



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2016)
16, Number 3, pp. 678–690

Versatile tests for comparing survival curves based on weighted log-rank statistics

Theodore G. Karrison
Department of Public Health Sciences
University of Chicago
Chicago, IL
tkarrison@health.bsd.uchicago.edu

Abstract. The log-rank test is perhaps the most commonly used nonparametric method for comparing two survival curves and yields maximum power under proportional hazards alternatives. While the assumption of proportional hazards is often reasonable, it need not hold. Several authors have therefore developed versatile tests using combinations of weighted log-rank statistics that are more sensitive to nonproportional hazards. [Fleming and Harrington \(1991, *Counting Processes and Survival Analysis*, Wiley\)](#) consider the family of G^p statistics and their supremum versions, while [Lee \(1996, *Biometrics* 52: 721–725\)](#) and [Lee \(2007, *Computational Statistics and Data Analysis* 51: 6557–6564\)](#) propose tests based on the more extended $G^{p,\gamma}$ family. In this article, I consider $Z_m = \max(|Z_1|, |Z_2|, |Z_3|)$, where Z_1 , Z_2 , and Z_3 are z statistics obtained from $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ tests, respectively. $G^{0,0}$ corresponds to the log-rank test, while $G^{1,0}$ and $G^{0,1}$ are more sensitive to early and late-difference alternatives. I conduct a simulation study to compare the performance of Z_m with the log-rank test, the more optimally weighted test, and Lee's (2007) tests, under the null hypothesis, proportional hazards, early difference, and late-difference alternatives. Results indicate that the method based on Z_m maintains the type I error rate, provides increased power relative to the log-rank test under early difference and late-difference alternatives, and entails only a small to moderate power loss compared with the more optimally chosen test. I apply the procedure to two datasets reported in the literature, both of which exhibit nonproportional hazards. Versatile tests such as Z_m may be useful in clinical trial settings where there is concern that the treatment effect may not conform to the proportional hazards assumption. I also describe the syntax for a Stata command, `verswlr`, to implement the method.

Keywords: st0449, verswlr, survival curves, log-rank test, nonproportional hazards, versatile tests, power

1 Introduction

In randomized clinical trials, as well as observational studies, interest is often focused on the comparison of two survival curves corresponding to different treatment arms or patient subgroups. The log-rank test is perhaps the most commonly used nonparametric procedure for performing such comparisons. The log-rank test is known to yield maximum power under proportional hazards (PH) alternatives, that is, alternatives in which the ratio of the death (or hazard) rates in the two groups is constant over time.

While many datasets exhibit proportional or nearly proportional hazard functions, it is not uncommon for curves to deviate from this assumption. For example, the effect of a treatment or covariate may wane over time, leading to a decreasing hazard ratio and a “closing up” of the two survival curves. Conversely, a treatment may have a delayed effect whereby the survival curves do not separate until a certain interval of time has elapsed. Several authors have therefore described versatile tests using combinations of weighted log-rank statistics that are more sensitive to nonproportional hazard (non-PH) alternatives. Tarone (1981) considers the maximum of the log-rank and generalized Wilcoxon statistics. Fleming and Harrington (1991) consider the family of G^ρ statistics and their supremum versions, while Lee (1996) and Lee (2007) propose tests based on the extended $G^{\rho,\gamma}$ family. The $G^{\rho,\gamma}$ family of test statistics provides a set of weighted log-rank tests, where the weights are governed by the ρ and γ parameters. $G^{0,0}$ specifies equal weights and is therefore equivalent to the log-rank test. $G^{1,0}$ places more weight on the earlier time points and corresponds to the Prentice–Wilcoxon statistic; $G^{0,1}$ places more weight on the later time points and hence is more sensitive to late-difference alternatives. Lee (2007) considers $|Z_1 + Z_2|/2$, $(|Z_1| + |Z_2|)/2$, and $\max(|Z_1|, |Z_2|)$, where Z_1 and Z_2 are Z statistics obtained from $G^{1,0}$ and $G^{0,1}$ tests, respectively. Lee (1996), on the other hand, evaluates the maximum over four Z statistics derived from $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, and $G^{2,2}$ tests, as well as their average. More recently, Yang and Prentice (2010) developed a test using adaptive weights that maintains optimality under PH and improves power under non-PH alternatives relative to the log-rank test.

In this article, I consider $\max(|Z_1|, |Z_2|, |Z_3|)$, where Z_1 , Z_2 , and Z_3 are Z statistics obtained from $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ tests. This particular combination will provide relatively good coverage across the range of likely possibilities: proportional hazards, early difference, and late-difference configurations. The user has the option, however, of specifying a different combination of tests. The test is straightforward to implement using Stata software and trivariate normal calculations.

2 Methods and formulas

The $G^{\rho,\gamma}$ family of weighted log-rank statistics can be expressed as

$$G^{\rho,\gamma} = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \int_0^\infty \{\hat{S}(t-)\}^\rho \{1 - \hat{S}(t-)\}^\gamma \frac{\bar{Y}_1(t) \bar{Y}_2(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)} \left\{ \frac{d\bar{N}_1(t)}{\bar{Y}_1(t)} - \frac{d\bar{N}_2(t)}{\bar{Y}_2(t)} \right\}$$

where $\hat{S}(t-)$ is the Kaplan–Meier (Kaplan and Meier 1958) estimate of the survival rate based on the pooled data (pooled over the two groups), $\bar{Y}_i(t)$ is the number of patients in group i at risk at time t , and $\bar{N}_i(t)$ is the number of failures in group i up to and including time t . More generally, if we let

$$W_{K_l} = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \int_0^\infty K_l(t) \frac{\bar{Y}_1(t) \bar{Y}_2(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)} \left\{ \frac{d\bar{N}_1(t)}{\bar{Y}_1(t)} - \frac{d\bar{N}_2(t)}{\bar{Y}_2(t)} \right\}$$

denote a weighted log-rank statistic with weight function $K_l(t)$, then $G^{0,0}$ corresponds to $K_l(t) = 1$, $G^{1,0}$ corresponds to $K_l(t) = \hat{S}(t-)$, and $G^{0,1}$ corresponds to $K_l(t) = 1 - \hat{S}(t-)$. Thus the $G^{0,0}$ or log-rank test has equal weights, whereas $G^{1,0}$ places more weight at early time points and $G^{0,1}$ at later ones.

Let Z_1 , Z_2 , and Z_3 denote Z statistics obtained from $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ tests, respectively. $\mathbf{Z} = (Z_1, Z_2, Z_3)'$ has an asymptotic, trivariate normal distribution with a variance-covariance matrix that can readily be computed using standard statistical software. To see this, note that the covariance between $W_{K_l}(t)$ and $W_{K_m}(t)$ is

$$\begin{aligned} \hat{\sigma}_{lm} = & \frac{n_1 + n_2}{n_1 n_2} \int_0^\infty K_l(t) K_m(t) \frac{\bar{Y}_1(t) \bar{Y}_2(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)} \\ & \left(1 - \frac{\Delta \bar{N}_1(t) + \Delta \bar{N}_2(t) - 1}{\bar{Y}_1(t) + \bar{Y}_2(t) - 1} \right) \left[\frac{d\{\bar{N}_1(t) + \bar{N}_2(t)\}}{\bar{Y}_1(t) + \bar{Y}_2(t)} \right] \end{aligned}$$

This is convenient because $\text{Cov}(G^{0,0}, G^{1,0}) = \text{Var}(G^{1/2,0})$, $\text{Cov}(G^{0,0}, G^{0,1}) = \text{Var}(G^{0,1/2})$, and $\text{Cov}(G^{1,0}, G^{0,1}) = \text{Var}(G^{1/2,1/2})$. In general,

$$\text{Cov}(G^{\rho_1, \gamma_1}, G^{\rho_2, \gamma_2}) = \text{Var}(G^{(\rho_1 + \rho_2)/2, (\gamma_1 + \gamma_2)/2})$$

Thus software routines that compute and store the variance of $G^{\rho, \gamma}$ statistics, such as the `sts test` routine in Stata, can be used to calculate the variance and covariance terms. Letting $Z_m = \max(|Z_1|, |Z_2|, |Z_3|)$, the p -value for Z_m can be derived by integrating under the trivariate normal density. I use the algorithm described by [Drezner \(1994\)](#), which yields errors on the order of 10^{-5} for matrix determinants ≤ 0.15 and provides errors less than 10^{-7} for determinants > 0.15 .

3 The verswlr command

3.1 Syntax

`verswlr varname [if] [in] [, options]`

varname should contain the group indicator variable and must be numeric.

See section 3.3 for the available *options*.

`by` is allowed; see `[D]` `by`.

You must `stset` your data before using `verswlr`; see `[ST]` `stset`.

3.2 Description

`verswlr` determines the significance level of the maximum of $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ weighted log-rank statistics or three other user-defined tests for the comparison of two survival curves. Output and returned scalars are the sample size, chi-squared statistic for each test, maximum Z statistic (absolute value), and two-sided p -value.

3.3 Options

`rho1(#)` and `gamma1(#)` specify the weights for the first test.

`rho2(#)` and `gamma2(#)` specify the weights for the second test.

`rho3(#)` and `gamma3(#)` specify the weights for the third test.

The default values are (0 0), (1 0), and (0 1), respectively.

3.4 Stored results

`verswlr` stores the following in `r()`:

Scalar	
<code>r(sampsize)</code>	sample size
<code>r(fh11)</code>	G^{ρ_1, γ_1} chi-squared statistic
<code>r(fh22)</code>	G^{ρ_2, γ_2} chi-squared statistic
<code>r(fh33)</code>	G^{ρ_3, γ_3} chi-squared statistic
<code>r(maxz)</code>	maximum Z statistic
<code>r(pval)</code>	two-sided p -value

4 Simulation results

I conducted a simulation study to compare the performance of the versatile method based on Z_m with the log-rank test and with the more “optimally” weighted test under the null hypothesis, proportional hazards, early difference, and late-difference alternatives. I also examined the three versatile procedures suggested by Lee (2007); that is, $|Z_1 + Z_2|/2$, $(|Z_1| + |Z_2|)/2$, and $\max(|Z_1|, |Z_2|)$, where Z_1 and Z_2 are Z statistics obtained from $G^{1,0}$ and $G^{0,1}$ tests, respectively. Data were generated from two Weibull distributions,

$$S_k(t) = \exp\{-(\lambda_k t)^{\gamma_k}\} \quad k = 1, 2$$

where λ_k and γ_k are the shape and scale parameters, respectively, for group k . Parameter values were set to represent four configurations for the treatment or group effect: no difference, proportional hazards, early difference, and late-difference alternatives (see figure 1). Clinical trials were simulated with 1) an accrual period of $a = 2$ years and follow-up period of $f = 3$ years, giving rise to uniform censoring between 3 and 5 years, and 2) an accrual period of $a = 3$ years and follow-up period of $f = 2$ years, giving rise to uniform censoring between 2 and 5 years. The observed survival time was taken as the minimum of randomly drawn survival and censoring times, with the indicator

variable set equal to 1 or 0 accordingly. $R = 5000$ replications were run for each configuration and sample size, providing a simulation error of no more than $\pm 1.4\%$. For the four scenarios, the median percent censored was 47%, 48%, 49%, and 41%, respectively, under (item 1) and 53%, 54%, 54%, and 49%, respectively, under (item 2).

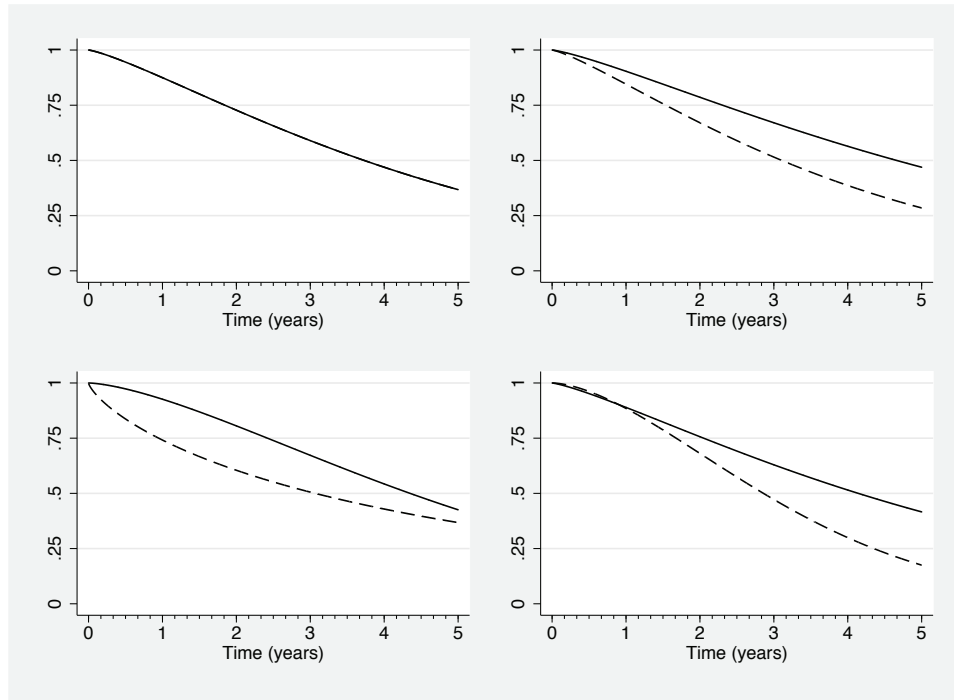


Figure 1. Survival configurations used in the simulation study (upper left: null, upper right: PH, lower left: early difference, lower right: late difference)

For each simulated dataset, $G^{0,0}$, $G^{1,0}$, $G^{0,1}$, Z_m , and the three tests proposed by Lee (2007) were performed. Under the null configuration, an “uncorrected” test, $Z_m(u)$, obtained by declaring statistical significance if $|Z_m| > 1.96$, was also conducted. The results are shown in tables 1 and 2. They indicate that the Z_m test maintains the type I error rate, while $Z_m(u)$ has an inflated error rate as high as 9%. All the remaining tests preserve the nominal alpha level. Under PH, the log-rank test has maximum power, as expected. However, the Z_m test comes close with a power decrease of only 2%–3%. $G^{1,0}$ also had a rather strong showing under PH. Under early and late-difference alternatives, the Z_m test provides increased power, ranging from 3% to 13% greater, relative to the log-rank test. The power loss vis-à-vis the more optimally chosen test is small to moderate: 2%–9% relative to $G^{1,0}$ under early difference alternatives and 1%–5% compared with $G^{0,1}$ under late-difference alternatives.

Lee’s (2007) two linear combination tests also do quite well under PH alternatives, but they have relatively low power for early difference configurations and were a little

less powerful than the maximum tests for late-difference alternatives. This behavior is similar to that observed in Lee's (2007) own simulations. The $\max(|Z_1|, |Z_2|)$ test of Lee (2007) fares just slightly poorer to Z_m under PH alternatives. This might be expected because it does not include $G^{0,0}$. Conversely, Lee's (2007) maximum test does slightly better than Z_m under early and late-difference alternatives, because it is not paying the additional "price" for incorporating a third test that is not optimal in those cases. Overall, however, the two maximum tests performed quite similarly.

Two additional observations are interesting. One would expect the power to be uniformly greater in table 1 than table 2 because the follow-up is longer with correspondingly lower censoring rates. This is true for PH and late-difference alternatives but not for the early difference alternatives. On further reflection, however, the shorter follow-up under the early difference scenario is actually beneficial because fewer patients approach the four to five year mark where the survival curves have nearly converged. The other observation is that the $G^{0,1}$ test can have very low power under early difference alternatives. This is because it places more weight not only where the difference between the curves is least but also where the variance is higher because of the censoring. Conversely, the $G^{1,0}$ test under late-difference alternatives exhibits an appreciable but less dramatic drop in power.

Table 1. Rejection rates (%) for log-rank, $G^{1,0}$, $G^{0,1}$, Z_m , Lee 1 ($(|Z_1 + Z_2|/2)$), Lee 2 ($(|Z_1| + |Z_2|)/2$), and Lee 3 ($\max(|Z_1|, |Z_2|)$) tests; $R = 5000$ simulations (accrual period: 2 years, follow-up period: 3 years)

Scenario	n^a	Log-rank	$G^{1,0}$	$G^{0,1}$	Z_m	$Z_m(u)$	Lee 1	Lee 2	Lee 3
Null ^b	50	4.8	5.1	5.1	4.8	8.3	5.1	5.1	4.8
Null	75	5.2	5.4	5.5	5.1	9.1	5.6	5.6	5.1
Null	100	5.4	5.2	5.3	4.9	8.5	5.3	5.3	4.9
Null	125	5.0	5.3	5.3	5.3	8.7	5.1	5.1	5.4
Null	150	5.3	5.4	4.9	5.2	8.8	5.1	5.1	5.2
PH ^c	50	43.9	41.8	35.7	41.4	—	43.2	43.2	41.1
PH	75	62.0	59.9	49.6	58.7	—	60.8	60.8	58.4
PH	100	73.6	72.0	61.8	71.3	—	72.4	72.4	70.9
PH	125	81.5	79.7	69.4	78.9	—	80.2	80.2	78.8
PH	150	89.0	87.2	78.8	87.5	—	88.3	88.3	87.0
Early ^d	50	36.4	51.4	6.2	43.2	—	23.5	23.6	43.8
Early	75	50.2	70.4	6.6	62.0	—	31.3	31.9	62.5
Early	100	63.8	82.4	7.3	77.2	—	40.7	43.3	77.8
Early	125	72.7	89.9	7.2	85.7	—	49.5	54.7	86.1
Early	150	80.2	94.4	7.8	91.7	—	55.2	63.6	92.0
Late ^e	50	51.7	38.4	62.2	57.5	—	56.8	56.8	57.5
Late	75	69.5	56.4	80.3	76.0	—	74.5	74.5	76.2
Late	100	81.9	68.3	90.0	87.8	—	86.3	86.3	87.7
Late	125	89.5	77.7	95.7	94.2	—	92.9	92.9	94.2
Late	150	94.2	84.4	97.7	97.0	—	96.2	96.2	97.0

^a Sample-size per group

^b $\lambda_1 = \lambda_2 = 0.20$, $\gamma_1 = \gamma_2 = 1.25$

^c $\lambda_1 = 0.16$, $\lambda_2 = 0.24$, $\gamma_1 = \gamma_2 = 1.25$

^d $\lambda_1 = 0.18$, $\lambda_2 = 0.20$, $\gamma_1 = 1.50$, $\gamma_2 = 0.75$

^e $\lambda_1 = 0.18$, $\lambda_2 = 0.28$, $\gamma_1 = 1.25$, $\gamma_2 = 1.65$

Table 2. Rejection rates (%) for log-rank, $G^{1,0}$, $G^{0,1}$, Z_m , and Lee 1 ($(|Z_1| + |Z_2|)/2$), Lee 2 ($(|Z_1| + |Z_2|)/2$), and Lee 3 ($\max(|Z_1|, |Z_2|)$) tests; $R = 5000$ simulations (accrual period: 3 years, follow-up period: 2 years)

Scenario	n^a	Log rank	$G^{1,0}$	$G^{0,1}$	Z_m	$Z_m(u)$	Lee 1	Lee 2	Lee 3
Null ^b	50	5.5	5.5	5.3	5.4	9.1	5.5	5.5	5.3
Null	75	5.4	5.3	5.5	5.4	8.8	5.6	5.6	5.4
Null	100	5.2	5.2	4.9	5.0	8.6	5.2	5.2	5.0
Null	125	4.5	4.5	4.8	4.7	8.0	4.6	4.6	4.7
Null	150	4.5	4.7	4.9	4.7	8.1	4.5	4.5	4.7
PH ^c	50	40.5	38.9	31.1	38.3	—	39.9	39.9	38.0
PH	75	55.7	54.3	44.1	53.2	—	54.7	54.7	52.9
PH	100	66.7	64.9	53.2	64.4	—	65.3	65.4	63.9
PH	125	76.7	74.8	63.6	74.6	—	76.0	76.0	74.3
PH	150	84.4	82.9	71.3	82.2	—	83.1	83.1	81.9
Early ^d	50	43.2	57.0	8.0	49.0	—	28.8	28.9	49.5
Early	75	58.7	74.8	9.3	67.3	—	39.9	40.5	67.7
Early	100	70.3	85.4	10.6	80.1	—	48.3	50.0	80.5
Early	125	80.8	92.2	10.8	89.2	—	58.2	62.2	89.4
Early	150	86.0	95.7	11.5	93.7	—	64.1	69.9	93.8
Late ^e	50	41.5	31.5	52.5	47.8	—	47.0	47.0	47.7
Late	75	57.5	44.8	70.0	65.7	—	64.2	64.3	65.4
Late	100	69.5	54.8	81.6	77.5	—	76.0	76.0	77.5
Late	125	79.5	64.7	88.9	86.4	—	85.3	85.3	86.5
Late	150	86.5	72.8	94.3	92.8	—	91.7	91.7	92.8

^a Sample-size per group

^b $\lambda_1 = \lambda_2 = 0.20$, $\gamma_1 = \gamma_2 = 1.25$

^c $\lambda_1 = 0.16$, $\lambda_2 = 0.24$, $\gamma_1 = \gamma_2 = 1.25$

^d $\lambda_1 = 0.18$, $\lambda_2 = 0.20$, $\gamma_1 = 1.50$, $\gamma_2 = 0.75$

^e $\lambda_1 = 0.18$, $\lambda_2 = 0.28$, $\gamma_1 = 1.25$, $\gamma_2 = 1.65$

5 Examples

We consider two examples. The first is from the Gastrointestinal Tumor Study Group (GTSG) trial, whose data were utilized by [Stablein, Carter, and Novak \(1981\)](#) as an example of non-PHs. The trial compared chemotherapy alone (chemo) with combination chemotherapy and radiation therapy (chemo+RT) for the treatment of advanced gastric cancer. Kaplan–Meier curves are shown in figure 2 and exhibit an early difference in survival in favor of the chemotherapy alone arm that diminishes by two years.

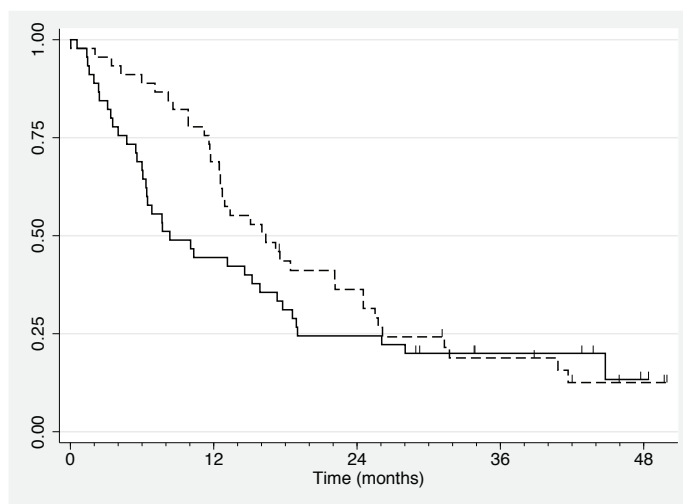


Figure 2. GTSG trial; solid: chemo+RT ($n = 45$), dashed: chemo ($n = 45$)

We read in the data and perform an `stset` (the survival time is contained in the variable `stime`, and the indicator variable is `indicos`). We then execute the `verswlr` command (group assignment is variable `trt`).

```
. use stablein
. stset stime, failure(indicos)
      failure event:  indicos != 0 & indicos < .
obs. time interval:  (0, stime]
exit on or before:   failure
```

```

90  total observations
0   exclusions

```

```

90  observations remaining, representing
74  failures in single-record/single-failure data
47791 total analysis time at risk and under observation
                                at risk from t =      0
                                earliest observed entry t = 0
                                last observed exit t = 1519

```

```
. verswlr trt
```

Versatile Logrank Test

Total Sample Size: $n = 90$

Rho-Gamma tests:

```
fh(rho1,gamma1) = fh(0,0): Chi-square = 1.3163575
fh(rho2,gamma2) = fh(1,0): Chi-square = 4.7309306
fh(rho3,gamma3) = fh(0,1): Chi-square = .26622297
```

Maximum Test:

```
Max abs z: 2.1750702
Two sided p-value: .05608832
```

For these data, the log-rank statistic yields $\chi^2 = 1.32$ and a nonsignificant p -value of 0.25, whereas the $G^{1,0}$ test is nominally significant at $p = 0.030$ ($\chi^2 = 4.73$). $G^{0,1}$ yields $\chi^2 = 0.27$ and a p -value of 0.61. However, one could question the validity of the $G^{1,0}$ result if the test was chosen after inspection of the survival curves. The Z_m test yields a borderline significant p -value of 0.056 ($Z_m = 2.18$).

The second example is from a head-and-neck cancer trial conducted by the Northern Oncology Group (NCOG); the data were analyzed by Efron (1988). Patients were randomized to receive radiation alone (RT) versus chemotherapy plus radiation (chemo+RT) and followed for up to 6 years. The Kaplan–Meier curves (figure 3) do not separate until about 4–6 months, although here the departure from PHs is not so strong as in the first example. The log-rank, $G^{1,0}$, and $G^{0,1}$ tests yield p -values of 0.022, 0.062, and 0.015, respectively; the significance level from the Z_m test is 0.029. The commands and output are as follows:

```
. use efron
. stset stime, failure(indicos)
      failure event:  indicos != 0 & indicos < .
obs. time interval:  (0, stime]
exit on or before:  failure
```

```
          96  total observations
           0  exclusions
```

```
          96  observations remaining, representing
           73  failures in single-record/single-failure data
1541.443    total analysis time at risk and under observation
                                     at risk from t =          0
                                     earliest observed entry t =      0
```

```
. verswlr trt
```

Versatile Logrank Test

Total Sample Size: n = 96

Rho-Gamma tests:

```
fh(rho1,gamma1) = fh(0,0): Chi-square = 5.2377665
fh(rho2,gamma2) = fh(1,0): Chi-square = 3.4765024
fh(rho3,gamma3) = fh(0,1): Chi-square = 5.9240772
```

Maximum Test:

```
Max abs z: 2.4339427
Two sided p-value: .02857177
```

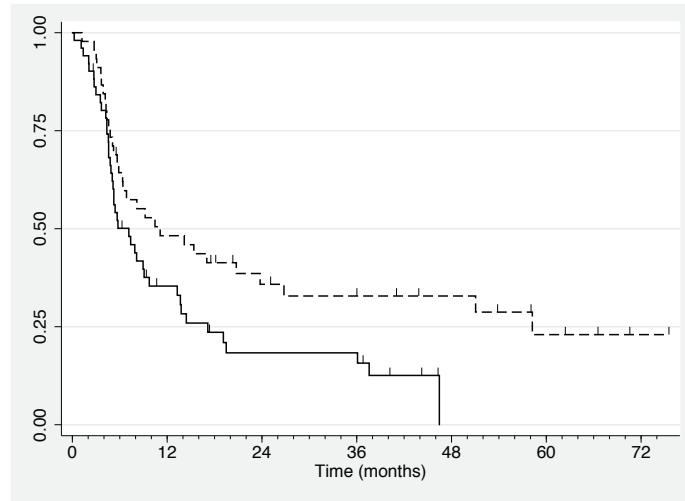


Figure 3. NCOG trial; solid: RT ($n = 51$), dashed: chemo+RT ($n = 45$)

6 Discussion

Several authors have proposed tests for comparing two survival curves based on the maximum or linear combination of weighted log-rank statistics. All of these tests yield a valid p -value for testing the null hypothesis between the curves, although they do not provide any clinically meaningful inference about the magnitude of the survival difference and so should be accompanied by the estimated survival curves. Simulation results indicate that the version adopted here, namely, the maximum of $G^{0,0}$, $G^{1,0}$, and $G^{0,1}$ tests, maintains the type I error rate and provides increased power relative to the log-rank test under early difference and late-difference alternatives; however, Z_m is associated with a small to moderate power loss relative to the more optimally chosen test. The `verswlr` command described in section 3 uses this test as its default but allows the user to specify other members from the $G^{\rho,\gamma}$ class of his or her choosing. However, a word of caution is warranted. Careful attention has been given to maintaining the alpha level of the versatile test at its nominal level. Consequently, the three tests should be designated a priori. If they are selected after inspection of the survival curves, inflation of the type I error can occur. The `verswlr` command was also constructed to allow users to carry out their own simulations at the design stage of a trial, and they are encouraged to do so.

Note that in the case of crossing survival curves, it is theoretically possible for $G^{1,0}$ and $G^{0,1}$ to both reject the null in opposite directions. The preferred treatment in such situations would depend on how one values the tradeoff between early versus late risks and their relative magnitudes. Also, some object to tests with increasing weights over time because they discount early deaths. Thus, as mentioned above, these are tests for survival “differences”, not necessarily “superiority”, and careful inspection of the curves is critical when interpreting the results.

From a design standpoint, there is a tradeoff if one chooses a versatile test. While the versatile procedure provides improved power for non-PH alternatives compared with the standard log-rank test, there is a reduction in power if PH does in fact hold. And of course the versatile test does not do as well as a test directed at the right non-PH alternative. Note also the substantial drop in power that can occur if an incorrectly weighted test is applied. The reduction in power of the max test relative to the more optimal test appears to be modest for powers in the range of 80%–90%, the typically desired range. For trials in which non-PH is considered plausible, the trial designer could therefore increase the sample size by a certain amount—say, set the power at 85% rather than 80% for a power calculation based on the log-rank test—and use the versatile procedure to provide some “insurance” against non-PHs. Alternatively, simulations could be conducted under different scenarios to determine the sample size needed for any desired level of power. Finally, we note that Lee’s (2007) maximum test and, in particular, the adaptive procedure developed by Yang and Prentice (2010) are strong competitors. Under non-PHs, the adaptive weights from Yang and Prentice’s (2010) test reflect the deviations from proportionality, leading to increased power. Moreover, the adaptively weighted statistics are asymptotically equivalent to the log-rank statistic under PHs, so the procedure maintains optimality when PH holds. Yang and Prentice (2010) performed simulations comparing the performance of their procedure with several different tests. Its versatility relative to Z_m under various non-PH scenarios is an area for further research.

7 Acknowledgment

The author thanks an anonymous reviewer for helpful comments.

8 References

- Drezner, Z. 1994. Computation of the trivariate normal integral. *Mathematics of Computation* 62: 289–294.
- Efron, B. 1988. Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association* 83: 414–425.
- Fleming, T. R., and D. P. Harrington. 1991. *Counting Processes and Survival Analysis*. New York: Wiley.
- Kaplan, E. L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.
- Lee, J. W. 1996. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* 52: 721–725.
- Lee, S.-H. 2007. On the versatility of the combination of the weighted log-rank statistics. *Computational Statistics and Data Analysis* 51: 6557–6564.

- Stablein, D. M., W. H. Carter, Jr., and J. W. Novak. 1981. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials* 2: 149–159.
- Tarone, R. E. 1981. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics* 37: 79–85.
- Yang, S., and R. Prentice. 2010. Improved logrank-type tests for survival data using adaptive weights. *Biometrics* 66: 30–38.

About the author

Theodore Karrison received his PhD in statistics from the University of Chicago. He is currently a research associate (professor) and director of the Biostatistics Laboratory in the Department of Public Health Sciences at the University of Chicago. His research interests include survival analysis methods and the design, conduct, and analysis of clinical trials.