Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(1):13-27

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2020-80208

Comparison of the Statistical Tests for Two Crossing Survival Functions

Kesişen İki Yaşam Fonksiyonu İçin Kullanılan İstatistiksel Testlerin Karşılaştırılması

Hülya ÖZEN[a], Cengiz BAL[a], Ertuğrul ÇOLAK[a]

[a]Department of Biostatistics, Eskişehir Osmangazi University Faculty of Medicine, Eskişehir, TURKEY

This study is partially presented as an orally in 21st National and 4th International Biostatistics Congress, 26-29 October 2019, Antalya, Turkey.

ABSTRACT Objective: Testing the equality of two survival functions is frequently used in survival analysis studies. The most preferred method in the literature is the log-rank (LR) test, but it provides misleading results in evaluating the crossing survival functions in which the proportional hazards assumption is violated. Therefore, different comparison tests have been proposed. In this study, currently proposed partitioned LR tests and weighted Lin-Wang tests were compared with LR and weighted LR tests. **Material and Methods:** Type I error rates and powers of comparison tests were compared in different sample sizes and censoring rates using the Monte Carlo simulation technique. The weaknesses and strengths of the comparison tests were examined in various scenarios. In addition to the simulation study, comparison tests were evaluated using open source real-life data set. **Results:** Simulation results showed that all tests provided reasonable Type I error rates. The most successful results in the scenarios of crossing survival functions belonged to partitioned LR tests. It was observed that the location of the crossing point affected the performance of the tests adversely. **Conclusion:** Proportional hazards assumption should be tested before comparing two survival functions. It is recommended to use the LR test, when the proportional hazards assumption is provided and the use of partitioned LR tests in the comparison of crossing survival functions.

Keywords: Log-rank test; partitioned log-rank tests; survival analysis; weighted lin-wang tests

ÖZET Amaç: İki yaşam fonksiyonunun benzerliğinin araştırılması yaşam analizi çalışmalarında sıklıkla kullanılmaktadır. Literatürde, en çok tercih edilen yöntem log-rank (LR) testidir ancak bu yöntem orantılı hazardlar varsayımının ihlal edildiği kesişen yaşam fonksiyonlarını değerlendirmede yanıltıcı sonuçlar sunmaktadır. Bu nedenle kesişen yaşam fonksiyonlarını değerlendirmek için literatürde farklı karşılaştırma testleri önerilmiştir. Bu çalışmada, literatürde önerilen güncel yöntemlerden parçalanmış LR testleri ve ağırlıklandırılmış Lin-Wang testlerinin, LR ve ağırlıklandırılmış LR testleri ile karşılaştırılması amaçlanmaktadır. **Gereç ve Yöntemler:** Monte Carlo simülasyon tekniği kullanılarak, karşılaştırma testlerinin Tip I hata oranları ve güçleri, farklı örneklem büyüklükleri ve sansür oranlarında karşılaştırıldı. Karşılaştırma testlerinin zayıf ve güçlü yönleri çeşitli senaryolarda incelendi. Simülasyon çalışmasına ek olarak, karşılaştırma testleri gerçek hayattaki bir veri seti üzerinde değerlendirildi. **Bulgular:** Simülasyon sonuçları, tüm testlerin makul Tip I hata oranları sağladığını gösterdi. Kesişen yaşam fonksiyonlarına ait senaryolarda en başarılı sonuçlar parçalanmış LR testlerine aitti. Kesişme noktasının konumunun, karşılaştırma testlerinin performansını olumsuz yönde etkilediği gözlendi. **Sonuç:** Orantılı hazardlar varsayımı, 2 yaşam fonksiyonu karşılaştırılmadan önce test edilmelidir. Orantılı hazardlar varsayımı sağlandığında LR testinin, kesişen yaşam fonksiyonlarının karşılaştırılmasında ise parçalanmış LR testlerinin kullanılması önerilmektedir.

Anahtar kelimeler: Log-rank testi; parçalanmış log-rank testleri; yaşam analizi; ağırlıklandırılmış lin-wang testleri

Survival analysis is frequently used in medical studies. A comparison of two survival functions with censored observations is one of the main goals in survival analysis. Log-rank (LR), the most commonly used test in the literature, has the greatest power while the proportional hazards assumption is valid during the follow-up period.[1] On the other hand, LR may lose power when two survival functions cross each other since the initially determined positive differences between survival functions are neutralized by later calculated

Correspondence: Hülya ÖZEN
Department of Biostatistics, Eskişehir Osmangazi University Faculty of Medicine, Eskişehir, TURKEY/TÜRKİYE
E-mail: hulya_ozen@yahoo.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 25 Nov 2020 ***Received in revised form:*** 04 Jan 2021 ***Accepted:*** 05 Jan 2021 ***Available online:*** 29 Apr 2021

negative differences.[2] The occurrence of crossing survival functions, which is seen frequently in the survival studies, is one of the situations where the proportional hazards assumption is violated. The location of the crossing point can be seen in the early or late period of follow-up studies. Therefore, the idea of using one comparison test for all situations seems to be inappropriate.[3] Unfortunately, researchers often use the LR test without testing if the proportional hazards assumption is valid.[4-6] Kristiansen examined 175 studies containing survival analysis in five reputable journals and found out 47% of the studies included crossing survival functions. In the review, 70% of those studies with crossing survival functions were evaluated with the LR test.[7] Suciu et al. reviewed 107 publications that include comparisons of survival functions in a major medical journal and found that 95 of these publications used the LR test, where 42 of them compared crossing survival functions. Researchers think that the inappropriate use of the LR test relies on its availability in many statistical packages.[8]

Various tests are recommended in the literature for comparison of the crossing survival functions. Most of these tests were obtained based on differences between the Nelson-Aalen cumulative hazard functions or Kaplan-Meier survival function estimators. Generalized Cramer-von Mises tests have been proposed by Koziol and Schumacher, especially for censored data sets.[9,10] Gill developed Renyi type tests with weighted tests and mathematical supremum function approach.[11] The modified Kolmogorov-Smirnov test is another proposed test for crossing survival functions.[12] Pepe and Fleming presented weighted Kaplan-Meier test to the literature.[13] Lin and Wang proposed a new test statistic based on squared differences at each time point.[14] Qiu and Sheng constructed a two-stage method to use in the case of crossing survival functions.[15] Kraus developed a group of tests using Neyman's smooth tests.[16] Lin and Xu obtained a new test on areas under two survival functions.[17] Koziol and Jia proposed weighted Lin-Wang tests with adding different weight functions to Lin and Wang's (LW) test.[18] Liu and Yin proposed partitioned LR tests with using weighted LR test statistics by partitioning the data set at each time point.[2]

Some of the aforementioned methods have been evaluated via simulation studies under different survival scenarios. The methods proposed by Lin and Wang, and by Lin and Xu performed better than LR and Wilcoxon tests in simulation studies under crossing survival scenarios.[14,17] Tubert-Bitter et al. and Li et al. conducted a simulation study in the cases of early, middle and late crossing survival functions.[19,20] Li et al. conducted a comprehensive simulation study and compared several proposed tests for crossing survival functions.[20] Morever, Liu et al. and Kraus performed simulation studies of various patterns of crossing hazard rates and concluded that some of the weighted LR tests lose power with respect to the methods they proposed.[15,16,21] Liu and Yin demonstrated that the methods they proposed performed better than some of the weighted LR tests Gehan-Wilcoxon (GW), Tarone-Ware (TW) and Peto-Peto (PP) and Qiu and Sheng's method under different crossing hazards scenarios.[2,15]

In this study, we performed simulation studies focusing on some particular crossing points since there are a few studies that pay attention to the location of crossing points. Partitioned LR tests and weighted LW tests were recently proposed as opposed to each other.[2,18] In addition to these tests, conventional weighted LR tests were included in the study. Monte Carlo simulation studies were conducted under various crossing survival functions scenarios, sample sizes and censoring rates to evaluate the power of these comparison tests. The aim of this study was to evaluate the performance of the comparison tests for various scenarios and to suggest an appropriate test that would be useful in the field of survival analysis. In Section 2; some notations, the comparison tests and details of simulation studies are explained. The simulation results and application of the comparison tests on a real data example are given in Section 3. Finally, we present a discussion in Section 4 and conclusions in Section 5.

MATERIAL AND METHODS

NOTATIONS AND TWO-SAMPLE HYPOTHESIS

Let $t_1 < t_2 < \ ... < \ t_m$ be the distinct failure times in the pooled sample of two groups, where t_m is the latest event time and $S_j(t) = P(T_j > t)$ be the survival function of group j. In this study, the main hypothesis that comparison tests deal with is shown below:

$$H_0: S_1(t) = S_2(t), \text{ for all } t \text{ versus } H_1: S_1(t) \neq S_2(t) \text{ , for some } t.$$

The Kaplan-Meier estimator of survival function is defined as $\hat{S}(t) = 1$ for $t < t_1$ and $\hat{S}(t) = \prod_{t_i \leq t}(1 - d_i/n_i)$ for $t \geq t_1$, where d_i and n_i indicates the number of observed failures and the number of subjects at risk at t_i time, respectively.

COMPARISON TESTS IN THE STUDY

There are basically three comparison test groups in the study: (1) weighted LR tests, (2) partitioned LR tests and (3) weighted LW tests. Among all, weighted LR tests are the most popular tests. It is seen that researchers add different weight functions to LR test in order to better reflect their alternative hypotheses such as GW, TW, PP, Modified Peto-Peto (MPP) and Fleming-Harrington (FH).[22-27] Both GW and TW give more weight to the difference between the number of events observed and expected at the beginning of the study. PP and MPP are alternative approaches for censored data sets. FH provides a wide test class that includes most of the tests considered as special cases with the weight function $\left[\hat{S}(t-)\right]^p\left[1-\hat{S}(t-)\right]^q$ $(p \geq 0, q \geq 0)$. To focus on the difference both in the beginning and end of the follow-up time, FH (p=1, q=1) situation is taken into consideration in this study. In order to obtain a result for two group comparisons, all weighted LR test statistics are evaluated with chi-square distribution with one degree of freedom. Even though different weight functions offer different approaches, weighted LR tests may fail to detect the difference in the case of comparing crossing survival functions due to some cancellation in numerator of the test statistic formula. Therefore some other testing strategies are proposed.

Liu and Yin proposed partitioned LR tests by using log-rank (SUP-LR) Gehan-Wilcoxon (SUP-GW) and Peto-Peto (SUP-PP) throughout the test statistics calculations.[2] Liu and Yin's partitioned log-rank tests take their name from partitioning the time axis throughout the calculations to solve the problem arising from the numerator of the weighted LR test statistic. In order to reduce power loss and prevent arbitrary selection of the partitioning point, distinct failure time points of the pooled data set are used. At each failure time point, the data set is partitioned into two parts. After obtaining test statistics for each part, they are summed and noted. Finally, the largest value of the summed test statistics is determined as partitioned log-rank test statistics. Liu and Yin used Brownian motion for the limit distribution of the partitioned log-rank test and concluded that it is not easy to calculate the p-value with this approach. For this reason, they proposed an algorithm using nonparametric bootstrap method to obtain the p-value.

The final comparison test group is weighted LW tests that are originated from LW test. Koziol and Jia obtained different comparison tests by adding the weight functions n_i (LW_{w1}), $\sqrt{n_i}$ (LW_{w2}) and $1/\sqrt{Var(d_{1i})}$ (LW_{w3}) where d_{1i} indicates the number of observed failures in group 1 at t_i time.[18] All tests in this group use the sum of the squared differences between the numbers of observed and expected failure observations over failure time points. With this approach, all positive and negative differences are considered, unlike weighted LR tests. Each test statistic in the group follows a standard normal distribution. Therefore, the assessment of the null hypothesis is done with a two-sided test based on the standard normal distribution.

SIMULATION DESIGN

Monte Carlo simulations are carried out to evaluate the performance of the tests in the study. All calculations during the study are conducted in R Studio v 1.1.463 with using R packages (dplyr, survival,survminer,

survMisc).[28-33] Comparisons are made through the statistical power and Type I error rate. Statistical power comparisons are obtained with four following scenarios: (1) two groups with proportional hazard rates, (2) two survival functions crossing at the point where S(t)>0.6, (3) two survival functions crossing at S(t) = 0.4∼0.6 and (4) two survival functions crossing at S(t) = 0.2∼0.4. Scenario (5) is created to obtain Type I error rates. During survival time generation process, Weibull $W(\alpha,\beta)$ and exponential $Exp(\lambda)$ distributions are used, where $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale parameter and $\lambda > 0$ is the rate parameter. To obtain crossing survival functions, survival time generation of group 2 is created by two steps. Detailed distribution and parameter values are shown in Table 1. Since survival time units are generated with various parameters, different values can be observed on survival time axes in Figure 1.

TABLE 1: Parameter values of distributions used in survival time generation.

	Scenarios				
	(1)	(2)	(3)	(4)	(5)
Group 1	$\text{Exp}(0.5)$	$W(2.5,30)$	$\text{Exp}(1/12)$	$W(1.5,5)$	$\text{Exp}(1)$
Group 2	$\text{Exp}(0.2)$	$\text{Exp}(0.125), t < 1$ $\text{Exp}(0.01), t \geq 1$	$\text{Exp}(1/4), t < 2$ $\text{Exp}(1/35), t \geq 2$	$\text{Exp}(1/4), t < 1.5$ $\text{Exp}(1/35), t \geq 1.5$	$\text{Exp}(1)$

t: Generated survival time.

Throughout the study, uniform (0,1) distribution is used in order to create randomly occurred censored observations across the groups. In all simulation studies, censoring rates are determined to be equal in each group and considered as 0%, 20%, 40%, 60%. Equal sample sizes ($n_1 = n_2 = 50, 100$) and unequal sample sizes ($n_1 = 40, n_2 = 80$; $n_1 = 50, n_2 = 100$) are determined in each group. To obtain p values for partitioned log-rank tests, 1000 bootstrap samples are employed for each two-group comparison. Type I error rates and powers for each scenario are calculated based on 2000 independent replications. The exact power of comparison tests is estimated by calculating the proportion of results where the null hypothesis is rejected at the $\alpha = 0.05$ significance level.

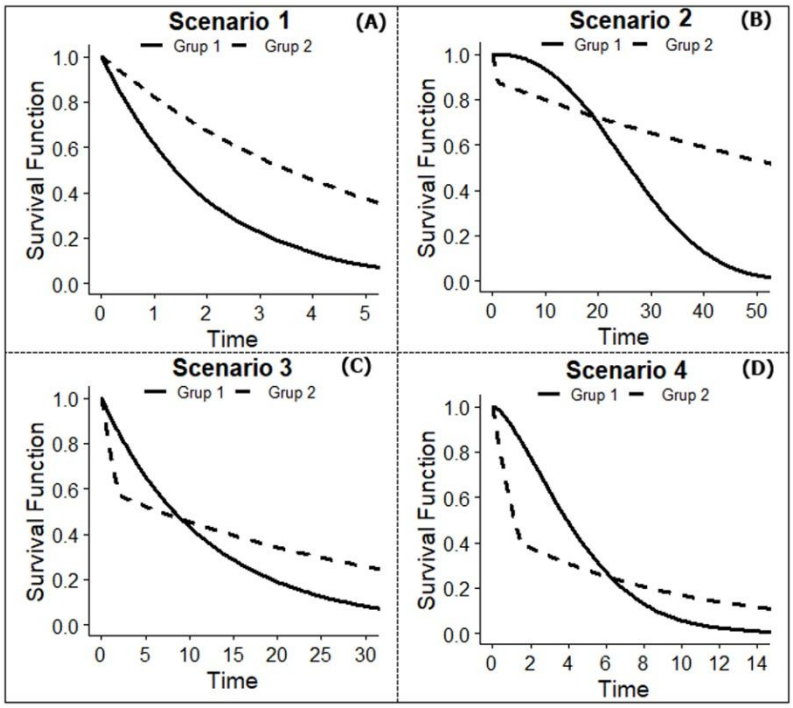

FIGURE 1: Configurations of survival functions for power simulation scenarios.

RESULTS

ESTIMATION OF STATISTICAL POWER AND TYPE I ERROR

Scenario 1. We consider a situation where two survival functions have proportional hazards (see Figure. 1 A). The estimated statistical power results are presented in Table 2. Power of comparison tests rise with an increase in sample size; however, the power of the tests declines with an increase in censoring rate. LR provides the highest powers among the others. Then it is followed by weighted LR tests. Partitioned log-rank tests and weighted LW tests also give high power results. However, they are not as successful as LR and other weighted LR tests. In each group, when the sample size is 50 or above and the censoring rate does not exceed 40%, the powers of tests are observed above 80%.

Scenario 2. The power estimations of the comparison tests obtained according to the early crossing survival functions (S(t)>0.6) scenario (see Figure. 1 B) are shown in Table 3. The increase in the censoring rate has the most adverse effect on the estimated power of the GW. When the sample sizes are equal to 50 and the censoring rate exceeds 40%, the power suddenly gets close to 52%. More emphasis on events at the beginning of the study makes this test insensitive to differences that occur at later times. FH presents the highest power among the other weighted LR tests. Partitioned log-rank tests and FH give the highest power in this scenario and followed by weighted LW test group. In most situations partitioned log-rank tests and FH provide powers equal to 100%.

Scenario 3. For middle crossing $(S(t) = 0.6 \sim 0.4)$ of the survival curves (see Figure. 1 C), it is observed that most of the comparison tests lose power. The results are demonstrated in Table 4. Powers of GW, TW, PP and MPP can barely exceed 10%. Even though FH provides the highest powers among the weighted LR tests, it is not as successful as partitioned log-rank tests. Though weighted LW tests address the crossing survival functions problem, their powers can exceed 80% when the sample sizes are equal to 100 with censoring rates lower than 40%. LW is slightly better when compared to weighted LW tests. Except for the sample sizes equal to 50 with a 60% censoring rate, partitioned log-rank tests provide powers higher than 80%.

Scenario 4. When considering late crossing $(S(t) = 0.2 \sim 0.4)$ of the survival curves (see Figure. 1 D), partitioned log-rank tests provide distinctly high powers. As presented in Table 5, for all sample sizes and censoring rates partitioned log-rank tests have results higher than 90%. Weighted LW tests can present powers higher than 80% when the sample sizes are equal to 100 with censoring rates lower than 40%. LW_{w1} is the most powerful among the LW tests under this scenario. Considering the previous scenarios, FH and LR provide smaller powers. LR can't exceed 10% for any sample sizes and censoring rates. GW presents the most powerful results among the weighted LR tests, followed by MPP, PP, and TW, respectively.

Scenario 5. We investigate Type I error rates of the comparison tests with generating two samples independently from an exponential distribution with a hazard rate equal to $\lambda_1 = \lambda_2 = 1$. For all sample sizes and censoring rates, the estimated Type I error rates are presented in Table 6. Partitioned log-rank tests slightly influenced by an increase in censoring rates. This influence is more obvious for unequal sample sizes situations. The rest of the comparison tests provide the Type I error rate estimations quite close to the nominal level 0.05.

TABLE 2: Estimated power results of comparison tests for Scenario 1.

(n_1, n_2)	c	LR	GW	TW	PP	MPP	FH	SUP-LR	SUP-GW	SUP-PP	LW	LW_{w1}	LW_{w2}	LW_{w3}
(50,50)	0%	0.9935	0.9755	0.9900	0.9755	0.9740	0.9870	0.9735	0.9755	0.9745	0.9305	0.9600	0.9575	0.8955
	20%	0.9745	0.9345	0.9655	0.9515	0.9485	0.9610	0.9465	0.9445	0.9465	0.8735	0.9225	0.9180	0.8200
	40%	0.9235	0.8410	0.8945	0.8870	0.8845	0.8675	0.8675	0.8765	0.8735	0.7625	0.8125	0.8095	0.6755
	60%	0.7775	0.6665	0.7355	0.7460	0.7460	0.6950	0.7540	0.7695	0.7560	0.5975	0.6575	0.6410	0.5310
(100,100)	0%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9995	0.9990	0.9995	0.9995	1.0000	0.9990
	20%	1.0000	0.9995	1.0000	0.9995	0.9995	0.9990	0.9995	1.0000	1.0000	0.9925	0.9975	0.9980	0.9860
	40%	0.9980	0.9895	0.9965	0.9955	0.9955	0.9935	0.9905	0.9930	0.9930	0.9650	0.9900	0.9880	0.9340
	60%	0.9750	0.9330	0.9645	0.9685	0.9685	0.9295	0.9390	0.9530	0.9435	0.8655	0.9355	0.9210	0.7985
(40,80)	0%	0.9985	0.9865	0.9935	0.9865	0.9865	0.9860	0.9895	0.9915	0.9910	0.9945	0.9890	0.9925	0.9880
	20%	0.9840	0.9525	0.9705	0.9610	0.9605	0.9590	0.9630	0.9675	0.9670	0.9690	0.9640	0.9730	0.9465
	40%	0.9355	0.8825	0.9115	0.9070	0.9040	0.8865	0.9175	0.9345	0.9285	0.9145	0.8980	0.9230	0.8675
	60%	0.7985	0.7330	0.7695	0.7745	0.7745	0.7080	0.7715	0.8025	0.7820	0.7670	0.7545	0.7765	0.7015
(50,100)	0%	1.0000	0.9975	0.9995	0.9975	0.9970	0.9990	0.9980	0.9975	0.9975	0.9995	0.9990	0.9995	0.9990
	20%	0.9955	0.9860	0.9910	0.9870	0.9870	0.9900	0.9860	0.9900	0.9900	0.9925	0.9890	0.9935	0.9825
	40%	0.9710	0.9320	0.9570	0.9535	0.9525	0.9380	0.9495	0.9620	0.9580	0.9535	0.9465	0.9600	0.9285
	60%	0.8700	0.8045	0.8505	0.8585	0.8585	0.7785	0.8580	0.8700	0.8610	0.8395	0.8455	0.8590	0.7760

Notations are presented as follows: n_1, n_2: sample sizes of two groups; c: censoring rate; LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang; LW_{w1}; LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{Var(d_{1i})}$

TABLE 3: Estimated power results of comparison tests for Scenario 2.

(n_1, n_2)	c	LR	GW	TW	PP	MPP	FH	SUP-LR	SUP-GW	SUP-PP	LW	LW_{w1}	LW_{w2}	LW_{w3}
(50,50)	0%	0.9995	0.8820	0.9850	0.8820	0.8710	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20%	0.9965	0.8190	0.9545	0.8775	0.8725	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	1.0000	1.0000
	40%	0.9735	0.6860	0.8850	0.8590	0.8550	1.0000	1.0000	1.0000	1.0000	0.9990	0.9985	0.9995	0.9970
	60%	0.8925	0.5195	0.7260	0.7595	0.7555	0.9960	0.9960	0.9945	0.9950	0.9805	0.9735	0.9795	0.9695
(100,100)	0%	1.0000	0.9910	0.9995	0.9910	0.9905	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20%	1.0000	0.9740	1.0000	0.9925	0.9920	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	40%	1.0000	0.9215	0.9965	0.9910	0.9905	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	60%	0.9935	0.8025	0.9555	0.9705	0.9695	1.0000	1.0000	1.0000	1.0000	0.9995	0.9995	0.9995	0.9975
(40,80)	0%	1.0000	0.9360	0.9930	0.9360	0.9355	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	1.0000	1.0000
	20%	0.9975	0.8365	0.9670	0.9075	0.9035	1.0000	1.0000	1.0000	1.0000	1.0000	0.9960	0.9995	1.0000
	40%	0.9830	0.7395	0.9025	0.8755	0.8700	1.0000	0.9995	0.9995	0.9995	0.9995	0.9825	0.9980	0.9990
	60%	0.8780	0.5780	0.7515	0.7825	0.7765	0.9945	0.9950	0.9950	0.9950	0.9775	0.9060	0.9590	0.9775
(50,100)	0%	1.0000	0.9765	0.9995	0.9765	0.9750	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20%	1.0000	0.9275	0.9935	0.9695	0.9685	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	40%	0.9920	0.8265	0.9550	0.9365	0.9355	1.0000	1.0000	1.0000	1.0000	0.9995	0.9945	0.9995	1.0000
	60%	0.9445	0.6700	0.8460	0.8725	0.8700	0.9985	0.9995	0.9995	0.9995	0.9945	0.9545	0.9850	0.9940

Notations are presented as follows: n_1, n_2: sample sizes of two groups; c: censoring rate; LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang; LW_{w1}; LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{\mathrm{Var}(d_{1i})}$

TABLE 4: Estimated power results of comparison tests for Scenario 3.

(n_1, n_2)	c	LR	GW	TW	PP	MPP	FH	SUP-LR	SUP-GW	SUP-PP	LW	LW_{w1}	LW_{w2}	LW_{w3}
(50,50)	0%	0.3075	0.0760	0.0795	0.0760	0.0770	0.6625	0.9930	0.9855	0.9850	0.7100	0.4765	0.6235	0.6790
	20%	0.2715	0.0885	0.0815	0.0790	0.0790	0.6825	0.9745	0.9620	0.9650	0.6175	0.4325	0.5555	0.5820
	40%	0.2095	0.0730	0.0630	0.0530	0.0545	0.6575	0.9070	0.8885	0.8980	0.5000	0.3240	0.4335	0.4705
	60%	0.1670	0.0665	0.0685	0.0715	0.0710	0.5520	0.7870	0.7615	0.7775	0.3655	0.2540	0.3215	0.3480
(100,100)	0%	0.5855	0.1050	0.1250	0.1050	0.1090	0.9005	1.0000	1.0000	1.0000	0.9590	0.9295	0.9585	0.9370
	20%	0.4880	0.0875	0.0925	0.0620	0.0610	0.9380	1.0000	0.9995	1.0000	0.9140	0.8285	0.8975	0.8735
	40%	0.3755	0.0715	0.0940	0.0665	0.0655	0.9185	1.0000	0.9980	0.9995	0.8110	0.6850	0.7895	0.7510
	60%	0.2875	0.0650	0.0810	0.1005	0.0970	0.8435	0.9700	0.9605	0.9665	0.6535	0.5010	0.6005	0.5935
(40,80)	0%	0.2385	0.0435	0.0510	0.0435	0.0455	0.6540	0.9960	0.9940	0.9940	0.6545	0.1040	0.2755	0.6740
	20%	0.2145	0.0420	0.0500	0.0335	0.0350	0.6715	0.9775	0.9720	0.9735	0.5450	0.0870	0.2415	0.5840
	40%	0.1730	0.0435	0.0565	0.0465	0.0460	0.6465	0.9305	0.9270	0.9290	0.4390	0.0865	0.2020	0.4705
	60%	0.1495	0.0340	0.0500	0.0630	0.0620	0.5595	0.8170	0.8305	0.8250	0.3415	0.0735	0.1720	0.3655
(50,100)	0%	0.3245	0.0430	0.0535	0.0430	0.0450	0.7600	1.0000	1.0000	0.9995	0.7515	0.1000	0.3470	0.7745
	20%	0.2695	0.0495	0.0640	0.0355	0.0345	0.7740	0.9945	0.9935	0.9945	0.6485	0.1095	0.3100	0.6795
	40%	0.1985	0.0395	0.0495	0.0390	0.0380	0.7580	0.9750	0.9705	0.9740	0.5335	0.0755	0.2290	0.5575
	60%	0.1755	0.0440	0.0630	0.0745	0.0725	0.6430	0.8765	0.8870	0.8850	0.4090	0.0755	0.1905	0.4295

Notations are presented as follows: n_1, n_2: sample sizes of two groups; c: censoring rate; LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang; LW_{w1}; LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{\mathrm{Var}(d_{1i})}$

TABLE 5: Estimated power results of comparison tests for Scenario 4.

(n_1, n_2)	c	LR	GW	TW	PP	MPP	FH	SUP-LR	SUP-GW	SUP-PP	LW	LW_{w1}	LW_{w2}	LW_{w3}
(50,50)	0%	0.0650	0.7925	0.4395	0.7925	0.8030	0.1340	1.0000	1.0000	1.0000	0.6065	0.6710	0.6090	0.6050
	20%	0.0665	0.6920	0.3575	0.5765	0.5850	0.2135	0.9980	0.9975	0.9970	0.5100	0.4940	0.4560	0.5055
	40%	0.0800	0.5720	0.2900	0.3540	0.3665	0.2860	0.9920	0.9840	0.9845	0.3895	0.3755	0.3500	0.4030
	60%	0.0705	0.4320	0.2280	0.1870	0.1975	0.3240	0.9275	0.9060	0.9135	0.2870	0.2650	0.2530	0.2975
(100,100)	0%	0.0830	0.9665	0.6975	0.9665	0.9685	0.1795	1.0000	1.0000	1.0000	0.9910	0.9995	0.9985	0.9850
	20%	0.0695	0.9315	0.6070	0.8515	0.8610	0.3210	1.0000	1.0000	1.0000	0.9200	0.9470	0.9255	0.9045
	40%	0.0660	0.8545	0.4995	0.6070	0.6175	0.4950	1.0000	1.0000	1.0000	0.7670	0.7735	0.7460	0.7620
	60%	0.0795	0.7015	0.3745	0.3130	0.3195	0.5375	0.9990	0.9985	0.9990	0.5900	0.5500	0.5450	0.5880
(40,80)	0%	0.0400	0.8225	0.4445	0.8225	0.8310	0.0610	1.0000	1.0000	1.0000	0.1640	0.6850	0.2355	0.2840
	20%	0.0355	0.7195	0.3650	0.5990	0.6095	0.1265	1.0000	0.9995	1.0000	0.1670	0.5910	0.1745	0.2660
	40%	0.0520	0.5800	0.2570	0.3375	0.3495	0.2105	0.9945	0.9920	0.9940	0.1415	0.4605	0.1500	0.2335
	60%	0.0405	0.3875	0.1855	0.1485	0.1570	0.2745	0.9385	0.9390	0.9375	0.1415	0.2930	0.1015	0.2105
(50,100)	0%	0.0330	0.9110	0.5390	0.9110	0.9135	0.0615	1.0000	1.0000	1.0000	0.2190	0.7700	0.2380	0.3945
	20%	0.0455	0.8230	0.4455	0.7085	0.7175	0.1345	0.9995	0.9995	0.9995	0.1985	0.6865	0.1985	0.3345
	40%	0.0535	0.6750	0.3335	0.4145	0.4235	0.2475	0.9995	0.9975	0.9975	0.1705	0.5430	0.1715	0.2785
	60%	0.0595	0.5050	0.2260	0.1895	0.1940	0.3180	0.9680	0.9675	0.9680	0.1550	0.3835	0.1235	0.2285

Notations are presented as follows: n_1, n_2: sample sizes of two groups; c: censoring rate; LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang; LW_{w1}: LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{\mathrm{Var}(d_{1i})}$

TABLE 6: Estimated Type I error of comparison tests.

(n_1, n_2)	c	LR	GW	TW	PP	MPP	FH	SUP-LR	SUP-GW	SUP-PP	LW	LW_{w1}	LW_{w2}	LW_{w3}
(50,50)	0%	0.0590	0.0600	0.0545	0.0600	0.0600	0.0580	0.0615	0.0560	0.0550	0.0435	0.0560	0.0555	0.0470
	20%	0.0550	0.0535	0.0485	0.0525	0.0530	0.0630	0.0580	0.0530	0.0530	0.0510	0.0515	0.0525	0.0485
	40%	0.0550	0.0535	0.0510	0.0505	0.0510	0.0595	0.0635	0.0580	0.0560	0.0465	0.0480	0.0495	0.0455
	60%	0.0480	0.0485	0.0510	0.0500	0.0505	0.0475	0.0590	0.0535	0.0560	0.0445	0.0490	0.0470	0.0440
(100,100)	0%	0.0505	0.0470	0.0525	0.0470	0.0470	0.0510	0.0645	0.0640	0.0620	0.0450	0.0465	0.0465	0.0425
	20%	0.0480	0.0480	0.0465	0.0455	0.0460	0.0505	0.0590	0.0535	0.0530	0.0570	0.0560	0.0560	0.0545
	40%	0.0525	0.0595	0.0540	0.0535	0.0535	0.0555	0.0585	0.0605	0.0575	0.0430	0.0480	0.0505	0.0390
	60%	0.0555	0.0585	0.0565	0.0570	0.0575	0.0475	0.0655	0.0525	0.0565	0.0415	0.0480	0.0490	0.0400
(40,80)	0%	0.0550	0.0485	0.0510	0.0485	0.0485	0.0585	0.0655	0.0655	0.0655	0.0555	0.0470	0.0525	0.0540
	20%	0.0555	0.0480	0.0525	0.0485	0.0490	0.0460	0.0755	0.0685	0.0690	0.0515	0.0525	0.0545	0.0475
	40%	0.0550	0.0530	0.0515	0.0530	0.0530	0.0490	0.0815	0.0755	0.0775	0.0565	0.0540	0.0550	0.0535
	60%	0.0450	0.0545	0.0505	0.0470	0.0490	0.0555	0.0690	0.0610	0.0655	0.0455	0.0460	0.0455	0.0435
(50,100)	0%	0.0555	0.0515	0.0530	0.0515	0.0505	0.0555	0.0745	0.0710	0.0685	0.0585	0.0470	0.0515	0.0560
	20%	0.0605	0.0550	0.0565	0.0545	0.0560	0.0525	0.0690	0.0640	0.0645	0.0595	0.0500	0.0560	0.0560
	40%	0.0505	0.0520	0.0515	0.0505	0.0510	0.0545	0.0720	0.0630	0.0635	0.0490	0.0510	0.0510	0.0480
	60%	0.0530	0.0530	0.0505	0.0510	0.0510	0.0585	0.0710	0.0660	0.0660	0.0515	0.0485	0.0510	0.0515

Notations are presented as follows: n_1, n_2: sample sizes of two groups; c: censoring rate; LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang;
LW_{w1}; LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{\mathrm{Var}(d_{1i})}$

REAL DATA EXAMPLE

Real data set belongs to the study, which was planned by Schein and supported by NIH-NIC N01-CM-67094, about gastric cancer patients whose tumors could not be removed locally by surgical operation. Randomly allocated 45 people in both groups were followed for eight years. The survival functions of patients with locally advanced gastric cancer who received chemotherapy alone (5 - FU and methyl - CCNU) or 5000 rad radiotherapy with the same chemotherapy were compared. The follow-up periods of the patients were recorded in days. The event of interest was determined as death. Further information can be obtained through the original study.[34]

As a result of the follow-up, the censoring rate was 4.44% in the chemotherapy group and 13.33% in the chemotherapy combined with the radiotherapy group. The data set can be found in the publication of Stablein and Koutrouvelis or from the book of Klein and Moeschberger.[35,36] The survival functions of the treatment groups are shown in Figure 2. The survival functions cross each other at the point S(t)=0.2, which is quite similar to Scenario 4 in the simulation studies. Also, the assumption of proportional hazards was evaluated with Grambsch and Therneau's goodness of fit test, which searches the relationship between Schoenfeld residuals and rank failure times.[37] According to the test result, the assumption of proportional hazards was violated ($\chi^2 = 13.1,\ p < 0.001$).

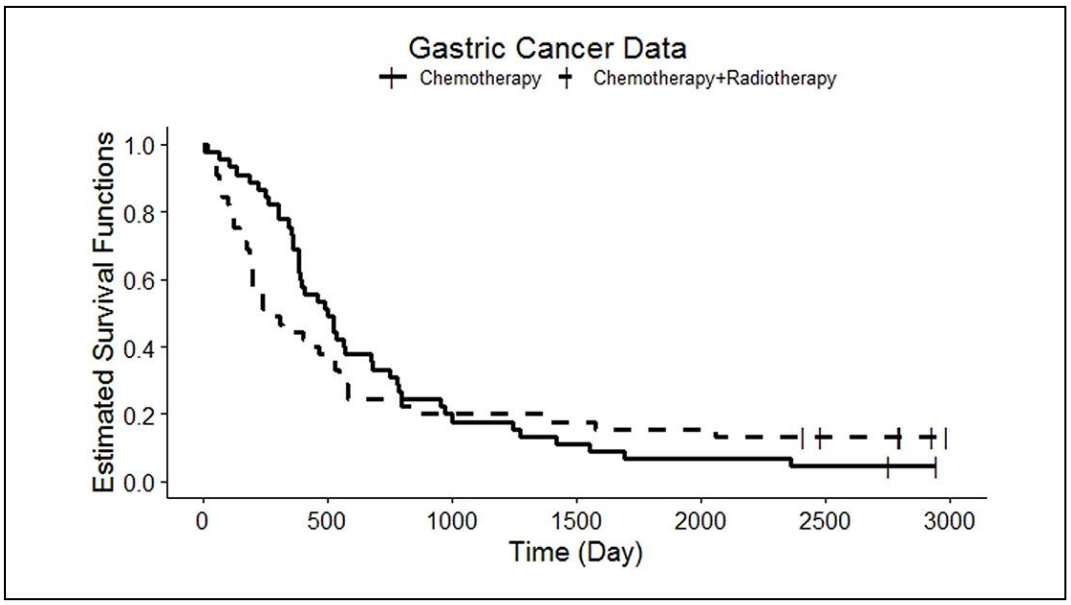

FIGURE 2: Kaplan-Meier estimated survival functions of the gastric cancer data.

The results of the comparison tests for gastric cancer patients are given in Table 7. The partitioned log-rank tests, GW, PP and MPP, which resulted in high power results in the corresponding simulation scenario, provided significant results as expected. LW, TW, FH and weighted LW tests failed to detect the difference. The lowest p-values were obtained by partitioned log-rank tests. It was observed that chemotherapy combined with the radiotherapy group provided a better prognosis for the late follow-up period.

TABLE 7: Results of comparison tests for gastric cancer data set.

Comparison Tests	Test Statistics	p value
LR	0.2319	0.6301
GW	3.9963	**0.0456**
TW	1.9269	0.1651
PP	4.0299	**0.0447**
MPP	4.1192	**0.0424**
FH	0.0111	0.9160
SUP-LR	17.2668	**0.0030**
SUP-GW	15.2706	**0.0040**
SUP-PP	15.2932	**0.0020**
LW	1.0855	0.2778
LW_{W1}	1.3310	0.1832
LW_{W2}	1.1853	0.2359
LW_{W3}	1.1340	0.2568

LR: Log-rank; GW: Gehan-Wilcoxon; TW: Tarone-Ware; PP: Peto-Peto; MPP: Modified Peto-Peto; FH: Fleming-Harrington(1,1); SUP-LR: Partitioned log-rank; SUP-GW: Partitioned Gehan-Wilcoxon; SUP-PP: Partitioned Peto-Peto; LW: Lin and Wang.
LW_{w1}: LW test weighted by n_i; LW_{w2}: LW test weighted by $\sqrt{n_i}$; LW_{w3}: LW test weighted by $1/\sqrt{\mathrm{Var}(d_{1i})}$

DISCUSSION

The crossing survival functions take place in cases where the proportional hazards assumption is violated. In this study, we aimed to suggest a comparison test suitable for various crossing survival function scenarios. The recently proposed partitioned log-rank tests and weighted LW tests were compared with each other and weighted LR tests. Comparisons were performed under certain scenarios with various sample sizes and censoring rates. Under the assumption of proportional hazards, the best results were provided by the LR test as expected. In the case of early, middle and late crossing survival functions, the partitioned log-rank tests exhibited superior power results. It was observed that the crossing point had a significant effect on the power of the comparison tests. The increase in the sample size positively affected the performance of the tests, but the increase in the censoring rate had a negative effect. In addition, the Type I error rates of the comparison tests were close to the nominal value of 0.05.

For the early crossing survival functions scenario, partitioned log-rank tests and weighted LW tests, which were developed for the crossing survival functions comparisons, presented powers higher than 90%. However, contrary to expectations, the LR test also showed high powers. Similar simulation results were found in the studies of Tubert-Bitter et. al., Li et. al. and Hsieh and Chen.[19,20,38]

Situations, where the middle or late crossing survival functions scenarios were used, LW tests could only provide powers higher than 80% when the sample size was equal to 100 with a low censoring rate. Even they are proposed for crossing survival functions, they could maintain high results for large sample sizes with low censoring rates. These results are supported by Koziol and Jia.[18] Moreover, LW test was observed to be affected by the increase in censoring rate in middle and late crossing scenarios, which is consistent with the results of Li et al.[20]

In Scenario 4, LR and FH tests gave the lowest power. GW, TW, PP and MPP tests, which were focused on the initial period of the study, provided higher powers with respect to the previous crossing scenar-

ios. Recently, other studies in the literature, which simulates the late crossing structure, have shown that LR and FH tests have lower powers than GW and TW tests.[20,38]

In scenarios where the proportional hazards assumption has been violated, assessments have been made considering different crossing points. The location of the crossing point was obtained to play an important role in the powers of the comparison tests. The LR was found to be the most affected method by the left-to-right movement of the crossing point. Weighted LR and LW tests were affected by the crossing point as well. The partitioned log-rank tests were marginally affected by this change. This is thought to be the result of the algorithm used in the calculation of test statistics. Partitioned log-rank tests make the calculations by evaluating the before and after of each time points where the event of interest is observed. In their study, Liu and Yin stated that if two survival functions cross at one of the time points of events, the powers of the tests will increase.[2]

One of the limitations of partitioned log-rank tests was that they could not maintain results during simulation runs when the sample sizes of both groups were less than 50 with a high censoring rate of 40% or above. Relevant weighted LR test statistics couldn't be calculated when drawn bootstrap samples contained censored observations with no event. The p-value calculation algorithm could not be manipulated due to the random selection necessity of the bootstrap technique. Therefore, the simulation runs stopped and could not be completed. Because of this problem, the smallest sample size for a balanced group design was chosen as 50 in the simulation studies. The determination of the maximum censoring rate limits for certain small sample sizes where the partitioned log-rank tests can provide results can be a future study subject.

A notable feature that distinguishes partitioned log-rank tests from the others is the use of the bootstrap technique during p-value calculations. A two-group comparison result with partitioned log-rank tests can present different p-values at different times. Although these obtained p-values of the same data set are different, they are mostly very close to each other.

CONCLUSION

To sum up, researchers that are comparing two survival functions should be aware that the implementation of the relevant comparison test bears some basic assumptions. The selection of the most appropriate test depends on the sample size, the censoring rate and the state of survival or hazard functions. Only one test does not fit equally in all situations. It is recommended to use partitioned log-rank tests for comparison of two crossing survival functions taking into consideration sample sizes and censoring rate.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Hülya Özen, Cengiz Bal, Ertuğrul Çolak;* ***Design:*** *Hülya Özen, Ertuğrul Çolak;* ***Control/Supervision:*** *Hülya Özen, Cengiz Bal, Ertuğrul Çolak;* ***Data Collection and/or Processing:*** *Hülya Özen;* ***Analysis and/or Interpretation:*** *Hülya Özen, Cengiz Bal, Ertuğrul Çolak;* ***Literature Review:*** *Hülya Özen;* ***Writing the Article:*** *Hülya Özen, Ertuğrul Çolak.*

REFERENCES

1. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966;50(3):163-70. [PubMed]
2. Liu Y, Yin G. Partitioned log-rank tests for the overall homogeneity of hazard rate functions. Lifetime Data Anal. 2017;23(3):400-25. [Crossref] [PubMed]
3. Lee CT. Statistical Methods for the Comparison of Crossing Survival Curves. In: Balakrishnan N, Rao CR, eds. Advances in survival analysis. Vol 23. 1st ed. Amsterdam: Elsevier; 2004. p. 277-89. [Crossref]
4. Krishnan A, Pasquini MC, Logan B, Stadtmauer EA, Vesole DH, Alyea E 3rd, et al; Blood Marrow Transplant Clinical Trials Network (BMT CTN). Autologous haemopoietic stem-cell transplantation followed by allogeneic or autologous haemopoietic stem-cell transplantation in patients with multiple myeloma (BMT CTN 0102): a phase 3 biological assignment trial. Lancet Oncol. 2011;12(13):1195-203. [Crossref] [PubMed] [PMC]
5. Park JH, Rivière I, Gonen M, Wang X, Sénéchal B, Curran KJ, et al. Long-Term Follow-up of CD19 CAR Therapy in Acute Lymphoblastic Leukemia. N Engl J Med. 2018;378(5):449-59. [Crossref] [PubMed] [PMC]
6. Weeda VB, Murawski M, McCabe AJ, Maibach R, Brugières L, Roebuck D, et al. Fibrolamellar variant of hepatocellular carcinoma does not have a better survival than conventional hepatocellular carcinoma--results and treatment recommendations from the Childhood Liver Tumour Strategy Group (SIOPEL) experience. Eur J Cancer. 2013;49(12):2698-704. [Crossref] [PubMed]
7. Kristiansen IS. PRM39 Survival curve convergences and crossing: a threat to validity of meta-analysis? Value in health. 2012;15(7):A652 [Crossref]
8. Suciu GP, Lemeshow S, Moeschberger M. Statistical Tests of the Equality of Survival Curves: Reconsidering the Options. In: Balakrishnan N, Rao CR, eds. Advances in survival analysis. Vol 23. 1st ed. Amsterdam: Elsevier; 2004. p. 251-61. [Crossref]
9. Koziol JA. Two sample cramér-von mises test for randomly censored data. Biometrical Journal. 1978;20(6):603-8. [Crossref]
10. Schumacher M. Two-Sample Tests of Cramér-von Mises and Kolmogorov-Smirnov-Type for Randomly Censored Data. Int Stat Rev. 1984;263-81. [Crossref]
11. Gill RD. Censoring and stochastic integrals. Stat Neerl. 1980;34(2):124. [Crossref]
12. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. Biometrics 1980;(36)4:607-625. [Crossref]
13. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics. 1989;45(2):497-507. [Crossref] [PubMed]
14. Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. Biometrical Journal. 2004;46(5):489-96. [Crossref]
15. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. J R Stat Soc Series B Stat Methodol. 2008;70(1):191-208. [Link]
16. Kraus D. Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. J Stat Plan Infer. 2009;139(10):3559-69. [Crossref]
17. Lin X, Xu Q. A new method for the comparison of survival distributions. Pharm Stat. 2010;9(1):67-76. [Crossref] [PubMed]
18. Koziol JA, Jia Z. Weighted Lin-Wang tests for crossing hazards. Comput Math Methods Med. 2014;2014:643457. [Crossref] [PubMed] [PMC]
19. Tubert-Bitter P, Kramar A, Chale JJ, Moreau T. Linear rank tests for comparing survival in two groups with crossing hazards. Comput Stat Data Anal. 1994;18(5):547-59. [Crossref]
20. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One. 2015;10(1):e0116774. [Crossref] [PubMed] [PMC]
21. Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. Stat Med. 2007;26(2):375-91. [Crossref] [PubMed]
22. Gehan EA. A Generalized wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965;52:203-23. [Crossref] [PubMed]
23. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. Biometrika. 1977;64(1):156-160. [Crossref]
24. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J R Stat Soc Ser A Stat Soc. 1972;135(2):185-207. [Crossref]
25. Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes: Nonparametric Hypothesis Testing. 1st ed. New York: Springer-Verlag; 1993. p. 393-5. [Crossref]
26. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study Ann Stat. 1982;10(4):1100-20. [Crossref]
27. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982;69(3):553-66. [Crossref]
28. R Studio Team. R Studio: Integrated Development for R. RStudio. PBC, Boston, MA. 2019. [Link]
29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. [Link]
30. Kassambara A, Kosinski M, Biecek P, Fabian S. Package 'survminer'. 2017. [Link]
31. Therneau TM, Lumley T. Package 'survival'. 2014. [Link]
32. Wickham H, Wickham MH. Package 'dplyr'. 2013. [Link]
33. Dardis C, Dardis MC. Package 'survMisc'. 2018. [Link]
34. A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Gastrointestinal Tumor Study Group. Cancer. 1982;49(9):1771-7. [Crossref] [PubMed]

35. Stablein DM, Koutrouvelis IA. A two-sample test sensitive to crossing hazards in uncensored and singly censored data. Biometrics. 1985;41(3):643-52. [Crossref] [PubMed]

36. Klein JP, Moeschberger ML. Hypothesis Testing. Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer-Verlag; 2003. p. 224-6. [Crossref]

37. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81(3):515-26. [Crossref]

38. Hsieh JJ, Chen HY. A testing strategy for two crossing survival curves. Commun Stat Simul Comput. 2017;46(8):6685-96. [Crossref]