Sample Size Calculation for the One-Sample Log-Rank Test

Rene Schmidt[1,*], Robert Kwiecien[1], Andreas Faldum[1], and Sandra Ligges[1]

[1] Institute of Biostatistics and Clinical Research, University of Münster

* E-mail: rene.schmidt@ukmuenster.de

Introduction

The one-sample log-rank test, first proposed by Breslow (1975), allows for the comparison of the survival curve of a new treatment arm with that of a historical control. Finkelstein *et al.* (2003) and Sun *et al.* (2011) both give a sample size formula for power requirements based on the number *d* of events to be observed.

We show that there is an alternative criterion that can be followed in order to achieve a desired power for the one-sample log-rank test. Both power approximations are asymptotically equivalent. But a small simulation study reveals that the two approximations perform differently well for small sample numbers.

General setting

The study time will be denoted by Latin *t*. We consider a phase II clinical trial with a time-to-event endpoint, where a new treatment is to be compared with a treatment of a historical control. We introduce the following notation:

- $\lambda_H(t)$: Known hazard function for the historical control,
- $\lambda(t)$: Unknown hazard function for new treatment,
- $R = \lambda(t)/\lambda_H(t)$: Corresponding hazard ratio,

We assume proportional hazards, i.e. R does not depend on t. Moreover, let:

- $\gamma_1(t)$: Bound for R below which treatment is considered useful.
- $\gamma_0(t)$: Bound for R above which treatment is discarded.

We consider testing the hypotheses:

$$H_0 : R \geq \gamma_0 \qquad (1)$$

The planning alternative hypothesis underlying power calculations is $H_1 : R = \gamma_1$for some $0 < \gamma_1 < \gamma_0$.

The test statistic

For each patient $i = 1, \dots, n$, let time-to-event data be given by:

- Δ_i: Censoring indicator. $\Delta_i = \begin{cases} 0 & , if\ patients\ i\ is\ censored \\ 1 & , if\ patient\ i\ had\ an\ event \end{cases}$
- X_i: Right-censored survival time.

The number of events observed in the new treatment arm is

$$O = \sum_{i=1}^{n} \Delta_i .$$

Let $\Lambda_H(x) = \int_0^x \lambda_H(t)dt$ denote the cumulative hazard function of the historical control, and let

$$E_H = \sum_{i=1}^{n} \Lambda_H(X_i) .$$

E_H may be interpreted as expected number of events if $R = 1$. Likewise, $\gamma_i E_H$ may be interpreted as the expected number of events in the new trial if hypothesis $H_i : R = \gamma_i, i = 0,1,$ holds true.

An one-sample test of $H_0 : R \geq \gamma_0$ on an one-sided significance level α can be defined by rejecting H_0 if and only if

$$Z := \frac{O - \gamma_0 E_H}{\sqrt{\gamma_0 E_H}} \leq \Phi^{-1}(\alpha) \qquad (2)$$

where Φ denotes the standard normal distribution function (Aalen et al., 2008). This test is referred to as ***one-sample log-rank test***.

Sample size acc. to Finkelstein et al. (2003)

According to Finkelstein et al. (2003), the schedule of the analysis may be based on the number D of observed events. In order to achieve approximately a desired power of at least $1 - \beta$ for the one-sample log-rank test (2) under the planning alternative $H_1 : R = \gamma_1$for allocated significance level α and effect $\theta := \gamma_1/\gamma_0$, the analysis of data is to be performed as soon as D reaches the critical value d given by

$$d \geq \theta \left(\frac{\Phi^{-1}(1-\alpha) + \sqrt{\theta} \cdot \Phi^{-1}(1-\beta)}{1-\theta} \right)^2 . \qquad (3)$$

Alternative criterion based on E_H

Alternatively, it can be shown that the power of the one-sample log-rank test is essentially determined by E_H: In order to reach an approximate power of at least $1 - \beta$ for the one-sample log-rank test (2) under the planning alternative hypothesis $H_1 : R = \gamma_1$for allocated significance level α and effect $\theta := \gamma_1/\gamma_0$, the analysis of data is to be performed as soon as E_H reaches the critical value e given by

$$e \geq \frac{1}{\gamma_0} \left(\frac{\Phi^{-1}(1-\alpha) + \sqrt{\theta} \cdot \Phi^{-1}(1-\beta)}{1-\theta} \right)^2 . \qquad (4)$$

Simulation

The power approximations of the one-sample log-rank test given by the two stopping criteria in (3) and (4) are asymptotically equivalent, but perform differently well for small sample size. Performance of both stopping criteria for small finite sample numbers is investigated by simulation.

For illustrative reasons, we assume the following setting: Assume that the estimated one-year EFS-rate $S_H(1)$ for the historical control population is 50%. Assume that superiority of new treatment with regard to EFS shall be proven while considering an one-year EFS-rate of at least 70% as clinically relevant improvement. So, we consider the null hypotheses (1) with $\gamma_0 = 1$ and apply the one-sample log-rank test (2) with planning alternative hypothesis $\gamma_1 = \log(0.7) / \log(0.5) \approx 0.515$. Assume we want to achieve a power of at least 85% for allocated one-sided significance level of $\alpha = 5\%$.

We conclude from (3) and (4) that the trial has to be stopped as soon as $e \geq 24.21$ or $d \geq 13$ for the first time, respectively. We assume exponentially distributed survival times with uniform annual accrual of 45 patients for one year. In particular, the hazard rate for the historical control is $\lambda_0 = -log(0.5)$.

In order to investigate whether the one-sample log-rank test fulfills the power requirements and holds the significance level when the two information criteria *e* and *d* are followed, we carried out simulations for eight values of the true treatment effect R ranging from 0.434 to 1. More precisely, R is the true hazard ratio of the new treatment arm to the historical control, whereas γ_1 is the assumed one. The corresponding true hazard rate for patients of the new treatment arm is $\lambda_1 = R\lambda_0$, respectively. In each of 10.000 runs for every constellation, the test statistic of the one-sample log-rank test was calculated twice, once when $e \geq 24.21$ and once when $d \geq 13$ was observed for the first time. The results are displayed in Figure 1.

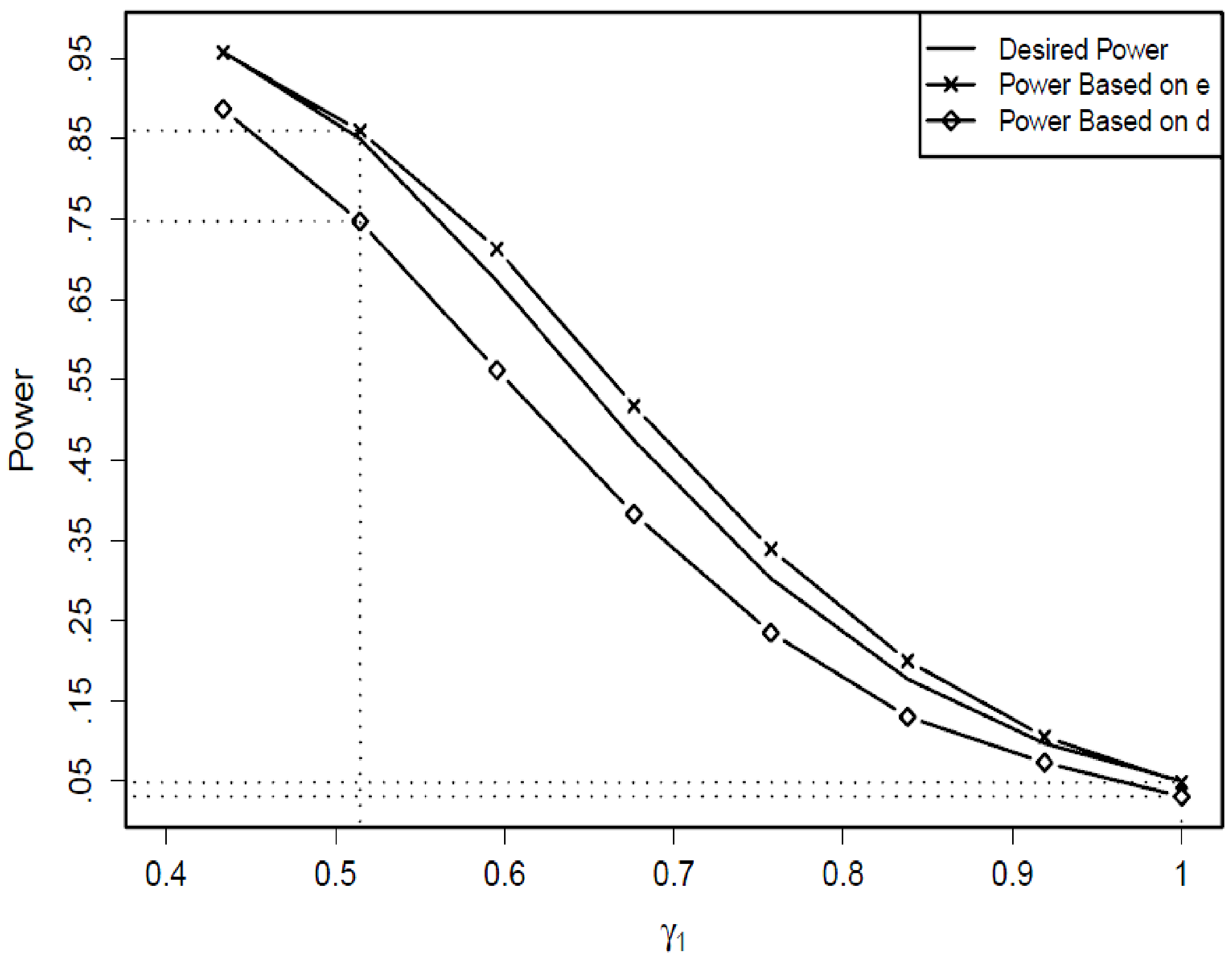

Figure 1. Comparison of Power for E_h and D.

Discussion

It turns out in this given scenario that the first approach based on the control of E_H leads to a power closed to the a-priori aspired power while the test procedure based on the control of D is conservative and the study is rather underpowered. For instance, in the scenario underlying the above simulation, a power of 85% is expected for $\gamma_1 \approx 0.515$. The test procedure based on following *e* reaches a power of 85.99%, while the one based on following *d* reaches a power of only 74.72% (see Figure 1). Besides, the test procedure based on following d is quite conservative here. Its significance level is estimated at 3.08%, while with a rejection rate of 4.83%, the test procedure based on following *e* exhausts the aspired significance level of 5% much better (see Figure 1). So, against the background of these simulation results, it appears more favorable to schedule the analyses according to the observed sum of the cumulative hazards instead of the number of events, despite of potentially higher logistic efforts. It will be contents of further research to assess robustness of this result in a wider range of simulation scenarios.

Bibliography

Aalen OO, Borgan O, and Gjessing HK (2008). *Survival and Event History Analysis*. New York: Springer Science + Business Med

Breslow NE (1975). Analysis of Survival Data under the Proportional Hazards Model. *International Statistics Review* 1975; **43**: 45-58

Finkelstein DM, Muzikansky A, Schoenfeld DA (2003). Comparing survival of a sample to that of a standard population. *Journal of the National Cancer Institute*; **95**: 1434–1439.

Sun X, Peng P, Tu D (2011). Phase II cancer clinical trials with a one-sample log-rank test and its corrections based on the Edgeworth expansion. *Contemporary Clinical Trials*; **32**: 108–113.