Published in final edited form as:
Stat Med. 2009 February 28; 28(5): 864–879. doi:10.1002/sim.3501.

Power and Sample Size Calculation for Log-rank Test with a Time Lag in Treatment Effect

Daowen Zhang[1,*] and **Hui Quan**[2]

1*Department of Statistics, North Carlina State University, Raleigh, NC 27695-8203*

2*Department of Biostatistics and Programming, Sanofi-Aventis, BX2-403A, 200 Crossing Blvd, P.O. Box 4890, Bridgewater, New Jersey 08807*

Summary

The log-rank test is the most powerful nonparametric test for detecting a proportional hazards alternative and thus is the most commonly used testing procedure for comparing time-to-event distributions between different treatments in clinical trials. When the log-rank test is used for the primary data analysis, the sample size calculation should also be based on the test to ensure the desired power for the study. In some clinical trials, the treatment effect may not manifest itself right after patients receive the treatment. Therefore, the proportional hazards assumption may not hold. Furthermore, patients may discontinue the study treatment prematurely and thus may have diluted treatment effect after treatment discontinuation. If a patient's treatment termination time is independent of his/her time-to-event of interest, the termination time can be treated as a censoring time in the final data analysis. Alternatively, we may keep collecting time-to-event data until study termination from those patients who discontinued the treatment and conduct an intent-to-treat (ITT) analysis by including them in the original treatment groups. We derive formulas necessary to calculate the asymptotic power of the log-rank test under this non-proportional hazards alternative for the two data analysis strategies. Simulation studies indicate that the formulas provide accurate power for a variety of trial settings. A clinical trial example is used to illustrate the application of the proposed methods.

Keywords

Censoring; Intent-to-treat analysis; Treatment termination

1 Introduction

In a clinical trial, especially a phase III clinical trial, where the primary endpoint is time to a specific event under study, investigators are often interested in comparing the distributions of this endpoint between a new treatment and a control. Before launching the trial, we need to know how much resource is required so that the study has enough power to detect the difference of interest. For time-to-event endpoints, the difference between their distributions is often characterized in terms of a constant hazard ratio, the so-called proportional hazards model [1]. It is well-known that the log-rank test is the asymptotically most powerful nonparametric test for detecting a (local) proportional hazards alternative [2], and thus, is the most commonly used testing procedure for comparing time-to-event distributions between treatments. When the log-rank test is pre-specified as the primary data analysis method, the design of a clinical

*email: dzhang2@stat.ncsu.edu

trial such as the sample size calculation and duration specification should consistently be based on this test to ensure the power of the final analysis.

It is well-known that the two-sample log-rank test statistic asymptotically has the standard normal distribution under the null hypothesis that the time to event endpoints have the same distribution in two treatment groups. Schoenfeld [3] studied the asymptotic properties of the two-sample log-rank test statistic for a general local alternative and showed that under some regularity conditions it has an asymptotic normal distribution with variance one and a non-centrality mean parameter in terms of hazard function difference in log-scale and the underlying censoring processes. For a proportional hazards alternative, Schoenfeld's [3] non-centrality mean parameter reduces to a quantity dependent on the expected total number of events from both treatment groups under some assumptions on the censoring processes, and the result has been widely used to design clinical trials with time-to-event endpoints. For practical use, it is often assumed that the time-to-event endpoints in both treatment groups have exponential distributions, a special case of the proportional hazards alternative, and that patients are enrolled at a constant accrual rate. Then one can manipulate the accrual period and the total study length to find a solution that best fits the current situation and the resource at hand.

Recently, a placebo-controlled clinical trial was designed to evaluate the treatment effect of an experimental drug on reducing cardiovascular risk. The trial used a fixed stopping time design. That is, even though patients enter the study at different calendar time, they will all complete the study at the same time. According to the past experience on similar drugs, it is expected that this treatment will not show its effect during the first year after patients receiving the treatment. Therefore, for design purpose, it may be reasonable to specify a special lag-time model for the treatment effect that assumes the underlying hazard ratio between treated and untreated patients is a two-step function with the first constant being 1 and the other constant being less than 1, reflecting certain treatment benefit. Note that these assumptions are used for power calculation only. They will not be used for the final data analysis. In fact, the log-rank test is pre-specified as the primary data analysis method. Therefore, whether or not these assumptions hold will have no impact on the validity of the final analysis results.

Other than the unavoidable end-of-study censoring, it is almost always the case that patients in a clinical trial may withdraw from the study and terminate the treatment prematurely. In the cardiovascular trial described above, patients who discontinue the treatments are planned to be followed until the end of the study and their event status will be ascertained. It raises a question on how we can use the information from those patients in data analysis. One strategy is to treat those patients' treatment termination times as censoring times in the data analysis if their treatment termination times are independent of the time-to-event endpoints. An alternative strategy is to keep all the patients in the original treatment groups as if no patient discontinued the treatment and use the pre-specified log-rank test to conduct an intent-to-treat (ITT) analysis. The first non-ITT strategy is valid only under the independence assumption between patients' treatment termination times and their time-to-event endpoints. A potential drawback associated with the ITT analysis is that patients who discontinued the treatments may take other available medication, causing bias in the treatment effect estimate. In the ideal situation where patients obtain no other treatment once they discontinue the assigned treatment, residual treatment benefit, if exists, may be diluted. In this case, the power of the ITT analysis using the log-rank test may compromise if the residual treatment effect is too low.

The analysis of time-to-event data when there is a time lag in the treatment effect has been studied by many researchers. Zucker and Lakatos [4] proposed a maximum efficiency robust statistic to test the treatment effect with a time lag. Luo et. al. [5] considered regression analysis when there is time lag in some covariate effects specified in terms of hazards ratio. In this

paper, our interest is to address the power and sample size calculation issue when the log-rank test is pre-specified for the non-ITT and ITT analyses.

Sample size calculation for time-to-event endpoints has been received considerable attention in statistics. Schoenfeld and Richter [6] developed nomograms to calculate the number of patients needed for a clinical trial under the exponential distributional assumption for the time-to-event endpoint. Lachin and Foulkes [7] discussed sample size and power calculation issues in practical settings with nonuniform patient entry, losses to follow-up, noncompliance, etc. They assumed the exponential distribution for the time-to-event endpoint and used maximum likelihood approach. Lakatos [8] derived a formula for calculating the power of the log-rank test for a general alternative. The formula involves iterative calculation of many quantities and hence can be computationally intensive. The method has been implemented in a computer program SIZE [9]. Lakatos and Lan [10] compared several methods for calculating sample size for the log-rank test.

In this paper, we investigate the asymptotic distribution of the two-sample log-rank test statistic under the lag-time model described above for the two strategies for handling treatment termination: non-ITT and ITT strategies. We consider the ideal situation for the ITT strategy assuming those discontinued patients obtain no other treatment once they terminate the assigned treatment. We derive formulas necessary to calculate the asymptotic power of the log-rank test for these two strategies for some commonly used design settings. Simulation studies are conducted to evaluate the accuracy of the proposed power formulas, and the cardiovascular clinical trial is used to illustrate its application. This paper is organized as follows. In Section 2, we review Schoenfeld's [3] result on the asymptotic distribution of the two-sample log-rank test statistic, then derive formulas specifically for the lag-time treatment effect model for the non-ITT and ITT strategies. In Section 3, we illustrate the proposed method using the cardiovascular clinical trial and present the simulation results based on the example. We conclude the paper by some discussion in Section 4.

2 Distributions of the log-rank test statistic

Denote by $\lambda_1(t)$ and $\lambda_0(t)$ the underlying true hazard functions for the treatment and control groups. Suppose there are n patients in the study, each of whom will be assigned to the new treatment with a constant probability π. At each event time x, denote by $n_1(x)$, $n_0(x)$ the number of patients at risk for the treatment and control groups and by $d_1(x)$, $d_0(x)$ the number of events from each group. Further denote $n(x) = n_0(x) + n_1(x)$ and $d(x) = d_0(x) + d_1(x)$. Then the standard two-sample log-rank test statistic used to test the null hypothesis $H_0 : \lambda_1(t) = \lambda_0(t)$ takes the following form

$$T = \frac{\sum_x \{d_1(x) - n_1(x)\, d(x)/n(x)\}}{\sqrt{\sum_x n_1(x)\, n_0(x)\, d(x)\{n(x) - d(x)\}/[n^2(x)\{n(x) - 1\}]}}. \quad (1)$$

It is well-known that under H_0, the test statistic T is asymptotically distributed as N(0, 1). Therefore, we will reject H_0 if $|T| \geq z_{\alpha/2}$ when the type I error rate is pre-specified to be α. Here $z_{\alpha/2}$ is the standard normal quantile corresponding to the right tail probability $\alpha/2$.

Let $f_k(t)$, $S_k(t)$ be the density and survival functions of the time-to-event endpoint and $H_k(t)$ be the cumulative distribution function of the censoring time for group $k = 0, 1$. Schoenfeld [3] showed that under the alternative $H_a : \lambda_1(t) \neq \lambda_0(t)$ satisfying $\log\{\lambda_1(t)/\lambda_0(t)\} = O(n^{-1/2})$, the log-rank test statistic T defined in (1) has an asymptotic distribution $T \overset{a}{\sim} \mathrm{N}(\phi, 1)$ as $n \to \infty$ with the following non-centrality mean parameter

$$\phi=\frac{\sqrt{n}\int_0^\infty \log\{\lambda_1(t)/\lambda_0(t)\}\,\pi(t)\{1-\pi(t)\}\,V(t)\,dt}{\left[\int_0^\infty \pi(t)\{1-\pi(t)\}\,V(t)\,dt\right]^{1/2}}, \tag{2}$$

where $V(t) = (1 - \pi)f_0(t)\{1 - H_0(t)\} + \pi f_1(t)\{1 - H_1(t)\}$ is the process of observing events and

$$\pi(t)=\frac{\pi S_1(t)\{1-H_1(t)\}}{(1-\pi)S_0(t)\{1-H_0(t)\}+\pi S_1(t)\{1-H_1(t)\}}.$$

If we assume the proportional hazards alternative

$$H_a:\lambda_1(t)=\lambda_0(t)\,e^{\gamma}$$

with $\gamma = O(n^{-1/2})$ and equal censoring process in both groups (*i.e.*, $H_0(t) = H_1(t)$), then the non-centrality parameter (2) can be simplified as $\phi=\gamma\sqrt{\pi(1-\pi)D}$, where D is the expected total number of events from the two groups combined. Therefore, for given type I error probability α and power $1 - \beta$, the required expected total number of events from the two groups combined is given by

$$D=\frac{\left(z_{\alpha/2}+z_\beta\right)^2}{\gamma^2\pi(1-\pi)},$$

which can be used to design a study.

2.1 Distribution of the log-rank test statistic under a lag-time model for the non-ITT analysis

For the cardiovascular trial described in Section 1, the lag-time treatment effect model can be expressed as

$$H_a:\frac{\lambda_1(t)}{\lambda_0(t)}=\begin{cases}1 & \text{if} \quad t\in[0,t_0)\\ e^{\gamma} & \text{if} \quad t\in[t_0,\infty)\end{cases}, \tag{3}$$

where t_0 is the time point (in the patient time scale) before which the new treatment has no effect relative to the control and γ is the full treatment effect we wish to detect.

If we directly apply Schoenfeld's [3] non-centrality parameter (2), the log-rank test statistic T under alternative (3) with $\gamma = O(n^{-1/2})$ and the equal censoring process assumption will have an asymptotic normal distribution $T \overset{a}{\sim} \mathrm{N}(\phi,1)$ with the following new non-centrality mean parameter

$$\phi=\gamma\sqrt{\pi(1-\pi)}\times\frac{\tilde{D}}{\sqrt{D}}, \tag{4}$$

where $\tilde{D}$ is the expected total number of events after t_0 from the two groups combined and D has the same definition as before.

The derivation of the non-centrality parameter ϕ in (2) depends on the assumption that $\gamma = O(n^{-1/2})$. Therefore, for the non-centrality parameter ϕ in (4) to be reasonably accurate, the treatment effect parameter γ in alternative (3) has to be very close to zero. When the treatment effect γ has a moderate size, the non-centrality parameter ϕ in (4) derived from (2) may not be accurate enough for practical use.

As shown using the iterative conditional expectation argument in Appendix, we can approximate the non-centrality mean parameter in the asymptotic normal distribution of the log-rank test statistic for alternative (3) by the following quantity

$$\phi = \sqrt{\pi(1-\pi)} \times \frac{(1-e^{-\gamma})\,\tilde{D}_1 + (e^{\gamma}-1)\,\tilde{D}_0}{\sqrt{D}}, \tag{5}$$

where $\tilde{D}_1$ and $\tilde{D}_0$ are the expected total number of events after t_0 from the treated and untreated groups respectively. If one uses the approximation $1 - e^{-\gamma} \approx \gamma$ and $e^{\gamma} - 1 \approx \gamma$ for γ close to 0, the non-centrality parameter in (5) reduces to the non-centrality parameter in (4) since $\tilde{D} = \tilde{D}_0 + \tilde{D}_1$. Alternatively, for the sake of computational convenience, $\tilde{D}_k$ can be expressed as $\tilde{D}_k = D_k - D_k^*$ (k=0,1), where D_0 and D_1 are the expected total number of events from the untreated and treated groups, D_0^* and D_1^* are the expected total number of events before t_0 from untreated and treated groups. The non-centrality parameter in (5) can be used to calculate the power of the log-rank test for different clinical design settings.

For a given design with allocation probability π and anticipated treatment effect γ, we only need to calculate D_1, D_1^*, D_0 and D_0^* in order to calculate the power of the log-rank test. We illustrate the calculation of these expected numbers for a situation we will encounter frequently in practice.

To accommodate the situation where patients may not enter the study in a steady rate, we assume that patients are enrolled in the accrual period $[0, A]$ with the following piece-wise constant accrual rate $a(x)$

$$a(x) = \begin{cases} a_1 & x \in [x_0, x_1) \\ a_2 & x \in [x_1, x_2) \\ \vdots & \\ a_K & x \in [x_{K-1}, x_K), \end{cases} \tag{6}$$

where $0 = x_0 < x_1 < \cdots < x_K = A$ is a partition of $[0, A]$, and $a_1, a_2, \cdots, a_K$ are K constants. Note that here a new variable x is used for the calendar time. Suppose the total study length is L ($> A$) with additional follow-up time $F = L - A$. Then the total number of patients in the study will be

$$n = \int_0^A a(x)\,dx = \sum_{i=1}^{K} a_i\,(x_i - x_{i-1}).$$

Denote by T_k the time-to-event random variable for a random patient (in the patient time scale) in treatment k (k = 0, 1). Further assume that the distribution of T_0 for the control group is exponential with constant hazard λ_0. Then according to alternative (3), the hazard rate of the patients in the treatment group is a two-piece step function

$$\lambda_1(t) = \begin{cases} \lambda_0 & \text{if} \quad t \in [0, t_0) \\ \lambda_1 = \lambda_0 e^{\gamma} & \text{if} \quad t \in [t_0, \infty). \end{cases} \tag{7}$$

In order to gain information regarding treatment effect γ from those patients entering the study near the end of $[0, A]$, the follow-up time F had better be chosen to be larger than t_0.

Other than the end-of-study censoring, we assume that patients may discontinue the treatments with a constant hazard rate τ. The treatment termination times will be assumed to be independent of the primary time-to-event endpoints as well as the treatment groups, and will be used as censoring times in this section. Denote by Z the time when a random patient discontinues the treatment. Then the probability that a patient entering at time $x \in [0, A]$ will be observed to have an event in the study before t_0 (in the patient time scale or $x + t_0$ in the calendar time scale) for the control group can be shown to be

$$p^*(x) = P[T_0 \le \min(L - x, t_0, Z)] = \begin{cases} \frac{\lambda_0}{\lambda_0+\tau}\left[1 - e^{-(\lambda_0+\tau)t_0}\right] & \text{if} \quad 0 \le x \le L - t_0 \\ \frac{\lambda_0}{\lambda_0+\tau}\left[1 - e^{-(\lambda_0+\tau)(L-x)}\right] & \text{if} \quad x > L - t_0, \end{cases} \tag{8}$$

which is also the same for the treatment group.

Since a patient is randomly assigned to treatment with probability π, D_0^*, the expected number of events before t_0 (in patient time scale) for control group, is equal to

$$D_0^* = \int_0^A (1 - \pi)\, a(x)\, p^*(x)\, dx.$$

By (8), it is easy to show that when $t_0 \le F$, D_0^* is equal to

$$D_0^* = \frac{(1 - \pi)\lambda_0}{\lambda_0 + \tau}\left[1 - e^{-(\lambda_0+\tau)t_0}\right] \sum_{i=1}^{K} a_i (x_i - x_{i-1}).$$

When $t_0 > F$ (not a common case in practice), we can artificially insert an x_J in the partition $[0, A] = \cup_{i=1}^{K} [x_{i-1}, x_i]$ such that $x_J = L - t_0$. Then D_0^* in this case can be shown to be

$$D_0^* = \frac{(1-\pi)\lambda_0}{\lambda_0+\tau}\left\{\left[1 - e^{-(\lambda_0+\tau)t_0}\right] \sum_{i=1}^{K} a_i (x_i - x_{i-1}) + \sum_{i=J+1}^{K} a_i (x_i - x_{i-1}) - \frac{e^{-(\lambda_0+\tau)L}}{\lambda_0+\tau} \sum_{i=J+1}^{K} a_i \left[e^{(\lambda_0+\tau)x_i} - e^{(\lambda_0+\tau)x_{i-1}}\right]\right\}.$$

Since $p^*(x)$ is the same for both the treatment and control groups, D_1^* will be the same as D_0^* except that $(1 - \pi)$ in D_0^* is replaced by π.

In order to calculate D_0 and D_1, we need the probability that a random patient entering at x will be observed to have an event during the study for both groups. For the control group, this probability can be shown to be

$$p_0(x) = P[T_0 \le \min(L-x, Z)] = \frac{\lambda_0}{\lambda_0+\tau} - \frac{\lambda_0}{\lambda_0+\tau} e^{-(\lambda_0+\tau)(L-x)}.$$

Therefore, D_0 can be calculated as

$$\begin{aligned} D_0 &= \int_0^A (1-\pi)\, a(x)\, p_0(x)\, dx \\ &= \frac{(1-\pi)\lambda_0}{\lambda_0+\tau} \left\{ \sum_{i=1}^{K} a_i (x_i - x_{i-1}) - \frac{e^{-(\lambda_0+\tau)L}}{\lambda_0+\tau} \sum_{i=1}^{K} a_i \left[e^{(\lambda_0+\tau)x_i} - e^{(\lambda_0+\tau)x_{i-1}} \right] \right\}. \end{aligned}$$

For the treatment group, the probability that a random patient entering at x is observed to have an event during the study can be shown to be

$$\begin{aligned} p_1(x) &= P[T_1 \le \min(L-x, Z)] \\ &= \begin{cases} \left(\frac{\tau}{\lambda_0+\tau} - \frac{\tau}{\lambda_1+\tau}\right) e^{-(\lambda_0+\tau)t_0} + \frac{\lambda_0}{\lambda_0+\tau} - \frac{\lambda_1}{\lambda_1+\tau} e^{(\lambda_1-\lambda_0)t_0 - (\lambda_1+\tau)(L-x)} & \text{if} \quad x \le L - t_0 \\ \frac{\lambda_0}{\lambda_0+\tau} - \frac{\lambda_0}{\lambda_0+\tau} e^{-(\lambda_0+\tau)(L-x)} & \text{if} \quad x > L - t_0. \end{cases} \end{aligned}$$

Denote the two functional expressions in $p_1(x)$ by $p_{11}(x)$ and $p_{12}(x)$ respectively. Then when $t_0 \le F$, $D_1 = \int_0^A \pi a(x)\, p_1(x)\, dx = \int_0^A \pi a(x)\, p_{11}(x)\, dx$ can be calculated as

$$\begin{aligned} D_1 = \pi \Bigg\{ & \left[\frac{\lambda_0}{\lambda_0+\tau} + \left(\frac{\tau}{\lambda_0+\tau} - \frac{\tau}{\lambda_1+\tau}\right) e^{-(\lambda_0+\tau)t_0} \right] \sum_{i=1}^{K} a_i (x_i - x_{i-1}) \\ & - \frac{\lambda_1 e^{(\lambda_1-\lambda_0)t_0 - (\lambda_1+\tau)L}}{(\lambda_1+\tau)^2} \sum_{i=1}^{K} a_i \left[e^{(\lambda_1+\tau)x_i} - e^{(\lambda_1+\tau)x_{i-1}} \right] \Bigg\}. \end{aligned}$$

When $t_0 > F$, D_1 is equal to

$$D_1 = \pi \left[\int_0^{L-t_0} a(x)\, p_{11}(x)\, dx + \int_{L-t_0}^{0} a(x)\, p_{12}(x)\, dx \right],$$

where the first integral is equal to

$$\begin{aligned} \int_0^{L-t_0} a(x)\, p_{11}(x)\, dx = & \left[\frac{\lambda_0}{\lambda_0+\tau} + \left(\frac{\tau}{\lambda_0+\tau} - \frac{\tau}{\lambda_1+\tau}\right) e^{-(\lambda_0+\tau)t_0} \right] \sum_{i=1}^{J} a_i (x_i - x_{i-1}) \\ & - \frac{\lambda_1 e^{(\lambda_1-\lambda_0)t_0 - (\lambda_1+\tau)L}}{(\lambda_1+\tau)^2} \sum_{i=1}^{J} a_i \left[e^{(\lambda_1+\tau)x_i} - e^{(\lambda_1+\tau)x_{i-1}} \right], \end{aligned}$$

and the second integral is equal to

$$\int_{L-t_0}^{A} a(x)\, p_{12}(x)\, dx = \frac{\lambda_0}{\lambda_0+\tau} \left\{ \sum_{i=J+1}^{K} a_i (x_i - x_{i-1}) - \frac{e^{-(\lambda_0+\tau)L}}{\lambda_0+\tau} \sum_{i=J+1}^{K} a_i \left[e^{(\lambda_0+\tau)x_i} - e^{(\lambda_0+\tau)x_{i-1}} \right] \right\}.$$

2.2 Distribution of the log-rank test statistic under a lag-time model for the ITT analysis

When the treatment termination times for those patients who discontinued the treatments are treated as censoring times as in the previous section, the hazard function of having an event

after the treatment termination is irrelevant to the calculation of the non-centrality parameter in (5), although the treatment termination process has to be used in calculating those expected numbers of events in (5). However, when we conduct an ITT analysis using the log-rank test, all patients including those early discontinued patients must be followed until study events or study termination, and all events occurring between treatment termination and study termination and the corresponding time to event have to be included in the analysis. Therefore, we need to derive the hazard functions of having study events for the patients in the two comparison groups during the entire study period. Denote by $\lambda_1^*(t)$ and $\lambda_0^*(t)$ the hazard functions for the treated and untreated groups. Then the power of the ITT log-rank test is calculated under the alternative specified by $\lambda_1^*(t)$ and $\lambda_0^*(t)$. Clearly, the hazard functions $\lambda_1^*(t)$ and $\lambda_0^*(t)$ are not the same as the hazard functions $\lambda_1(t)$ and $\lambda_0(t)$ specified previously, but will depend on $\lambda_1(t)$, $\lambda_0(t)$, the treatment termination process and the residual treatment effect.

Lakatos [8] derived an approximation to the non-centrality parameter of the log-rank test statistic for any alternatives. Suppose the new hazard functions $\lambda_1^*(t)$ and $\lambda_0^*(t)$ can be calculated, we can then use Lakatos' [8] formula to find out the approximate non-centrality parameter associated with hazard functions $\lambda_1^*(t)$ and $\lambda_0^*(t)$.

Partition the patient time $[0, L) = \cup[t_i, t_{i+1})$ into many small sub-intervals with equal length Δ. Lakatos' [8] formula then leads to the following approximate non-centrality parameter for the ITT log-rank test statistic

$$\phi \approx \frac{\sum D_i \left(\frac{\xi_i p_i}{1+\xi_i p_i} - \frac{p_i}{1+p_i} \right)}{\left\{ \sum D_i \frac{p_i}{(1+p_i)^2} \right\}^{1/2}}, \tag{9}$$

where $D_i = \left\{ n_1(t_i)\lambda_1^*(t_i) + n_0(t_i)\lambda_0^*(t_i) \right\} \Delta$ is the estimated expected total number of events in $[t_i, t_{i+1})$ from the two groups combined, $\xi_i = \lambda_1^*(t_i)/\lambda_0^*(t_i)$ is the hazard ratio at t_i between the treated and untreated groups, and $p_i = n_1(t_i)/n_0(t_i)$ is the ratio of number of patients at risk at t_i between the treated and untreated groups. Lakatos [8] calculated $n_0(t_i)$ and $n_1(t_i)$ iteratively assuming a constant accrual rate. For the non-constant accrual rate $a(x)$ given in (6), Lakatos' [8] formula can be modified to calculate the number of patients at risk at t_{i+1} as follows

$$n_k(t_{i+1}) = \begin{cases} n_k(t_i)\left\{1 - \lambda_k^*(t_i)\Delta\right\} & \text{if} \quad t_i < F \\ n_k(t_i)\left\{1 - \lambda_k^*(t_i)\Delta - \frac{a(L-t_i)}{a(0)+a(\Delta)+\cdots+a(L-t_i)}\right\} & \text{if} \quad t_i \geq F \end{cases}, \quad k=0,1.$$

If we assume the ideal situation where discontinued patients obtain no other medication, it may be reasonable to assume that the hazard for discontinued patients in the control group will remain the same after they terminated treatment. Under the assumption given in the previous section that the patients in the control group has a constant hazard λ_0 of having study events, we have $\lambda_0^*(t) = \lambda_0$.

In order to derive $\lambda_1^*(t)$, we need to specify the hazard of having study events after those patients discontinue the study treatment in the treated group. Since the new treatment does not have an effect before t_0 relative to the control, it may be reasonable to assume that the hazard of those patients discontinued before t_0 will remain to be λ_0, and the hazard of those discontinued after t_0 is a constant given by $\tilde{\lambda}_1 \in [\lambda_1, \lambda_0]$ (for example, $\tilde{\lambda}_1$ may take $\tilde{\lambda}_1 = w\lambda_1 + (1-w)\lambda_0$ for some

$w \in [0, 1]$). Then the conditional survival function for those patients who discontinue at time Z is

$$S_1(t|Z) = e^{-\lambda_0 t}, t \in [0,\infty), \quad \text{if} \quad Z<t_0,$$

and if $Z \geq t_0$

$$S_1(t|Z) = e^{-\Lambda_1(t|Z)} = \begin{cases} e^{-\lambda_0 t} & t \in [0,t_0) \\ e^{-\lambda_0 t_0 - \lambda_1 (t-t_0)} & t \in [t_0,Z) \\ e^{-\lambda_0 t_0 - \lambda_1 (Z-t_0) - \tilde{\lambda}_1 (t-Z)} & t \in [Z,\infty). \end{cases}$$

The overall survival function $S_1^*(t)$ for the patients in the treatment group can then be calculated as

$$\begin{aligned} S_1^*(t) = & \quad \mathrm{E}[I(T \geq t)] = \mathrm{E}\{\mathrm{E}[I(T \geq t)|Z]\} = \mathrm{E}[S_1(t|Z)] \\ = & \begin{cases} e^{-\lambda_0 t} & \text{if} \quad 0 \leq t \leq t_0 \\ (1 - e^{-\tau t_0}) e^{-\lambda_0 t} + e^{(\lambda_1 - \lambda_0) t_0 - (\lambda_1 + \tau) t} & \\ + \frac{\tau}{\lambda_1 - \tilde{\lambda}_1 + \tau} \left[e^{\left(\tilde{\lambda}_1 - \lambda_0 - \tau\right) t_0 - \tilde{\lambda}_1 t} - e^{(\lambda_1 - \lambda_0) t_0 - (\lambda_1 + \tau) t} \right] & \text{if} \quad t>t_0. \end{cases} \end{aligned}$$

From this, we can calculate the density function $f_1^*(t) = -dS^*(t)/dt$ for the treatment group and hence the hazard function

$$\lambda_1^*(t) = \frac{f_1^*(t)}{S_1^*(t)},$$

and the non-centrality parameter in (9) and hence the power of the ITT log-rank test.

3 An example and simulation studies

We now consider the cardiovascular clinical trial introduced in Section 1. It is expected that the yearly hazard rate to have a cardiovascular event for the untreated patients under study is $\lambda_0 = 0.03$. The new treatment under investigation is not expected to have an effect relative to the placebo in the first year after patients receive the treatment (so $t_0 = 1$ year), and then is expected to reduce the hazard rate of cardiovascular risk by 25% so that $\lambda_1 = \lambda_0 \times 0.75 = 0.0225$, which the investigators would like to detect with 90% power using the standard two-sample log-rank test at the significance level 0.05. It is anticipated that 12000 patients per year will be available for this trial (so $a(x) = 12000$) and the patients will discontinue the treatment at a yearly constant hazard rate 10% (so $\tau = 0.1$). The investigators plan a 50 month trial (so $L = 50/12 = 4.17$ years). Assume equal probability ($\pi = 0.5$) for treatment assignment. The investigators want to know for how long they should enroll patients to ensure the desired power to detect the anticipated treatment effect.

For illustration, we treat treatment termination times as censoring times first. That is, we use the methodology developed in Section 2.1. For the given design characteristics (except the specified power), the non-centrality parameter (5) is a function of accrual period A only, so is the power of the standard two-sample log-rank test. Figure 1 presents the power of the log-rank test in the range [1,2] of accrual periods using non-centrality parameter (5). In comparison, the log-rank test power using Schoenfeld's [3] non-centrality parameter (4) is superimposed

in Figure 1. To evaluate the accuracy of the log-rank test power using Schoenfeld's [3] non-centrality parameter (4) and the proposed non-centrality parameter (5), simulation using the given design characteristics is used to calculate the empirical powers and 95% confidence intervals (CI's) at the following accrual periods $A = 1.0, 1.1, 1.2, \cdots, 1.9, 2.0$, where each empirical power is obtained using 10,000 simulation runs. It is seen from Figure 1 that 2 out of 11 powers calculated using Schoenfeld's [3] non-centrality parameter (4) are outside the corresponding CI's, while all the powers calculated using the proposed new non-centrality parameter (5) are contained in the corresponding CI's. For further comparison, we calculated the residual sum-of-squares (RSS) for each setting. The new non-centrality parameter (5) produced 8 out of 11 smaller RSS's. It is also noticed from Figure 1 that Schoenfeld's [3] non-centrality parameter (4) slightly over-estimates the log-rank power. These results indicate that the proposed non-centrality parameter (5) produces more accurate power than Schoenfeld's [3] formula.

From Figure 1, we can solve for the accrual period A that yields 90% power for Schoenfeld's [3] non-centrality parameter (4) and our non-centrality parameter (5). The former gives $A_1 = 1.313$ years and the later gives $A_2 = 1.385$ years. The difference between these two estimates is not negligible. Of course, using Schoenfeld's [3] non-centrality parameter (4) leads to a smaller sample size (1313 vs. 1385), and hence a power lower than the one specified by the design.

If we decided to fix the accrual period A at some value, then the non-centrality parameters (4) and (5) are functions of the study length L, so are the log-rank powers. We can then calculate the powers at different L's and solve for L that yields the desired power. Figure 2 presents the powers calculated at $A = 1.42$ years (17 months) in the range [3.5, 4.5] of L using non-centrality parameters (4) and (5), as well as the empirical powers and CI's using 10,000 simulation runs for 11 settings. Although all powers calculated using both non-centrality parameter formulas are within the corresponding CI's, our new non-centrality parameter (5) produced 9 out of 11 smaller RSS's, again indicating the better accuracy of the newly proposed formula. The study length L that yields 90% log-rank test power using our new non-centrality parameter is $L = 4.13$ years.

Similarly, we can use ITT non-centrality parameter (9) to calculate the log-rank test power to determine the accrual period A if ITT analysis is to be conducted. We assumed the ideal setting where discontinued patients obtain no other medication. In this case, it is required to have the knowledge about $\tilde{\lambda}_1$, the hazard rate of cardiovascular risk after treatment termination for withdrawn patients. Or equivalently, the weight w has to be specified if $\tilde{\lambda}_1 = w\lambda_1 + (1 - w)\lambda_0$ is used. To illustrate the impact of w, we set $A = 1.42$ years, $L = 4.17$ years and calculate the power using ITT non-centrality parameter (9) as a function of w, which is presented in Figure 3. One thousand sub-intervals per year is used to calculate non-centrality parameter (9). It is seen from Figure 3 that in order to achieve at least the same power (90.4%) as the non-ITT analysis treating treatment termination times as censoring times, the weight w has to be greater than 0.7. This may or may not be reasonable. To evaluate the accuracy of the power calculated using the non-centrality parameter (9), we superimpose the empirical powers and CI's obtained from 10,000 simulation runs for 11 equally spaced weights (w) in [0, 1]. It can be seen that the calculated powers using non-centrality parameter (9) are all contained in the corresponding CI's and agree very well with the empirical powers.

Last, we compare in Figure 4 the powers as functions of accrual period A for non-ITT analysis using non-centrality parameter (5) and ITT analysis using non-centrality parameter (9) for three different weights $w = 0$, 0.5 and $w = 1$. It is evident from this figure that for the cardiovascular clinical trial the non-ITT analysis may be more powerful than ITT analysis if the residual treatment benefit is low. This indicates that if the independence assumption between treatment

termination times and time-to-event endpoints is reasonable, we may prefer to use the non-ITT analysis for the actual data analysis. Otherwise, the ITT analysis should be considered. For example, an HDL increasing drug may at the same time increase blood pressure. Patients withdrawn from the study due to high blood pressure may have a higher chance to have cardiovascular events had they be followed until the end of the study. Thus, the non-informative censoring assumption may not hold for the non-ITT analysis and the ITT analysis may be more reasonable.

4 Discussion

In this paper, we discussed the asymptotic distribution of the two-sample log-rank test statistics under a special lag-time model for the treatment effect. We proposed a new approximation to the non-centrality parameter in this asymptotic normal distribution when patients' treatment termination times are independent of time-to-event endpoints of interest and are treated as censoring times. For given design characteristics such as accrual period, study length and accrual rate, this non-centrality parameter can be calculated easily. Simulation studies indicate that the proposed new non-centrality parameter gives more accurate power than Shoenfeld's [3] formula, which is commonly used in practice.

We also discussed the calculation of the non-centrality parameter of the log-rank test if ITT strategy is used for the data analysis. We posed a model for the treatment termination process so that the distributions of time to event endpoints for each group can be derived. We then modified Lakatos' [8] result to accommodate the possible situation where the accrual rate is a piece-wise step function. Simulation studies indicate that the calculated power is reasonably accurate if the number of sub-intervals used to calculate the non-centrality parameter is sufficiently large, although with some computational cost.

We illustrated the proposed power calculation using a cardiovascular risk clinical trial, and compared the power difference between the non-ITT and ITT strategies. If we can assume that the treatment termination process is independent of the primary time-to-event endpoint, the non-ITT strategy is usually more powerful. However, if this independence assumption does not satisfy, the non-ITT analysis may be invalid and the power is irrelevant. In this case, the best strategy is to conduct the ITT long-rank test.

The standard log-rank test puts equal weights on all time points. A weighted log-rank test may be pre-specified as the primary data analysis method. In the case when the patients' termination times are independent of the time-to-event endpoints, the maximum efficiency robust test statistics proposed by Zucker and Lakatos [4] can be used. Their test statistics are robust to the mis-specification of the lag-time treatment effect model. However, the non-centrality parameter of their test statistic in general will not have a nice form similar to (5), even under the alternative (3). In this case, Lakatos' [8] method as described in Section 2.2 can be used to calculate the non-centrality parameter and hence the power. In the case when the patients' termination times are not independent of the time-to-event endpoints so that the ITT strategy is the only valid analysis, search for an efficient yet robust test statistic similar to Zucker and Lakatos' [4] may be difficult, since it depends on the patients' treatment termination process, residual treatment effect, etc. More research is needed for this problem.

Acknowledgment

The research of Zhang was partly supported by an NIH grant R01 CA85848-08.

Appendix: Derivation of the non-centrality parameter (5)

Denote by S the numerator of the log-rank test statistic T in (1). Consider a fine partition of $[0, \infty) = \cup[x, x + \Delta x)$ with small and equal length Δx for each interval. Then S can be interpreted as the sum of

$$S(x) = d_1(x) - n_1(x)\frac{d(x)}{n(x)}$$

over this partition, where

- $d_1(x)$, $d_0(x)$ are the total number of events in $[x, x + \Delta x)$ from the treatment and control groups respectively;
- $n_1(x)$ and $n_0(x)$ are the number of patients at risk at x for the treatment and control groups respectively;
- $d(x) = d_0(x) + d_1(x)$ is the total number of events in $[x, x + \Delta x)$ from the two groups combined;
- $n(x) = n_0(x) + n_1(x)$ is the total number of patients at risk at x from the two groups combined.

Re-write $S(x)$ as follows:

$$S(x) = \frac{n_0(x)\,d_1(x)}{n_1(x) + n_0(x)} - \frac{n_1(x)\,d_0(x)}{n_1(x) + n_0(x)}.$$

Denote by $H(x)$ all the information available up to x. Then under the independence assumption between patients' treatment termination time and their time-to-event data, we can show that $d_1(x)$ and $d_0(x)$ have the following approximate conditional distributions

$$d_1(x)\,|\,H(x) \sim \text{Bin}\{n_1(x), \lambda_1(x)\Delta x\}, d_0(x)\,|\,H(x) \sim \text{Bin}\{n_0(x), \lambda_0(x)\Delta x\}.$$

Therefore,

$$\begin{aligned} \mathrm{E}[S(x)\,|\,H(x)] &\approx \frac{n_0(x)n_1(x)\lambda_1(x)\Delta x}{n_1(x)+n_0(x)} - \frac{n_1(x)n_0(x)\lambda_0(x)\Delta x}{n_1(x)+n_0(x)} \\ &= \frac{n_1(x)n_0(x)}{n_1(x)+n_0(x)}[\lambda_1(x) - \lambda_0(x)]\Delta x. \end{aligned}$$

By the alternative (3), $\mathrm{E}[S(x)|H(x)] \approx 0$ if $x < t_0$. When $x \geq t_0$, $\lambda_1(x) = e^{\gamma}\lambda_0(x)$ and $\lambda_0(x) = e^{-\gamma}\lambda_1(x)$. In this case,

$$\begin{aligned} \mathrm{E}[S(x)\,|\,H(x)] &\approx \frac{n_1(x)n_0(x)}{[n_1(x)+n_0(x)]^2}[n_1(x)+n_0(x)][\lambda_1(x) - \lambda_0(x)]\Delta x. \\ &= \frac{n_1(x)n_0(x)}{[n_1(x)+n_0(x)^2}\{n_1(x)[\lambda_1(x) - \lambda_0(x)] + n_0(x)[\lambda_1(x) - \lambda_0(x)]\}\Delta x \\ &= \frac{n_1(x)n_0(x)}{[n_1(x)+n_0(x)]^2}\left[n_1(x)\lambda_1(x)\Delta x(1 - e^{-\gamma}) + n_0(x)\lambda_0(x)\Delta x(e^{\gamma} - 1)\right] \\ &\approx \frac{n_1(x)n_0(x)}{[n_1(x)+n_0(x)]^2}\{\mathrm{E}[d_1(x)\,|\,H(x)](1 - e^{-\gamma}) + \mathrm{E}[d_0(x)\,|\,H(x)](e^{\gamma} - 1)\}. \end{aligned}$$

If the censoring processes in both groups are the same and the treatment effect is close to zero (*i.e.*, $\gamma \approx 0$), then

$$\frac{n_1(x)\,n_0(x)}{[n_1(x)+n_0(x)]^2} \approx \pi(1-\pi).$$

This leads to

$$\begin{aligned} \mathrm{E}[S(x)] &= \mathrm{E}\{\mathrm{E}[S(x)|H(x)]\} \\ &\approx \begin{cases} 0 & \text{when} \quad x<t_0 \\ \pi(1-\pi)\{\mathrm{E}[d_1(x)](1-e^{-\gamma})+\mathrm{E}[d_0(x)](e^{\gamma}-1)\} & \text{when} \quad x \geq t_0 \end{cases} . \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(S) &= \sum_x \mathrm{E}[S(x)] = \sum_{x<t_0} \mathrm{E}[S(x)] + \sum_{x\geq t_0} \mathrm{E}[S(x)] \\ &\approx 0+\pi(1-\pi)\left\{\sum_{x\geq t_0} \mathrm{E}[d_1(x)](1-e^{-\gamma}) + \sum_{x\geq t_0} \mathrm{E}[d_0(x)](e^{\gamma}-1)\right\} \\ &= \pi(1-\pi)\left\{(1-e^{-\gamma})\sum_{x\geq t_0} \mathrm{E}[d_1(x)] + (e^{\gamma}-1)\sum_{x\geq t_0} \mathrm{E}[d_0(x)]\right\} \\ &= \pi(1-\pi)\left[(1-e^{-\gamma})\tilde{D}_1 + (e^{\gamma}-1)\tilde{D}_0\right]. \end{aligned} \tag{10}$$

Now consider the random variable inside the square-root of the denominator of the log-rank test statistic (1). Since $[x, x + \Delta x)$ is an interval with small Δx, it is reasonable to assume that there is at most one event in $[x, x + \Delta x)$ from the two groups combined. That is, $d(x)$ is either 0 or 1. So

$$\frac{d(x)\{n(x)-d(x)\}}{n(x)-1} = d(x),$$

and

$$\mathrm{E}\left\{\sum_x \frac{n_1(x)\,n_0(x)\,d(x)[n(x)-d(x)]}{n^2(x)[n(x)-1]}\right\} \approx \pi(1-\pi)\sum_x \mathrm{E}[d(x)] = \pi(1-\pi)D. \tag{11}$$

Combining (10) and (11) yields the non-centrality parameter (5).

REFERENCES

1. Cox D. Regression models and life-tables. Journal of the Royal Statistical Society-B 1972;34:187–220.
2. Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. John Wiley & Sons; New York: 1991.
3. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981;68:316–319.
4. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. Biometrika 1990;77:853–864.
5. Luo X, Turnbull BW, Cai H, Clark LC. Regression for censored survival data with lag effects. Communications in Statistics-Theory and Methods 1994;23:3417–3438.

6. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics 1982;38:163–170. [PubMed: 7082758]

7. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 1986;42:507–519. [PubMed: 3567285]

8. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical-trials. Biometrics 1988;44:229–241. [PubMed: 3358991]

9. Shih JH. Sample size calculation for complex clinical trials with survival endpoints. Controlled clinical trials 1995;16:395–407. [PubMed: 8720017]

10. Lakatos E, Lan KKG. A comparison of sample size methods for the logrank statistic. Statistics in Medicine 1992;11:179–191. [PubMed: 1579757]