

mean of W under the null is $E_{H_0}(W) = 0$. Therefore, by the counting-process central limit theorem (Fleming and Harrington, 1991), under the null hypothesis, L_1 is asymptotically standard normal distributed. Hence, we reject the null hypothesis H_0 with one-sided type I error α if $L_1 = W/\hat{\sigma} < -z_{1-\alpha}$, where $z_{1-\alpha}$ is the $100(1 - \alpha)$ percentile of the standard normal distribution. Simulation results showed, however, that the OSLRT L_1 is conservative, even when the sample size is relatively large (Kwak and Jung, 2013; Sun et al., 2011; and Wu, 2015).

As $n^{-1}E \rightarrow E_{H_0}(\Lambda_0(X))$ and $n^{-1}O \rightarrow E_{H_0}(\cdot)$, and $E_{H_0}(\Lambda_0(X)) = E_{H_0}(\cdot) = \text{Var}_{H_0}(W)$, as shown in the Appendix, to correct the conservativeness of the OSLRT L_1 , we propose the MOSLRT, which is defined as

$$L_2 = \frac{O - E}{\sqrt{(O + E)/2}}.$$

The counting-process formulation of the MOSLRT is given by $L_2 = W/\hat{\sigma}$, where

$$W = n^{-1/2} \sum_{i=1}^n \int_0^\infty \{dN_i(t) - Y_i(t)d\Lambda_0(t)\},$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty \{dN_i(t) + Y_i(t)d\Lambda_0(t)\}/2.$$

Under the null hypothesis, $n^{-1}O \rightarrow E_{H_0}(\cdot)$ and

$$n^{-1} \sum_{i=1}^n \int_0^\infty Y_i(t)d\Lambda_0(t) \rightarrow \int_0^\infty G(t)S_0(t)d\Lambda_0(t) = E_{H_0}(\Lambda_0(X)).$$

Thus, from equation (A1) in the Appendix, $\hat{\sigma}^2 \rightarrow \sigma_0^2 = \int_0^\infty G(t)S_0(t)d\Lambda_0(t)$. Therefore, under the null hypothesis, L_2 is asymptotically standard normal distributed. Hence, we reject the null hypothesis H_0 with one-sided type I error α if $L_2 = W/\hat{\sigma} < -z_{1-\alpha}$.

3 Sample Size Calculation

To design the study, we must calculate the sample size to detect a specified survival difference at the alternative $H_1: S(t) = S_1(t)(> S_0(t))$, given the type I error α and power of $1 - \beta$. The exact variance of W has been derived for the sample size calculation (Wu, 2015).

Let the exact mean and variance of W at the alternative be $E_{H_1}(W) = \sqrt{n}\omega$ and $\text{Var}_{H_1}(W) = \sigma^2$, respectively, where ω and σ^2 are given below. By the central limit theorem, $(W - \sqrt{n}\omega)/\sigma$ is approximately standard normal distributed under H_1 . Under the alternative hypothesis,

$$\hat{v}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty Y_i(t)d\Lambda_0(t) \rightarrow v_0 = \int_0^\infty G(t)S_1(t)d\Lambda_0(t).$$

Thus, the power of the OSLRT $L_1 = W/\hat{\sigma}$ should satisfy the following equations:

$$1 - \beta = P(L_1 < -z_{1-\alpha} | H_1) \simeq \Phi \left(-\frac{\sqrt{v_0}}{\sigma} z_{1-\alpha} - \frac{\sqrt{n}\omega}{\sigma} \right).$$

Therefore, the required sample size for the test statistic L_1 is given by

$$n = \frac{(\sqrt{v_0} z_{1-\alpha} + \sigma z_{1-\beta})^2}{\omega^2}, \quad (2)$$

where $\omega = v_1 - v_0$ and $\sigma^2 = v_1 - v_1^2 + 2v_{00} - v_0^2 - 2v_{01} + 2v_0 v_1$, with v_0 , v_1 , v_{00} and v_{01} being given by the following equations (the derivation is given by Wu (2015) and also see equations (A3) to (A6) in the Appendix):

$$v_0 = \int_0^\infty G(t) S_1(t) d\Lambda_0(t),$$

$$v_1 = \int_0^\infty G(t) S_1(t) d\Lambda_1(t),$$

$$v_{00} = \int_0^\infty G(t) S_1(t) \Lambda_0(t) d\Lambda_0(t),$$

$$v_{01} = \int_0^\infty G(t) S_1(t) \Lambda_0(t) d\Lambda_1(t).$$

Similarly, for the MOSLRT $L_2 = W/\hat{\sigma}^2$, under the alternative, $\hat{\sigma}^2 \rightarrow \bar{\sigma}^2 = (v_1 + v_0)/2$ (see Appendix); thus, the power of the MOSLRT L_2 should satisfy the following equations:

$$1 - \beta = P(L_2 < -z_{1-\alpha} | H_1) \simeq \Phi \left(-\frac{\bar{\sigma}}{\sigma} z_{1-\alpha} - \frac{\sqrt{n}\omega}{\sigma} \right).$$

Therefore, the required sample size for the test statistic L_2 is given by

$$n = \frac{(\bar{\sigma} z_{1-\alpha} + \sigma z_{1-\beta})^2}{\omega^2}, \quad (3)$$

where $\bar{\sigma}^2$, σ^2 , and ω are the same as given above.

To calculate the sample size based on formulae (2) and (3), we have to calculate the quantities v_0 , v_1 , v_{00} , and v_{01} . Assume that subjects were recruited with a uniform distribution over the accrual period t_a and were followed for a period of t_f and that no subjects were lost to follow-up. Thus, the censoring distribution was a uniform distribution over $[t_f, t_a+t_f]$. Then for any parametric survival distributions determined by the null and alternative hypotheses as described above, the quantities v_0 , v_1 , v_{00} , and v_{01} can be calculated by numerical integrations. Therefore, the study design can be conducted by calculating the sample size with formula (2) for the OSLRT and formula (3) for the MOSLRT.

4 Comparison

In this section, five methods for designing single-arm phase II trials with time-to-event endpoints (the naive approach, a non-parametric test based on the Nelson-Aalen estimator, a parametric MLE test, the OSLRT and the MOSLRT) are compared under the exponential model.

To introduce the naive approach, suppose that n patients were enrolled on the study and followed to a landmark time point x . The event status of all n patients determined whether or not the patient experienced an event or not. Let d be the total number of events observed in n patients. Then d follows a binomial distribution $B(n, p)$, where $p = S(x)$ is the probability of a patient experiencing an event before or at x . The hypothesis (1) is then the same as the following hypothesis:

$$H_0: p \leq p_0 \text{ vs. } H_a: p > p_0, \quad (4)$$

where $p_0 = S_0(x)$, and the study is powered at the alternative $p_1 = S_1(x) (> p_0)$. Thus, the study design can be carried out based on the exact binomial test given as follows: For a given type I error α , a positive integer r exists such that

$$\sum_{k=r}^n b(k, p_0; n) \leq \alpha \text{ and } \sum_{k=r-1}^n b(k, p_0; n) > \alpha,$$

where $b(k, p, n)$ is the binomial probability. We reject the null hypothesis H_0 if $d \geq r$. Thus, given the power $1 - \beta$ under the alternative hypothesis $p = p_1$, the sample size required for the study can be obtained by solving for the smallest integer n that satisfies by the following equation:

$$\sum_{k=r}^n b(k, p_1; n) \geq 1 - \beta.$$

Under the exponential model, hypothesis (1) is also equivalent to the following hypothesis for the cumulative hazard function:

$$H_0: \Lambda(x) \geq \Lambda_0(x) \text{ vs. } H_a: \Lambda(x) < \Lambda_0(x), \quad (5)$$

where x is the landmark time point, $\Lambda_0(x) = \lambda_0 x$, and the study is powered at the alternative $\Lambda_1(x) = \lambda_1 x$, with $\lambda_1 < \lambda_0$. To test this hypothesis, Lin et al. (1999) proposed a nonparametric test based on the Nelson-Aalen estimate of the cumulative hazard function, which is given by

$$Z_x = \frac{n^{1/2} \{\log \hat{\Lambda}(x) - \log \Lambda_0(x)\} \hat{\Lambda}(x)}{\hat{\sigma}(x)},$$

where $\hat{\Lambda}(x) = \sum_{\{i: X_i \leq x\}} \frac{\Delta_i}{R_i}$, with $R_i = \sum_{j=1}^n I\{X_j \geq X_i\}$, and $\hat{\sigma}^2(x) = \sum_{\{i: X_i \leq x\}} \frac{\Delta_i}{R_i^2/n}$ is an estimate of $\sigma^2(x) = \int_0^x \frac{d\Lambda_1(t)}{P_{H_1}(X \geq t)}$, which is the asymptotic variance of $\hat{\Lambda}(x)$ at the alternative. Thus, the sample size required for the study based on the test statistic Z_x with type I error α and power of $1 - \beta$ is given by

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2(x)}{(\log \Lambda_0(x) - \log \Lambda_1(x))^2 \Lambda_1^2(x)}. \quad (6)$$

This test statistic has been used to design single-arm phase II survival trials with a two-stage or multistage sequential monitoring procedure by Lin et al. (1999), Case and Morgan (2003), and Huang et al. (2010).

Hypothesis (1) is also equivalent to the hypothesis for the hazard parameter

$$H_0: \lambda \geq \lambda_0 \text{ vs. } H_a: \lambda < \lambda_0, \quad (7)$$

and the study is powered at the alternative $\lambda = \lambda_1 (< \lambda_0)$. Thus, the following parametric MLE test statistic (Sprott, 1973; and Lawless, 1982) can be used to test the hypothesis:

$$Z = (\hat{\phi} - \phi_0) / (\hat{\phi}^2 / 9d)^{1/2},$$

where $\phi_0 = \lambda_0^{1/3}$, $\hat{\phi} = \hat{\lambda}^{1/3}$, and $\hat{\lambda} = d/U$, with $d = \sum_{i=1}^n \Delta_i$ and $U = \sum_{i=1}^n X_i$. The sample size n required for the test statistic Z can be calculated by

$$n = \frac{(\sigma_1 R^{1/3} z_{1-\alpha} + \sigma_0 z_{1-\beta})^2}{9(R^{1/3} - 1)^2 \sigma_0^2 \sigma_1^2}, \quad (8)$$

where $\sigma_0^2 = E_{H_0}(\Delta)$, $\sigma_1^2 = E_{H_1}(\Delta)$, and $R = \lambda_0/\lambda_1$.

To calculate the sample size using formulae (2), (3), (6), and (8), we assume that subjects were recruited with a uniform distribution over the accrual period t_a , were followed for t_f and that the study period τ is $t_a + t_f$. We further assume that no subjects were lost to follow-up. Then the censoring distribution is uniform over the interval $[t_f, t_a + t_f]$. Thus, σ_i^2 in the sample size formula (8) can be calculated by

$$\sigma_i^2 = 1 - \frac{1}{\lambda_i t_a} (e^{-\lambda_i t_f} - e^{-\lambda_i \tau}), i=0, 1$$

and $\sigma^2(x)$ in the sample size formula (6) can be calculated by

$$\sigma^2(x) = \int_0^x \frac{\lambda_1}{\min\{(\tau - t)/t_a, 1\} e^{-\lambda_1 t}} dt, x < \tau.$$

Quantities v_0 , v_1 , v_{00} , and v_{01} in the sample size formulae (2) and (3) can then be calculated as follows:

$$v_0 = \frac{\lambda_0}{\lambda_1} \left\{ 1 - \frac{1}{\lambda_1 t_a} (e^{-\lambda_1 t_f} - e^{-\lambda_1 \tau}) \right\},$$

$$v_1 = 1 - \frac{1}{\lambda_1 t_a} (e^{-\lambda_1 t_f} - e^{-\lambda_1 \tau}),$$

$$v_{00} = \lambda_0^2 \int_0^{t_f} t e^{-\lambda_1 t} dt + \frac{\lambda_0^2}{t_a} \int_{t_f}^{\tau} t(\tau - t) e^{-\lambda_1 t} dt,$$

$$v_{01} = \lambda_0 \lambda_1 \int_0^{t_f} t e^{-\lambda_1 t} dt + \frac{\lambda_0 \lambda_1}{t_a} \int_{t_f}^{\tau} t(\tau - t) e^{-\lambda_1 t} dt.$$

To compare the study designs based on the five methods, we first calculated the sample sizes of each method for various design scenarios under the exponential model. The design parameters were set as follows: the landmark time point $x = 1, 2$; accrual time $t_a = 1, 3$; follow-up time $t_f = 1, 2, 3$; type I error $\alpha = 0.05$; and power $1 - \beta = 80\%$. The survival probability under the null $S_0(x)$ was 0.2 to 0.7, and under the alternative $S_1(x)$ was 0.35 to 0.8. Under these design scenarios, sample sizes were calculated for each of the five methods (Table 1). The empirical type I error and power for the corresponding sample size were simulated based on 100,000 simulation runs (Table 1), except for the naive method for which we used the exact binomial test.

The results can be summarized as follows. The naive method had the largest sample size. Thus, it is the least efficient method. There is little gain in efficacy with the non-parametric test Z_x compared to the naive method. The OSLRT was more efficient than the naive method and the nonparametric test but less efficient than the parametric test Z and the MOSLRT. Furthermore, the empirical type I errors of the OSLRT were always smaller than the nominal level (indicating conservativeness), and the sample sizes were overestimated (indicating inaccuracy). The parametric test Z and the MOSLRT had almost identical sample sizes, which were much smaller than these for the other three methods. Therefore, both the parametric test Z and the MOSLRT were more efficient. The saving in sample size for the parametric MLE test and the MOSLRT could be more than 50% compared to those for the naive approach and the nonparametric test. This is partly because the naive approach only used information about whether an event had occurred prior to the landmark time point, not the actual time-to-event data, and the nonparametric test used events that occurred before the landmark time point only; events that occurred after the landmark time point were not used. Furthermore, the parametric test and the MOSLRT preserved the type I error well and provided adequate power for the study design. The limitation of the parametric test is that it applies to the exponential distribution only and is not robust against mis-specification of the underlying distribution. In contrast, the MOSLRT can be applied to any parametric survival distribution.

5 Simulation Studies

To further study the performance of the MOSLRT under various survival models (Table 2), we calculated sample sizes for each of these distributions under various design scenarios. Simulation studies were conducted to assess the accuracy of the sample size estimation and the performance of the MOSLRT with small sample sizes. Under each of these distributions, the shape parameter was set to 0.5, 1, and 2 to reflect the different types of hazard function; the landmark time point x was set to 2, and the survival probabilities under the null $S_0(x)$ and alternative $S_1(x)$ were set to 0.2 to 0.7 and 0.35 to 0.8, respectively. We assumed that subjects were recruited with a uniform distribution over the accrual period t_a of 3 years and followed for a period t_f of 1 year. The censoring distribution was a uniform distribution on the interval $[t_f, t_a + t_f]$. Thus, the quantities v_0 , v_1 , v_{00} , and v_{01} can be calculated by numeric integrations. Therefore, given the nominal significance level of 0.05 and the power of 80%, the required sample sizes for each design scenario and each distribution were calculated. The empirical type I error and the power of the corresponding design were simulated based on 100,000 runs (Table 3).

The simulation results can be summarized as follows. First, the MOSLRT preserved the type I error well when $S_0(x) < 0.5$ and was slightly liberal when $S_0(x) \geq 0.5$, except in the case of the Gompertz distribution with $\gamma = 2$, where the type I error was controlled very well. Second, the MOSLRT provided adequate power in all scenarios, even when the sample size was small. Third, the calculated sample sizes were close for the different values of the shape parameters. Thus, mis-specification of the shape parameter did not have a substantial impact on the study design, particularly when $S_0(x) < 0.5$. To further investigate the accuracy of the normal approximation for the OSLRT and MOSLRT, we conducted 100,000 simulation runs under the Weibull model to simulate the empirical distribution functions of L_1 and L_2 under

the null hypothesis with sample size $n = 30, 50, 100$, and 200 for the same design parameters as discussed above (Table 4). The simulation results showed that the distribution of L_1 had a lighter left tail than the standard normal distribution function, while that of L_2 had a slightly heavier left tail. These results explained the observations from previous simulations that L_1 was conservative and L_2 was slightly liberal when $S_0(x)$ was relatively large.

Overall, the MOSLRT performed better than the OSLRT; it preserved type I error well and provided adequate power for designing single-arm phase II survival trials. The MOSLRT can be applied to any parametric survival distribution and thus has advantages over the other methods. Therefore, we recommend using the MOSLRT for designing single-arm phase II trials with time-to-event endpoints.

6 Example

Rhabdoid tumors are aggressive pediatric malignancies with a poor prognosis. Over the past 5 years, St. Jude Children's Research Hospital has enrolled 14 pediatric patients with recurrent or refractory non-CNS rhabdoid tumors that were treated with conventional chemotherapy. All 14 patients had events within 3 years, where an event was defined as disease relapse or death. The Weibull model was fitted to the data by using R, which resulted in an estimate (standard error) of the shape parameter κ of 1.37 (0.28) and a median EFS time of 0.936 years. The Kaplan-Meier estimate of 1-year EFS was 43% (se=12%). For comparison, the exponential model was also fitted to the data, and the Kaplan-Meier curve and the fitted exponential and Weibull survival curves were plotted (Figure 1). The log-likelihood for the Weibull model was -13.60 , whereas for the exponential model it was -14.60 . The likelihood ratio test statistic was $2[-13.60 - (-14.60)] = 2.0$, which was not significant compared with a chi-square percentile with one degree of freedom. However, both the log-likelihood value and curve fitting suggest that the Weibull model provides a more satisfactory model than does the exponential model.

Now, suppose that we wish to design a new trial and we consider that the molecular agent alisertib is not worthy of further evaluation if the 1-year EFS is at most 43% but is promising if the 1-year EFS is at least 55%, with 80% power and 5% type I error. Furthermore, assume that the accrual period will be 3 years and the follow-up period will be 1 year. Then under the assumption of the Weibull model with shape parameter κ of 1.37, uniform entry, and no loss to follow-up, the required sample size calculated from formula (3) is 61 patients by using the MOSLRT; the empirical type I error is 0.05; and the power is 81%, based on 100,000 simulations for this design. Using the naive approach, 111 patients are required for the study. Thus, using the MOSLRT reduces the number of patients required by 45% compared to the naive approach.

7 Conclusion

An MOSLRT has been proposed for designing single-arm phase II trials with time-to-event endpoints. The proposed test is simple to use, and the sample size formula is easy to compute. Simulation results showed that the proposed test preserves type I error well and provides adequate power for study design when the sample size is within a range that is

typical for single-arm phase II trials. The study design can be conducted under any parametric survival distribution which can be obtained by fitting the historical data and using model selection criteria to select the best model, or under a spline version of the survival distribution, which can be fitted using the R function **oldlogspline** without a specific assumption of a parametric form of the underlying distribution. The MOSLRT uses the actual censored time-to-event data; it is much more efficient than the naive method, the nonparametric test Z_X , and the OSLRT, and it is as efficient as the parametric MLE test Z under the correct model. Thus, the proposed MOSLRT provides a new method for designing single-arm phase II survival trials with a flexible choice of survival distribution to meet the requirements of different types of historical data. Because the MOSLRT has an independent increment structure, trial monitoring using the MOSLRT can also be developed based on the well-known error spending function methodology (Lan and DeMets, 1983).

Appendix

First, we calculate the mean and variance of W under the null hypothesis H_0 by noting that $E_{H_0}(O) = nE_{H_0}(\cdot)$ and $E_{H_0}(E) = nE_{H_0}(\Lambda_0(X))$.

Let $f_0(t)$, $S_0(t)$, and $\Lambda_0(t)$ be the density, survival, and cumulative hazard functions of failure time T under the null and $g(t)$ and $G(t)$ be the density and survival functions of censoring time C . Then, by exchange of integrations, we have

$$E_{H_0}(\Delta) = \int_0^\infty \left(\int_0^y f_0(t)g(y)dt \right) dy = \int_0^\infty \left(f_0(t) \int_t^\infty g(y)dy \right) dt = \int_0^\infty G(t)S_0(t)d\Lambda_0(t).$$

Let $S_X(t)$ be the survival distribution of $X = T \wedge C$ under the null, then $S_X(t) = S_0(t)G(t)$ and, by integration by parts, we have

$$E_{H_0}(\Lambda_0(X)) = - \int_0^\infty \Lambda_0(t)dS_X(t) = \int_0^\infty S_X(t)d\Lambda_0(t) = \int_0^\infty G(t)S_0(t)d\Lambda_0(t) = E_{H_0}(\Delta). \quad (A1)$$

Therefore, the mean of W under the null is $E_{H_0}(W) = \sqrt{n}\{E_{H_0}(\Delta) - E_{H_0}(\Lambda_0(X))\} = 0$. By a similar calculation, we have

$$E_{H_0}(\Delta\Lambda_0(X)) = \int_0^\infty \left(\int_0^y f_0(t)g(y)\Lambda_0(t)dt \right) dy = \int_0^\infty \left(f_0(t)\Lambda_0(t) \int_t^\infty g(y)dy \right) dt = \int_0^\infty G(t)S_0(t)\Lambda_0(t)d\Lambda_0(t),$$

and

$$E_{H_0}(\Lambda_0^2(X)) = - \int_0^\infty \Lambda_0^2(t)dS_X(t) = 2 \int_0^\infty S_X(t)\Lambda_0(t)d\Lambda_0(t) = 2 \int_0^\infty G(t)S_0(t)\Lambda_0(t)d\Lambda_0(t).$$

We have shown that $E_{H_0}(\cdot) = E_{H_0}(\Lambda_0(X))$ and $E_{H_0}(\Lambda_0^2(X)) = 2E_{H_0}(\Delta\Lambda_0(X))$. Therefore,

$$\text{Var}_{H_0}(W) = E_{H_0}(\Delta - \Lambda_0(X))^2 = E_{H_0}(\Delta) - 2E_{H_0}(\Delta\Lambda_0(X)) + E_{H_0}(\Lambda_0^2(X)) = E_{H_0}(\Delta) = \int_0^\infty G(t)S_0(t)d\Lambda_0(t).$$

(A2)

Thus,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty \{dN_i(t) + Y_i(t)d\Lambda_0(t)\} / 2$$

is a consistent estimate of $\text{Var}_{H_0}(W)$ under the null and

$$W/\hat{\sigma} \rightarrow N(0, 1).$$

Now we derive the exact variance of W under the alternative. Let $f_1(t)$, $S_1(t)$, and $\Lambda_1(t)$ be the density, survival, and cumulative hazard functions, respectively, of failure time T under the alternative. Then by similar calculation, we have

$$E_{H_1}(\Delta) = \int_0^\infty \left(\int_0^y f_1(t)g(y)dt \right) dy = \int_0^\infty \left(f_1(t) \int_t^\infty g(y)dy \right) dt = \int_0^\infty G(t)S_1(t)d\Lambda_1(t) = v_1.$$

(A3)

Let $S_X(t)$ be the survival distribution of $X = T \wedge C$ under the alternative, then $S_X(t) = G(t)S_1(t)$ and, by integration by parts, we have

$$E_{H_1}(\Lambda_0(X)) = - \int_0^\infty \Lambda_0(t)dS_X(t) = \int_0^\infty S_X(t)d\Lambda_0(t) = \int_0^\infty G(t)S_1(t)d\Lambda_0(t) = v_0. \quad (\text{A4})$$

Thus, $E_{H_1}(W) = \sqrt{n}\{E_{H_1}(\Delta) - E_{H_1}(\Lambda_0(X))\} = \sqrt{n}(v_1 - v_0)$. Similarly, we have

$$E_{H_1}(\Delta\Lambda_0(X)) = \int_0^\infty \left(\int_0^y f_1(t)g(y)\Lambda_0(t)dt \right) dy = \int_0^\infty \left(f_1(t)\Lambda_0(t) \int_t^\infty g(y)dy \right) dt = \int_0^\infty G(t)S_1(t)\Lambda_0(t)d\Lambda_1(t) = v_{01},$$

(A5)

and

$$E_{H_1}(\Lambda_0^2(X)) = - \int_0^\infty \Lambda_0^2(t) dS_X(t) = 2 \int_0^\infty S_X(t) \Lambda_0(t) d\Lambda_0(t) = 2 \int_0^\infty G(t) S_1(t) \Lambda_0(t) d\Lambda_0(t) = 2v_{00}.$$

(A6)

Therefore, the exact variance of W under the alternative is given by

$$\begin{aligned} \text{Var}_{H_1}(W) &= \text{Var}_{H_1}(\Delta) + \text{Var}_{H_1}(\Lambda_0(X)) - 2\text{Cov}_{H_1}(\Delta, \Lambda_0(X)) \\ &= E_{H_1}(\Delta) - E_{H_1}^2(\Delta) + E_{H_1}(\Lambda_0^2(X)) - E_{H_1}^2(\Lambda_0(X)) \\ &\quad - 2E_{H_1}(\Delta\Lambda_0(X)) + 2E_{H_1}(\Delta)E_{H_1}(\Lambda_0(X)) \\ &= v_1 - v_1^2 + 2v_{00} - v_0^2 - 2v_{01} + 2v_0v_1 = \sigma^2. \end{aligned}$$

Under the alternative H_1 ,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty \{dN_i(t) + Y_i(t)d\Lambda_0(t)\} / 2 \rightarrow \bar{\sigma}^2 = (v_0 + v_1) / 2; \quad (\text{A7})$$

thus, $W - \sqrt{n}\omega \rightarrow N(0, \sigma^2)$ and by Slutsky's theorem, it follows that

$$\frac{W}{\hat{\sigma}} - \frac{\sqrt{n}\omega}{\bar{\sigma}} \rightarrow N\left(0, \frac{\sigma^2}{\bar{\sigma}^2}\right) \text{ as } n \rightarrow \infty. \quad (\text{A8})$$

References

- Breslow NE. Analysis of survival data under the proportional hazards model. *International Statistics Review*. 1975; 43:44–58.
- Brown, BW.; Brauner, C.; Chan, A.; Gutierrez, D.; Herson, J.; Lovato, J.; Polsley, J.; Russell, K.; Venier, J. Method used in STPLAN version 4.5. The University of Texas and M.D. Anderson Cancer Center; 2010.
- Case LD, Morgan TM. Design of phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology*. 2003; 3:1–12. [PubMed: 12515580]
- Fleming, TR.; Harrington, DP. *Counting processes and survival analysis*. New York: John Wiley and Sons; 1991.
- George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*. 1977; 27:15–24.
- Huang B, Talukder E, Thomas N. Optimal two-stage phase II designs with long-term endpoints. *Statistics in Biopharmaceutical Research*. 2010; 2:51–61.
- Kwak MJ, Jung SH. Phase II clinical trials with time-to-event endpoints: Optimal two-stage designs with one-sample log-rank test. *Statistics in Medicine*. 2013; 33:2004–2016. [PubMed: 24338995]
- Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*. 1981; 2:93–114. [PubMed: 7273794]
- Lakatos E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials*. 1986; 7:189–199. [PubMed: 3802834]

- Lan K, DeMets D. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
- Lawless, JF. Statistical methods for lifetime data. New York: John Wiley and Sons; 1982.
- Lin DY, Yao Q, Ying ZL. A general theory on stochastic curtailment for censored survival data. *Journal of the American Statistical Association*. 1999; 94:510–521.
- Owzar K, Jung SH. Designing phase II studies in cancer with time-to-event endpoints. *Clinical Trials*. 2008; 28:209–221.
- Rubenstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases*. 1981; 34:469–479. [PubMed: 7276137]
- Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983; 39:499–503. [PubMed: 6354290]
- Sprott DA. Normal likelihoods and relation to a large sample theory of estimation. *Biometrika*. 1973; 60:457–465.
- Sun XQ, Peng P, Tu DS. Phase II cancer clinical trial with a one-sample log-rank test and its corrections based on the Edgeworth expansion. *Contemporary Clinical Trials*. 2011; 32:108–113. [PubMed: 20888929]
- Woolson RF. Rank-tests and a one-sample log-rank test for comparing observed survival-data to a standard population. *Biometrics*. 1981; 37:687–696.
- Wu J. Sample size calculation for the one-sample log-rank test. *Pharmaceutical Statistics*. 2015; 14:26–33. [PubMed: 25339496]

Figure 1.
Step functions are the Kaplan-Meier survival curve and its 90% confidence boundaries.
Solid and dotted curves are the fitted Weibull and exponential survival distributions,
respectively.

Table 1

Sample size, type I error, and power comparisons of five methods (the naive approach, the parametric MLE test Z , the nonparametric test Z_x , L_1 , and L_2) under the exponential model for various designs with nominal type I error of 0.05 and power of 80%.

Design			Sample size *						Type I error					Power				
(t_a, t_b, x)	S_0	S_1	n_B	n_{Z_x}	n_Z	n_{L_1}	n_{L_2}	Z_x	Z	L_1	L_2	Z_x	Z	L_1	L_2			
(1, 1, 1)	0.2	0.35	56	50	42	45	40	.059	.046	.041	.051	.811	.815	.823	.809			
		0.4	35	30	25	27	24	.061	.047	.039	.052	.823	.823	.824	.817			
	0.3	0.45	67	64	50	54	48	.045	.047	.041	.052	.796	.813	.822	.807			
		0.5	0.65	69	66	50	55	48	.054	.050	.043	.054	.812	.805	.818	.802		
	0.6	0.75	62	63	42	49	41	.041	.051	.041	.055	.791	.808	.819	.797			
		0.7	0.8	127	120	82	90	78	.043	.052	.043	.056	.791	.800	.810	.794		
(3, 1, 1)	0.2	0.35	56	50	38	42	37	.059	.046	.039	.050	.811	.815	.831	.814			
		0.4	35	30	23	25	22	.061	.045	.037	.049	.823	.826	.832	.823			
	0.3	0.45	67	64	44	47	42	.045	.046	.041	.050	.796	.817	.821	.810			
		0.5	0.65	69	66	40	43	38	.054	.048	.041	.052	.812	.814	.820	.809		
	0.6	0.75	62	63	33	36	31	.041	.049	.040	.053	.791	.809	.824	.806			
		0.7	0.8	127	120	58	63	55	.043	.049	.043	.055	.791	.807	.812	.799		
(3, 2, 2)	0.2	0.35	56	50	40	44	39	.059	.047	.040	.051	.811	.813	.827	.811			
		0.4	35	30	24	26	23	.061	.045	.039	.050	.823	.818	.826	.818			
	0.3	0.45	67	64	47	51	45	.045	.047	.042	.051	.796	.809	.822	.807			
		0.5	0.65	69	66	46	50	44	.054	.049	.042	.054	.812	.809	.816	.802		
	0.6	0.75	62	63	40	44	37	.041	.050	.040	.054	.791	.810	.820	.797			
		0.7	0.8	127	120	73	80	70	.043	.050	.043	.056	.791	.803	.811	.798		
(3, 3, 2)	0.2	0.35	56	50	38	41	37	.059	.046	.040	.049	.811	.818	.825	.814			
		0.4	35	30	23	25	22	.061	.044	.038	.049	.823	.828	.835	.824			
	0.3	0.45	67	64	44	47	42	.045	.047	.042	.049	.796	.818	.824	.810			
		0.5	0.65	69	66	40	44	38	.054	.048	.041	.053	.812	.808	.821	.802		
	0.6	0.75	62	63	34	37	32	.041	.050	.041	.053	.791	.814	.819	.806			
		0.7	0.8	127	120	61	66	58	.043	.049	.043	.054	.791	.805	.810	.800		

* Sample sizes calculated for the naive approach, the non-parametric test Z_X , the parametric test Z , the OSLRT, and the MOSLRT were recorded as n_B , n_Z , n_{Z_X} , n_{OSLRT} , and n_{MOSLRT} , respectively. Sample sizes calculated for Z_X based on formula (6) were overestimated. Thus, sample sizes recorded in this table for Z_X were reduced by simulation runs to achieve an empirical power close to the nominal level. The sample sizes of Z_X do not depend on the accrual time t_f and follow-up time x because the landmark time point x was set as $x = t_f$.

Table 2

Various parametric distributions used for single-arm phase II trial designs.

	Surv. function	Density	Parameter	Cumu. hazar	Hazard	
Dist.	$S(t)$	$f(t)$	Scale	Shape	$\Lambda(t)$	$\lambda(t)$
Exponential	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	λ	1	λt	λ
Weibull	$e^{-\lambda t^\kappa}$	$\kappa \lambda t^{\kappa-1} e^{-\lambda t^\kappa}$	λ	κ	λt^κ	$\kappa \lambda t^{\kappa-1}$
Gamma	$1 - I_k(\lambda t)$	$\frac{\lambda t^{k-1} e^{-\lambda t^k}}{\Gamma(k)}$	λ	k	$-\log S(t)$	$\frac{f(t)}{S(t)}$
Log-normal	$1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$	$\frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}}$	μ	σ	$-\log S(t)$	$\frac{f(t)}{S(t)}$
Gompertz	$e^{-\frac{\theta}{\gamma}(e^{\gamma t} - 1)}$	$\theta e^{\gamma t} e^{-\frac{\theta}{\gamma}(e^{\gamma t} - 1)}$	θ	γ	$\frac{\theta}{\gamma}(e^{\gamma t} - 1)$	$\theta e^{\gamma t}$
Log-logistic	$\frac{1}{1 + \lambda t^p}$	$\frac{p \lambda t^{p-1}}{(1 + \lambda t^p)^2}$	λ	p	$\log(1 + \lambda t^p)$	$\frac{p \lambda t^{p-1}}{1 + \lambda t^p}$

Abbreviations: Dist., Distribution; Cumu., Cumulative; Surv., Survival

Sample size, simulated empirical type I error (α), and power ($1 - \beta$) for the MOSLRT based on 100,000 simulation runs for various distributions with nominal type I error of 0.05 and power of 80%. The censoring distribution is uniform over $[t_f, t_d + t_f]$, where $t_d = 3$ and $t_f = 1$.

Table 3

Dist.	Design	n	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$
$W(\lambda, \kappa)$		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 2$		
	0.2 vs 0.35	44	.052	.804	44	.052	.809	44	.052	.813
	0.2 vs 0.4	26	.053	.803	26	.051	.810	26	.049	.818
	0.3 vs 0.45	55	.053	.804	53	.052	.803	51	.051	.806
	0.5 vs 0.65	58	.054	.796	55	.054	.800	49	.052	.806
	0.6 vs 0.75	52	.057	.797	48	.058	.797	41	.053	.803
	0.7 vs 0.8	100	.056	.792	91	.055	.794	75	.055	.797
$G(\lambda, k)$		$k = 0.5$			$k = 1$			$k = 2$		
	0.2 vs 0.35	47	.053	.806	44	.051	.810	42	.050	.805
	0.2 vs 0.4	28	.052	.815	26	.052	.811	25	.052	.810
	0.3 vs 0.45	56	.052	.804	53	.052	.802	54	.051	.803
	0.5 vs 0.65	57	.052	.795	55	.053	.799	54	.053	.804
	0.6 vs 0.75	50	.055	.795	48	.057	.798	47	.055	.804
	0.7 vs 0.8	97	.056	.798	91	.057	.795	87	.056	.795
$LN(\mu, \sigma)$		$\sigma = 2$			$\sigma = 1$			$\sigma = 0.5$		
	0.2 vs 0.35	36	.052	.804	37	.051	.810	39	.052	.814
	0.2 vs 0.4	22	.051	.808	22	.052	.806	24	.052	.822
	0.3 vs 0.45	48	.052	.803	49	.052	.805	52	.052	.807
	0.5 vs 0.65	56	.054	.797	57	.053	.799	60	.052	.803
	0.6 vs 0.75	51	.056	.789	52	.055	.801	54	.053	.802
	0.7 vs 0.8	101	.055	.793	100	.056	.796	102	.054	.799
$LL(\lambda, p)$		$p = 0.5$			$p = 1$			$p = 2$		
	0.2 vs 0.35	35	.052	.799	36	.051	.807	37	.050	.803
	0.2 vs 0.4	22	.052	.813	22	.051	.811	23	.050	.812

Dist.	Design	α	$1 - \beta$	n	α	$1 - \beta$	n	α	$1 - \beta$	n
	0.3 vs 0.45	.49	.051	.801	.50	.052	.804	.52	.052	.802
	0.5 vs 0.65	.58	.054	.797	.58	.055	.794	.61	.053	.801
	0.6 vs 0.75	.53	.057	.797	.52	.056	.795	.52	.055	.793
	0.7 vs 0.8	.102	.055	.791	.99	.056	.795	.96	.052	.795
<hr/>										
$GZ(\theta, \gamma)$		$\gamma = 0.5$			$\gamma = 1$			$\gamma = 2$		
	0.2 vs 0.35	.43	.051	.808	.43	.050	.811	.45	.050	.814
	0.2 vs 0.4	.26	.051	.819	.26	.049	.818	.27	.047	.824
	0.3 vs 0.45	.51	.052	.807	.50	.050	.809	.51	.048	.814
	0.5 vs 0.65	.50	.054	.807	.46	.051	.813	.44	.049	.819
	0.6 vs 0.75	.42	.054	.803	.37	.052	.813	.33	.049	.817
	0.7 vs 0.8	.77	.054	.796	.64	.054	.803	.53	.050	.814

$W(\lambda, \kappa)$, $G(\lambda, k)$, $LN(\mu, \sigma)$, $LL(\lambda, p)$, and $GZ(\theta, \gamma)$ are the Weibull, gamma, log-normal, log-logistic and Gompertz distributions, respectively.

Simulated distribution functions of OSLRT L_2 compared to the standard normal distribution function $\Phi(x)$ based on 100,000 simulation runs for the Weibull distribution.

Table 4

		x							
κ	n	Test	-3.0	-1.96	-0.67	0.0	0.67	1.96	3.0
0.5	30	L_1	.0003	.0169	.2428	.4949	.7352	.9632	.9959
		L_2	.0021	.0285	.2504	.4949	.7440	.9783	.9993
	50	L_1	.0006	.0190	.2446	.4958	.7412	.9669	.9964
		L_2	.0021	.0283	.2506	.4958	.7477	.9771	.9991
1	100	L_1	.0008	.0210	.2470	.4974	.7430	.9692	.9977
		L_2	.0019	.0280	.2512	.4974	.7475	.9770	.9989
	200	L_1	.0008	.0210	.2480	.4969	.7447	.9702	.9978
		L_2	.0016	.0259	.2512	.4969	.7479	.9758	.9988
2	30	L_1	.0005	.0167	.2374	.4870	.7334	.9628	.9961
		L_2	.0019	.0266	.2440	.4870	.7412	.9771	.9994
	50	L_1	.0005	.0192	.2427	.4908	.7367	.9668	.9969
		L_2	.0018	.0271	.2480	.4908	.7430	.9770	.9991
3	100	L_1	.0008	.0199	.2460	.4956	.7415	.9695	.9977
		L_2	.0020	.0256	.2499	.4956	.7456	.9767	.9990
	200	L_1	.0009	.0214	.2484	.4958	.7423	.9712	.9979
		L_2	.0016	.0251	.2513	.4958	.7453	.9760	.9988
4	30	L_1	.0005	.0167	.2308	.4789	.7256	.9626	.9960
		L_2	.0016	.0255	.2373	.4789	.7329	.9765	.9992
	50	L_1	.0006	.0180	.2344	.4834	.7297	.9656	.9970
		L_2	.0016	.0250	.2395	.4834	.7351	.9757	.9991

Wu									
x									
3.0	961	0.67	0.0	0.67	961	0.3	0.67	1.7	2
9977	6694	7374	4899	2398	0192	8000	0.0008	0.0013	0.0014
9990	9762	5142	4899	2437	0247	9100	0.0014	0.0014	0.0014
9979	9713	7415	4947	2445	0206	8000	0.0008	0.0013	0.0014
9988	9759	7444	4947	2470	0244	9100	0.0014	0.0014	0.0014
$\Phi(x)$									
9987	9750	7486	5000	2514	0250	9100	0.0014	0.0014	0.0014