HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2017 August 30.

Published in final edited form as:

J Biopharm Stat. 2017 ; 27(1): 124–134. doi:10.1080/10543406.2016.1148711.

Sample Size Calculation for Testing Differences Between Cure Rates with the Optimal Log-rank Test

Jianrong Wu

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

Abstract

In this article, sample size calculations are developed for use when the main interest is in the differences between the cure rates of two groups. Following the work of Ewell and Ibrahim, the asymptotic distribution of the weighted log-rank test is derived under the local alternative. The optimal log-rank test under the proportional distributions alternative is discussed, and sample size formulae for the optimal and standard log-rank tests are derived. Simulation results show that the proposed formulae provide adequate sample size estimation for trial designs and that the optimal log-rank test is more efficient than the standard log-rank test, particularly when both cure rates and percentages of censoring are small.

Keywords

clinical trial; cure model; log-rank test; optimal test; sample size

1 Introduction

When survival data include a portion of cured patients or long-term survivors, cure models are useful for analyzing the data and designing clinical trials. Recently, various parametric and semiparametric cure models have been proposed by Farewell (1982), Peng et al. (1998), and Kuk and Chen (1992). A maximum-likelihood expectation-maximization (EM) algorithm for parametric and semiparametric cure models has been proposed by Peng and Dear (2000) and Sy and Taylor (2000). A SAS macro PSPMCM, developed by Corbiere and Joly (2007), is available to fit both parametric and semiparametric cure models. Thus, survival data in which a portion of patients are cured can be analyzed using these methods for the purpose of designing clinical trials using the selected cure models.

In a cancer clinical trial in which a portion of patients experience long-term survival, the main interest is often in the differences between cure rates. Examples from the Children's Cancer Group trials are given by Lee and Sather (1995). To develop an appropriate test for testing the differences between cure rates in a two-arm randomized trial, Gray and Tsiatis (1989) proposed a family of cure models with a proportional distributions alternative. The optimal log-rank test was discussed under the proportional distributions alternative, which has the form of a G^{ρ} test where $\rho = -1$ (Harrington and Fleming, 1982), and its efficacy relative to that of the standard log-rank test was also investigated. Ewell and Ibrahim (1997) extended the work of Gray and Tsiatis by deriving the large sample distribution of the

weighted log-rank test under a more general sequence of local alternatives that allows for treatment effects on both short- and long-term survival. They also derived a power calculation for the weighted log-rank test assuming exponential failure times.

In this article, we focus on the situation where the main interest is in the differences between the cure rates of two groups. Following the work of Ewell and Ibrahim, sample size formulae are derived for both the standard log-rank test and the optimal weighted log-rank test. The relative efficacy of the two tests is also discussed.

The rest of the paper is organized as follows. A mixture cure model is introduced in Section 2. The sample size formula for the weighted log-rank test is derived in Section 3. The optimal log-rank test and its sample size formula are obtained in Section 4. In section 5, comparisons of the efficiency and robustness of the two tests are presented, and simulations are conducted to study the performance of the proposed sample size formulae. Section 6 illustrates clinical trial design using the proposed methods. Conclusions are presented in section 7.

2 Cure Models

The failure time, T^*, is assumed to be $vT+(1-v)\infty$, where v is an indicator of whether a subject will eventually ($v = 1$) or never ($v = 0$) experience treatment failure, and T denotes the failure time if the subject is not cured, with a survival distribution $S(t)$, which is the conditional distribution for patients who will experience treatment failure and is often called the latency distribution. Thus, the unconditional survival distribution of T^* is a mixture model of a cure rate $\pi = P(v = 0)$ and a latency distribution $S(t)$ given by

$$S*(t)=\pi+(1-\pi)S(t).$$

Let $\lambda^*(t)$ and $\lambda(t)$ be the hazard functions of T^* and T, respectively. We then have the following relation between the two hazard functions:

$$\lambda*(t)=\frac{(1-\pi)S(t)}{\pi+(1-\pi)S(t)}\lambda(t).$$

For a two-arm randomized survival trial, let T_{ij}^* and C_{ij} denote the survival and censoring times, respectively, of patient i in the j^{th} group, where $j = 1, 2$ (1 for the control group and 2 for the treatment group). The observed data then consist of $\{X_{ij}, \Delta_{ij}; i = 1, \ldots, n_j, j = 1, 2\}$, where $X_{ij}=T_{ij}^* \wedge C_{ij}$ and $\Delta_{ij}=I(T_{ij}^* \leq C_{ij})$. It is commonly assumed that $\left\{T_{ij}^*, C_{ij}, i=1,\ldots,n_j\right\}$ are independent and identically distributed samples of (T_j, C_j) for control ($j = 1$) and treatment ($j = 2$) and that T_{ij} is independent of C_{ij}. Let $S_j^*(t)$ denote the unconditional survival distribution and let $\lambda_j^*(t)$ denote its hazard function for the j^{th} group. When the main interest is in testing for differences between cure rates, it is reasonable to assume that the conditional survival distributions are the same for the two groups and are denoted by $S(t)$, with the hazard function and cumulative hazard function being denoted by

$\lambda(t)$ and $\Lambda(t)$, respectively. The cure rate for the j^{th} group is defined by π_j, where $0 \leq \pi_j < 1$. Then, the survival distribution of the mixture cure model for the j^{th} group is given by

$$S_j^*(t)=\pi_j+(1-\pi_j)S(t), \quad (1)$$

and the hazard function for the j^{th} group is given by

$$\lambda_j^*(t)=\frac{(1-\pi_j)S(t)}{\pi_j+(1-\pi_j)S(t)}\lambda(t).$$

We are interested in testing the following hypothesis:

$$H_0{:}\pi_1=\pi_2 \ \ vs. \ \ H_1{:}\pi_1 \neq \pi_2. \quad (2)$$

Furthermore, we define the parameters γ and π_0 as follows:

$$\gamma=\frac{1}{2}log\frac{1-\pi_2}{1-\pi_1},$$

$$\pi_0=1-\left[(1-\pi_1)(1-\pi_2)\right]^{1/2},$$

where γ is the half-log ratio of the failure rates, and π_0 is the proportion of cured patients under the null hypothesis. Then, hypothesis (2) is equivalent to the following hypothesis:

$$H_0{:}\gamma=0 \ \ vs. \ \ H_1{:}\gamma \neq 0. \quad (3)$$

The mixture cure model (1) can be written as

$$S_j^*(t)=1-e^{(-1)^j\gamma}(1-\pi_0)\left\{1-S(t)\right\}, \quad (4)$$

and the corresponding hazard function is given by

$$\lambda_j^*(t)=\frac{e^{(-1)^j\gamma}(1-\pi_0)S(t)}{1-e^{(-1)^j\gamma}(1-\pi_0)+e^{(-1)^j\gamma}(1-\pi_0)S(t)}\lambda(t). \quad (5)$$

This alternative implies that the unconditional failure distributions for two groups are proportional; it is called a proportional distributions alternative by Gray and Tsiatis (1989).

To test hypothesis (2) or (3), or the difference in the unconditional failure distributions, a weighted score test can be used, which is given by

$$U_w = n^{-1/2} \int_0^\infty W(t) \left\{ \frac{Y_1(t)}{Y(t)} dN_2(t) - \frac{Y_2(t)}{Y(t)} dN_1(t) \right\},$$

where $n = n_1 + n_2$ is the total sample size of two groups, $W(t)$ is a weight function that converges in probability to $w(t)$, $N_j(t)$ is the number of observed failures by time t, $Y_j(t)$ is the number of subjects at risk just prior to t in groups $j = 1, 2$, and $Y(t) = Y_1(t) + Y_2(t)$. By the martingale central limit theorem (Fleming and Harrington, 1991), under the null hypothesis, U_w converges in distribution to a normal variable with a mean of zero and variance

$$\sigma_w^2 = p(1-p)(1-\pi_0) \int_0^\infty w^2(t) G(t) S(t) d\Lambda(t), \quad (6)$$

where $p = \lim_{n \to \infty} n_1/n$, and $G(t)$ is the common survival distribution of the censoring time of two groups (see appendix). The variance σ_w^2 in (6) can be estimated by

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

where $N(t) = N_1(t) + N_2(t)$. Therefore, under the null hypothesis, the weighted log-rank test $L_w = U_w / \hat{\sigma}_w$ is asymptotically standard normal distributed. Thus, given a significance level α, we reject the null hypothesis if $|L_w| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution.

3 Sample Size Formula

To derive the sample size formula, we need to know the asymptotic distribution of the weighted log-rank test under the alternative hypothesis. Consider a sequence of local alternatives

$$S_j^{*(n)}(t) = 1 - e^{(-1)^j \gamma_n} (1 - \pi_0) \{1 - S(t)\},$$

where $n^{1/2} \gamma_n = \gamma_a$. Under the local alternatives, as shown in the appendix, the weighted log-rank test $L_w = U_w / \hat{\sigma}_w$ converges in distribution to a normal variable with unit variance and mean $\mu(w, \gamma_a)/\sigma_w$, where σ_w^2 is given by (6), and

$$\mu(w, \gamma_a) = 2p(1-p)(1-\pi_0)\gamma_a \int_0^\infty w(t) \{S_0^*(t)\}^{-1} G(t) S(t) d\Lambda(t), \quad (7)$$

for which $S_0^*(t) = \pi_0 + (1-\pi_0) S(t)$.

Therefore, on the basis of the limiting distribution of L_w under the local alternative, given a type I error of α, to achieve a power of $1-\beta$, the total sample size n of two groups must approximately satisfy the following equation:

$$1-\beta \simeq \Phi\{\mu(w,\gamma_a)/\sigma_w - z_{1-\alpha/2}\}.$$

For a local alternative γ, we replace γ_a by $n^{1/2}\gamma$. Then, the sample size required to detect a local alternative γ can be determined by

$$n=\frac{(z_{1-\alpha/2}+z_{1-\beta})^2\sigma_w^2}{\mu(w,\gamma)^2}. \quad (8)$$

Substituting equations (6) and (7) into (8), the total sample size for the weighted log-rank test can be calculated by

$$n=\frac{(z_{1-\alpha/2}+z_{1-\beta})^2\int_0^\infty w^2(t)G(t)S(t)d\Lambda(t)}{4p(1-p)(1-\pi_0)\gamma^2[\int_0^\infty w(t)\{S_0^*(t)\}^{-1}G(t)S(t)d\Lambda(t)]^2}. \quad (9)$$

4 Optimal Log-rank Test

It is well known that the log-rank test is optimal against the proportional hazards model. However, the cure model (1) does not satisfy the proportional hazards assumption; thus, the log-rank test is not an optimal test, and a study design based on the log-rank test is not fully efficient. Therefore, it is desirable to find an optimal test for the cure model (1) under the local proportional distributions alternative. As the mean of the weighted log-rank test is proportional to

$$\int_0^\infty w(t)\{S_0^*(t)\}^{-1}h(t)dt,$$

where $h(t) = G(t)S(t)\lambda(t)$, by using the Cauchy-Schwartz inequality, we obtain the following inequality:

$$\int_0^\infty w(t)\{S_0^*(t)\}^{-1}h(t)dt \le \left\{\int_0^\infty w^2(t)h(t)dt\int_0^\infty \{S_0^*(t)\}^{-2}h(t)dt\right\}^{1/2},$$

with equality if only if $w(t)$ is proportional to $\{S_0^*(t)\}^{-1}$. That is, the optimal weight function $w(t)$ is proportional to $\{S_0^*(t)\}^{-1}$, which minimizes the sample size given by formula (9). Thus, taking the weight function $W(t) = \{K(t^-)\}^{-1}$, where $K(t^-)$ is the left-continuous version of the Kaplan-Meier estimate computed from the pooled sample of two groups, gives the asymptotically optimal test for the proportional distributions alternative. Hence, by substituting $w(t)=\{S_0^*(t)\}^{-1}$ into formula (9), the sample size for the optimal log-rank test L_K is given by

$$n_K = \frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{4p(1-p)(1-\pi_0)\gamma^2 \int_0^\infty \{S_0^*(t)\}^{-2} G(t)S(t)d\Lambda(t)}, \quad (10)$$

and by substituting $w(t) = 1$ into formula (9), the sample size for the standard log-rank test L is given by

$$n = \frac{(z_{1-\alpha/2}+z_{1-\beta})^2 \int_0^\infty G(t)S(t)d\Lambda(t)}{4p(1-p)(1-\pi_0)\gamma^2 [\int_0^\infty \{S_0^*(t)\}^{-1} G(t)S(t)d\Lambda(t)]^2}. \quad (11)$$

The asymptotic relative efficiency $\rho = n/n_K$ (Randales and Wolfe, 1979) of the optimal test compared to the standard log-rank test is given by

$$\rho = \frac{\int_0^\infty \{S_0^*(t)\}^{-2} G(t)S(t)d\Lambda(t) \int_0^\infty G(t)S(t)d\Lambda(t)}{[\int_0^\infty \{S_0^*(t)\}^{-1} G(t)S(t)d\Lambda(t)]^2}. \quad (12)$$

In the special case when there is no censoring, that is, when $G(t) = 1$, the asymptotic relative efficiency ρ in (12) is reduced to

$$\rho = \frac{(1-\pi_0)^2}{\pi_0 [log(\pi_0)]^2}.$$

5 Comparison

We investigated three important issues. First, we studied the relative efficiency of the optimal log-rank test versus the standard log-rank test. Second, we evaluated the robustness of the optimal and standard log-rank tests when the hazard parameter was misspecified in the trial design. Third, we investigated the performance of the two sample size formulae under various design scenarios.

The relative efficiency ρ given in equation (12) was calculated for selected cure rates under the exponential cure model with an uncured hazard parameter $\lambda = 1$. Assume a uniform accrual over $[0, \tau]$ and no follow-up period, where τ is determined by the percentage of censoring ranging from 0% to 50%. The results (Table 1) showed that when the cure rate π_0 was at most 10% and there was no censoring, the gain in efficiency of the optimal log-rank test versus the standard log-rank test was more than 50%, whereas if the cure rate π_0 was at least 50%, the gain in efficiency was less than 5%. If the percentage of censoring was more than 50%, then the gain in efficiency was less than 10%, regardless of the cure rate. We also investigated the relative efficiency through the sample size calculations. Under the same assumptions, sample sizes were calculated under various combinations of the cure rates of two groups. Similarly, the largest gain in efficiency was achieved when both the cure rate and percentage of censoring were small (Table 2).

To evaluate the robustness of the two tests, sample sizes (n) were calculated under exponential models with hazard parameters λ = 0.1 and 1. Cure rates were set to $\pi_1 = 0.1$ and $\pi_2=\pi_1 e^{\gamma_0}/(1-\pi_1+\pi_1 e^{\gamma_0})$, where γ_0 ranged from 1.5 to 2.0, accrual time $t_a = 1$, and follow-up time $t_f = 2$. Sample sizes (n^*) were also calculated under misspecification of the hazard parameter within a range of λ±20%λ. The %diff = $100(n^*-n)/n$ was calculated for the evaluation of robustness. The results showed that both tests were sensitive to the misspecification of the hazard parameter. However, the %diff was similar for both tests, and the optimal test was slightly more sensitive than the standard log-rank test (Table 3).

To investigate the performance of the sample size formulae for the optimal and standard log-rank tests, we calculated sample sizes under the cure model (1), where cure rates were set as in Table 3, and the conditional survival distribution was Weibull, $S(t)=e^{-\lambda t^{\kappa}}$, or log-logistic, $S(t)=\frac{1}{1+\lambda t^{\kappa}}$. The scale parameter λ was set to 0.4, and the shape parameter κ was set to 0.5, 1, or 2, reflecting a decreasing, constant, and increasing hazard function, respectively, for the Weibull distribution; and a decreasing and single-mode hazard function for the log-logistic distribution. We assumed that subjects were recruited with a uniform distribution over the accrual period $t_a = 1$, with a follow-up period $t_f = 2$. We further assumed that no subject was lost to follow-up during the study. Then, the censoring time was uniformly distributed over the interval $[t_f, t_a + t_f]$, that is, the censoring survival distribution $G(t) = 1$ if $t \le t_f$, $= (t_a + t_f - t)/t_a$ if $t_f \le t \le t_a + t_f$, $= 0$ otherwise. Therefore, given a two-sided nominal significance level of 0.05 and power of 90%, the required sample sizes were calculated for each design scenario under each distribution. The empirical type I errors and powers of the corresponding designs were simulated based on 100,000 runs. The simulation results presented in Table 4 can be summarized as follows. First, the empirical powers of both the optimal and standard log-rank tests were close to the nominal level of 90%. Thus, the sample sizes were adequately estimated. Second, the empirical type I errors of both tests were close to the nominal level of 5%. Thus, both tests preserved type I error well. Third, the sample sizes calculated from the optimal test were smaller than those calculated for the standard log-rank test.

Overall, the results showed that the derived sample size formulae provide adequate sample size estimation for trial design if the main interest is to detect the differences between the cure rates of two groups and that the optimal test is more efficient than the standard log-rank test, particular when both cure rates and percentage censoring are small.

6 Example

We illustrate study design under a parametric cure model by using the data from the Eastern Cooperative Oncology Group (ECOG) trial e1684. The ECOG trial e1684 was a two-arm phase III clinical trial to compare the relapse-free survival (RFS) of patients with melanoma who were treated with high-dose interferon alpha-2b or placebo as postoperative adjuvant therapy. The trial accrued patients between 1984 and 1990 and remained blinded under analysis until 1993 (Kirkwood, et al., 1996). Researchers have studied this dataset extensively using cure models (Corbiere and Joly, 2007). There were 92 deaths among the 146 patients in the treatment group. The SAS macro PSPMCM was applied to this data to fit

the treatment arm data under the Weibull cure model (Figure 1), with an estimated shape parameter κ of 1.018, scale parameter λ of 0.836, and a cure rate of 35%. Suppose we wish to design a two-arm randomized phase III trial to detect a 20% difference between the cure rate in the arm that receives the new treatment and that in the control arm that receives the same therapy as the treatment arm of the ECOG trial, with a two-sided type I error of 0.05, power of 90% at the alternative, a uniform accrual with a 5-year accrual period and 5-year of follow-up, no loss to follow-up, and equal allocation between the two groups. Then, the required sample sizes calculated using formulae (10) and (11) under the Weibull cure model are 266 and 280 patients, respectively. The corresponding simulated empirical type I error and power are 0.05 and 91.4% for the optimal log-rank test, and 0.05 and 90.7% for the standard log-rank test. As the cure rate is relatively high, the gain in efficiency is only approximately 5% in this example.

7 Conclusion

For cancer clinical trials in which a portion of patients are cured, the main interest is in demonstrating the differences between the cure rates in the two treatment groups. In this article, sample size formulae are derived for both the optimal and standard log-rank tests. Because the proposed cure model is not a proportional hazards model, the standard log-rank test is not fully efficient. Thus, a sample size calculation derived under the optimal test can ensure the efficacy of the study design. The optimal log-rank test is implemented in the standard statistical software R by using the **survdiff** function with the option **rho** = −**1**. The simulation results demonstrated that the sample size formula for the optimal test provides adequate sample size estimation and is more efficient than the formula for the standard log-rank test. Finally, if trials are planned to include interim analyses to enable them to be halted early if futility or efficacy is demonstrated, then the group sequential methods developed by Lee and Sather (1995) can be used.

Acknowledgments

The author acknowledges an anonymous reviewer for his/her valuable comments that improved an earlier version of the paper. This work was supported in part by the National Cancer Institute support grant CA21765 and ALSAC.

Appendix: Derivation of the asymptotic distribution of the weighted log-rank test

The weighted score test is given by

$$U_w = n^{-1/2} \int_0^\infty W(t) \left\{ \frac{Y_1(t)}{Y(t)} dN_2(t) - \frac{Y_2(t)}{Y(t)} dN_1(t) \right\},$$

where $n = n_1 + n_2$ is the total sample size of two groups, $W(t)$ is a weight function that converges in probability to $w(t)$, $N_j(t)$ is the number of observed failures by time t, $Y_j(t)$ is the number of subjects at risk just prior to t in groups $j = 1, 2$, and $Y(t) = Y_1(t) + Y_2(t)$. If we define martingale processes such that $M_j(t) = N_j(t) - \int_0^t \lambda_j^*(t) Y_j(t) dt$, $j = 1, 2$, where $\lambda_j^*(t)$ is given in equation (5), then the weighted score test can be written as

$$U_w=n^{-1/2}\int_0^\infty W(t)\left\{\frac{Y_1(t)}{Y(t)}dM_2(t)-\frac{Y_2(t)}{Y(t)}dM_1(t)\right\}+\int_0^\infty W(t)\frac{Y_1(t)Y_2(t)}{nY(t)}n^{1/2}\{\lambda_2^*(t)-\lambda_1^*(t)\}\,dt.$$

Under the null hypothesis $H_0: \gamma = 0$, we have $\lambda_1^*(t)=\lambda_2^*(t)=\lambda_0^*(t)$, where

$$\lambda_0^*(t)=\frac{(1-\pi_0)S(t)}{\pi_0+(1-\pi_0)S(t)}\lambda(t).$$

Hence, by the martingale property, the mean of U_w is 0 and the variance of U_w is given by

$$\mathrm{Var}(U_w)=n^{-1}E\int_0^\infty W^2(t)\frac{Y_1(t)Y_2(t)}{Y(t)}d\Lambda_0^*(t),$$

where $\Lambda_0^*(t)=\int_0^t\lambda_0^*(u)du$. As

$$n^{-1}\frac{Y_1(t)Y_2(t)}{Y(t)}=\frac{n_1n_2}{n^2}\frac{\{Y_1(t)/n_1\}\{Y_2(t)/n_2\}}{Y(t)/n}\to p(1-p)\frac{\pi_1(t)\pi_2(t)}{\pi(t)},$$

where $p=\lim_{n\to\infty}n_1/n, \pi_j(t)=P(T_{ij}^*>t)$ and $\pi(t) = p\pi_1(t) + (1 - p)\pi_2(t)$. Thus, by the martingale central limit theorem (Fleming and Harrington, 1991), $U_w\to N(0,\sigma_w^2)$, where

$$\sigma_w^2=p(1-p)\int_0^\infty w^2(t)G(t)S_0^*(t)\lambda_0^*(t)dt,$$

for which $S_0^*(t)=\pi_0+(1-\pi_0)S(t)$ and $G(t)$ is the common survival distribution of the censoring time of the two groups. By noting that $S_0^*(t)\lambda_0^*=(1-\pi_0)S(t)\lambda(t)$, we have

$$\sigma_w^2=p(1-p)(1-\pi_0)\int_0^\infty w^2(t)G(t)S(t)\lambda(t)dt. \quad (13)$$

The variance σ_w^2 can be estimated by

$$\hat\sigma_w^2=n^{-1}\int_0^\infty W^2(t)\frac{Y_1(t)Y_2(t)}{Y(t)}d\hat\Lambda_0^*(t),$$

where $d\hat\Lambda_0^*(t)=dN(t)/Y(t)$ and $N(t) = N_1(t) + N_2(t)$. Therefore, the weighted log-rank test $L_w=U_w/\hat\sigma_w$ is asymptotically standard normal distributed under the null hypothesis.

To derive the asymptotic distribution of the weighted log-rank test under the alternative, consider a sequence of local alternatives $H_1^{(n)}:S_j^{*(n)}(t)=1-e^{(-1)^j\gamma_n}(1-\pi_0)\{1-S(t)\}$, or

$$\lambda_j^{*(n)}(t)=\frac{e^{(-1)^j\gamma_n}(1-\pi_0)S(t)}{1-e^{(-1)^j\gamma_n}(1-\pi_0)+e^{(-1)^j\gamma_n}(1-\pi_0)S(t)}\lambda(t),$$

where $n^{1/2}\gamma_n = \gamma_a < \infty$, and define martingale processes as $M_j^{(n)}(t)=N_j(t) - \int_0^t Y_j(u)\lambda_j^{*(n)}(u)\,du$. Then, $U_w = U_{1w}+U_{1w}+U_{2w}$, where

$$U_{1w}=n^{-1/2}\int_0^\infty W(t)\left\{\frac{Y_2(t)}{Y(t)}dM_1^{(n)}(t) - \frac{Y_1(t)}{Y(t)}dM_2^{(n)}(t)\right\},$$

and

$$U_{2w}=n^{-1/2}\int_0^\infty W(t)\frac{Y_1(t)Y_2(t)}{Y(t)}\left\{\lambda_1^{*(n)}(t) - \lambda_2^{*(n)}(t)\right\}dt.$$

As $\gamma_n \to 0$, $H_1^{(n)} \to H_0$ and, $\lambda_j^{*(n)}(t) \to \lambda_0^*(t)$, and by the martingale central limiting theorem, U_{1w} converges to a normal variable with mean $EU_{1w} = 0$ and variance

$$\begin{aligned}EU_{1w}^2=&n^{-1}E\int_0^\infty W^2(t)\left\{\frac{Y_2^2(t)}{Y^2(t)}Y_1(t)\lambda_1^{*(n)}(t)+\frac{Y_1^2(t)}{Y^2(t)}Y_2(t)\lambda_2^{*(n)}(t)\right\}du \to p(1\\&-p)\int_0^\infty w^2(t)\left\{(1-p)\frac{\pi_2^2(t)\pi_1(t)}{\pi^2(t)}\lambda_0^*(t)+p\frac{\pi_1^2(t)\pi_2(t)}{\pi^2(t)}\lambda_0^*(t)\right\}dt\\=&p(1\\&-p)\int_0^\infty w^2(t)\frac{\pi_1(t)\pi_2(t)}{\pi(t)}\lambda_0^*(t)du\\=&p(1-p)\int_0^\infty w^2(t)G(t)S_0^*(t)\lambda_0^*(t)dt=\sigma_w^2.\end{aligned}$$

By Taylor's expansion of $\lambda_j^*(t)$ at $\gamma_n = 0$, we have

$$\lambda_j^*(t) \simeq \frac{(1-\pi_0)S(t)}{\pi_0+(1-\pi_0)S(t)}\lambda(t)+\frac{(1-\pi_0)S(t)}{\{\pi_0+(1-\pi_0)S(t)\}^2}\lambda(t)(-1)^j\gamma_n.$$

It then follows that

$$\lim_{n\to\infty} n^{1/2}\{\lambda_2^*(t) - \lambda_1^*(t)\} = \frac{2\gamma_a(1-\pi_0)S(t)\lambda(t)}{\{\pi_0+(1-\pi_0)S(t)\}^2}.$$

By substituting this into U_{2w}, we have shown that U_{2w} converges in probability to $\mu(w, \gamma_a)$, where

$$\mu(w,\gamma_a)=2p(1-p)(1-\pi_0)\gamma_a\int_0^\infty w(t)\{S_0^*(t)\}^{-1}G(t)S(t)\lambda(t)dt.$$

Thus, under the local alternatives $H_1^{(n)}$, the weighted log-rank test is asymptotically normal distributed with mean μ_w/σ_w and unit variance, that is,

$$L_w=U_w/\hat{\sigma}_w \to N(\mu(w,\gamma_a)/\sigma_w,1).$$

References

Corbiere F, Joly P. A SAS macro for parametric and semiparametric mixture cure models. Computer Methods and Programs in Biomedicine. 2007; 85:173–180. [PubMed: 17157948]

Ewell M, Ibrahim JG. The large sample distribution of the weighted log rank statistic under general local alternatives. Lifetime Data Analysis. 1997; 3:5–12. [PubMed: 9384622]

Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics. 1982; 38:1041–1046. [PubMed: 7168793]

Fleming, TR., Harrington, DP. Counting processes and survival analysis. John Wiley and Sons; New York: 1991.

Gray RJ, Tsiatis AA. A linear rank test for use when the main interest is in differences in cure rates. Biometrics. 1989; 45:899–904. [PubMed: 2790128]

Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982; 69:553–566.

Kirkwood JM, Straderman MH, Ernstoff MS, Smith TJ, Borden EC, Blum RH. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. Journal of Clinical Oncology. 1996; 14:7–17. [PubMed: 8558223]

Kuk AYC, Chen CH. A mixture model combining logistic regression with proportional hazards regression. Biometrika. 1992; 79:531–541.

Lee JW, Sather HN. Group sequential methods for comparison of cure rates in clinical trials. Biometrics. 1995; 51:756–763. [PubMed: 7662857]

Peng Y, Dear KBG. A nonparametric mixture model for cure rate estimation. Biometrics. 2000; 56:237–243. [PubMed: 10783801]

Peng Y, Dear KBG, Denham JW. A generalized F mixture model for cure rate estimation. Statistics in Medicine. 1998; 17:813–830. [PubMed: 9595613]

Randales, RH., Wolfe, DA. Introduction to the theory of nonparametric statistics. John Wiley & Sons; New York: 1979.

Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. Biometrics. 2000; 56:227–236. [PubMed: 10783800]

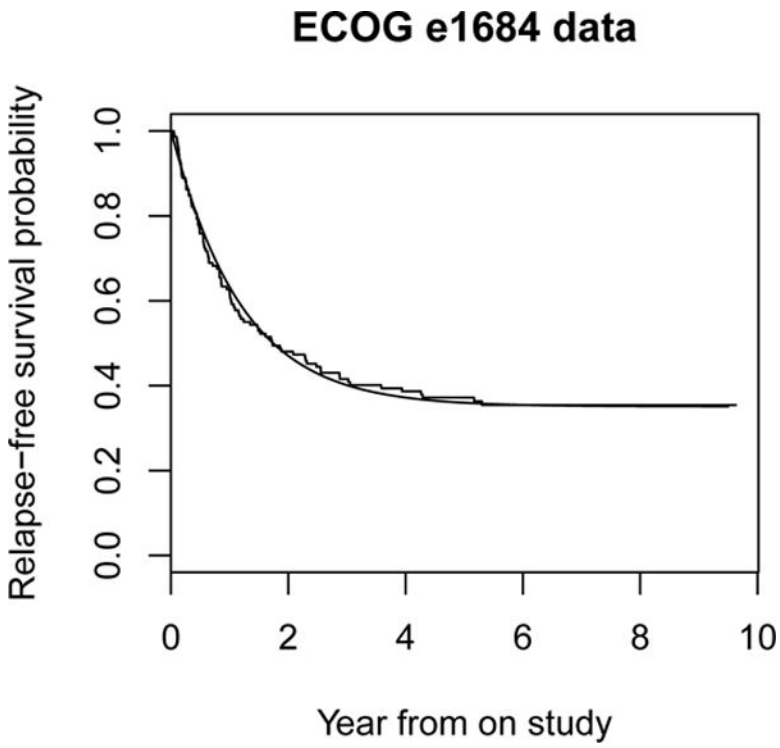

Figure 1.
Relapse-free survival for ECOG e1864 data. The step function is the Kaplan-Meier survival curve. The solid curve is the fitted Weibull cure model.

Table 1

The relative efficiency ρ of the optimal log-rank test compared to the standard log-rank test under the exponential model with a hazard parameter $\lambda = 1$ and a uniform accrual over the interval $[0, \tau]$, where τ is determined by the percentage of censoring.

	Cure rate π_0								
Cens	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
None	1.528	1.235	1.127	1.072	1.041	1.022	1.011	1.004	1.001
10%	1.490	1.221	1.120	1.068	1.039	1.021	1.010	1.004	1.001
20%	1.399	1.190	1.105	1.061	1.035	1.019	1.009	1.004	1.001
30%	1.272	1.144	1.084	1.050	1.029	1.016	1.008	1.003	1.001
40%	1.166	1.099	1.061	1.037	1.022	1.012	1.006	1.002	1.001
50%	1.095	1.061	1.040	1.026	1.016	1.009	1.005	1.002	1.000

Censoring time was uniformly distributed over $[0, \tau]$, with the value of τ being chosen so that the probability of the failure time being censored for a subject who was not cured was the specified censoring percentage. Abbreviation: Cens: censoring.

J Biopharm Stat. Author manuscript; available in PMC 2017 August 30.

Table 2

Sample sizes for the optimal and standard log-rank tests for various cure rates in two groups with a nominal type I error of 5% and power of 90%. Here, sample sizes were calculated under the exponential model, with a hazard parameter $\lambda = 1$ and a uniform accrual over the interval $[0, \tau]$, where τ is determined by the percentage of censoring.

		Cure rate (π_1, π_2)						
		(.05, .15)	(.05, .2)	(.1, .2)	(.1, .3)	(.2, .4)	(.3, .5)	(.4, .6)
Test	**Cens**	**Sample size**						
L	None	598	301	738	217	257	279	281
	10%	766	379	916	263	304	323	321
	20%	1067	513	1218	338	375	388	379
	30%	1566	730	1697	453	481	483	460
	40%	2323	1058	2415	623	632	616	573
	50%	3479	1577	3509	881	860	813	740
L_K	None	394	215	554	177	230	261	270
	10%	517	275	698	217	272	303	310
	20%	766	391	964	286	341	367	367
	30%	1233	597	1423	398	445	461	448
	40%	1993	926	2144	569	597	594	562
	50%	3180	1437	3262	831	828	794	729

Censoring time was uniformly distributed over $[0, \tau]$, with the value of τ being chosen so that the probability of the failure time being censored for a subject who was not cured was the specified censoring percentage. Abbreviations: Cens: censoring; L: standard log-rank test; L_K: optimal log-rank test.

Table 3

Sample sizes for the exponential cure models under misspecification of the hazard parameter λ, with cure rates $\pi_1 = 0.1$ and $\pi_2 = \pi_1 e^{\gamma_0}/(1-\pi_1 + \pi_1 e^{\gamma_0})$, uniform accrual with accrual time $t_a = 1$ and follow-up time $t_f = 2$, and nominal type I error of 5% and power of 90%.

	True λ		Misspecified λ			
		λ = 0.1	λ = 0.08		λ = 0.12	
Test	γ_0	n	n^*	%diff	n^*	%diff
L	1.5	2282	2885	26.4	1880	−17.6
	1.6	1873	2367	26.4	1545	−17.5
	1.7	1554	1963	26.3	1283	−17.4
	1.8	1302	1643	26.2	1075	−17.4
	1.9	1100	1387	26.1	909	−17.4
	2.0	938	1181	25.9	776	−17.3
L_K	1.5	2274	2879	26.6	1872	−17.7
	1.6	1868	2363	26.5	1538	−17.7
	1.7	1550	1959	26.4	1278	−17.5
	1.8	1298	1640	26.3	1071	−17.5
	1.9	1097	1385	26.3	906	−17.4
	2.0	935	1179	26.1	773	−17.3
		λ = 1	λ = 0.8		λ = 1.2	
Test	γ_0	n	n^*	%diff	n^*	%diff
L	1.5	219	259	18.3	197	−10.0
	1.6	185	218	17.8	167	−9.7
	1.7	158	185	17.1	144	−8.9
	1.8	137	159	16.1	124	−9.5
	1.9	119	138	16.0	109	−8.4
	2.0	105	121	15.2	96	−8.6
L_K	1.5	193	233	20.7	170	−11.9
	1.6	164	198	20.7	146	−11.0

J Biopharm Stat. Author manuscript; available in PMC 2017 August 30.

	λ = 1		λ = 0.8		λ = 1.2	
Test	γ_0	n	n^*	**%diff**	n^*	**%diff**
	1.7	142	169	19.0	127	−10.6
	1.8	123	146	18.7	111	−9.8
	1.9	108	128	18.5	98	−9.3
	2.0	96	112	16.7	87	−9.4

%diff: change in sample size through misspecified hazard parameter λ, i.e., %diff =100 × (n^* − n)/n, where n is the sample size calculated under the true λ and n^* is the sample size calculated under the misspecified λ. Abbreviations: L: standard log-rank test; L_K: optimal log-rank test.

J Biopharm Stat. Author manuscript; available in PMC 2017 August 30.

Table 4

Sample sizes (n) and corresponding simulated empirical type I errors $(\hat{\alpha})$ and powers $(1-\hat{\beta})$ for the optimal and standard log-rank tests under the Weibull and log-logistic distributions, with a scale parameter λ = 0.4, cure rates π_1 = 0.1 and $\pi_2=\pi_1 e^{\gamma_0}/(1-\pi_1+\pi_1 e^{\gamma_0})$, nominal type I error of 0.05, power of 90%, and uniform accrual with accrual time t_a = 1 and follow-up time t_f = 2.

			κ = 0.5			κ=1			κ = 2		
Dist	**Test**	γ_0	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$
WB	L	1.5	841	.048	.905	510	.053	.905	222	.052	.914
		1.6	695	.049	.900	424	.050	.899	188	.052	.914
		1.7	580	.051	.901	355	.045	.906	161	.050	.922
		1.8	488	.050	.903	301	.050	.906	139	.051	.924
		1.9	415	.049	.905	258	.048	.907	121	.052	.921
		2.0	356	.051	.907	222	.053	.906	106	.050	.925
	L_K	1.5	827	.049	.901	490	.053	.904	195	.051	.919
		1.6	683	.045	.901	408	.048	.910	166	.055	.919
		1.7	571	.051	.902	343	.051	.902	143	.048	.925
		1.8	481	.049	.905	291	.050	.906	125	.052	.928
		1.9	410	.053	.904	250	.052	.909	110	.052	.926
		2.0	351	.052	.906	216	.047	.910	97	.052	.932
LG	L	1.5	1112	.048	.900	762	.052	.908	404	.048	.906
		1.6	916	.050	.907	630	.049	.903	337	.049	.908
		1.7	763	.047	.908	526	.053	.904	284	.051	.908
		1.8	641	.047	.905	443	.050	.907	241	.050	.916
		1.9	544	.049	.903	377	.050	.907	207	.049	.915
		2.0	465	.048	.900	324	.054	.907	180	.050	.912
	L_K	1.5	1100	.048	.902	746	.051	.903	382	.053	.907
		1.6	907	.049	.906	617	.045	.897	319	.053	.909
		1.7	755	.050	.908	516	.052	.898	270	.051	.914
		1.8	635	.048	.903	436	.051	.906	230	.050	.911

J Biopharm Stat. Author manuscript; available in PMC 2017 August 30.

			κ= 0.5			κ=1			κ= 2		
Dist	**Test**	γ_0	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$
		1.9	539	.049	.908	371	.056	.903	198	.051	.910
		2.0	461	.053	.904	319	.049	.910	172	.050	.916

Abbreviations: Cens: censoring; Dist: distribution; WB: Weibull; LG: log-logistic; L: standard log-rank test; L_K: optimal log-rank test.