VFAST Transactions on Mathematics
http://vfast.org/journals/index.php/VTM@ 2024 ISSN(e): 2309-0022, ISSN(p): 2411-6343
Volume 12, Number 1, January-June, 2024 *pp: 349-365*

Nonparametric test for multiple crossing Survival Curves

Qamruz Zaman[1*], Nisar Ullah[1], Syed Habib Shah[2], Muhammad Ali[3], Muhammad Irshad[1], Sumayyia Azam[1]

[1]Department of Statistics, University of Peshawar, Pakistan; [2]Institute of Numerical Sciences Kohat University of Science and Technology; [3]Department of Statistics, Government Post Graduate College Charsadda, Pakistan

Keywords: *Log-rank test, Wilcoxon test, Survival analysis, Failure Time, Statistical power* ***Subject Classification***: *62D05*

Journal Info:
Submitted:
May 10, 2024
Accepted:
June 25, 2024
Published:
June 30, 2024

Abstract Log-rank, Wilcoxon and Tarone-Ware tests are most commonly used tests for testing the overall homogeneity of survival curves, but in certain situation it appears that they have a significant loss of statistical testing power. One such case is the more than one time crossing of survival curves. The problem considered often occurs in medical research. To overcome this problem, in this article, we present and study a nonparametric test procedure based on a new weight. The proposed new weighted test has greater power to detect overall differences between more than one time crossing survival curves. Simulation studies are performed to compare the proposed method with existing methods. Furthermore, the advantage of the new test is finally exemplified in the analysis of a -thalassaemia major data.

***Correspondence Author Email Address:**
cricsportsresearchgroup@gmail.com
DOI: 10.21015/vtm.v12i1.1839

1 Introduction

The role of censoring is the key to survival analysis procedure Csaldori et. al, [1]. The subject of survival analysis is divided into three main categories namely; Parametric, Semi-parametric, and Non-parametric procedures, Li X et. al, [2]. Out of these three non-parametric is considered the backbone of survival analysis Srujana et. al, [3]. Tests for the comparison of survival curves are considered for the comparison Wang [4]. The comparison of different treatments of a particular disease or comparing the performance of a particular treatment on sex is very common in medical research. Several tests are used for this purpose,

out of which the most commonly used rank-based tests are the log-rank Mantel et.al, [5], Peto and Peto [6], Wilcoxon and Tarone-Ware test, Gehan[7], Gehan [8], and Tarone and Ware [9]. The log-rank test is more appropriate and powerful as compared to other tests in a situation where two or more survival curves do not cross or there occurs a major difference between the risk of an event in two groups i.e. whose hazard functions are proportional Bland et. al, [10] and Fleming et.al, [11], while in case of crossing curves weighted tests are more powerful e.g. Wilcoxon test, Tarone-Ware test.

These tests are influenced by early differences Sposto[12] and neither of these tests is developed to detect differences in multiple crossing situations. When these tests are applied to survival data in which the survival curves cross more than once, these tests are unable to detect the differences or may have little power, You and Wang [13]. To handle the problem, we propose a new weighted testing procedure that has greater power than log-rank, Wilcoxon and Tarone-Ware tests. The new weighted test is applied to the -thalassaemia major data set.

In Section 2, we briefly review concepts of log-rank test, Wilcoxon test, Tarone-Ware test and introduce the new weighted test. Simulation studies are performed in Section 3. In Section 4 we apply methods to a data set and discussions are given in Section 5. For the analysis R- package is used.

2 Material and Methods

Consider a failure time study where k distinct ordered failure times are observed for two groups (labeled I and II). At time t_i, let $d_1(t_i)$ and $d_2(t_i)$ be the number of observed failures in the two groups respectively. Similarly, $n_1(t_i)$ and $n_2(t_i)$ denote the number of persons at risk just before time t_i for groups I and II, respectively. Consequently, the total number of failures and number of persons at and prior to time t_i are $R_1(t_i)$, $R_2(t_i)$, $N_1(t_i)$, and $N_2(t_i)$ respectively.

In survival analysis, Log-rank and other weighted tests are used for comparing the survival functions of two or more groups testing:

$$H_0 : S_1(t) = S_2(t) \quad \text{vs} \quad H_A : S_1(t) \neq S_2(t)$$

where $S_1(t)$ and $S_2(t)$ are the survival functions of group I and II, respectively.

Furthermore, all information can be summarized in a contingency Table 1.

Table 1. Contingency Table for Failure Time Study

Time t_i	$d_1(t_i)$	$d_2(t_i)$	**Total at Risk**
t_1	$d_1(t_1)$	$d_2(t_1)$	$n_1(t_1) + n_2(t_1)$
t_2	$d_1(t_2)$	$d_2(t_2)$	$n_1(t_2) + n_2(t_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_k	$d_1(t_k)$	$d_2(t_k)$	$n_1(t_k) + n_2(t_k)$

Under the null hypothesis assumption that survival is independent of group membership, both the number of failures for two groups and the number of persons surviving can be determined from the value of d_{1i} alone. When the marginal totals of Table 1 are fixed, d_{1i} follows a Hypergeometric distribution with mean $E(d_{1i}) = \frac{n_{1i}d_i}{n_i}$ and variance $\text{Var}(d_{1i}) = \frac{n_{1i}n_{2i}d_i(n_i-d_i)}{n_i^2(n_i-1)}$.

For the overall measure of deviation between the observed and expected failure, sum their differences over the total number of death times to get the statistic

$$U = \sum_{i=1}^{k} (d_{1i} - E(d_{1i})). \tag{1}$$

The variance of U is

$$\mathrm{Var}(U) = \sum_{i=1}^{k} \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}. \tag{2}$$

Furthermore, according to Collett [14],

$$\frac{U}{\sqrt{\mathrm{Var}(U)}} \sim N(0, 1), \tag{3}$$

$$\frac{U^2}{\mathrm{Var}(U)} \sim \chi^2_{(1)}. \tag{4}$$

Therefore, the test statistic for the log-rank test is

$$\frac{(U)^2}{\mathrm{Var}(U)} \sim \chi^2_{(1)}. \tag{5}$$

The log-rank test is more appropriate, powerful, and reliable compared to other tests in situations where two or more survival curves do not cross, Muse et.al, [15]. In case of crossing curves, we use the weighted tests. The most commonly used weighted test is the Wilcoxon test, which assigns more weight to initial failures. The test statistic is

$$\frac{\left(\sum_{i=1}^{k} n_i (d_{1i} - E(d_{1i}))\right)^2}{\mathrm{Var}_w(U)} \sim \chi^2_{(1)}, \tag{6}$$

where n_i is the total number of persons prior to time t_i.

The other commonly used weighted test is the Tarone-Ware test, where weight suggested is the square root of the total number of persons prior to time t_i, i.e.,

$$w_i = \sqrt{n_i}. \tag{7}$$

The Tarone-Ware test statistic is

$$\frac{\left(\sum_{i=1}^{k} \sqrt{n_i} (d_{1i} - E(d_{1i}))\right)^2}{\mathrm{Var}_{tw}(U)} \sim \chi^2_{(1)}, \tag{8}$$

where

$$\mathrm{Var}_{tw}(U) = \sum_{i=1}^{k} \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i (n_i - 1)}. \tag{9}$$

A weighted test based on quartiles was proposed by Fernández et. al, [16] . All these weighted tests provide reasonable results when survival curves cross once, but in real-life scenarios, there are many cases where it happens more than once. In such cases, the log-rank test and weighted tests may fail to detect differences, Dormuth et. al, [17].

In this article, an attempt is made to overcome this problem . One of the main reasons for multiple crossings is the increasing number of tied events between deaths of different groups. More ties result in

more crossing points, which reduces the power of the log-rank and weighted tests as they do not account for ties. While survival time is often considered continuous, real-world situations involve tied events.

Since the log-rank and weighted tests assume independent survival and consider the failure cases of only one group in case of ties, they may lose power. To address this, a new weighted test is proposed that combines characteristics of the Wilcoxon and Tarone-Ware tests. This new weight is defined as

$$W_{nw} = \begin{cases} \sqrt{n_i} & \text{in the absence of tie,} \\ \sqrt{n_i} & \text{in the presence of tie.} \end{cases} \tag{10}$$

Unlike the Wilcoxon and Tarone-Ware tests, this new weight varies at each tie time.

The test statistic based on the new weight is

$$\frac{\left(\sum_{i=1}^{k} W_{nw}(d_{1i} - E(d_{1i}))\right)^2}{\mathrm{Var}_{nw}(U)} \sim \chi^2_{(1)}, \tag{11}$$

where

$$\mathrm{Var}_{nw}(U) = \sum_{i=1}^{k} W_{nw}^2 \mathrm{Var}(d_{1i}). \tag{12}$$

3 Simulation

To compare the performance of the new weighted test, several simulation studies were designed with failure times generated under different conditions. Popular continuous distributions such as the Exponential distribution and Weibull distribution were used throughout the simulation studies. Continuous distributions are commonly used as they are easy to handle and interpret, often by considering decimal points [18**?**].

The new modified method is compared with the log-rank test, Wilcoxon test, and Tarone-Ware test in the simulations. Sample sizes of 50, 60, 70, 80, 100, and 120 were selected for each group, and each simulation study was repeated 5000 times.

To obtain the number of ties between the events of two groups, we considered 1, 2, 3 decimal points in both the survival and censoring distributions. The average number of ties per iteration is also shown in tables. In this discussion, we present the results for 1 decimal point.

3.1 Performance of tests when Survival curves cross at more than one point

For obtaining a situation where survival curves cross at more than one point (Figure 1), survival times for Group I and Group II were generated from a Weibull distribution $S(t) = e^{-(\lambda t)^p}$ with $\lambda = 1$ and $p = 1.5$. The censoring distribution for Group I was $U(0, 6.5)$, resulting in a 13% censoring rate. Similarly, Group II had a censoring distribution of $U(0, 3.2)$, resulting in a 26% censoring rate.

The statistical power of the new weighted test compared with the log-rank test, Wilcoxon test, and Tarone-Ware test is summarized in Table 2. Since the log-rank test assigns equal weight to all events, while the Wilcoxon and Tarone-Ware tests assign more weight to initial events, the results show that the conventional tests fail in this situation, leading to a considerable loss of power. The proposed method

demonstrates greater power even with both equal and unequal sample sizes. Given the crossing of two curves, the log-rank test naturally has low power, but Table 2 reveals that the power of the Wilcoxon and Tarone-Ware tests is also considerably smaller and nearly the same as that of the log-rank test.

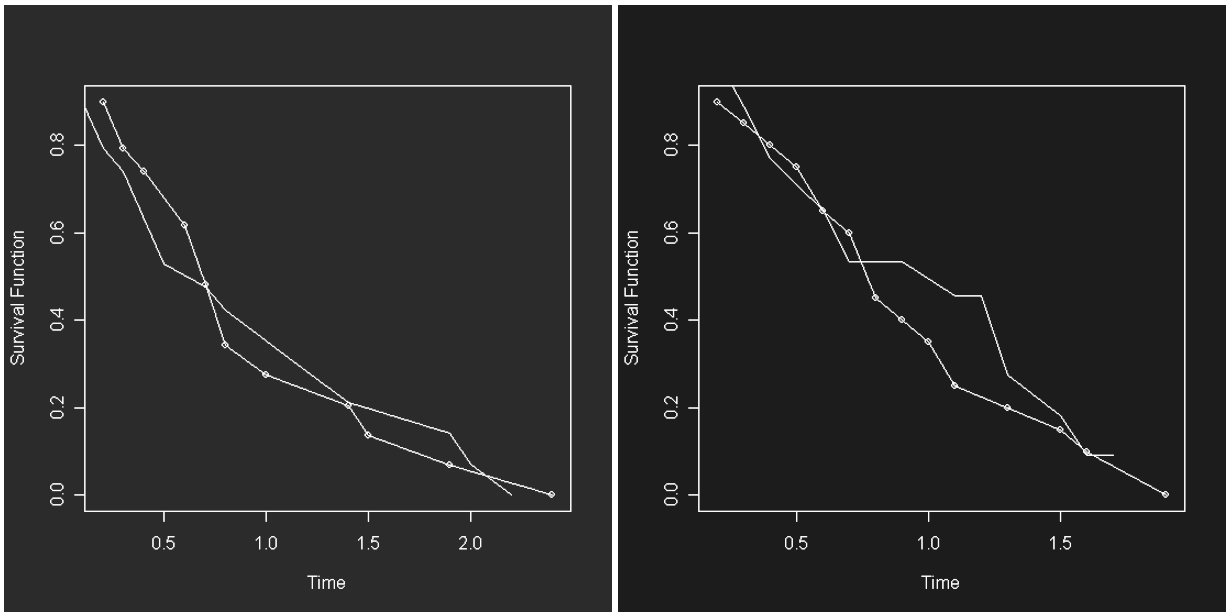

Figure 1. Survival curves where curves cross at multiple points.

Table 2. Power of log-rank, Wilcoxon, Tarone-Ware, and new tests for multiple crossing

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
50, 50	12.8	0.045	0.049	0.049	0.043
60, 60	14.1	0.046	0.052	0.052	0.048
70, 70	15.1	0.049	0.048	0.048	0.047
80, 80	15.9	0.049	0.053	0.053	0.043
100, 100	17.1	0.045	0.049	0.049	0.046
120, 120	18.1	0.043	0.044	0.044	0.048
50, 60	13.5	0.042	0.040	0.047	0.041
50, 70	14.0	0.044	0.046	0.046	0.048
50, 80	14.4	0.048	0.050	0.049	0.045
50, 100	15.0	0.053	0.049	0.047	0.053
50, 120	15.4	0.042	0.047	0.053	0.047
60, 50	13.4	0.045	0.047	0.050	0.043
60, 70	14.6	0.047	0.052	0.052	0.048
60, 80	15.0	0.050	0.049	0.049	0.047
60, 100	15.6	0.041	0.046	0.046	0.043
60, 120	16.1	0.048	0.052	0.048	0.046
70, 50	13.7	0.042	0.041	0.045	0.046

Table 2. (continued)

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
70, 60	14.5	0.041	0.045	0.048	0.045
70, 80	15.5	0.045	0.049	0.049	0.046
70, 100	16.2	0.048	0.044	0.044	0.048
70, 120	16.6	0.046	0.045	0.045	0.046
80, 50	14.0	0.044	0.046	0.041	0.047
80, 60	14.8	0.046	0.041	0.044	0.046
80, 70	15.4	0.041	0.045	0.046	0.045
80, 100	16.6	0.044	0.048	0.048	0.048
80, 120	17.1	0.045	0.045	0.046	0.052
100, 50	14.4	0.049	0.049	0.047	0.050
100, 60	15.2	0.047	0.047	0.045	0.047
100, 70	15.9	0.049	0.045	0.046	0.045
100, 80	16.4	0.048	0.048	0.048	0.043
100, 120	17.7	0.052	0.046	0.052	0.050
120, 50	14.7	0.041	0.043	0.046	0.043
120, 60	15.5	0.045	0.045	0.041	0.046
120, 70	16.2	0.046	0.046	0.044	0.048
120, 80	16.7	0.042	0.044	0.044	0.049
120, 100	17.5	0.045	0.045	0.045	0.046

3.2 Performance of tests when Survival curves cross at the middle/ lower part

We considered a situation where two curves cross at the middle/lower part of curves (Figure 2). Again, Weibull distribution with shape = 1.5 and scale = 1 was used to generate survival times for Group I, and $U(0, 6.5)$ was used to obtain a 13% censoring rate for the group. In Group II, survival times follow a Weibull distribution with shape = 10 and scale = 1. For obtaining a 20% censoring rate, $U(0, 4.5)$ was used. Table 3 is reserved to show the statistical power of the log-rank test, Wilcoxon test, Tarone-Ware test, and proposed test. From the table, we conclude that the proposed test has considerably greater power in each combination of sample size (equal or unequal). Since the situation considered here covered two aspects, i.e., middle and lower part crossing, for small or any combination with a small sample size, the power of the Wilcoxon test is smaller than the new test. The Wilcoxon test is sensitive to the initial difference and not so sensitive to the middle as well as to the differences between groups that occur at the middle/last observed times. The power of the Wilcoxon test increases with the increase of sample size, but it does not approach the power of the new weighted test. We observed the same behavior from the Tarone-Ware test, although its results are much better than the log-rank test but worse than the Wilcoxon and our proposed test. Since the two curves cross, this is not an ideal situation for the log-rank test. The power of the log-rank test is much smaller than the other tests.

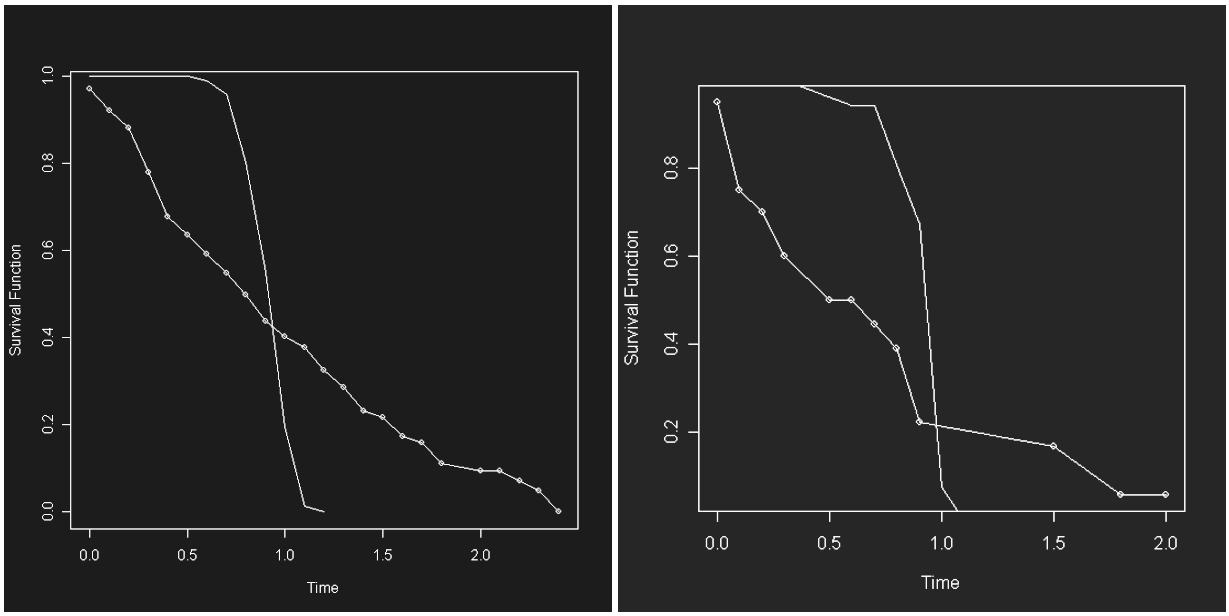

Figure 2. Survival curves crosses at middle

Table 3. Power of log-rank, Wilcoxon, Tarone-Ware, and new tests for middle/lower part crossing

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
50, 50	5.3	0.056	0.668	0.258	0.987
60, 60	5.7	0.060	0.751	0.349	0.995
70, 70	6.0	0.056	0.800	0.391	0.999
80, 80	6.2	0.058	0.857	0.437	0.999
100, 100	6.5	0.057	0.915	0.499	0.999
120, 120	6.7	0.055	0.956	0.578	0.999
50, 60	5.5	0.080	0.717	0.311	0.989
50, 70	5.6	0.096	0.740	0.380	0.991
50, 80	5.8	0.109	0.759	0.405	0.991
50, 100	6.0	0.131	0.796	0.446	0.995
50, 120	6.1	0.163	0.808	0.484	0.999
60, 50	5.5	0.049	0.704	0.309	0.994
60, 70	5.8	0.071	0.781	0.359	0.995
60, 80	6.0	0.085	0.805	0.422	0.997
60, 100	6.2	0.112	0.827	0.370	0.999
60, 120	6.4	0.132	0.869	0.401	0.999
70, 50	5.6	0.031	0.730	0.310	0.998
70, 60	5.8	0.050	0.779	0.365	0.999
70, 80	6.1	0.071	0.823	0.400	0.999

Table 3. (continued)

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
70, 100	6.3	0.093	0.862	0.490	0.999
70, 120	6.5	0.104	0.887	0.508	0.999
80, 50	5.7	0.027	0.736	0.310	0.999
80, 60	5.9	0.037	0.794	0.365	0.999
80, 70	6.0	0.049	0.818	0.400	0.999
80, 100	6.4	0.078	0.883	0.490	0.999
80, 120	6.6	0.103	0.895	0.508	0.999
100, 50	5.8	0.015	0.793	0.311	0.999
100, 60	6.0	0.030	0.835	0.374	0.999
100, 70	6.1	0.031	0.871	0.420	0.999
100, 80	6.2	0.049	0.893	0.452	0.999
100, 120	6.7	0.079	0.931	0.562	0.999
120, 50	5.8	0.010	0.827	0.319	0.999
120, 60	6.0	0.016	0.866	0.364	0.999
120, 70	6.1	0.021	0.893	0.421	0.999
120, 80	6.3	0.027	0.915	0.473	0.999
120, 100	6.5	0.039	0.943	0.524	0.999

3.3 Performance of tests when Survival curves cross at the earlier stage

For handling the present case, a censoring distribution was considered for both groups. For Group I, Weibull distribution with survival function, i.e. $S(t) = \exp(-\lambda t^{\rho})$ and $(\lambda, \rho) = (0.1, 0.5)$, was used for generating the survival times. Similarly, for Group II, survival times obtained from the other form of Weibull function, i.e. $S(t) = \exp(-\lambda t^{\rho})$ and $(\lambda, \rho) = (0.1, 1.5)$. Due to the different values of the shape parameter of the two groups, we got 13% and 22% censoring rates for Group I and II, respectively. The power of the log-rank test due to the assignment of equal weight and not being sensitive to the crossing situation is smaller as compared to other tests. It decreases with the increasing sample size. The power of our test is higher than the other competitors and, unlike the log-rank test, it increases with the increase of sample size. The power of the Wilcoxon test, which is smaller than the proposed test power, increases with the sample size. Again, the simulation results of the Tarone-Ware test are not satisfactory; in some cases, the log-rank test's performance is better than, shown in Figure 3 and Table 4.

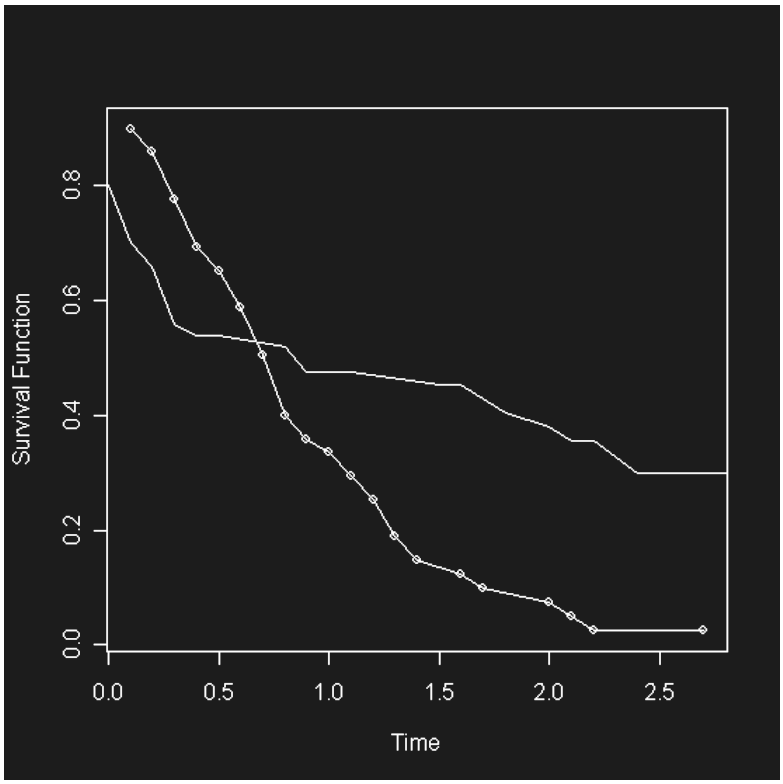

Figure 3. Survival curves cross at earlier stage

Table 4. Power of log-rank, Wilcoxon, Tarone-Ware, and new tests for early stage crossing

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
50, 50	10.0	0.178	0.218	0.106	0.304
60, 60	12.2	0.176	0.285	0.112	0.367
70, 70	13.3	0.173	0.354	0.118	0.432
80, 80	14.3	0.156	0.422	0.129	0.472
100, 100	16.1	0.152	0.550	0.161	0.483
120, 120	17.4	0.138	0.662	0.191	0.451
50, 60	11.6	0.141	0.246	0.088	0.345
50, 70	12.3	0.100	0.294	0.078	0.396
50, 80	12.9	0.068	0.329	0.080	0.458
50, 100	13.8	0.041	0.386	0.083	0.516
50, 120	14.5	0.029	0.410	0.085	0.539
60, 50	11.2	0.208	0.253	0.126	0.344
60, 70	12.8	0.147	0.320	0.102	0.419
60, 80	13.5	0.105	0.360	0.104	0.460
60, 100	14.5	0.071	0.436	0.102	0.491

Table 4. (continued)

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
60, 120	15.3	0.038	0.481	0.102	0.511
70, 50	11.6	0.223	0.305	0.138	0.398
70, 60	12.5	0.204	0.315	0.128	0.409
70, 80	14.0	0.140	0.394	0.111	0.441
70, 100	15.0	0.093	0.468	0.126	0.483
70, 120	15.8	0.064	0.527	0.117	0.475
80, 50	11.8	0.238	0.344	0.166	0.452
80, 60	12.8	0.223	0.352	0.160	0.459
80, 70	13.6	0.194	0.374	0.136	0.465
80, 100	15.5	0.113	0.499	0.138	0.481
80, 120	16.3	0.081	0.546	0.134	0.468
100, 50	12.3	0.251	0.390	0.178	0.546
100, 60	13.3	0.240	0.429	0.180	0.525
100, 70	14.2	0.220	0.437	0.164	0.502
100, 80	14.9	0.209	0.470	0.165	0.502
100, 120	17.0	0.128	0.591	0.164	0.457
120, 50	12.5	0.247	0.442	0.191	0.612
120, 60	13.6	0.250	0.474	0.196	0.592
120, 70	14.5	0.251	0.471	0.190	0.577
120, 80	15.2	0.215	0.527	0.186	0.548
120, 100	16.5	0.185	0.591	0.181	0.505

3.4 Performance of tests, under proportional hazard model (Non overlapping)

For the proportional hazard model, Weibull distribution with shape parameter = 1.5 and scale = 1, and Weibull distribution with shape parameter = 2 and scale = 1.5 were considered for generating survival times for Group I and II respectively. To obtain equal censoring rate (here 15%), censoring times of two groups were drawn from $U(0, 6.5)$ and $U(0, 2)$ respectively (Figure 4). Table 5 shows the statistical power of the four tests. As in the proportional hazard model, survival curves do not cross and the log-rank test is the most suitable choice giving optimal power. One can see this fact from the table that the power of the log-rank test is higher than the other tests. The power of the proposed test is smaller than the power of the log-rank and Tarone-Ware tests but increases with the increase of sample size. The performance of the proposed test compared with the Tarone-Ware test is not clear; at some points it gives better results, but not always.

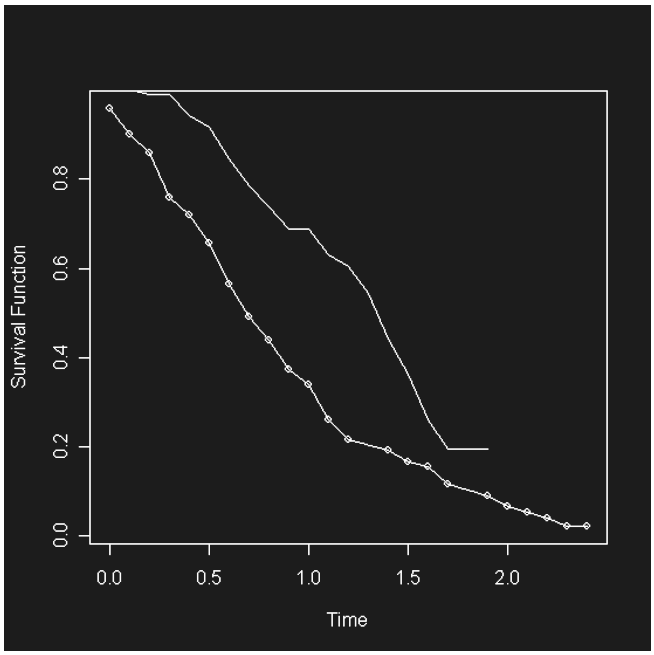

Figure 4. Survival curves do not cross

Table 5. Power of log-rank, Wilcoxon, Tarone-Ware, and new tests for proportional hazards model

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
50, 50	8.7	0.996	0.988	0.875	0.883
60, 60	9.4	0.999	0.997	0.928	0.931
70, 70	9.9	0.999	0.995	0.954	0.955
80, 80	10.3	0.999	0.994	0.976	0.980
100, 100	11.0	0.999	0.996	0.993	0.993
120, 120	11.5	0.999	0.996	0.997	0.996
50, 60	9.1	0.999	0.995	0.906	0.895
50, 70	9.4	0.999	0.995	0.926	0.910
50, 80	9.7	0.999	0.995	0.941	0.912
50, 100	10.1	0.999	0.996	0.959	0.911
50, 120	10.4	0.999	0.999	0.972	0.915
60, 50	9.0	0.998	0.995	0.947	0.907
70, 70	9.7	0.999	0.997	0.895	0.936
60, 80	9.9	0.999	0.998	0.960	0.938
60, 100	10.4	0.999	0.998	0.976	0.943
60, 120	10.8	0.999	0.998	0.984	0.942
70, 50	9.1	0.999	0.997	0.908	0.912
70, 60	9.5	0.999	0.998	0.942	0.949

Table 5. (continued)

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
70, 80	10.2	0.999	0.998	0.968	0.960
70, 100	10.7	0.999	0.998	0.983	0.967
70, 120	11.0	0.999	0.999	0.986	0.965
80, 50	9.2	0.999	0.998	0.917	0.921
80, 60	9.7	0.999	0.998	0.945	0.952
80, 70	10.1	0.999	0.998	0.968	0.968
80, 100	10.8	0.999	0.998	0.986	0.980
80, 120	11.2	0.999	0.998	0.992	0.984
100, 50	9.3	0.999	0.997	0.936	0.918
100, 60	9.8	0.999	0.997	0.963	0.964
100, 70	10.2	0.999	0.997	0.976	0.978
100, 80	10.5	0.999	0.997	0.984	0.987
100, 120	11.4	0.999	0.997	0.995	0.990
120, 50	9.4	0.999	0.998	0.950	0.910
120, 60	9.9	0.999	0.998	0.971	0.960
120, 70	10.2	0.999	0.998	0.983	0.982
120, 80	10.6	0.999	0.998	0.991	0.989
120, 100	11.1	0.999	0.998	0.996	0.995

3.5 Size of the test/ Estimated Type I error

For comparing the size of three tests, two random samples are generated independently from an exponential distribution with parameter λ = 1. The censoring distribution is $U(0, 6.5)$ which results in about 15% censoring in each group. To obtain the number of ties, we round both censoring and survival distributions up to 3 decimal points. We also tried rounding to 1 and 2 decimal points and obtained nearly similar results. In this simulation, we performed 5000 iterations. The sizes were calculated as the proportions of rejections based on 5000 iterations. The results of the simulation for different sample sizes are summarized in Table 6. The simulation result shows that the size of each test is approximated by the significance level α = 0.05. In other words, all statistics behave reasonably well in terms of sizes.

Table 6. Size of log-rank, Wilcoxon, and new tests

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
50, 50	1.0	0.053	0.053	0.053	0.053
60, 60	1.5	0.055	0.055	0.051	0.057
70, 70	2.0	0.058	0.052	0.052	0.055
80, 80	2.6	0.059	0.058	0.052	0.059
100, 100	3.5	0.045	0.058	0.058	0.059

Table 6. (continued)

Sample Size	Average Tie	Log-rank test	Wilcoxon test	Tarone-Ware Test	New test
120, 120	5.6	0.053	0.050	0.048	0.053
50, 60	1.2	0.053	0.050	0.051	0.051
50, 70	1.4	0.057	0.052	0.052	0.052
50, 80	1.6	0.050	0.051	0.051	0.056
50, 100	2.0	0.053	0.052	0.054	0.059
50, 120	2.4	0.052	0.050	0.053	0.059
60, 50	1.3	0.056	0.047	0.050	0.053
70, 70	1.7	0.053	0.046	0.050	0.057
60, 80	1.9	0.058	0.053	0.052	0.055
60, 100	2.4	0.054	0.051	0.054	0.059
60, 120	2.9	0.048	0.045	0.057	0.053
70, 50	1.4	0.056	0.052	0.057	0.057
70, 60	1.7	0.053	0.047	0.059	0.059
70, 80	2.2	0.058	0.051	0.051	0.059
70, 100	2.8	0.059	0.053	0.051	0.058
70, 120	3.3	0.058	0.053	0.054	0.058
80, 50	1.6	0.058	0.052	0.058	0.052
80, 60	2.0	0.054	0.052	0.053	0.054
80, 70	2.3	0.056	0.051	0.056	0.051
80, 100	3.2	0.058	0.053	0.057	0.056
80, 120	3.8	0.049	0.058	0.058	0.052
100, 50	2.0	0.049	0.049	0.050	0.050
100, 60	2.5	0.055	0.051	0.057	0.049
100, 70	2.8	0.052	0.049	0.053	0.052
100, 80	3.2	0.054	0.057	0.054	0.054
100, 120	4.8	0.057	0.053	0.058	0.058
120, 50	2.4	0.055	0.057	0.056	0.056
120, 60	2.8	0.056	0.055	0.059	0.055
120, 70	3.4	0.053	0.051	0.052	0.052
120, 80	3.8	0.052	0.052	0.053	0.053
120, 100	4.5	0.052	0.048	0.049	0.052

3.6 Example

SALAHUDDIN and ZAMAN (2005) compared on the basis of sex the survival times of 320 -thalassaemia major patients at the Fatimid Foundation Peshawar (enrolled between 1990 and 2003) using the log-rank test, which favored the null hypothesis. The data is also referenced by SYED ALMAS, QAMRUZ ZAMAN, DANISH WASIM, SOFIA MANSOOR (2024). The group sizes were 191 (57 censored) for the male group

and 129 (44 censored) for the female group, with 40 occurrences of ties. The survival comparison of the two groups is depicted in Figure 5. This demonstrates the presence of multiple crossings, and literature suggests that the log-rank test has limited power to detect differences in such cases. For this dataset, the results of the log-rank test, Wilcoxon test, and Tarone-Ware test are shown in Table 7.

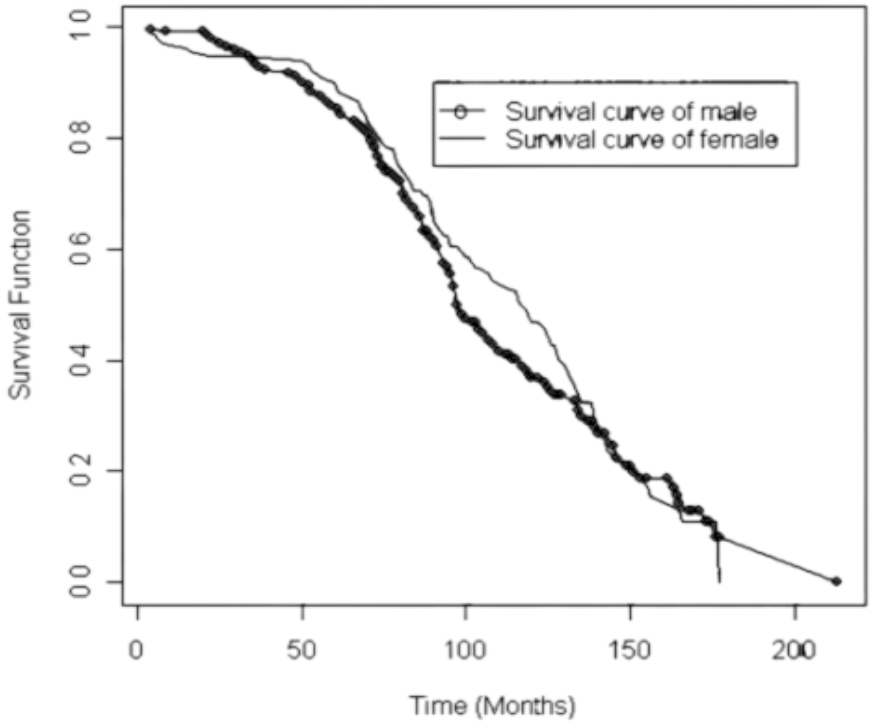

Figure 5. Survival curves do not cross

Table 7. Critical values and p-values from the application of the log-rank and Weighted tests for the Thalassaemia data set

Statistical test	Critical value	p-value
Log-rank	0.692	0.406
Wilcoxon	1.468	0.226
Tarone-Ware	1.305	0.253

From the Figure, we can tell that these two curves are different. However, the results of commonly used tests shown in Table 7 suggest no significant differences between the male and female survivals. At a 5% significant level, the new weighted test statistic gives a statistic value of 4.95 with a p-value of 0.0261 indicating the considerable difference between the survivals. Here the new test performs better than other tests.

4 Conclusion

There are several tests available for the comparison of two or more survival curves (Wilcoxon, Tarone-Ware, Peto and Peto etc); all these tests are useful for single crossing of curves. In case of more than one time crossing of survival curves, these tests may have little power. A new weight-based test, which offers a bridge between Wilcoxon and Tarone-Ware weight, is proposed to handle the situation. The method is used for general situations, as it is not based on any specific assumption on the underlying distribution. The results of an extensive simulation study show that the new weighted test is more powerful in different circumstances, where the curves cross.

Recommendations

It is recommended that this study can be extended by developing new estimators utilizing other sampling schemes.

Data Availability

The data used to support the findings of the study are available within this article.

Acknowledgment

The authors express appreciation to the editor and reviewers for their valuable and positive comments/suggestions which certainly have improved the presentation and quality of the paper.

Author Contributions

Qamruz Zaman: Conceptualization, Supervision. **Nisar Ullah**: analyzed and interpreted data, Methodology. **Syed Habib Shah**: Investigation, Data curation, **Muhammad Ali**: removed all the grammatical mistakes. **Muhammad Irshad and Sumayyia Azam** : Improved language and Editing, software use.

Compliance with Ethical Standards

The authors declare that they have no conflicts of interest. Additionally, this article does not involve studies with human participants or animals conducted by any of the authors. Furthermore, informed consent was taken from all individual participants involved in the study.

Author Information

ORCID:
Qamruz Zaman: 0000-0002-9557-4036
Nisar Ullah: 0009-0009-1317-1208
Muhammad Ali: 0009-0002-7548-5199
Muhammad Irshad: 0009-0000-4656-7772
Sumayyia Azam: 0009-0006-1350-0225

References

[1] Csalódi, R., Bagyura, Z., Vathy-Fogarassy, Á., & Abonyi, J. (2024). Time-dependent frequent sequence mining-based survival analysis. Knowledge-Based Systems, 296, 111885.

[2] Li, X., Marcus, D., Russell, J., Aboagye, E. O., Ellis, L. B., Sheeka, A., ... & Rockall, A. G. (2024). Weibull parametric model for survival analysis in women with endometrial cancer using clinical and T2-weighted MRI radiomic features. *BMC Medical Research Methodology*, *24*(1), 107.

[3] Srujana, B., Verma, D., & Naqvi, S. (2024). Machine learning vs. survival analysis models: a study on right censored heart failure data. Communications in Statistics-Simulation and Computation, 53(4), 1899-1916.

[4] Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, *51*(6), 1-36.

[5] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep, 50(3), 163-170.

[6] Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society: Series A (General), 135(2), 185-198.

[7] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, 52(1-2), 203-224.

[8] Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. Biometrika, 52(3/4), 650-653.

[9] Tarone, R. E., & Ware, J. (1977). On distribution-free tests for equality of survival distributions. Biometrika, 64(1), 156-160.

[10] Bland, J. M., & Altman, D. G. (2004). The log-rank test. Bmj, 328(7447), 1073.

[11] Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., & Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. Biometrics, 607-625.

[12] SPOSTO, R., STABLEIN, D., & CARTER-CAMPBELL, S. H. E. L. L. Y. (1997). A partially grouped logrank test. Statistics in medicine, 16(6), 695-704.

[13] You, N., He, X., Dai, H., & Wang, X. (2023). Ball divergence for the equality test of crossing survival curves. Statistics in medicine, 42(29), 5353-5368.

[14] Collett, D. (1994). Modelling survival data. In Modelling survival data in medical research (pp. 53-106). Springer US.

[15] Muse, A. H., Chesneau, C., Ngesa, O., & Mwalili, S. (2022). Flexible parametric accelerated hazard model: Simulation and application to censored lifetime data with crossing survival curves. Mathematical and Computational Applications, 27(6), 104.

[16] Fernández, T., Gretton, A., Rindt, D., & Sejdinovic, D. (2023). A kernel log-rank test of independence for right-censored data. *Journal of the American Statistical Association*, *118*(542), 925-936.

[17] Dormuth, I., Liu, T., Xu, J., Pauly, M., & Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. Contemporary clinical trials, 128, 107165.

[18] Brookmeyer, R., & Curriero, F. C. (2002). Survival curve estimation with partial non-random exposure information. Statistics in medicine, 21(18), 2671-2683.

[19] Adebayo, S. B., & Fahrmeir, L. (2005). Analysing child mortality in Nigeria with geoadditive discrete-time survival models. Statistics in medicine, 24(5), 709-728.