STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 311–328, DOI 10. 21307

EXAMINING TESTS FOR COMPARING SURVIVAL CURVES WITH RIGHT CENSORED DATA

Pinar Gunel Karadeniz[1], **Ilker Ercan**[2]

ABSTRACT

Background and objective: In survival analysis, estimating the survival probability of a population is important, but on the other hand, investigators want to compare the survival experiences of different groups. In such cases, the differences can be illustrated by drawing survival curves, but this will only give a rough idea. Since the data obtained from survival studies contains frequently censored observations some specially designed tests are required in order to compare groups statistically in terms of survival. **Methods:** In this study, Logrank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto tests and tests belonging to Fleming-Harrington test family with (p, q) values; (1, 0), (0.5, 0.5), (1, 1), (0, 1) ve (0.5, 2) are examined by means of Type I error rate obtained from a simulation study, which is conducted in the cases where the event takes place with equal probability along the follow-up time. **Results:** As a result of the simulation study, Type I error rate of Logrank test is equal or close to the nominal value. **Conclusions:** When survival data were generated from lognormal and inverse Gaussian distribution, Type I error rate of Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and Fleming-Harrington (1,0) tests were close to the nominal value.

Key words: survival analysis, survival curves, comparison of survival curves, right censored observations.

1. Introduction

In survival analysis, investigators frequently want to determine if individuals from one population live longer than individuals from a second population. When all individuals in the population are observed, it is easy to estimate and compare the survival functions of different populations. However, most clinical research is completed in a prespecified time period, and it is not always possible to observe

[1] Uludag University, Faculty of Medicine, Department of Biostatistics Gorukle Campus Bursa/TURKEY 16059. E-mail: gunelpinar@yahoo.com.
[2] Uludag University, Faculty of Medicine, Department of Biostatistics Gorukle Campus Bursa/TURKEY 16059. E-mail: iercan@msn.com.

all individuals in this period until they experience the event. In such cases, censored data are encountered.

Since time-to-event data often include censored observations, some specialized methods are needed to compare the survival experiences of two groups (Dawson and Trapp, 2001). Several methods for testing the equality of two survival curves in the presence of censored data have been proposed.

First, Cox (1953) showed that an F test can be used to test the difference between two groups (whether or not the data include censored observations) when the survival data follow the exponential distribution. Then, the original (unweighted) log-rank test, which extended this test, was proposed by Mantel and Haenszel (1959) with Mantel (1966). Then, the generalized Wilcoxon tests, Gehan-Wilcoxon test (Gehan, 1965), the Peto-Peto test (Peto and Peto, 1972), and the Tarone-Ware test (Tarone and Ware, 1977) were proposed, respectively. Another useful subfamily within the class of weighted log-rank statistics is also proposed by Fleming and Harrington (1981) and Harrington and Fleming (1982).

There are studies in the literature that compare the performances of survival comparison tests. Lee et al. (1975) compared the size and power of the tests using small samples from the exponential and Weibull distributions with and without censoring. They arranged their simulation study with censoring rates and sample sizes of the groups being the same. Latta (1981) extended the simulations to include log-normal distributions, unequal sample sizes and censoring of only one group. Fleming et al. (1987) examined the properties of the tests based on linear rank statistics. Beltangady and Frankowski (1989) focused on the effect of unequal censoring by using various combinations of censoring proportions. Leton and Zuluaga (2001; 2005) compared the performance of various versions of generalized Wilcoxon and log-rank tests under scenarios of early and late hazard differences. Akbar and Pasha (2009) compared the performances of the log-rank and generalized Wilcoxon tests with low and high censoring rates for small and large sample sizes. Jurkiewicz and Wycinka (2011) compared the log-rank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto and F-H tests when the sample size is small.

Log-rank test is proposed in order to give equal weight to all failures among the follow-up (Lee and Wang, 2003). However, for the log-rank test there is an assumption that the hazard ratio of the groups should be proportional along the follow-up period (Fleming et al., 1987; Lee, 1996; Buyske et al., 2000). Only in this situation is the log-rank test powerful. When the hazard ratio is non-constant, the Gehan-Wilcoxon and Tarone-Ware tests can be more powerful than the log-rank test (Tarone and Ware, 1977; Pepe and Fleming, 1989). The Peto-Peto test is also efficient when proportional hazard assumption is violated (Kleinbaum and Klein, 2005). F-H tests, which are the most flexible tests for choosing weights, are focused on crossing the hazard ratios of groups (Pepe and Fleming, 1989).

The log-rank test, which compares outcomes over the whole time interval, may not adequately detect important differences between groups which occur either early or late in the interval (Klein et al., 2001). In some situations, a

treatment will decrease the hazard for some initial period, but its effect on the hazard becomes negligible later on (Pepe and Fleming, 1989). Therefore, the need to use tests that give more weight to early failures arises. In such cases, the Gehan-Wilcoxon and Tarone-Ware tests, which give more weight to the events that occur earlier, can be used. Likewise, the Peto-Peto and F-H (1,0) tests give more weight to early events as well.

When survival comparison tests are examined in the literature in terms of censoring, the Gehan-Wilcoxon test is powerful if the censoring rate is low (Stevenson, 2009; Martinez and Naranjo, 2010). Nevertheless, if the censoring rate is high, the Gehan-Wilcoxon test has less power. In addition, both the Gehan-Wilcoxon and the Peto-Peto tests have the assumption that censoring distributions of two groups should be same. When this assumption is violated, Efron stated that the Peto-Peto test has better performance than the Gehan-Wilcoxon test. For the log-rank test, it is more efficient when the censoring distribution of groups is different (Wang et al., 2010). This property is an advantage of the log-rank test over the others.

In this study, type I error rates were considered in examining the tests. Weibull, log-normal, exponential and inverse Gaussian distributions with different shape and scale parameters were used in order to generate survival times. The aim of this study is to examine the survival comparison tests in regard to type I error rates with right-censored data in some defined particular cases with events spread equally during the follow-up time.

2. Materials and Methods

2.1. Survival Comparison Tests

In survival analysis, estimating the survival probability of a population is important and investigators also want to compare the survival experiences of different groups. In such cases, the differences between groups can be illustrated by drawing survival curves obtained from the Kaplan-Meier (K-M) method, but this will only give a rough comparison and does not reveal whether the differences are statistically significant or not (Lee and Wang, 2003; Kim and Dailey, 2008).

When there are no censored observations, standard independent sample tests can be used to compare two survival distributions. However, in practice, censored data are frequently encountered. In such cases, in order to analyze the difference between two groups statistically, specially designed tests are used (Lee and Wang, 2003).

In this study, survival comparison tests (log-rank, Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto, and Fleming-Harrington test family (with (p, q) values: (1, 0), (0.5, 0.5), (1, 1), (0, 1) and (0.5, 2), respectively), which are used to compare survival curves from two groups in cases of right-censored data, were compared in regard to type I error rates in the specific case that events

occurred at equal rates throughout the follow-up time and when the follow-up time fits some specific distributions. The nominal value was considered as 0.05 for type I error rates. When type I error rates were close to the nominal value the false positivity was close to the desired value so that the probability of making a wrong decision when there was not a real difference was at the desired value.

Suppose we have survival data as in Table 1. In order to obtain the general test statistic, which compares survival curves, Table 2 can be generated from Table 1.

Table 1. Sample survival data set

Individual (Patient)	Survival Time (t_j)	Status Variable (1: Event occurred 0: Censored observation)	Group
1	t_1	1	1
2	t_2	1	1
3	t_3	0	2
4	t_4	1	2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
29	t_{29}	0	1
30	t_{30}	1	2

Table 2. Summary of observations at t_j time period

Group	1	2	Total
Number of events	d_{1j}	d_{2j}	d_j
Number of individuals at risk	r_{1j}	r_{2j}	r_j

Table 2 is generated repeatedly in all time periods in which the event of interest occurs (Bland and Altman, 2004; Kleinbaum and Klein, 2005). That is, by taking Table 1 as reference, at t_1, t_2, t_4, ..., t_{30} time periods in which the event of interest occurred, 2 by 2 tables are obtained. The observed and expected events in each group are considered from these tables. The general test statistic is obtained as the sum of the squared differences of the observed and expected counts scaled by the expected counts (Fisher and Belle, 1993; Klein et al., 2001). The test statistic is as in Equation 1 (Altman, 1991; Stevenson, 2009).

$$Test\ statistic = \frac{\left(\sum_1^k w_j\left(O_{ij} - E_{ij}\right)\right)^2}{var\left(\sum_1^k w_j\left(O_{ij} - E_{ij}\right)\right)} \quad (1)$$

Here,

i, denotes the group; j denotes the time that the event occurred,

O_{ij}, number of observed events in the i^{th} group at the j^{th} time period,

E_{ij}, number of expected events in the i^{th} group at the j^{th} time period.

O_{ij} and E_{ij} are computed as in Equation 2 and Equation 3, respectively (Leton and Zuluaga, 2005):

$$O_{ij} = \sum_{j=1}^{k} d_{ij} \quad (2)$$

$$E_{ij} = \sum_{j=1}^{k} d_j \frac{r_{ij}}{r_j} \quad (3)$$

When Equation 2 and Equation 3 are replaced in Equation 1, the general test statistic equals Equation 4 (Leton and Zuluaga, 2005):

$$Test\ statistic = \frac{\left(\sum_j w_j\left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} w_j^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \quad (4)$$

In Equation 4;

d_{ij}, number of individuals who experience the event in group i at time j

d_j, total number of individuals in both groups who experience the event

r_{ij}, number of individuals at risk in group i at time j

r_j, total number of individuals at risk at time j

r_{1j}, number of individuals at risk in group *1*

r_{2j}, number of individuals at risk in group *2*.

The test statistic is compared to a chi-square table with 1 degree of freedom (Altman, 1991; Dawson and Trapp, 2001; Stevenson, 2009). The survival comparison tests are designated according to weight w_j, which is given in Equation 4.

Hypotheses for the survival comparison test are as below (Lee and Wang, 2003; Kleinbaum and Klein, 2005).

H_0: $S_1(t) = S_2(t)$ (survival probability of two groups is equal)

H_1: $S_1(t) \neq S_2(t)$ (survival probability of two groups is different) or

H_1: $S_1(t) < S_2(t)$ (survival probability of the first group is less than the survival probability of second group) or

H_1: $S_1(t) > S_2(t)$ (survival probability of the first group is greater than the survival probability of the second group)

2.1.1. Log-rank Test

The log-rank test, which is also known as the Mantel Log-rank Test, is the most commonly used test for comparing survival curves. It gives equal weight to early and late failures (Stevenson, 2009; Allison, 2010). The test statistic is based on the ranks of the time period in which the event occurred (Lee and Wang, 2003).

It takes $w_j=1$ as the weight in Equation 4. The test statistic turns into Equation 5:

$$Logrank\ test\ statistic = \frac{\left(\sum_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \quad (5)$$

The log-rank test assumes that the hazard functions for the two groups are parallel meaning that the hazard ratios of two groups are constant among all time periods (Dawson and Trapp, 2001; Stevenson, 2009).

Survival curves can be used to visualize whether the hazard functions of the two groups are parallel or not (Martinez and Naranjo, 2010).

2.1.2. Gehan Generalized Wilcoxon Test

The Gehan Generalized Wilcoxon Test is a distribution-free two-sample test and it is a generalization of the Wilcoxon test that samples right-censored observations (Gehan, 1965; Lee et al., 1975; Kim and Dailey, 2008).

The Gehan-Wilcoxon test uses the number of individuals at risk at time period t_j as the weight; thus, in Equation 4, $w_j=r_j$.

Since the weight is the number of individuals at risk, the Gehan-Wilcoxon test places more emphasis on the information at the beginning of the survival curve, where the number at risk is larger, allowing early failures to receive more weight than later failures (Tarone and Ware, 1977; Fisher and Belle, 1993; Kleinbaum and Klein, 2005).

The test statistic is as in Equation 6.

$$Gehan - Wilcoxon\ test\ statistic = \frac{\left(\sum_j r_j \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} r_j^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{6}$$

In comparison with the log-rank test, the Gehan-Wilcoxon test does not have the assumption that the hazard functions of two groups are parallel making it a powerful test (Dawson and Trapp, 2001; Stevenson, 2009).

2.1.3. Tarone-Ware Test

The Tarone-Ware test places heavy weight on hazards in the early periods, just as the Gehan-Wilcoxon test does. It uses the square root of the number of individuals at risk as weight $w_j=\sqrt{r_j}$ (Tarone and Ware, 1977; Klein et al, 2001; Kleinbaum and Klein, 2005; Allison, 2010).

The weight used in the Tarone-Ware test is greater than the weight used in the log-rank test but less than the weight used in the Gehan-Wilcoxon test.

The Tarone-Ware test statistic is as in Equation 7.

$$Tarone - Ware\ test\ statistics = \frac{\left(\sum_j \sqrt{r_j} \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} r_j \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{7}$$

2.1.4. Peto-Peto Test

The Peto-Peto test assigns weights that depend on the estimated percentile of the failure time distribution. Failures occurring early, when the estimated survivor function is large, receive larger weights, while those in the right tail of the failure time distribution receive smaller weights (Prentice and Marek, 1979). This test is used when the hazard ratio between groups is not constant (Stevenson, 2009).

The Peto-Peto test uses the estimation of survival function as weight $w_j=\tilde{S}(t)$. The survival function here is a modified version of the K-M estimator (Allison, 2010). The test statistic is given in Equation 8.

$$Peto - Peto\ test\ statistic = \frac{\left(\sum_j \tilde{S}(t_j) \left(d_{ij} - d_j \frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k} \tilde{S}(t_j)^2 \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{8}$$

Here,

$$\tilde{S}(t) = \prod\nolimits_{t_j<t} \left(1 - \frac{d_j}{r_j + 1}\right) \tag{9}$$

2.1.5. Modified Peto-Peto Test

The Modified Peto-Peto test is an extension of the Peto-Peto test (Allison, 2010). It provides even greater weight to the early events as the Peto-Peto test (Hintze, 2007).

The modified Peto-Peto test uses survival function and the number of individuals at risk as weight $w_j=\tilde{S}(t_j)r_j/(r_j+1)$ (Hintze, 2007).

The test statistic is given in Equation 10.

$$Modified\ Peto-Peto\ test\ statistic$$

$$= \frac{\left(\sum_j \tilde{S}(t_j)r_j/(r_j+1)\left(d_{ij} - d_j\frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k}\left[\tilde{S}(t_j)r_j/(r_j+1)\right]^2 \frac{r_{1j}r_{2j}d_j(r_j-d_j)}{r_j^2(r_j-1)}} \tag{10}$$

2.1.6. Fleming-Harrington Test Family

Fleming-Harrington (F-H) test family comprises weighted log-rank tests. This family was designed in order to test the hypothesis of whether the survival curves of groups are equal or not equal, just as log-rank and other survival comparison tests do (Logan et al., 2008).

F-H tests use $w_j = \hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q$ equality as weight when $p \geq 0$ and $q \geq 0$ (Oller and Gomez, 2010). Here, $\hat{S}(t)$ is an estimation of the Kaplan-Meier survival function. The test statistic is as below in Equation 11.

$$F - H\ test\ statistic$$

$$= \frac{\left(\sum_j \hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q \left(d_{ij} - d_j\frac{r_{ij}}{r_j}\right)\right)^2}{\sum_{j=1}^{k}\left[\hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q\right]^2 \frac{r_{1j}r_{2j}d_j(r_j-d_j)}{r_j^2(r_j-1)}} \tag{11}$$

Here, Kaplan-Meier survival function is obtained as follows:

$$\hat{S}(x) = \prod\nolimits_{t_j\leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{12}$$

In F-H tests, the choice of p and q determines what weight is given to middle and late occurring events (Gomez et al., 2009; Oller and Gomez, 2012). For example, if it is accepted that a treatment has an impact in the earlier periods, then

q is chosen as 0 with increasing values of p to ensure weight is given to earlier events. When p and q are equal, it ensures weight is given to events occurring in the middle of the whole time period. When p equals 0, increasing values of q ensure that more weight is placed on late events (Lee, 1996; Gomez et al., 2009). When p and q are both 0, the test is equivalent to the log-rank test. If p=1 and q=0, the test will be approximately equal to the Peto-Peto test (Harrington and Fleming, 1982). The choice of the weight function in F-H test must be made before evaluating the data and based on clinical expectations for the outcome (Klein et al., 2001; Gomez et al., 2009).

The summary of survival comparison tests and their weights are given in Table 3 (Kleinbaum and Klein, 2005; Jurkiewicz and Wycinka, 2011).

Table 3. Survival comparison tests and their weights

TEST	**WEIGHT (w_j)**	
LOGRANK	*1*	Equal weights throughout the whole time period
GEHAN-WILCOXON	r_j	Places very heavy weight on hazards at the beginning of the study
TARONE-WARE	$\sqrt{r_j}$	Places heavy weight on hazards at the beginning of the study
PETO-PETO	$\tilde{S}(t) = \prod_{t_j<t}\left(1 - \frac{d_j}{r_j + 1}\right)$	Places slightly more weight on hazards at the beginning of the study
MODIFIED PETO-PETO	$\tilde{S}(t) = \prod_{t_j<t}\left(1 - \frac{d_j}{r_j + 1}\right)$	Places slightly more weight on hazards at the beginning of the study
F-H (1,0)	$\hat{S}(t_{j-1})$	Places slightly more weight on hazards at the beginning of the study
F-H (0.5,0.5)	$\hat{S}(t_{j-1})^{0.5}\left[1 - \hat{S}(t_{j-1})\right]^{0.5}$	Places weight on hazards in the middle of the study
F-H (1,1)	$\hat{S}(t_{j-1})\left[1 - \hat{S}(t_{j-1})\right]$	Places weight on hazards in the middle of the study
F-H (0,1)	$1 - \hat{S}(t_{j-1})$	Places weight on hazards at the end of the study
F-H (0.5,2)	$\hat{S}(t_{j-1})^{0.5}\left[1 - \hat{S}(t_{j-1})\right]^{2}$	Places weight on hazards at the end of the study

3. Theory/calculation

3.1. Simulation Study

In this study, in order to examine survival comparison tests, a simulation study with 500 replicates was conducted, and type I error rates were obtained.

Survival times for two groups with sample sizes of n=10, 30, 50, and 100 were generated from the Weibull, log-normal, exponential and inverse Gaussian distributions with different shape and scale parameters. The status variable was generated from the binomial distribution with a probability of p=0.50.

While generating the survival data, other simulation studies in the literature were reviewed and most frequently used distributions with their most frequently used parameters were considered for our simulation study. Additionally, various parameters of the distributions were included. The reason for this choice is that in survival analysis follow-up time data fit generally the aforementioned distributions.

For the exponential distribution, the scale parameter was selected as β= 0.5, 1, 1.5; for the Weibull distribution, the shape parameter was α= 1, 2, 3 and the scale parameter was β= 1.5, 2.5, 3.5; for the log-normal distribution, the shape parameter was σ= 1, 2, 3 and the scale parameter was m= 0; for the inverse Gaussian distribution, the location parameter was μ= 0.5 and the scale parameter was λ= 1, 2, 3.

The data were generated using R software version 3.0.3 and the data were analyzed using NCSS package program with 500 replicates. Winautomation program is used for replicates.

4. Results

Type I error rates according to the simulation study for the sample sizes of n=10, 30, 50 and 100 are given in Table 4.

Table 4. Type I error rates of tests

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Exponential (0.5)				Exponential (1)				Exponential (1.5)			
Logrank	0.0420	0.0560	0.0560	0.0520	0.0760	0.0420	0.0520	0.0620	0.0600	0.0720	0.0480	0.0500
Gehan-Wilcoxon	0.0360	0.0540	0.0620	0.0320	0.0600	0.0480	0.0460	0.0520	0.0560	0.0520	0.0500	0.0540
Tarone-Ware	0.0440	0.0560	0.0600	0.0460	0.0620	0.0460	0.0460	0.0520	0.0540	0.0680	0.0500	0.0520
Peto-Peto	0.0400	0.0560	0.0560	0.0440	0.0680	0.0440	0.0480	0.0520	0.0500	0.0620	0.0520	0.0520
Mod. Peto-Peto	0.0420	0.0560	0.0580	0.0440	0.0660	0.0460	0.0460	0.0520	0.0500	0.0600	0.0520	0.0520
F-H (1, 0)	0.0400	0.0540	0.0580	0.0440	0.0680	0.0440	0.0500	0.0520	0.0500	0.0640	0.0500	0.0520
F-H (0.5, 0.5)	0.0480	0.0460	0.0520	0.0620	0.0860	0.0360	0.0600	0.0640	0.0440	0.0500	0.0500	0.0480
F-H (1, 1)	0.0500	0.0480	0.0620	0.0560	0.0800	0.0420	0.0640	0.0620	0.0500	0.0480	0.0460	0.0420
F-H (0, 1)	0.0580	0.0760	0.0640	0.0620	0.0860	0.0580	0.0700	0.0680	0.0580	0.0620	0.0420	0.0540
F-H (0.5, 2)	0.0640	0.0860	0.0620	0.0680	0.0840	0.0620	0.0700	0.0640	0.0560	0.0800	0.0480	0.0500

Table 4. Type I error rates of tests (cont.)

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Weibull (1, 1.5)				Weibull (1, 2.5)				Weibull (1, 3.5)			
Logrank	0.0440	0.0520	0.0440	0.0520	0.0620	0.0600	0.0520	0.0480	0.0620	0.0520	0.0620	0.0480
Gehan-Wilcoxon	0.0320	0.0480	0.0340	0.0520	0.0520	0.0580	0.0440	0.0500	0.0480	0.0440	0.0680	0.0400
Tarone-Ware	0.0340	0.0520	0.0280	0.0460	0.0540	0.0480	0.0420	0.0540	0.0560	0.0420	0.0760	0.0400
Peto-Peto	0.0340	0.0520	0.0280	0.0460	0.0580	0.0500	0.0440	0.0500	0.0500	0.0400	0.0720	0.0440
Mod. Peto-Peto	0.0340	0.0540	0.0260	0.0460	0.0580	0.0480	0.0440	0.0520	0.0520	0.0400	0.0760	0.0440
F-H (1, 0)	0.0320	0.0460	0.0280	0.0460	0.0560	0.0480	0.0440	0.0500	0.0540	0.0380	0.0720	0.0440
F-H (0.5, 0.5)	0.0560	0.0520	0.0560	0.0560	0.0800	0.0700	0.0540	0.0480	0.0640	0.0620	0.0480	0.0500
F-H (1, 1)	0.0660	0.0420	0.0560	0.0620	0.0820	0.0580	0.0560	0.0480	0.0700	0.0560	0.0480	0.0580
F-H (0, 1)	0.0740	0.0580	0.0580	0.0620	0.0860	0.0540	0.0660	0.0680	0.0760	0.0600	0.0400	0.0580
F-H (0.5, 2)	0.0740	0.0700	0.0600	0.0560	0.0920	0.0540	0.0580	0.0640	0.0740	0.0580	0.0460	0.0560
	Weibull (2, 1.5)				Weibull (2, 2.5)				Weibull (2, 3.5)			
Logrank	0.0660	0.0600	0.0600	0.0480	0.0480	0.0560	0.0500	0.0400	0.0540	0.0560	0.0440	0.0500
Gehan-Wilcoxon	0.0520	0.0500	0.0520	0.0420	0.0280	0.0520	0.0460	0.0400	0.0640	0.0700	0.0420	0.0580
Tarone-Ware	0.0640	0.0620	0.0640	0.0480	0.0380	0.0560	0.0380	0.0400	0.0580	0.0640	0.0480	0.0560
Peto-Peto	0.0640	0.0620	0.0620	0.0420	0.0400	0.0560	0.0400	0.0400	0.0580	0.0720	0.0480	0.0560
Mod. Peto-Peto	0.0620	0.0580	0.0620	0.0420	0.0360	0.0560	0.0400	0.0400	0.0560	0.0720	0.0480	0.0540
F-H (1, 0)	0.0620	0.0620	0.0620	0.0420	0.0380	0.0560	0.0400	0.0400	0.0580	0.0700	0.0480	0.0560
F-H (0.5, 0.5)	0.0760	0.0740	0.0540	0.0560	0.0480	0.0560	0.0480	0.0520	0.0600	0.0620	0.0580	0.0620
F-H (1, 1)	0.0880	0.0640	0.0580	0.0600	0.0460	0.0560	0.0520	0.0400	0.0660	0.0480	0.0580	0.0540
F-H (0, 1)	0.0980	0.0700	0.0660	0.0540	0.0540	0.0660	0.0680	0.0460	0.0760	0.0560	0.0620	0.0540
F-H (0.5, 2)	0.0920	0.0760	0.0620	0.0560	0.0560	0.0720	0.0680	0.0520	0.0820	0.0560	0.0580	0.0540
	Weibull (3, 1.5)				Weibull (3, 2.5)				Weibull (3, 3.5)			
Logrank	0.0540	0.0460	0.0540	0.0500	0.0500	0.0560	0.0600	0.0560	0.0480	0.0560	0.0440	0.0380
Gehan-Wilcoxon	0.0600	0.0440	0.0540	0.0520	0.0440	0.0400	0.0500	0.0500	0.0480	0.0520	0.0340	0.0440
Tarone-Ware	0.0580	0.0420	0.0500	0.0540	0.0380	0.0500	0.0360	0.0420	0.0500	0.0480	0.0280	0.0360
Peto-Peto	0.0580	0.0440	0.0540	0.0520	0.0400	0.0480	0.0400	0.0420	0.0480	0.0460	0.0340	0.0360
Mod. Peto-Peto	0.0580	0.0420	0.0500	0.0520	0.0420	0.0480	0.0400	0.0400	0.0480	0.0500	0.0340	0.0360
F-H (1, 0)	0.0580	0.0420	0.0540	0.0500	0.0400	0.0500	0.0400	0.0420	0.0440	0.0460	0.0340	0.0360
F-H (0.5, 0.5)	0.0540	0.0440	0.0580	0.0420	0.0580	0.0720	0.0600	0.0580	0.0580	0.0460	0.0520	0.0380
F-H (1, 1)	0.0580	0.0420	0.0600	0.0400	0.0680	0.0680	0.0500	0.0600	0.0660	0.0460	0.0400	0.0480
F-H (0, 1)	0.0640	0.0560	0.0580	0.0480	0.0820	0.0880	0.0580	0.0500	0.0800	0.0580	0.0580	0.0560
F-H (0.5, 2)	0.0740	0.0580	0.0540	0.0520	0.0820	0.0800	0.0560	0.0460	0.0700	0.0580	0.0680	0.0520

Table 4. Type I error rates of tests (cont.)

Tests	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100	n=10	n=30	n=50	n=100
	Lognormal (0, 1)				Lognormal (0, 2)				Lognormal (0, 3)			
Logrank	0.0580	0.0560	0.0540	0.0620	0.0700	0.0640	0.0640	0.0540	0.0520	0.0520	0.0440	0.0560
Gehan-Wilcoxon	0.0420	0.0500	0.0480	0.0520	0.0640	0.0460	0.0560	0.0520	0.0520	0.0480	0.0540	0.0400
Tarone-Ware	0.0500	0.0580	0.0540	0.0540	0.0700	0.0600	0.0620	0.0500	0.0460	0.0540	0.0480	0.0520
Peto-Peto	0.0480	0.0500	0.0520	0.0540	0.0680	0.0600	0.0600	0.0500	0.0480	0.0540	0.0460	0.0500
Mod. Peto-Peto	0.0500	0.0500	0.0500	0.0520	0.0700	0.0600	0.0620	0.0500	0.0480	0.0540	0.0440	0.0520
F-H (1, 0)	0.0440	0.0520	0.0480	0.0560	0.0680	0.0600	0.0620	0.0500	0.0520	0.0540	0.0480	0.0500
F-H (0.5, 0.5)	0.0600	0.0600	0.0540	0.0520	0.0540	0.0560	0.0600	0.0400	0.0520	0.0520	0.0480	0.0500
F-H (1, 1)	0.0580	0.0580	0.0540	0.0520	0.0620	0.0600	0.0540	0.0380	0.0600	0.0540	0.0500	0.0520
F-H (0, 1)	0.0640	0.0660	0.0720	0.0760	0.0720	0.0660	0.0480	0.0320	0.0700	0.0640	0.0500	0.0380
F-H (0.5, 2)	0.0720	0.0680	0.0680	0.0740	0.0700	0.0600	0.0460	0.0420	0.0680	0.0780	0.0620	0.0320
	Inverse Gaussian (0.5, 1)				Inverse Gaussian (0.5, 2)				Inverse Gaussian (0.5, 3)			
Logrank	0.0400	0.0560	0.0600	0.0560	0.0600	0.0460	0.0520	0.0480	0.0540	0.0740	0.0360	0.0380
Gehan-Wilcoxon	0.0420	0.0580	0.0500	0.0600	0.0440	0.0480	0.0360	0.0520	0.0580	0.0620	0.0400	0.0420
Tarone-Ware	0.0360	0.0580	0.0520	0.0520	0.0540	0.0520	0.0520	0.0500	0.0580	0.0660	0.0340	0.0400
Peto-Peto	0.0380	0.0540	0.0520	0.0560	0.0500	0.0500	0.0480	0.0500	0.0540	0.0680	0.0340	0.0400
Mod. Peto-Peto	0.0400	0.0520	0.0500	0.0540	0.0500	0.0540	0.0480	0.0480	0.0540	0.0680	0.0360	0.0400
F-H (1, 0)	0.0360	0.0520	0.0500	0.0560	0.0500	0.0500	0.0480	0.0480	0.0540	0.0680	0.0320	0.0400
F-H (0.5, 0.5)	0.0460	0.0480	0.0600	0.0580	0.0640	0.0500	0.0460	0.0420	0.0620	0.0640	0.0380	0.0300
F-H (1, 1)	0.0500	0.0560	0.0520	0.0600	0.0660	0.0500	0.0460	0.0400	0.0660	0.0560	0.0320	0.0320
F-H (0, 1)	0.0600	0.0420	0.0560	0.0460	0.0700	0.0580	0.0520	0.0500	0.0720	0.0680	0.0420	0.0500
F-H (0.5, 2)	0.0640	0.0480	0.0540	0.0460	0.0720	0.0640	0.0500	0.0520	0.0720	0.0680	0.0460	0.0460

In the case that the event occurs with equal probability along the follow-up time, the type I error rate of the log-rank test is equal or too close to the nominal value (0.05) for all distributions.

5. Discussion

In this study, a simulation was conducted in order to examine the performance of survival comparison tests under various scenarios, and the type I error rates were evaluated.

As a result, in the case that the event occurs with equal probability along the follow-up time, the type I error rate of the log-rank test is equal or too close to the nominal value. This result is in agreement with Lee and Wang (2003), who state that the "log-rank test gives equal weight to all failures." In addition, when the sample size gets larger, the type I error rate approaches the nominal value for all

tests. For the exponential distributions, the best results for all tests were obtained when the scale parameter was 1.5. When the scale parameter was 0.5, the best result was obtained for log-rank test; and the results farthest from the nominal value were obtained for the F-H tests, which give more weight to middle and late events (F-H (0.5,0.5), (1,1), (0,1), (0.5,2)). When the scale parameter was 1 for the exponential distribution, the closest type I error rates to the nominal value were obtained for the tests that give more weight to early events, namely, the Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, Modified Peto-Peto and F-H (1,0). For F-H tests, which give more weight to middle and late events (F-H (0.5,0.5), (1,1), (0,1), (0.5,2)), the type I error rate tended to be greater than the nominal value when the scale parameters of the exponential distribution were 0.5 and 1. When survival data were generated from the Weibull distribution for all parameters of the distribution, the type I error rate of the log-rank test was equal or close to the nominal value. When the shape parameter of the Weibull distribution is 1, the type I error rates obtained were very close to the type I error rates of the exponential distribution. This result supports information found in the literature that a Weibull distribution with a shape parameter of 1 is equivalent to the exponential distribution (Kalbfleisch and Prentice, 2002). When the shape parameter of a Weibull distribution was 2, for almost all tests, the type I error rates were close to the nominal value. When the shape parameter of the Weibull distribution was 3 (which means the distribution is close to a normal distribution), the type I error rate for all tests were found to be close to 0.05. The error rate tended to be smaller than 0.05 only for a Weibull distribution with a shape parameter of 3 and a scale parameter of 3.5. In their study, Lee et al. (1975) demonstrated that if it is known that the survival data fit the exponential or Weibull distributions, the log-rank test has the best result; our simulation results further support this result.

When survival data were generated from the log-normal distribution, type I error rates of the Gehan-Wilcoxon and the Peto-Peto were equal or very close to the nominal value. In his study, Latta (1981) stated that the Gehan-Wilcoxon and the Peto-Peto tests best perform when the survival data fit log-normal distribution; therefore, our result agrees with Latta's result. The type I error rates for the Tarone-Ware, Modified Peto-Peto and F-H (1,0) tests were also close to the nominal value. For the log-normal distribution, the type I error rate of the log-rank test tended to be larger than the nominal value. The Gehan-Wilcoxon, Tarone-Ware and Peto-Peto tests showed suitable results in terms of type I error rate of an inverse Gaussian distribution that is similar to a log-normal distribution in its probability density function and hazard function.

In addition to all these results, it is stated in the literature that while comparing survival curves of two different groups, the hazard ratio should be examined. There have been several graphical methods for assessing the proportional hazards assumption (Martinez and Naranjo, 2010). If hazard ratios are parallel, the log-rank test is more efficient; if the hazard ratio of one group tends to differ more than the other as time progresses, the Tarone-Ware, Peto-Peto

and Gehan-Wilcoxon tests are more efficient (Peto and Peto, 1972; Lee et al., 1975; Harrington and Fleming, 1982). Furthermore, in the case that the hazard ratios of two groups cross, F-H tests are advantageous because the weight of the test may be specified accordingly.

Limitation of this study is that we exceedingly stick to the literature with regards to choosing distributions and their parameters. Although various distributions with various parameters were included in this study, it would be better to evaluate more distributions with more parameters in order to evaluate more different situations that are encountered in practice.

6. Conclusions

As a consequence, when making a choice of methods to compare survival curves, one must pay particular attention to the proportional hazards assumption, the proportion of censoring, the size of the sample under consideration and/or the distribution of the survival data. Besides, as mentioned in the discussion section in detail, when we encountered specific circumstances (specified distribution with specified parameter) that we indicate the type I error rate is close to nominal value, it is suggested to use the stated survival comparison tests.

Once these are taken into account, it is possible to make a more informed decision about the type of test that should be used to compare survival curves.

REFERENCES

AKBAR, A, PASHA, G. R., (2009). Properties of Kaplan-Meier estimator: group comparison of survival curves. European Journal of Scientific Research, 32 (3), pp. 391–397. https://www.researchgate.net/profile/Atif_Akbar2/publication/255648672_Properties_of_Kaplan-Meier_Estimator_Group_Comparison_of_Survival_Curves/links/549a82f00cf2b80371359dd2.pdf.

ALLISON, P. D., (2010). Survival analysis using SAS: a practical guide, 2nd edition, SAS Press, North Carolina.

ALTMAN, D. G., (1991). Practical statistics for medical research, Chapman&Hall, London.

BLAND, J. M., ALTMAN, D. G., (2004). The logrank test. British Medical Journal, 328, pp. 1073. http://www.bmj.com/content/328/7447/1073.long.

BELTANGADY, M. S., FRANKOWSKI, R. F., (1989). Effect of unequal censoring on the size and power of the logrank and Wilcoxon types of tests for survival data. Statistics in Medicine, 8 (8), pp. 937–945. https://www.ncbi.nlm.nih.gov/pubmed/2799123.

BUYSKE, S., FAGERSTROM, R., YING, Z., (2000). A class of weighted log-rank tests for survival data when the event is rare. Journal of the American Statistical Association, 95 (449), pp. 249–258. https://www.jstor.org/stable/2669542?seq=1#page_scan_tab_contents.

COX, D. R., (1953). Some simple approximate tests for poisson variates. Biometrika, 40, pp. 354–360. https://www.jstor.org/stable/2333353?seq=1#page_scan_tab_contents.

DAWSON, B., TRAPP, R. G., (2001). Basic&Clinical Biostatistics. Boston: McGraw Hill.

FISHER, L. D., BELLE, G. V., (1993). Biostatistics, a methodology for the health sciences, John Wiley&Sons Inc, New York.

FLEMING, T. R., HARRINGTON, D. P., (1981). A class of hypothesis tests for one and two samples censored survival data. Communications in Statistics, 10, pp. 763–794. http://www.tandfonline.com/doi/abs/10.1080/03610928108828073?journalCode=lsta20.

FLEMING, T. R., HARRINGTON, D. P., O'Sullivan, M., (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. Journal of the American Statistical Association, 82 (397), pp. 312–320. https://www.jstor.org/stable/2289169?seq=1#page_scan_tab_contents.

GEHAN, E. A., (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika, 52, pp. 203–223. https://academic.oup.com/biomet/article-abstract/52/1-2/203/359447/A-generalized-Wilcoxon-test-for-comparing.

GOMEZ, G., CALLE, M. L., OLLER, R., LANGOHR, K., (2009). Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling, 9 (4), pp. 259–297. http://journals.sagepub.com/doi/abs/10.1177/1471082X0900900402.

HARRINGTON, DP., FLEMING, T. R., (1982). A class of rank test procedures for censored survival data. Biometrika, 69 (3), pp. 553–566. https://www.jstor.org/stable/2335991.

HEINZE, G., GNANT, M., SCHEMPER, M., (2003). Exact log-rank tests for unequal follow-up. Biometrics, 59, pp. 1151–1157. https://www.ncbi.nlm.nih.gov/pubmed/14969496.

HINTZE, J. L., (2007). NCSS user guide V tabulation, item analysis, proportions, diagnostic tests, and survival / reliability, Published by NCSS, Kaysville, Utah.

JURKIEWICZ, T., WYCINKA, E., (2011). Significance tests of differences between two crossing survival curves for small samples. Acta Universitatis Lodziensis Folia Oeconomica, 255, pp. 114. http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.hdl_11089_690?printView=true.

KALBFLEISCH, J. D., PRENTICE, R. L., (2002). The Statistical Analysis of Failure Time Data. New York: John Wiley&Sons Inc.

KIM, J., KANG, D. R., NAM, C. M., (2006). Logrank-type tests for comparing survival curves with interval-censored data. Computational Statistics & Data Analysis, 50 (11), pp. 3165–3178. http://www.sciencedirect.com/science/article/pii/S0167947305001441

KIM, J. S., DAILEY, R. J., (2008). Biostatistics for oral healthcare, Blackwell Publishing Company, Iowa, pp. 287–291.

KLEIN, J. P., RIZZO, J. D., ZHANG, M. J., KEIDING, N., (2001). Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part I: Unadjusted analysis. Bone Marrow Transplantation, 28, pp. 909–915. https://www.ncbi.nlm.nih.gov/pubmed/11753543.

KLEINBAUM, D. G., KLEIN, M., (2005). Survival Analysis a Self-Learning Text. New York: Springer.

LATTA, R. B., (1981). A monte carlo study of some two-sample rank tests with censored data. Journal of American Statistical Association, 76 (375), pp. 713–719. https://www.jstor.org/stable/2287536?seq=1#page_scan_tab_contents .

LEE, E. T., DESU, M. M., GEHAN, E. A., (1975). A Monte Carlo study of the power of some two-sample tests. Biometrika, 62 (2), pp. 425–432. https://www.jstor.org/stable/2335383?seq=1#page_scan_tab_contents.

LEE, J. W., (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. Biometrics, 52 (2), pp. 721–725. http://www.jstor.org/stable/2532911?seq=1#page_scan_tab_contents.

LEE, E. T., WANG, J. W., (2003). Statistical Methods for Survival Data Analysis. New Jersey: John Wiley&Sons Inc.

LETON, E., ZULUAGA, P., (2001). Equivalence between score and weighted tests for survival curves. Communications in Statistics - Theory and Methods, 30 (4), pp. 591–608. http://www.tandfonline.com/doi/abs/10.1081/STA-100002138.

LETON, E., ZULUAGA, P., (2005). Relationships among tests for censored data. Biometrical Journal, 47 (3), pp. 377–387. http://onlinelibrary.wiley.com/doi/10.1002/bimj.200410115/abstract.

LOGAN, B. R., KLEIN, J. P., ZHANG, M. J., (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics, 64 (3), pp. 733–740. https://www.ncbi.nlm.nih.gov/pubmed/18190619.

MANTEL, N., HAENSZEL, W., (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of National Cancer Institute, 22 (4), pp. 719–748. https://www.ncbi.nlm.nih.gov/pubmed/13655060.

MANTEL, N., (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50 (3), pp. 163–170. https://www.ncbi.nlm.nih.gov/pubmed/5910392.

MARTINEZ, RLMC, NARANJO, J. D., (2010). A pretest for choosing between logrank and Wilcoxon tests in the two-sample problem. Metron: International Journal of Statistics, 68 (2), pp. 111–125. https://link.springer.com/article/10.1007/BF03263529.

OLLER, R., GOMEZ, G., (2012). A generalized Fleming and Harrington's class of tests for interval-censored data. The Canadian Journal of Statistics, 40 (3), pp. 501–516. http://onlinelibrary.wiley.com/doi/10.1002/cjs.11139/abstract.

PEPE, M. S., FLEMING, T. R., (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics, 45, pp. 497–507. https://www.jstor.org/stable/2531492?seq=1#page_scan_tab_contents.

PETO, R, PETO, J., (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, 135 (2), pp. 185–207. https://www.jstor.org/stable/2344317?seq=1#page_scan_tab_contents.

PRENTICE, R. L., MAREK, P., (1979). A qualitative discrepancy between censored data rank tests. Biometrics, 35 (4), pp. 861–867. https://www.jstor.org/stable/2530120?seq=1#page_scan_tab_contents.

STEVENSON, M., (2009). An Introduction to Survival Analysis, EpiCentre, IVABS. Massey Massey University. http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson_survival_analysis_195_721.pdf.

TARONE, R. E., WARE, J., (1977). On distribution-free tests for equality of survival distributions. Biometrika, 64, pp. 156–160. https://www.jstor.org/stable/2335790?seq=1#page_scan_tab_contents.

WANG, R., LAGAKOS, S. W., GRAY, R. J., (2010). Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. Biostatistics, 11 (4), pp. 676–692. https://www.ncbi.nlm.nih.gov/pubmed/20439258.

XIE, J., LIU, C., (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in Medicine, 24 (20), pp. 3089–3110.
http://onlinelibrary.wiley.com/doi/10.1002/sim.2174/abstract.