Published in final edited form as:

J Biopharm Stat. 2017 ; 27(4): 639–658. doi:10.1080/10543406.2016.1167073.

Optimal Two-Stage Logrank Test for Randomized Phase II Clinical Trials

Minjung Kwak^a and Sin-Ho Jung^{b,c,*}

^aDepartment of Statistics, Yeungnam University, Gyeongbuk, 712-749, ROK

^bDepartment of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA

^cBiostatistics and Clinical Epidemiology Center, Samsung Medical Center, Seoul, ROK

Summary

Randomized controlled clinical trials are conducted to determine whether a new treatment is safe and efficacious compared to a standard therapy. We consider randomized clinical trials with right censored time to event endpoint, called survival time here. The two-sample logrank test is popularly used to test if the experimental therapy has a longer survival distribution than the control therapy or not. We consider an early stopping for futility only or for both futility and efficacy. For planning such clinical trials this paper presents two-stage designs that are optimal in the sense that either the maximal sample size or the expected sample size when the experimental therapy is futile or superior is minimized under the given type I and II error rates. Optimal designs for a range of design parameters are tabulated and evaluated using simulations.

Keywords

Expected Sample Size; Futility; Minimax Design; Optimal Design; Survival Distribution

1 Introduction

While binary endpoints, such as tumor response, are popularly used as the primary outcome of phase II cancer clinical trials, we sometimes use time to event (such as progression or recurrence), called survival time hereafter, as the primary outcome as well. When the study endpoint is survival time, the maximum likelihood estimator (MLE) for exponential survival distributions may be used to compare survival distributions between treatment arms. Sample size calculation methods for test statistic based on the MLE of exponential distributions have been proposed by Pasternack and Gilbert (1971), George and Desu (1973) and Lachin (1981). Rubinstein et al. (1981) proposed to use the sample size formula derived for the MLE test for the nonparametric logrank test by showing that this formula provides a reasonable power even for the logrank test through simulations. Their simulations are limited to balanced designs only, and this approximation does not hold under an unbalanced allocation setting.

*Correspondence to: Sin-Ho Jung, sinho.jung@duke.edu.

Because of their robustness, nonparametric rank tests are generally preferred to parametric MLE tests in survival analysis. The logrank test (Peto and Peto, 1972) has been widely used for testing the equality of two survival distributions in the presence of censoring. The asymptotic normality of the logrank test can be found in Andersen et al. (1982) and Fleming and Harrington (1991). Numerous methods have been proposed for sample size estimation including Lakatos (1977), Schoenfeld (1983) and Yateman and Skene (1992). These sample size methods are for single-stage trials.

In randomized controlled clinical trials where the primary objective is to compare efficacy of the new treatment with the standard regimen, we can save time and resources by stopping the trial early for futility when we have enough evidence that the new treatment is unlikely to beat the control regimen. On the other hand, if the new regimen is overwhelmingly better than control regimen, then we may want to stop the trial for efficacy to prevent patients from being assigned to an inferior arm. We consider early termination for futility only or for both efficacy and futility. In this paper we propose optimal design and analysis methods for two-stage log-rank tests that can be useful for randomized phase II clinical trials. We optimize the two-stage trial design to minimize the maximal sample size or the expected sample size under the null hypothesis with a futility stopping only. When stopping for both futility and efficacy, we minimize the average expected sample sizes under the null and alternative hypotheses.

Section 2 reviews a sample size calculation for single-stage randomized controlled clinical trials with a survival endpoint as the primary outcome. We discuss asymptotic theory for two-stage design and analysis methods for randomized controlled clinical trials in Section 3. In Section 4, we propose optimal two-stage designs and present single-stage and optimal two-stage designs under various design parameter settings. In Section 5, we evaluate the finite sample performance of the designed presented in Section 4 using simulations. We conclude this article with some discussion on the implications of the optimal two-stage design in Section 6. At the cost of heavier computations, these methods can be easily extended to randomized trials with more than two stages.

2 Two-Sample Log-Rank Test (Review)

2.1 Test Statistic

We assume that arm 1 is a control arm and arm 2 is an experimental arm. Suppose that n_k patients are randomized to arm k ($k = 1, 2$) and the survival times from the n_k patients, are independent and identically distributed with cumulative hazard function $\Lambda_k(t)$ and hazard function $\lambda_k(t) = \Lambda_k'(t)$. Under the proportional hazards assumption, $\lambda_1(t)/\lambda_2(t) = \lambda(t)$ denotes the hazard ratio. We want to test $H_0: \lambda = 1$ against $H_1: \lambda > 1$.

Let T_{ki} denote the survival time for patient i in arm k , $1 \leq i \leq n_k$, $k = 1, 2$. Then we usually observe (X_{ki}, δ_{ki}) , where X_{ki} is the minimum of T_{ki} and censoring time C_{ki} , and δ_{ki} is an event indicator taking 1 if the patient had an event and 0 otherwise. Within each arm, the censoring times are independent of the survival times. Let $N_k(t) = \sum_{i=1}^{n_k} \delta_{ki} I(X_{ki} \leq t)$ and $Y_k(t) = \sum_{i=1}^{n_k} I(X_{ki} \geq t)$ denote the event and the at-risk processes for arm k , respectively.

Let $n = n_1 + n_2$, the maximal sample size, $N(t) = N_1(t) + N_2(t)$ and $Y(t) = Y_1(t) + Y_2(t)$. Then, the log-rank test statistic is given as

$$W = \frac{1}{\sqrt{n}} \int_0^\infty \frac{Y_1(t)Y_2(t)}{Y(t)} \{d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)\},$$

where $\hat{\Lambda}_k(t) = \int_0^t Y_k(s)^{-1} dN_k(s)$ is the Nelson-Aalen (Nelson, 1969; Aalen, 1978) estimator of $\Lambda_k(t)$. Under H_0 , $W/\hat{\sigma}$ is asymptotically standard normal with

$$\hat{\sigma}^2 = \frac{1}{n} \int_0^\infty \frac{Y_1(t)Y_2(t)}{Y(t)^2} dN(t)$$

by Fleming and Harrington (1991). Hence, we reject H_0 , in favor of H_1 , if $W/\hat{\sigma} > z_{1-\alpha}$ with one-sided type I error rate α .

2.2 Sample Size Calculation

Let $p_k = n_k/n$ ($p_1 + p_2 = 1$) denote the allocation proportion for arm k . We assume that patients are accrued with a constant accrual rate, r , during an accrual period and followed during an additional follow-up period b after the last patient is entered. Let $S_k(t) = \exp\{-\Lambda_k(t)\} = P(T_{ki} > t)$ denote the survivor function for arm k , and $G(t) = P(C_{ki} > t)$ denote the survivor function of the censoring distribution which is common between two arms.

We note that $Y_k(t)/n$ uniformly converges to $p_k G(t) S_k(t)$. By the lemma 4.1 of Fleming and Harrington (1991), under H_1 , $\hat{\sigma}^2$ converges to

$$\sigma_0^2 = p_1^2 p_2^2 \int_0^\infty \frac{G(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1 p_2^2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t)$$

and the variance of W is given as

$$\sigma_1^2 = p_1 p_2^2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1^2 p_2 \int_0^\infty \frac{G(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t).$$

Furthermore, under H_1 , we can show that $E(W) = \sqrt{n}\omega$, where

$$\omega = p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)}{p_1 S_1(t) + p_2 S_2(t)} \{d\Lambda_1(t) - d\Lambda_2(t)\}.$$

Hence, given n , the power is given as

$$1 - \beta = P\left(\frac{W}{\hat{\sigma}} \geq z_{1-\alpha} | H_1\right) = P\left(\frac{W - \sqrt{n}\omega}{\sigma_1} \geq \frac{\sigma_0}{\sigma_1} z_{1-\alpha} - \frac{\sqrt{n}\omega}{\sigma_1} | H_1\right) = \Phi\left(\frac{\sigma_0}{\sigma_1} z_{1-\alpha} - \frac{\sqrt{n}\omega}{\sigma_1}\right),$$

where $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Given power $1 - \beta$, the required sample size is given as

$$n = \left(\frac{\sigma_0 z_{1-\alpha} + \sigma_1 z_{1-\beta}}{\omega} \right)^2. \quad (1)$$

Note that this formula is derived without a parametric assumption for survival distributions or a nearby alternative assumption hypothesis.

By modifying George and Desu's (1973) formula, Rubinstein et al. (1981) propose to approximate the sample size for the log-rank test by that of the exponential MLE test, i.e.

$$\left(\frac{\log \Delta_1}{z_{1-\alpha} + z_{1-\beta}} \right)^2 = D_1^{-1} + D_2^{-1} \quad (2)$$

under a balanced allocation ($p_1 = p_2 = 1/2$), where Δ_1 denotes the hazard ratio under H_1 and D_k denotes the number of events from arm k .

Using a nearby alternative hypothesis approximation (i.e. $\Delta_1 \approx 1$), Schoenfeld (1983) derive the total number of events required for the logrank test as following:

$$D = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{p_1 p_2 (\log \Delta_1)^2} \quad (3)$$

Noting that, for $\Delta_1 \approx 1$ or $S_1(t) \approx S_2(t)$, we have $\log \Delta_1 \approx \Delta_1 - 1$ and the probability of an event for arm k is

$$d_k = P(T_k \leq C) \approx - \int_0^\infty G(t) dS_k(t),$$

we can show that our formula (1) can be approximated by the latter two formulas (2) and (3) under the balanced allocation.

Under Exponential Survival and Uniform Censoring Distributions—Suppose that patients are accrued at a constant rate during accrual period a and all patients are followed for an additional follow-up period b after completion of accrual. Then, $C_{ki} \sim U(b, a + b)$ with survivor function $G(t) = 1$ if $t \leq b$; $= 1 - (t - b)/a$ if $b < t \leq a + b$; $= 0$ if $t > a + b$.

Furthermore, suppose that the survival times have an exponential distribution with hazard rate λ_k for arm $k = 1, 2$. Then, we have $S_k(t) = \exp(-\lambda_k t)$ and $\Lambda_k(t) = \lambda_k t$.

Under these distributional assumptions, we have

$$\sigma_0^2 = p_1^2 p_2 \lambda_1 \left\{ \int_0^{a+b} \frac{e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt - \frac{1}{a} \int_b^{a+b} \frac{(t-b)e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt \right\} \\ + p_1 p_2^2 \lambda_2 \left\{ \int_0^{a+b} \frac{e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt - \frac{1}{a} \int_b^{a+b} \frac{(t-b)e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt \right\} \quad (4)$$

$$\sigma_1^2 = p_1 p_2^2 \lambda_1 \left\{ \int_0^{a+b} \frac{e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt - \frac{1}{a} \int_b^{a+b} \frac{(t-b)e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt \right\} \\ + p_1^2 p_2 \lambda_2 \left\{ \int_0^{a+b} \frac{e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt - \frac{1}{a} \int_b^{a+b} \frac{(t-b)e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt \right\} \quad (5)$$

and

$$\omega = p_1 p_2 (\lambda_1 - \lambda_2) \left\{ \int_0^{a+b} \frac{e^{-(\lambda_1 + \lambda_2)t}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}} dt - \frac{1}{a} \int_b^{a+b} \frac{(t-b)e^{-(\lambda_1 + \lambda_2)t}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}} dt \right\}. \quad (6)$$

We calculate these integrals using a numerical method. By plugging these in (1), we can calculate the sample size for given input values of $(\alpha, 1 - \beta, \lambda_1, \lambda_2, a, b, p_1)$.

The required number of events at the analysis is calculated by $D = n(p_1 d_1 + p_2 d_2)$, where

$$d_k = - \int_0^\infty G(t) dS_k(t) = 1 - \frac{e^{-b\lambda_k}}{a\lambda_k} (1 - e^{-a\lambda_k}).$$

When Accrual Rate is Specified instead of Accrual Period—Now we consider a sample size calculation when accrual rate r is given instead of accrual period a . Given $(\alpha, 1 - \beta, \lambda_1, \lambda_2, r, b, p_1)$, $\sigma_1^2 = \sigma_1^2(a)$, $\sigma_2^2 = \sigma_2^2(a)$ and $\omega = \omega(a)$ are functions of a from (4)–(6). Also, under a constant accrual rate assumption, we have $n = a \times r$ approximately. So, by replacing n with $a \times r$ in (1), we obtain an equation on a ,

$$a \times r = \left\{ \frac{\sigma_0(a) z_{1-\alpha} + \sigma_1(a) z_{1-\beta}}{\omega(a)} \right\}^2.$$

We solve this equation using a numerical method, such as the bisection method. Let a^* denote the solution to this equation. Then, the required sample size is obtained as $n = a^* \times r$.

Example 1: Suppose that the control arm is known to have 20% of 1-year progression-free survival (PFS). We want to show that the experimental arm is expected to increase 1-year PFS to 40%. Assuming an exponential PFS model, the annual hazard rates for the two arms are $\lambda_1 = 1.609$ and $\lambda_2 = 0.916$ with an hazard ratio of $\lambda_1 = 1.756$. Assuming a monthly

accrual of 5 patients ($r = 60$ per year) and $b = 1$ year of additional follow-up period, the required sample size for the log-rank test with 1-sided $\alpha = 10\%$ and 90% of power with balanced allocation ($p_1 = p_2 = 1/2$) is given as $n = 102$ (51 per arm), requiring an accrual period of about 20 months ($a = 102/5$). At the final analysis, we expect $D = 89$ events with 48 subjects and 41 subjects for arms 1 and 2, respectively, under H_1 .

3 Two-Stage Log-Rank Test

Multi-stage clinical trial design for the two-sample log-rank has been widely investigated, e.g. Slud and Wei (1982) and Tsiatis (1982). For randomized phase II trials, two-stage design will be most appropriate due to its small size and relatively short study period compared to large scale phase III trials. With a survival endpoint, it is important to find a reasonable interim analysis time point. If an interim analysis is scheduled for an early stage of study, we may not have enough number of events for a reasonable probability to stop the study early for futility or superiority of the experimental therapy. On the other hand, if it is scheduled for a late stage of the study, we may have most of the planned patient accrual already so that the interim analysis may not be able to save resources even when the analysis result indicates to stop the trial. This is likely to happen for a phase II trials with a fast patient accrual. If such is the case, we may consider a single-stage design that is discussed in the previous section.

3.1 Statistical Testing

We conduct an interim analysis at time τ which may be determined in terms of number of events or calendar time. We assume that τ is smaller than the planned accrual period a , so that we can save the number of patients if the experimental therapy does not show efficacy compared to the control.

For patient $i = 1, \dots, n_k$ in arm k , let T_{ki} denote the survival time with survivor distribution $S_k(t)$ and cumulative hazard function $\Lambda_k(t)$, and e_{ki} denote the entering time ($0 \leq e_{ki} \leq a$). C_{ki} denotes the censoring time at the final analysis with survivor function $P(C_{ki} \geq t) = G(t)$ that is defined by the accrual and missing trends and additional follow up period. For a patient who is accrued during stage 1 (i.e. $e_{ki} < \tau$), $\tilde{C}_{ki}$ has a survivor function $G_1(t) = P\{\min(\tau - e_{ki}, C_{ki}) \geq t\}$. We observe $(\tilde{X}_{ki}, \tilde{\delta}_{ki})$ at the interim analysis and (X_{ki}, δ_{ki}) at the final analysis, where $\tilde{X}_{ki} = \min(T_{ki}, \tilde{C}_{ki})$, $\tilde{\delta}_{ki} = I(T_{ki} \leq \tilde{C}_{ki})$, $X_{ki} = \min(T_{ki}, C_{ki})$, and $\delta_{ki} = I(T_{ki} \leq C_{ki})$. We define at-risk processes $\tilde{Y}_{ki}(t) = I(\tilde{X}_{ki} \geq t)$ and $Y_{ki}(t) = I(X_{ki} \geq t)$, and event processes $\tilde{N}_{ki}(t) = \tilde{\delta}_{ki} I(\tilde{X}_{ki} \leq t)$ and $N_{ki}(t) = \delta_{ki} I(X_{ki} \leq t)$. Define $\tilde{Y}_k(t) = \sum_{i=1}^{n_k} \tilde{Y}_{ki}(t)$, $\tilde{Y}(t) = \tilde{Y}_1(t) + \tilde{Y}_2(t)$, $Y_k(t) = \sum_{i=1}^{n_k} Y_{ki}(t)$, $Y(t) = Y_1(t) + Y_2(t)$, $\tilde{N}_k(t) = \sum_{i=1}^{n_k} \tilde{N}_{ki}(t)$, $\tilde{N}(t) = \tilde{N}_1(t) + \tilde{N}_2(t)$, $N_{ki}(t) = \sum_{i=1}^{n_k} N_{ki}(t)$, and $N(t) = N_1(t) + N_2(t)$.

Let $\tilde{n} = \sum_{i=1}^n I(e_{ki} \leq \tau)$ denote the number of patients who are entered before the interim analysis ($\tilde{n} \leq n$). For an accrual rate r , we have $\tau \approx \tilde{n}/r$. Test statistics at the interim and final analyses are calculated as

$$W_1 = \frac{1}{\sqrt{\tilde{n}}} \int_0^\infty \frac{\tilde{Y}_1(t) \tilde{Y}_2(t)}{\tilde{Y}(t)} \{d\tilde{\Lambda}_1(t) - d\tilde{\Lambda}_2(t)\}$$

and

$$W = \frac{1}{\sqrt{n}} \int_0^\infty \frac{Y_1(t) Y_2(t)}{Y(t)} \{d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)\},$$

respectively. Here, $\tilde{\Lambda}_k(t) = \int_0^t \tilde{Y}_k(s)^{-1} d\tilde{N}_k(s)$ and $\hat{\Lambda}_k(t) = \int_0^t Y_k(s)^{-1} dN_k(s)$ are the Nelson-Aalen (Nelson, 1969, Aalen, 1978) estimate of $\Lambda_k(t)$ from the data at the interim analysis and the final analysis, respectively. The censoring time at the interim analysis is denoted as $\tilde{C}_{ki} = \max\{\min(\tau - e_{ki}, C_{ki}), 0\}$.

For large sample sizes at the interim and final analyses, the null distribution of (W_1, W) is approximately bivariate normal with means 0, variances and covariance that can be approximated by

$$\text{var}(W_1) = \hat{\sigma}_1^2 = \frac{1}{\tilde{n}} \int_0^\infty \frac{\tilde{Y}_1(t) \tilde{Y}_2(t)}{\tilde{Y}(t)^2} d\tilde{N}(t),$$

$$\text{var}(W) = \hat{\sigma}^2 = \frac{1}{n} \int_0^\infty \frac{Y_1(t) Y_2(t)}{Y(t)^2} dN(t),$$

and $\text{cov}(W) = \hat{\sigma}_1^2$, respectively, see e.g. Tsiatis (1982).

For the patients who enter the study after τ , i.e. $e_{ki} > \tau$, their survival times are censored at time 0 at the interim analysis (i.e. $\tilde{X}_{ki} = 0$ and $\tilde{\delta}_{ki} = 0$), so that they make no contributions to W_1 and $\hat{\sigma}_1^2$. A two-stage trial using the log-rank test is conducted as follows.

Two-Stage Phase II Trial

Design stage: Specify α and an interim analysis time and an early stopping values c_f and c_u for futility and efficacy, respectively. For a two-stage design with a futility stopping only, we set $c_u = \infty$.

Stage 1: If $W_1/\hat{\sigma}_1 < c_f$, then reject the experimental therapy (arm 2) and stop the trial for futility. If $W_1/\hat{\sigma}_1 > c_u$, then reject the standard therapy (arm 1) and stop the trial for superiority. Otherwise, proceed to Stage 2.

Stage 2: If $W/\hat{\sigma} < c$, then accept the experimental therapy. Here, critical value c satisfies

$$\alpha = P\left(\frac{W_1}{\hat{\sigma}_1} > c_u \mid H_0\right) + P\left(c_f < \frac{W_1}{\hat{\sigma}_1} \leq c_u, \frac{W}{\hat{\sigma}} > c \mid H_0\right),$$

Two-Stage Phase II Trial

which can be approximated by

$$\alpha = \bar{\Phi}(c_u) + \int_c^\infty \phi(z) \left\{ \Phi\left(\frac{c_u - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) - \Phi\left(\frac{c_l - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) \right\} dz,$$

where $\hat{\rho} = \sqrt{\hat{\sigma}_1 / \hat{\sigma}}$.

When the study proceeds to the second stage, we do not need to save the raw data at the interim analysis, but only $\hat{\sigma}_1$. Wieand, Schroeder and O'Fallon (1994) propose a two-stage design with an interim futility test when 50% of the events that are expected at the final analysis are observed. They assume that the accrual period is long enough, compared to the median survival time, so that the interim analysis can be conducted during accrual period. They propose an early termination when the estimated hazard rate for the experimental arm is larger than that of the control arm. This is approximately equivalent to using $c_l = 0$ and $c_u = \infty$ in our two-stage design. Readers may read Pampallona and Tsiatis (1994) and Lachin (2005) about general group sequential futility testing methods.

3.2 Sample Size Calculation

Let $p_k = n_k/n$ ($p_1 + p_2 = 1$) denote the allocation proportion for arm k . At first we derive a power function given τ , c_k and c_u together with accrual period a , follow-up period b , $\Lambda_k(t)$ for $k = 1, 2$ under H_1 and $(\alpha, 1 - \beta)$. An interim analysis time τ may be determined in terms of calendar time or observed number of events, but at the design stage we assume that it is determined as a calendar time. If we want to specify it in terms of number of events, we can convert it to a calendar time based on the expected accrual rate and specified survival distributions at the design stage. We choose values for c_l and c_u depending on how aggressively we want to screen out an ineffective or very effective experimental therapy at the interim analysis.

The power function is given as

$$1 - \beta = P\left(\frac{W_1}{\hat{\sigma}_1} > c_u | H_1\right) + P\left(c_l < \frac{W_1}{\hat{\sigma}_1} \leq c_u, \frac{W}{\hat{\sigma}} > c | H_1\right).$$

In order to derive a power function, we have to calculate c for a specified type I error rate α , i.e.

$$\alpha = P\left(\frac{W_1}{\hat{\sigma}_1} > c_u | H_0\right) + P\left(c_l < \frac{W_1}{\hat{\sigma}_1} \leq c_u, \frac{W}{\hat{\sigma}} > c | H_0\right),$$

although it may be recalculated at the final analysis using the collected survival data. Hence, for a power calculation, we need to derive the limits of $\hat{\sigma}_1^2$ and $\hat{\sigma}^2$ under both H_0 and H_1 , and $E(W_1)$, $E(W)$, $\text{var}(W_1)$ and $\text{var}(W)$ under H_1 . By the independent increment of the log-rank

test statistic, the correlation coefficient between W_1 and W is $\sqrt{\text{var}(W_1)/\text{var}(W)}$ under both H_0 and H_1 .

We derive following asymptotic results using Lemma 4.1 of Fleming and Harrington (1991).

Under H_0 , we have $E(W_1) = E(W) = 0$. Furthermore, for large n , we can show that $\hat{\sigma}_1^2$ and $\hat{\sigma}^2$ converge

$$v_1 = -p_1 p_2 \int_0^\infty G_1(t) dS_2(t)$$

and

$$v = -p_1 p_2 \int_0^\infty G(t) dS_2(t),$$

respectively, under H_0 . Note also that $\text{var}(W_1) = v_1$ and $\text{var}(W) = v$ under H_0 . Hence, by independent increment of the log-rank statistic, $\text{corr}(W_1, W)$ is $\rho_0 = \sqrt{v_1/v}$ under H_0 . We need these asymptotic results under H_0 to calculate c .

Similarly we derive following asymptotic results under H_1 . Under H_1 , we have

$E(W_1) = \sqrt{n}\omega_1$ and $E(W) = \sqrt{n}\omega$, where

$$\omega_1 = p_1 p_2 \int_0^\infty \frac{G_1(t) S_1(t) S_2(t)}{p_1 S_1(t) + p_2 S_2(t)} \{d\Lambda_1(t) - d\Lambda_2(t)\}$$

and

$$\omega = p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)}{p_1 S_1(t) + p_2 S_2(t)} \{d\Lambda_1(t) - d\Lambda_2(t)\}.$$

Furthermore, $\hat{\sigma}_1^2$ and $\hat{\sigma}^2$ converge to

$$\sigma_{01}^2 = p_1^2 p_2 \int_0^\infty \frac{G_1(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1 p_2^2 \int_0^\infty \frac{G_1(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t)$$

and

$$\sigma_0^2 = p_1^2 p_2 \int_0^\infty \frac{G(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1 p_2^2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t)$$

respectively. The variances of W_1 and W are given as

$$\sigma_{11}^2 = p_1 p_2^2 \int_0^\infty \frac{G_1(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1^2 p_2 \int_0^\infty \frac{G_1(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t)$$

and

$$\sigma_1^2 = p_1 p_2^2 \int_0^\infty \frac{G(t) S_1(t) S_2(t)^2}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_1(t) + p_1^2 p_2 \int_0^\infty \frac{G(t) S_1(t)^2 S_2(t)}{\{p_1 S_1(t) + p_2 S_2(t)\}^2} d\Lambda_2(t)$$

respectively under H_1 . By independent increment of the log-rank statistic, $\text{corr}(W_1, W)$ is given as $\rho_1 = \sigma_{11}/\sigma_1$.

In summary, $(W_1/\hat{\sigma}_1, W/\hat{\sigma})$ is asymptotically distributed as $N(0, \Sigma_0)$ under H_0 and $N(\mu, \Sigma_1)$ under H_1 , where

$$\mu = \begin{pmatrix} \sqrt{\hat{n}}\omega_1/\sigma_{11} \\ \sqrt{\hat{n}}\omega/\sigma_1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} \sigma_{11}^2/\sigma_{01}^2 & \rho_1\sigma_{11}\sigma_1/\sigma_{01}\sigma_0 \\ \rho_1\sigma_{11}\sigma_1/\sigma_{01}\sigma_0 & \sigma_1^2/\sigma_0^2 \end{pmatrix}.$$

If (X, Y) is a bivariate normal random vector with means μ_x and μ_y , variances σ_x^2 and σ_y^2 , and correlation coefficient ρ , then it is well known that the conditional distribution of X given $Y = y$ is normal with mean $\mu_x + (\rho\sigma_x/\sigma_y)(y - \mu_y)$ and variance $\sigma_x^2(1 - \rho^2)$. This result simplifies the calculation of type I error rate and power below. For example, given design parameters $(\alpha, 1 - \beta, p_1, r, b, \Lambda_1(t), \Lambda_2(t), \tau, c_l, c_u)$, $(X, Y) = (W_1/\hat{\sigma}_1, W/\hat{\sigma})$ is asymptotically $N(0, \Sigma_0)$ under H_0 . So, in this case, $Y \sim N(0, 1)$ and the conditional distribution of X given $Y = y$ is $N(\rho_0 y, 1 - \rho_0^2)$. Using this result, we obtain the stage 2 critical value c by solving the equation

$$\alpha = P(X \geq c_u) + P(c_l < X < c_u, Y \geq c) = \bar{\Phi}(c_u) + \int_c^\infty \phi(y) \left\{ \Phi\left(\frac{c_u - \hat{\rho}y}{\sqrt{1 - \hat{\rho}^2}}\right) - \Phi\left(\frac{c_l - \hat{\rho}y}{\sqrt{1 - \hat{\rho}^2}}\right) \right\} dy.$$

If the interim analysis time τ and the stopping values (c_l, c_u) are reasonably chosen, the power of a two-stage design is not be much different from that of the corresponding single-stage design. So, when searching for the required accrual period (or sample size) of a two-stage design, we may start from that of the corresponding single-stage design. Assuming an accrual pattern with a constant accrual rate r , the design procedure of a two-stage design can be summarized as follows.

Design of a Two-Stage Trial

- 1 Given $(\alpha, 1 - \beta, p_1, r, b, \Lambda_1(t), \Lambda_2(t))$, calculate the sample size n and accrual period a_0 required for a single-stage design.

Design of a Two-Stage Trial

2 Determine an interim analysis time τ during the accrual period a_0 of the chosen single-stage design (i.e. $\tau < a_0$) and the stopping values c_l and c_u at the interim analysis.

3 Then, the accrual period required for a two-stage design is obtained around a_0 as follows:

A. At $a = a_0$ (note that $\tilde{n} = r\tau$ and $n = ra_0$),

- Obtain c by solving equation

$$\alpha = \bar{\Phi}(c_u) + \int_c^\infty \phi(z) \left\{ \Phi\left(\frac{c_u - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) - \Phi\left(\frac{c_l - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) \right\} dz.$$

- Given $(\tilde{n}, n, c_l, c, \alpha)$, calculate

$$\text{power} = \bar{\Phi}(\bar{c}_u) + \int_{\bar{c}}^\infty \phi(z) \left\{ \Phi\left(\frac{\bar{c}_u - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) - \Phi\left(\frac{\bar{c}_l - \hat{\rho}z}{\sqrt{1 - \hat{\rho}^2}}\right) \right\} dz,$$

where

$$\bar{c}_l = \frac{\sigma_{01}}{\sigma_{11}}(c_l - \frac{\omega_1 \sqrt{\tilde{n}}}{\sigma_{01}}), \bar{c}_u = \frac{\sigma_{01}}{\sigma_{11}}(c_u - \frac{\omega_1 \sqrt{\tilde{n}}}{\sigma_{01}}) \text{ and } \bar{c} = \frac{\sigma_0}{\sigma_1}(c - \frac{\omega \sqrt{n}}{\sigma_0})$$

B. If the power is smaller than $1 - \beta$, increase a slightly, and repeat (A) until the power is close enough to $1 - \beta$. We may change the interim analysis time τ at this step too.

At the design stage, we may want to calculate the probabilities of early termination (PET) under H_0 and under H_1 by

$$\text{PET}_0 = P(W_1/\hat{\sigma}_1 \leq c_l | H_0) + P(W_1/\hat{\sigma}_1 > c_u | H_0) = \Phi(c_l) + \bar{\Phi}(c_u)$$

and

$$\text{PET}_1 = P(W_1/\hat{\sigma}_1 \leq c_l | H_1) + P(W_1/\hat{\sigma}_1 > c_u | H_1) = \Phi(\bar{c}_l) + \bar{\Phi}(\bar{c}_u).$$

PET_0 and PET_1 should not be too small in order for an interim futility test not to be trivial.

When we consider the two-stage design for futility only, i.e., $c_u = \infty$, the probabilities of early termination become $\text{PET}_0 = \Phi(c_l)$ and $\text{PET}_1 = \Phi(\bar{c}_l)$. In such case, while PET_0 should not be too small in order for an interim futility test to be of worth, PET_1 should not be too large to avoid early rejection of an efficacious therapy with immature data.

The expected sample size (EN) under H_h ($h = 0, 1$) is given as

$$\text{EN}_h = \tilde{n} \times \text{PET}_h + n \times (1 - \text{PET}_h).$$

The average expected sample size is also given as $\overline{\text{EN}} = (\text{EN}_0 + \text{EN}_1)/2$.

Under Uniform Accrual and Exponential Survival Models—Suppose that the survival distributions are exponential with hazard rates λ_1 and λ_2 in arms 1 and 2, respectively. If patients are accrued at a constant rate during period a and followed for an additional period of b , and the interim analysis is taken place before completion of accrual, i.e. $\tau < a$, then the censoring distribution at the interim analysis is $U(0, \tau)$ and that after the second stage is $U(b, a + b)$ with survivor functions

$$G_1(t) = \begin{cases} 1 & \text{if } t \leq 0 \\ 1 - t/\tau & \text{if } 0 < t \leq \tau \\ 0 & \text{if } t > \tau \end{cases}$$

and

$$G(t) = \begin{cases} 1 & \text{if } t \leq b \\ 1 - (t - b)/a & \text{if } b < t \leq a + b \\ 0 & \text{if } t > a + b \end{cases}$$

respectively. Since $\tau < a$, $G_1(t)$ is free of a . Note that we only assume administrative censoring. If loss to follow-up is expected, then we may incorporate it in the calculation if its distribution is given, or we may increase the final sample size by the expected proportion of loss to follow-up.

Under these distributional assumptions, σ_0^2 , σ_1^2 and ω are the same as those in (4), (5) and (6), respectively, and

$$v_1 = p_1 p_2 \left\{ 1 - \frac{1}{\tau \lambda_1} (1 - e^{-\lambda_1 \tau}) \right\}$$

$$v = p_1 p_2 \left\{ 1 - \frac{1}{a \lambda_1} e^{-b \lambda_1} (1 - e^{-\lambda_1 a}) \right\}$$

$$\sigma_{01}^2 = \frac{p_1^2 p_2 \lambda_1}{\tau} \int_0^\tau \frac{(\tau - t) e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt + \frac{p_1 p_2^2 \lambda_2}{\tau} \int_0^\tau \frac{(\tau - t) e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt$$

$$\sigma_{11}^2 = \frac{p_1 p_2^2 \lambda_1}{\tau} \int_0^\tau \frac{(\tau - t) e^{-(\lambda_1 + 2\lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt + \frac{p_1^2 p_2 \lambda_2}{\tau} \int_0^\tau \frac{(\tau - t) e^{-(2\lambda_1 + \lambda_2)t}}{(p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t})^2} dt$$

and

$$\omega_1 = \frac{p_1 p_2 (\lambda_1 - \lambda_2)}{\tau} \int_0^\tau \frac{(\tau - t) e^{-(\lambda_1 + \lambda_2)t}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}} dt.$$

We use a numerical method to calculate these integrals.

Example 2: From Example 1, a single-stage randomized controlled trial requires $n = 102$ under the design setting. $(\alpha, 1 - \beta, \lambda_1, \lambda_2, r, b, p_1) = (0.1, 0.9, 1.609, 0.916, 60, 1, 1/2)$. Under the same design setting, a two-stage trial with an interim analysis at almost one year with $(c_l, c_u) = (-0.88, 2.18)$ requires a maximal sample size of $n = 102$, which is the same number of patients with that for the single-stage design. At the interim and final analyses, we expect 20 and 82 events, respectively, under H_1 . The probabilities of early termination are given as $PET_0 = 0.2041$ and $PET_1 = 0.4173$. From $B = 5,000$ simulations, this two-stage design has an empirical type I error of 10.4% and power of 92.5%, which are very close to the nominal $\alpha = 10\%$ and $1 - \beta = 90\%$, respectively.

4 Optimal Two-Stage Designs

We propose some optimal two-stage designs for given $(\alpha, 1 - \beta, r, b, \Lambda_1(t), \Lambda_2(t))$. Given (r, b) , a candidate two-stage design specified by $(n, \tilde{n}, c_l, c_u, c)$ has a type I error rate of α for $\Lambda_1(t) = \Lambda_2(t)$ and a power no smaller than $1 - \beta$ for $\Lambda_1(t) > \Lambda_2(t)$. We consider two optimality criteria, one to minimize the expected sample size and the other to minimize the maximal sample size. By specifying an accrual rate r , we assume a uniform accrual trend, but we can extend the following results to any accrual pattern.

4.1 Minimax and Optimal Designs

Among the candidate two-stage designs with futility stopping only, we define the *optimal* design as the one minimizing the expected sample size under H_0 , EN_0 . Simon (1989) used this criterion to find an optimal two-stage design with a binary outcome. For two-stage designs with both futility and efficacy stopping, we propose to choose the *optimal* design by minimizing the average expected sample size, $\overline{EN}$. Chang et al. (1987) use a similar criterion for three-stage phase II trials with a binary outcome. We also define the *minimax* design as the one minimizing the maximal sample size n . For a given n , there may be multiple two-stage designs satisfying the $(\alpha, 1 - \beta)$ condition. The minimax design has the smallest expected sample size among them.

Through our experience from numerical studies, we have found that the final sample size of the minimax design is not very different from the sample size of the corresponding single-stage design, a reasonable interim analysis should be conducted when the survival data are somewhat matured, and the rejection value for the futility bound c_l at the first stage is not very different from 0. Given $(\alpha, 1 - \beta, p_1, r, b, \Lambda_1(t), \Lambda_2(t))$, let n_0 denote the sample size of the single-stage design. An efficient computational procedure to identify the minimax and optimal designs with interim stopping for both futility and efficacy can be summarized as follows. Calculation to identify the minimax and optimal designs with interim stopping for futility can be obtained by setting $c_u = \infty$.

Search for Optimal Designs with both Futility and Efficacy Stopping Values

(A) Specify $(\alpha, 1 - \beta, p_1, r, b, \Lambda_1(t), \Lambda_2(t))$.

(B) Find the sample size n_0 for the single-stage design.

(C) For $n(n_0 - 3)$, while changing the values of $(\tilde{n}, c_f, c_u)$, calculate c and power.

- i. If the power is smaller than $1 - \beta$, then go to the next combination of $(\tilde{n}, c_f, c_u)$.
- ii. If the power is larger than or equal to $1 - \beta$, then calculate $\overline{EN}$.
 - If this $\overline{EN}$ is smaller than the minimum of EN (or $\overline{EN}$) among the all candidate designs we have gone through so far, then save the current $(n, \tilde{n}, c_f, c_u, c)$ together with $\overline{EN}$.
 - Otherwise, go to the next combination of $(\tilde{n}, c_f, c_u)$.

(ii) If we have gone through all possible combinations of $(\tilde{n}, c_f, c_u)$, then save $(\tilde{n}, c_f, c_u, \overline{EN})$ of the design with the smallest $\overline{EN}$, and continue to the next n .

(D) Among the designs saved during procedure (C), the one with the smallest n is the minimax design, and the one with the smallest $\overline{EN}$ is the optimal design.

The search for the optimal designs with a futility stopping value only (i.e. $c_u = \infty$) is conducted using the above algorithm with $\overline{EN}$ replaced by EN_0 .

Table 1 lists the single-stage, and minimax and optimal two-stage designs with a futility stopping only under various design settings. We consider 1-1 allocation (i.e. $p_1 = p_2 = 1/2$), an annual accrual rate of $r = 60$ or 90 patients, a follow-up period of $b = 1$ year and an exponential survival distribution with an annual hazard rate $\lambda_2 = 0.9$ for the experimental arm, corresponding to a median survival of about nine months. The interim analysis time is given by $\tau = \tilde{n}/r$. Under H_1 , we assume a hazard rate of $\lambda_1/\lambda_2 = 1.4, 1.5, 1.6$ or 1.7 . We also consider $(\alpha, 1 - \beta) = (0.05, 0.8), (0.1, 0.85)$, or $(0.1, 0.9)$. For all of these 2-stage designs, PET_1 is small because they are not intended to stop the trial early when the experimental treatment is efficacious. For given $(\alpha, 1 - \beta)$, c is quite free of λ_1/λ_2 , whereas c_f changes in λ_1/λ_2 . The maximal sample size for a minimax 2-stage design is almost identical to the sample size of the corresponding single-stage design, but the expected sample size under H_0 for the former is smaller than the sample size of the latter. Compared to the minimax designs, the optimal designs conduct the interim analysis earlier (i.e. with a smaller $\tilde{n}$ or $\tilde{D}$) more aggressively (i.e. using a larger c_f), so that they have a smaller EN_0 in spite of a larger n . We can compare the study periods between the single-stage design and the minimax and optimal two-stage designs for a given design setting. The study period of the single-stage design is $a + b$. Since the maximal sample size for a minimax two-stage design is almost identical to the sample size of the corresponding single-stage design, the expected study period of the minimax design under H_0 is $\tau \times PET_0 + (a + b) \times (1 - PET_0)$. Hence, the difference in (expected) study period between the single-stage design and the minimax 2-stage design is $(a + b - \tau) \times PET_0$. Since the optimal design has a smaller expected sample size than the minimax design, we save more expected study period by using the optimal 2-stage design.

Table 2 reports the minimax and optimal 2-stage designs with both futility and efficacy stopping values under the same design settings as in Table 1. The maximal sample size for a

minimax 2-stage design is similar to the sample size of the corresponding single-stage design, but the former can be smaller than the latter by as many as 8 when $(\gamma, \alpha, 1 - \beta, \delta) = (60, 0.1, 0.9, 1.5)$. Even though the optimal designs with both futility and efficacy stopping values have larger maximal sample sizes than those with a futility stopping value only, their $\overline{EN}$ is smaller than EN_0 of the latter. As in Table 1, the optimal designs tend to conduct the interim analysis earlier and more aggressively (i.e. using a larger c_f and a smaller c_u) than the minimax designs. PET_0 is larger than PET_1 if $\alpha < \beta$, but they are similar if $\alpha = \beta$. Given $(\alpha, 1 - \beta)$, c is quite free of δ , while (c_f, c_u) change in δ .

4.2 Implementation of Two-Stage Designs

In the previous section, the minimax and optimal designs are selected for a specified design setting including accrual rate. Once the study is open, however, the realized accrual pattern may be different from that specified at the design. In this case, a two-stage analysis based on the prespecified times τ and $a + b$ may result in the variances of W_1 and W and their correlation coefficient different from those specified at the design stage. This will possibly leads to change in the performance, like power, the two-stage log-rank test. Noting that the power of the log-rank test depends on the number of events rather than the number of patients, we propose to choose the interim and the final analysis times based on the expected numbers of events that are calculated based on the design setting. Let $\tilde{D}$ and D denote the expected number of events under H_1 at the interim and final analysis times, respectively, based on the design parameter values, i.e.

$$\tilde{D} = \tilde{n} \{1 - (1 - e^{-\lambda_1 \tau}) / \lambda_1 \tau\}$$

and

$$D = n \{1 - e^{-\lambda_1 b} (1 - e^{-\lambda_1 a}) / \lambda_1 a\},$$

where $\tau = \tilde{n}/r$ and $a = n/r$. If $\lambda_1 \approx \lambda_2$, the asymptotic distribution of $(W_1/\hat{\sigma}_1, W/\hat{\sigma})$ is a bivariate normal with means 0, variances 1 and correlation coefficient

$\hat{\rho} = \sqrt{\hat{\sigma}_1/\hat{\sigma}} = \sqrt{n\tilde{D}/\tilde{n}D}$. Hence, as far as the real accrual trend is close to the expected one at the design, the two-stage design specified by $(\alpha, \lambda_1, \lambda_2, c_f, c_u, c, r, b, \tilde{n})$ will be identical to that specified by $(\alpha, \lambda_1, \lambda_2, c_f, c_u, c, \tilde{D}, D)$. Recall that c_f and c_u are fixed by the design and c is recalculated for a type I error rate of α at the analysis reflecting the realized accrual trend. So, design parameters $(\tilde{n}, c, b)$ will be used as just references during the trial.

In summary we propose to conduct a study with a two-stage design as follows. We explain for the case when we stop for both futility and efficacy. Conducting study with a two-stage design with interim stopping for futility can be obtained by setting $c_u = \infty$.

A Two-Stage Design based on the Number of Events

- Specify $(\lambda_1, \lambda_2, \alpha, 1 - \beta)$.

A Two-Stage Design based on the Number of Events

- Choose a two-stage design $(\tilde{n}, n, c_l, c_u, c, b)$.
- Calculate $\tilde{D}$ and D based on the selected two-stage design.
- Accrue n patients and follow them until D events are observed unless the study is stopped early by the stage 1 analysis.

Stage 1: When $\tilde{D}$ events are observed, conduct the stage 1 analysis by calculating W_1 and $\hat{\sigma}_1^2$, and stop the trial either for futility if $W_1/\hat{\sigma}_1 < c_l$ or for superiority if $W_1/\hat{\sigma}_1 > c_u$. Otherwise, proceed to stage 2.

Stage 2: Conduct the final analysis when D events are observed by calculating W , $\hat{\sigma}^2$ and critical value c' based on $\hat{p} = \hat{\sigma}_1/\hat{\sigma}$. We reject the study therapy if $W/\hat{\sigma} > c'$.

4.3 Simulation Study

The two-sample log-rank test is based on large sample approximation. To demonstrate the small sample performance of our two-stage design, we conduct simulations under the minimax and the optimal two-stage designs listed in Tables 1–2. Given (λ_h, r, b) , we generate 5,000 simulation samples of size n , and conduct two-stage analysis defined by $(\tilde{D}, D, c_l, c_u)$ as presented in the previous section. By calculating the p-value for testing, we do not have to calculate c for each simulated sample. We consider the same design settings of Tables 1 and 2, but assume that the real accrual rate could be 60 or 90. So, a real accrual rate of 60 (90) with $r = 90$ ($r = 60$) means an overestimation (underestimation) of the accrual rate at the design stage. Tables 3–4 report the empirical size and power of the two-stage designs in Tables 1 and 2. Note that the empirical size and power of the two-stage trials based on the number of events are closely match with their nominal counterparts whether the accrual rate is exactly specified or not. We observe similar results whether we use an early stopping for futility only (Table 3) or one for both futility and efficacy (Table 4).

5 Discussion

A simple approach to evaluating results of a clinical trial is to plan just one statistical analysis at the end of the study using a fixed-sample size design. Its planning and conduction are easy and the methods for estimation are well established. However, this single stage approach is less appropriate when data become available sequentially. Such is the case in studies on chronic diseases, like cancer, in which recruitment lasts many years, so that outcomes are observed from earlier patients while the accrual is still ongoing. In such situations there might be ethical, practical and economic reasons for monitoring the data during the study period.

In this paper we have considered randomized controlled clinical trials where the primary endpoint is a right censored failure time. We restricted our attention to two-stage designs that can be useful for phase II clinical trials because more than two stages are difficult to manage in practice with no additional gain. At first, we have proposed two optimality criteria for two-stage designs with an early stopping only due to futility: one to minimize the maximal sample size for a minimax design and the other to minimize the estimated sample size when the study therapy is not efficacious for an optimal design. These criteria were considered by Simon (1989) for binary outcomes, such as tumor response. We also have investigated two-

stage designs with an early stopping due to inefficacy or superiority for which the optimal design is to minimize the average of the expected sample sizes under null and alternative hypotheses. Chang et al (1987) have proposed this criteria for binary outcomes. Compared to single-stage designs, these two-stage designs not only save the expected sample size, but shorten the total study period by curtailing the followup period.

R programs to search for minimax and optimal designs will be provided upon request.

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989; 10:1–10. [PubMed: 2702835]
2. Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics*. 1987; 43:865–874. 1987. [PubMed: 3427171]
3. Aalen OO. Nonparametric inference for a family of counting processes. *Annals of Statistics*. 1978; 6:701–726.
4. Andersen PK, Borgan O, Gill RD, Kidding N. Linear nonparametric tests for comparison of counting processes with application to censored survival data (with discussion). *International Statistical Review*. 1982; 50:219–258.
5. Case LD, Morgan TM. Duration of accrual and follow-up for two-stage clinical trials. *Lifetime Data Analysis*. 2001; 7:21–37. [PubMed: 11280845]
6. Fleming, TR., Harrington, DP. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
7. George SL, Desu MM. Planning the size and duration of a trial studying the time to some critical event. *Journal of Chronic Disease*. 1973; 27:15–24.
8. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*. 1981; 2:93–113. [PubMed: 7273794]
9. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*. 2005; 24:2747–2764. [PubMed: 16134130]
10. Lakatos E. Sample Sizes based on the Log-Rank Statistic in Complex Clinical Trials. *Biometrics*. 1977; 64:156–160.
11. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*. 1988; 44:229–241. [PubMed: 3358991]
12. Nelson W. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*. 1969; 1:27–52.
13. Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*. 1994; 42:19–35.
14. Pasternack BS, Gilbert HS. Planning the duration of long-term survival time studies designed for accrual by cohorts. *Journal of Chronic Disease*. 1971; 24:13–24.
15. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*. 1972; 135:185–206.
16. Rubinstein L, Gail M, Santner T. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease*. 1981; 27:15–24.
17. Schoenfeld DA. Sample size formula for the proportional hazards regression model. *Biometrics*. 1983; 39:499–503. [PubMed: 6354290]
18. Schoenfeld DA, Tsiatis AA. A modified log rank test for highly stratified data. *Biometrika*. 1987; 74:167–175.
19. Slud EV, Wei LJ. Two-Sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of American Statistical Society*. 1982; 77:862–868.
20. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of American Statistical Society*. 1982; 77:855–861.

21. Wieand S, Schroeder G, O'Fallon JR. Stopping when the experimental regimen does not appear to help. *Statistics in Medicine*. 1994; 13:1453–1458. [PubMed: 7973224]
22. Yateman NA, Skene AM. Sample size for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine*. 1992; 11:1103–1113. [PubMed: 1496198]

Table 1

Single-stage, and minimax and optimal two-stage designs for futility stopping with 1-1 allocation, $(\alpha, 1 - \beta) = (0.05, 0.8)$, $(0.1, 0.85)$ or $(0.1, 0.9)$, $\lambda_2 = 0.9$, hazard ratio $= \lambda_1/\lambda_2 = 1.5, 1.6$ or 1.7 , accrual rate $r = 60$ or 90 patients per year, and the follow-up time $b = 1$ year

Single-Stage			Minimax Design								Optimal Design									
<i>n</i>	<i>D</i>	<i>n</i>	<i>ñ</i>	<i>c_t</i>	<i>c</i>	<i>D</i> ₁	<i>D</i>	PET ₀	PET ₁	EN ₀	<i>n</i>	<i>ñ</i>	<i>c_t</i>	<i>c</i>	<i>D</i> ₁	<i>D</i>	PET ₀	PET ₁	EN ₀	
When the accrual rate is 60 patients per year:																				
α = 0.05, 1 − β = 0.8																				
1.5	173	155	174	112	−0.320	1.641	65	156	.374	.029	160.7	187	95	0.140	1.621	50	169	.556	.104	155.8
1.6	133	117	133	87	−0.460	1.641	44	117	.323	.025	124.5	139	81	0.050	1.621	40	123	.520	.084	120.9
1.7	107	93	107	72	−0.455	1.641	34	93	.325	.028	100.5	111	68	−0.030	1.621	31	97	.488	.076	98.5
α = 0.1, 1 − β = 0.85																				
1.5	153	136	154	115	−0.495	1.279	67	137	.310	.018	146.8	157	93	−0.365	1.270	49	139	.358	.042	143.4
1.6	118	103	118	104	−0.180	1.270	59	102	.429	.028	114.0	120	75	−0.485	1.270	35	104	.314	.034	111.4
1.7	95	82	95	80	−0.385	1.270	40	81	.350	.024	91.6	97	63	−0.500	1.270	27	83	.309	.035	90.5
α = 0.1, 1 − β = 0.9																				
1.5	185	167	185	159	−0.495	1.279	107	167	.310	.006	180.1	190	120	−0.165	1.270	71	171	.434	.034	172.8
1.6	142	127	142	125	−0.495	1.279	77	125	.310	.007	138.6	145	94	−0.325	1.270	50	129	.373	.027	134.1
1.7	114	99	115	101	−0.190	1.270	57	100	.425	.018	111.0	117	74	−0.495	1.270	35	102	.310	.023	108.9
When the accrual rate is 90 patients per year:																				
α = 0.05, 1 − β = 0.8																				
1.5	179	155	180	131	−0.495	1.641	66	155	.310	.018	171.0	190	104	−0.060	1.621	46	165	.476	.082	165.0
1.6	137	116	138	110	−0.495	1.641	51	116	.310	.018	132.5	145	79	−0.240	1.621	29	123	.405	.071	128.4
1.7	110	92	111	89	−0.500	1.641	37	92	.309	.021	106.7	114	74	−0.265	1.621	27	95	.396	.057	104.2
α = 0.1, 1 − β = 0.85																				
1.5	159	136	159	148	−0.195	1.270	79	136	.423	.025	155.9	162	105	−0.495	1.270	46	138	.310	.034	151.6
1.6	122	102	122	115	−0.395	1.270	54	102	.346	.019	119.9	125	88	−0.495	1.270	36	104	.310	.033	117.9
1.7	98	81	98	94	−0.485	1.270	41	81	.314	.018	96.8	100	78	−0.500	1.270	30	82	.309	.031	95.7
α = 0.1, 1 − β = 0.9																				
1.5	191	167	192	156	−0.500	1.279	87	167	.309	.010	185.4	195	131	−0.370	1.270	66	169	.356	.025	182.0
1.6	147	125	147	129	−0.290	1.270	65	125	.386	.017	143.1	149	103	−0.495	1.270	46	127	.310	.021	141.1
1.7	118	99	118	111	−0.400	1.270	53	99	.345	.013	116.3	120	91	−0.495	1.270	38	101	.310	.020	114.8

Table 2

Minimax and optimal two-stage designs for futility and superiority stopping with 1-1 allocation, $(\alpha, 1 - \beta) = (0.05, 0.8)$, $(0.1, 0.9)$, $\lambda_2 = 0.9$, hazard ratio $= \lambda_1/\lambda_2 = 1.5, 1.6$ or 1.7 , accrual rate $r = 60$ or 90 patients per year, and the follow-up time $b = 1$ year

Minimax Design										Optimal Design									
n	$\tilde{n}$	(c_b, c_u)	c	D_1	D	PET_0	PET_1	$\overline{EN}$	n	$\tilde{n}$	(c_b, c_u)	c	D_1	D	PET_0	PET_1	$\overline{EN}$		
When the accrual rate is 60 patients per year:																			
$\alpha = 0.05, 1 - \beta = 0.8$																			
1.5	172	145	(0.240, 2.500)	1.641	94	155	.601	.343	158.7	191	106	(0.235, 2.165)	1.680	60	173	.608	.371	148.4	
1.6	133	111	(-0.075, 2.580)	1.641	65	117	.475	.274	123.9	145	85	(0.130, 2.200)	1.680	43	129	.566	.340	117.0	
1.7	107	95	(-0.095, 2.595)	1.641	52	93	.467	.273	101.6	113	68	(-0.035, 2.515)	1.641	31	99	.492	.222	96.0	
$\alpha = 0.1, 1 - \beta = 0.85$																			
1.5	151	148	(0.030, 2.295)	1.279	97	134	.523	.408	149.0	167	98	(-0.130, 1.835)	1.328	52	149	.482	.414	135.5	
1.6	117	115	(-0.260, 2.200)	1.289	68	101	.411	.410	115.2	126	81	(-0.195, 1.955)	1.309	40	110	.448	.367	106.7	
1.7	95	90	(-0.480, 2.470)	1.279	47	81	.322	.273	92.6	101	66	(-0.340, 2.005)	1.309	29	87	.389	.323	87.6	
$\alpha = 0.1, 1 - \beta = 0.9$																			
1.5	177	170	(0.215, 2.095)	1.279	117	159	.603	.562	172.1	204	126	(0.070, 1.610)	1.387	78	186	.582	.611	157.3	
1.6	137	135	(0.030, 1.985)	1.289	87	121	.536	.588	135.0	153	92	(-0.235, 1.855)	1.328	49	136	.439	.443	125.3	
1.7	112	110	(-0.025, 1.865)	1.309	65	97	.521	.613	110.2	124	77	(-0.300, 1.810)	1.348	38	109	.417	.450	103.1	
When the accrual rate is 90 patients per year:																			
$\alpha = 0.05, 1 - \beta = 0.8$																			
1.5	176	174	(-0.165, 2.295)	1.660	102	152	.445	.414	174.7	195	117	(0.095, 2.200)	1.680	56	171	.552	.328	160.7	
1.6	138	127	(-0.055, 2.555)	1.641	64	116	.483	.279	132.8	149	94	(-0.055, 2.240)	1.680	40	127	.491	.288	126.6	
1.7	110	109	(-0.010, 2.520)	1.641	51	92	.502	.297	109.1	116	74	(-0.240, 2.565)	1.641	27	97	.410	.175	103.0	
$\alpha = 0.1, 1 - \beta = 0.85$																			
1.5	158	153	(-0.045, 2.020)	1.289	84	134	.504	.463	154.9	172	105	(-0.210, 2.090)	1.289	46	148	.435	.297	146.8	
1.6	122	121	(-0.330, 1.965)	1.309	60	102	.395	.455	121.2	130	91	(-0.435, 1.925)	1.328	38	109	.359	.342	115.6	
1.7	100	92	(-0.655, 2.015)	1.318	39	82	.278	.364	96.7	105	71	(-0.595, 1.975)	1.328	25	87	.300	.285	94.4	
$\alpha = 0.1, 1 - \beta = 0.9$																			
1.5	187	181	(-0.185, 1.875)	1.309	108	163	.457	.598	183.1	205	127	(-0.285, 1.870)	1.328	63	180	.419	.424	171.5	
1.6	147	136	(-0.490, 2.460)	1.279	71	125	.319	.322	142.7	156	100	(-0.480, 1.925)	1.328	44	133	.343	.373	135.2	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Optimal Design									
Minimax Design									
n	$\tilde{n}$	$(c^{(1)}, c_u)$	c	D_1	D	PET_0	PET_1	$\overline{EN}$	
118	117	(−0.285, 2.585)	1.270	57	99	.393	.292	117.0	125
									79
									(−0.660, 1.975)
									1.328
									30
									105
									.279
									.315
									110.7

Table 3

Simulation results for the two-stage designs with a futility stopping only as given in Table 1: Empirical type I error rate and power $(\hat{\alpha}, 1 - \hat{\beta})_r$ when the realized accrual rate $\hat{r}$ is possibly different from the specified accrual rate r and the interim and final analyses are conducted based on the number of events (D_1, D) using the stage I rejection value c_1 as determined at design

Minimax Design						Optimal Design					
$(\alpha, 1 - \beta)$	n	(D_1, D)	c_1	$(\hat{\alpha}, 1 - \hat{\beta})_{60}$	$(\hat{\alpha}, 1 - \hat{\beta})_{90}$	n	(D_1, D)	c_1	$(\hat{\alpha}, 1 - \hat{\beta})_{60}$	$(\hat{\alpha}, 1 - \hat{\beta})_{90}$	
When the specified accrual rate is $r = 60$ patients per year											
$(.05, .8)$	1.5	174	(65, 156)	-0.320	(.049, .806)	(.052, .803)	187	(50, 169)	0.140	(.047, .792)	(.052, .803)
	1.6	133	(44, 117)	-0.460	(.054, .805)	(.049, .808)	139	(40, 123)	0.050	(.055, .795)	(.047, .796)
	1.7	107	(34, 93)	-0.455	(.048, .809)	(.043, .815)	111	(31, 97)	-0.030	(.047, .804)	(.051, .807)
$(.1, .85)$	1.5	154	(67, 137)	-0.495	(.099, .869)	(.102, .856)	157	(49, 139)	-0.365	(.102, .850)	(.096, .856)
	1.6	118	(59, 102)	-0.180	(.100, .848)	(.105, .866)	120	(35, 104)	-0.485	(.104, .851)	(.103, .858)
	1.7	95	(40, 81)	-0.385	(.101, .843)	(.097, .864)	97	(27, 83)	-0.500	(.101, .850)	(.094, .864)
$(.1, .9)$	1.5	185	(107, 167)	-0.495	(.101, .909)	(.106, .894)	190	(71, 171)	-0.165	(.096, .910)	(.096, .905)
	1.6	142	(77, 125)	-0.495	(.102, .900)	(.099, .899)	145	(50, 129)	-0.325	(.109, .909)	(.110, .906)
	1.7	115	(57, 100)	-0.190	(.103, .915)	(.099, .914)	117	(35, 102)	-0.495	(.101, .900)	(.099, .910)
When the specified accrual rate is $r = 90$ patients per year											
$(.05, .8)$	1.5	180	(66, 155)	-0.495	(.052, .806)	(.049, .806)	190	(46, 165)	-0.060	(.050, .787)	(.045, .805)
	1.6	138	(51, 116)	-0.495	(.055, .807)	(.047, .809)	145	(29, 123)	-0.240	(.046, .802)	(.049, .794)
	1.7	111	(37, 92)	-0.500	(.055, .808)	(.051, .809)	114	(27, 95)	-0.265	(.055, .804)	.052, .814)
$(.1, .85)$	1.5	159	(79, 136)	-0.195	(.099, .855)	(.110, .858)	162	(46, 138)	-0.495	(.104, .850)	(.103, .846)
	1.6	122	(54, 102)	-0.395	(.095, .861)	(.105, .866)	125	(36, 104)	-0.495	(.095, .854)	(.102, .849)
	1.7	98	(41, 81)	-0.485	(.095, .855)	(.103, .861)	100	(30, 82)	-0.500	(.104, .861)	(.095, .858)
$(.1, .9)$	1.5	192	(87, 167)	-0.500	(.105, .902)	(.104, .901)	195	(66, 169)	-0.370	(.100, .904)	(.111, .907)
	1.6	147	(65, 125)	-0.290	(.105, .908)	(.100, .905)	149	(46, 127)	-0.495	(.109, .905)	(.105, .905)
	1.7	118	(53, 99)	-0.400	(.103, .910)	(.111, .909)	120	(38, 101)	-0.495	(.096, .909)	(.106, .906)

Table 4

Simulation results for the two-stage designs with utility and superiority stopping values as given in Table 2: Empirical type I error rate and power ($\hat{\alpha}, 1 - \hat{\beta}$) _{$\hat{r}$} when the realized accrual rate $\hat{r}$ is possibly different from the specified accrual rate r and the interim and final analyses are conducted based on the number of events (D_1, D) using the stage I rejection value ψ as determined at design

$(\alpha, 1 - \beta)$	Minimax Design					Optimal Design				
	n	(D_1, D)	(c, c_u)	$(\hat{\alpha}, 1 - \hat{\beta})_{60}$	$(\hat{\alpha}, 1 - \hat{\beta})_{90}$	n	(D_1, D)	(c, c_u)	$(\hat{\alpha}, 1 - \hat{\beta})_{60}$	$(\hat{\alpha}, 1 - \hat{\beta})_{90}$
When the specified accrual rate is $r = 60$ patients per year										
(.05, .8)	1.5	172 (94, 155)	(0.240, 2.500)	(.045, .797)	(.057, .799)	191	(60, 173)	(0.235, 2.165)	(.051, .799)	(.055, .798)
	1.6	133 (65, 117)	(-0.075, 2.580)	(.052, .814)	(.049, .810)	145	(43, 129)	(0.130, 2.200)	(.053, .801)	(.052, .810)
(.1, .85)	1.7	107 (52, 93)	(-0.095, 2.595)	(.051, .811)	(.049, .820)	113	(31, 99)	(-0.035, 2.515)	(.047, .808)	(.050, .814)
	1.5	151 (97, 134)	(0.030, 2.295)	(.094, .849)	(.098, .859)	167	(52, 149)	(-0.130, 1.835)	(.099, .857)	(.102, .844)
(1, .9)	1.6	117 (68, 101)	(-0.260, 2.200)	(.102, .850)	(.096, .853)	126	(40, 110)	(-0.195, 1.955)	(.100, .852)	(.099, .853)
	1.7	95 (47, 81)	(-0.480, 2.470)	(.099, .856)	(.100, .858)	101	(29, 87)	(-0.340, 2.005)	(.110, .855)	(.096, .857)
(1, .9)	1.5	177 (117, 159)	(0.215, 2.095)	(.097, .897)	(.096, .890)	204	(78, 186)	(0.070, 1.610)	(.102, .886)	(.096, .895)
	1.6	137 (87, 121)	(0.030, 1.985)	(.103, .899)	(.100, .898)	153	(49, 136)	(-0.235, 1.855)	(.100, .908)	(.107, .901)
(1, .9)	1.7	112 (65, 97)	(-0.025, 1.865)	(.101, .903)	(.104, .894)	124	(38, 109)	(-0.300, 1.810)	(.095, .909)	(.102, .913)
When the specified accrual rate is $r = 90$ patients per year										
(.05, .8)	1.5	176 (102, 152)	(-0.165, 2.295)	(.047, .803)	(.050, .794)	195	(56, 171)	(0.095, 2.200)	(.054, .800)	(.058, .790)
	1.6	138 (64, 116)	(-0.055, 2.555)	(.053, .807)	(.049, .807)	149	(40, 127)	(-0.055, 2.240)	(.053, .802)	(.054, .804)
(.1, .85)	1.7	110 (51, 92)	(-0.010, 2.520)	(.049, .809)	(.056, .805)	116	(27, 97)	(-0.240, 2.565)	(.050, .806)	(.045, .809)
	1.5	158 (84, 134)	(-0.045, 2.020)	(.102, .858)	(.104, .856)	172	(46, 148)	(-0.210, 2.090)	(.105, .857)	(.100, .843)
(1, .9)	1.6	122 (60, 102)	(-0.330, 1.965)	(.107, .849)	(.100, .854)	130	(38, 109)	(-0.435, 1.925)	(.093, .864)	(.101, .865)
	1.7	100 (39, 82)	(-0.655, 2.015)	(.095, .865)	(.103, .861)	105	(25, 87)	(-0.595, 1.975)	(.105, .855)	(.107, .859)
(1, .9)	1.5	187 (108, 163)	(-0.185, 1.875)	(.101, .892)	(.103, .894)	205	(63, 180)	(-0.285, 1.870)	(.098, .903)	(.097, .909)
	1.6	147 (71, 125)	(-0.490, 2.460)	(.096, .906)	(.102, .904)	156	(44, 133)	(-0.480, 1.925)	(.096, .902)	(.096, .901)
(1, .9)	1.7	118 (57, 99)	(-0.285, 2.585)	(.094, .909)	(.102, .908)	125	(30, 105)	(-0.660, 1.975)	(.102, .901)	(.100, .901)