Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(10): 17646–17660.
DOI: 10.3934/mbe.2023784
Received: 04 July 2023
Revised: 28 August 2023
Accepted: 30 August 2023
Published: 15 September 2023

Research article

Comparison of methods to testing for differential treatment effect under non-proportional hazards data

María del Carmen Pardo [1,2] **and Beatriz Cobo** [3,*]

[1] Department of Statistics and O.R., Complutense University of Madrid, Plaza de Ciencias 3, Madrid 28040, Spain.

[2] Institute of Interdisciplinary Mathematics, Complutense University of Madrid, Plaza de Ciencias 3, Madrid 28040, Spain.

[3] Department of Quantitative Methods for Economics and Business, University of Granada, Paseo de Cartuja 7, Granada 18011, Spain.

* **Correspondence:** Email: beacr@ugr.es; Tel: +34958244111.

Abstract: Many tests for comparing survival curves have been proposed over the last decades. There are two branches, one based on weighted log-rank statistics and other based on weighted Kaplan-Meier statistics. If we carefully choose the weight function, a substantial increase in power of tests against non-proportional alternatives can be obtained. However, it is difficult to specify in advance the types of survival differences that may actually exist between two groups. Therefore, a combination test can simultaneously detect equally weighted, early, late or middle departures from the null hypothesis and can robustly handle several non-proportional hazard types with no a priori knowledge of the hazard functions. In this paper, we focus on the most used and the most powerful test statistics related to these two branches which have been studied separately but not compared between them. Through a simulation study, we compare the size and power of thirteen test statistics under proportional hazards and different types of non-proportional hazards patterns. We illustrate the procedures using data from a clinical trial of bone marrow transplant patients with leukemia.

Keywords: Survival analysis; non-proportional hazards; weighted log-rank statistics; weighted Kaplan-Meier statistics; power

1. Introduction

The most commonly used two-sample nonparametric test statistic is the log-rank (LR) statistic, which is locally most powerful against proportional hazards (PH) alternatives. However, rank-based

log-rank statistics may not be sensitive to certain types of stochastic ordering alternatives, particularly if the hazard functions cross or a delayed effect is observed. In reality, the PH does often not hold true as for example in some immuno-oncology therapies [1]. As an alternative to generalized linear rank statistics, [2] proposed nonparametric weighted Kaplan-Meier (WKM) statistics, which are non-rank-based statistics and therefore sensitive to the magnitude and duration of the difference in survival curves over time. These test statistics seems to compare favorably with the popular log-rank test statistic across a broad range of stochastic ordering alternatives.

As a particular case of WKM, you get the restricted mean survival time (RMST) which is an alternative robust and clinically interpretable summary measure that does not rely on the proportional hazard (PH) assumption. [3] concluded that the RMST-based test has better performance than the log-rank test when the truncation time is reasonably close to the tail of the observed curves under non-PH scenarios where late separation of survival curves is observed.

Under non-proportional alternatives, the effect of a treatment may wane over time, leading to a decreasing hazard ratio and a closing up of the two survival curves or may have a delayed effect whereby they do not separate until a certain interval of time has elapsed. As another more powerful alternative over the unweighted log-rank statistic, [4] proposed a class of adaptive weighted log-rank statistics. These tests may assign more weight on either early, middle or late survival differences with a suitable choice of weight function. A particularly useful family in the class of weighted log-rank statistics was introduced in [5] and [6]. However, in many situations, it is difficult to know in advance the kind of survival differences that will actually occur. Therefore, [7] proposed the maximum of a finite collection of these weighted log-rank statistics trying to overcome the above problem.

Later, [8, 9] proposed tests based on a more extended family of test statistics where the weights are governed by two parameters and combinations of them that are more sensitive to nonproportional hazard alternatives. Furthermore, [10] considered a new combination of these test statistics. On the other hand, [11] considered the same family of weighted functions for the WKM test statistic and evaluated the maximum of nine of these test statistics. [12] proposed a different weighted functions for these test statistics. Recently, [13] investigated weighted log-rank tests, the supremum log-rank test and composite tests (mainly based on log-rank tests). On the other hand, [14] studied the performance of nine tests. Seven out of nine are based on weighted log-rank and the other two are WKM and RMST. Both papers came to the conclusion that: there is not a single most powerful test across all scenarios but the maximum of several weighted log-rank tests seems to be the best choice. The former proposed the so-called MaxCombo test and the later the proposal studied by [10]. Recently, [15] have performed additional simulations to evaluate the MaxCombo test under some extreme scenarios which are unlikely in real life. Futhermore, they have provided design and analysis considerations based on a combination test under different non-PH types and present a straw man proposal for practitioners.

To our knowledge, the class of maximum weighted log-rank tests and that of maximum weighted Kaplan-Meier tests have not been compared to each other. Therefore, we try to fill this gap as well as to compare them with the RMST which is easily interpretable. The paper is organized as follows: In Section 2, we describe the tests to be compared and our approach to get the critical points. A simulation study of the performance of the tests is presented in Section 3. In Section 4, we apply all the methods to a real data set from a clinical trial of marrow transplant patients with leukemia. We finish with a discussion in Section 5.

2. Methods

We assume the two-sample general random censorship model. Consider two censored samples with sample sizes n_1 from the treatment group and n_2 from the control group. Let T_{ij}, $i = 1, 2$, $j = 1, ..., n_i$, be independent, positive random variables and C_{ij} be independent censoring variables that are also independent of the survival variables T_{ij}. The data really observed are $\left(X_{ij}, \delta_{ij}\right)$, $i = 1, 2$, $j = 1, ..., n_i$, where $X_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I\left(T_{ij} \leq C_{ij}\right)$, where I is the indicator function. Define survival functions $S_i(t) = P(T_{ij} \geq t)$ and $C_i(t) = P(C_{ij} \geq t), i = 1, 2$, which are assumed to be absolutely continuous throughout the paper. The null hypothesis to test is that the survival distributions between two groups are the same, that is,

$$H_0 : S_1(t) = S_2(t) \tag{2.1}$$

2.1. Weigthed Log-rank tests and combinations

The so-called weighted log-rank statistics are given by

$$WLR = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \int_0^{t_c} \widehat{w}(t) \frac{\overline{Y}_1(t)\overline{Y}_2(t)}{\overline{Y}_1(t) + \overline{Y}_2(t)} \left\{ \frac{d\overline{N}_1(t)}{\overline{Y}_1(t)} - \frac{d\overline{N}_2(t)}{\overline{Y}_2(t)} \right\} \tag{2.2}$$

where $\overline{N}_i(t)$ is the number of failures in group i before or at time t and $\overline{Y}_i(t)$ the number at risk in group i at time $t-$, $i = 1, 2$, $\widehat{w}(t)$ is a bounded nonnegative weight function and $t_c = \sup\{t/\hat{C}_1(t) \wedge \hat{C}_2(t) > 0\}$ where $\widehat{C}_i$ denotes the Kaplan-Meier estimator of the censoring survival function in group i. As the weight function is sensitive to the alternative hypothesis, it should be chosen carefully.

[7] proposed

$$\widehat{w}(t) = \left\{\widehat{S}(t-)\right\}^{\lambda} \left\{1 - \widehat{S}(t-)\right\}^{\gamma} \text{ for } \lambda \geq 0 \text{ and } \gamma \geq 0 \tag{2.3}$$

with $\widehat{S}(t-)$ being the left-continuous version of the product-limit estimator ([16]) for the survival function based on the pooled survival data. We consider the family of test statistic given in (2.2) when the class of functions (2.3) are chosen, $\text{WLR}^{\lambda,\gamma}$. In particular, $\lambda = \gamma = 0$ corresponds to the log-rank statistic, $WLR^{0,0}$, which specifies equal weights and $\lambda = 1$ and $\gamma = 0$ corresponds to the Prentice-Wilcoxon statistic, $WLR^{1,0}$, which places more weight on the earlier time points. In contrast, $WLR^{0,1}$ places more weight on the later time points and $WLR^{1,1}$ will emphasize the middle differences. [8] studied the maximum over the statistics $WLR^{0,0}$, $WLR^{2,0}$, $WLR^{0,2}$ and $WLR^{2,2}$ as well as their average. [9] studied the test statistics $\left|WLR^{1,0} + WLR^{0,1}\right|/2$, $\left(\left|WLR^{1,0}\right| + \left|WLR^{0,1}\right|\right)/2$ and $\max\left(\left|WLR^{1,0}\right|, \left|WLR^{0,1}\right|\right)$. [10] and [13] considered $WLR_{max3} = \max\left(\left|WLR^{0,0}\right|, \left|WLR^{1,0}\right|, \left|WLR^{0,1}\right|\right)$. [14] focused on the so called MaxCombo test $WLR_{max4} = \max(\left|WLR^{0,0}\right|, \left|WLR^{0,1}\right|, \left|WLR^{1,0}\right|, \left|WLR^{1,1}\right|)$.

According to the conclusions of the papers that have studied these test statistics and combinations, we focus on maximum combination tests. To conduct inference on these test statistics, i.e. to calculate the p-value for testing (2.1), there are different methods. One is to obtain the asymptotic distribution of these test statistics as the maximum of a multivariate normal distribution with mean zero and covariance that can be consistently estimated by

$$\widehat{\sigma}_{ij} = \frac{n_1 + n_2}{n_1 n_2} \int_0^{t_c} \hat{w}_i(t)\hat{w}_j(t) \frac{\overline{Y}_1(t)\overline{Y}_2(t)}{\overline{Y}_1(t) + \overline{Y}_2(t)} \left(1 - \frac{\Delta\overline{N}_1(t) + \Delta\overline{N}_2(t) - 1}{\overline{Y}_1(t) + \overline{Y}_2(t) - 1}\right) \left\{ \frac{d\left(\overline{N}_1(t) + \overline{N}_2(t)\right)}{\overline{Y}_1(t) + \overline{Y}_2(t)} \right\}$$

for $i, j = 1, 2, 3, 4$, $\hat{w}_i(t), i = 1, 2, 3, 4$ are the weight functions with $\lambda = \gamma = 0; \lambda = 1, \gamma = 0; \lambda = 0, \gamma = 1$ and $\lambda = \gamma = 1$, respectively and $\Delta\overline{N}_i(t) = \overline{N}_i(t) - \overline{N}_i(t-)$, the number of events at time t in group i. However, in this paper, to account for small and moderate sample sizes, bootstrap is used to obtain p-values as is explained in Section 2.3.

2.2. Weigthed Kaplan-Meier tests and combinations

[2] defined the weighted Kaplan-Meier statistics as

$$WKM = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \int_0^{t_c} \widehat{u}(t) \left(\widehat{S}_1(t) - \widehat{S}_2(t)\right) dt.$$

They proposed using $\widehat{u}(t) = \widehat{C}_1(t-)\widehat{C}_2(t-)/(p_1\widehat{C}_1(t-) + p_2\widehat{C}_2(t-))$ to stabilize the test statistic and down weight its variance toward the end of the observation period if censoring is heavy.

[11] proposed

$$\widehat{u}(t) = \widehat{w}(t)\frac{\widehat{C}_1(t-)\widehat{C}_2(t-)}{p_1\widehat{C}_1(t-) + p_2\widehat{C}_2(t-)} \text{ for } \lambda \geq 0 \text{ and } \gamma \geq 0 \tag{2.4}$$

with $p_i = n_i/(n_1 + n_2)$ and $\widehat{C}_i$ denotes the Kaplan-Meier estimator of the censoring random variables in group i and $\hat{w}(t)$ is defined in (2.3). We shall call these test statistics $WKM^{\lambda,\gamma}$. In particular, $\lambda = \gamma = 0$, $WKM^{0,0}$, corresponds to the Pepe-Fleming's WKM test. [11] studied $\max_{i,j=0,1,2}\left(WKM^{i,j}\right)$.

As a particular case, we have the difference in RMST ([3])

$$D = \int_0^{t_c} \left(\widehat{S}_1(t) - \widehat{S}_2(t)\right) dt.$$

To be fair the comparison of these test statistics with that based on the log-rank test, we will focus on their maximum combination tests counterparts. In particular, $WKM_{max4} = \max(WKM^{0,0}, WKM^{0,1}, WKM^{1,0}, WKM^{1,1})$ and $WKM_{max3} = \max(WKM^{0,0}, WKM^{0,1}, WKM^{1,0})$. To conduct inference on these test statistics, i.e. to calculate the p-value for testing (2.1), the asymptotic distribution of these test statistics can be obtained as the maximum of a multivariate normal distribution with mean zero and covariance that can be consistently estimated by

$$\widehat{\sigma}_{ij} = -\int_0^{t_c} \frac{\left[\int_t^{t_c} \hat{u}_i(s)S(s)ds\right]\left[\int_t^{t_c} \hat{u}_j(s)S(s)ds\right]}{\widehat{S}(t)\widehat{S}(t-)} \frac{p_1\widehat{C}_1(t-) + p_2\widehat{C}_2(t-)}{\widehat{C}_1(t-)\widehat{C}_2(t-)} d\widehat{S}(t)$$

for $i, j = 1, 2, 3, 4$, and $\hat{u}_i(t), i = 1, 2, 3, 4$ are the weight functions given in (2.4) for $\hat{w}_i(t)$ defined in the previous section. However, in this paper, to account for small and moderate sample sizes, bootstrap is used to obtain p-values as is explained in Section 2.3.

2.3. Computing p-values using bootstrap

The p-value of our test statistics can be obtained using their asymptotic distributions either by numerical integration or Monte Carlo estimation of the multivariate Gaussian distribution. [11] pointed out the variance-covariance matrix may not have a close form solution and requires intense computation. Therefore, our proposal is to use Bootstrap instead of asymptotic distributions that has

less computational cost and improve the precision of the asymptotic approximations in small and moderate samples as well as deal with analytically challenging problems in some cases.

We propose to calculate the p-values as follows:

- Step 1: To obtain the value of the test statistic for the original sample, T_0
- Step 2: Draw B bootstrap samples from the original data, so $n_1 + n_2$ observations are obtained and consider the first n_1 as controls and the other n_2 as treatments then to obtain the value of the test statistic for the bootstrap sample, $T_B(i), i = 1, ..., B$.
- Step 3: Estimate the p-value by counting the number of $T_B(i)$ greater than T_0.

We also offer R-code facilitating the implementation of the evaluation of the p-value of the test statistics in Supplementary.

3. Simulation study

We carry out a simulation study to compare the performance of these thirteen test statistics: $WLR^{0,0}$, $WLR^{0,1}$, $WLR^{1,0}$, $WLR^{1,1}$, WLR_{max4}, WLR_{max3}, $WKM^{0,0}$, $WKM^{0,1}$, $WKM^{1,0}$, $WKM^{1,1}$, WKM_{max4}, WKM_{max3} and D. The simulation design described in [9] has been considered but it is similar to that included in the other cited references such as [17] and [13]. We used piecewise exponential models to generate simulated data with parameters $\lambda_i(t)$ from groups 1 and 2 set to represent common alternatives that might arise in real data. To be specific, in Figure 1, (a) null case $\lambda_1 = 2, \lambda_2 = 2$, (b) proportional hazards $\lambda_1 = 1, \lambda_2 = 2$, (c) early survival differences, $t < 0.3$, $\lambda_1 = 3, \lambda_2 = 0.75$; $0.3 \leq t < 0.6$, $\lambda_1 = 0.75, \lambda_2 = 3$; $t \geq 0.6$, $\lambda_1 = 1, \lambda_2 = 1$, (d) late survival differences, $t < 0.5$, $\lambda_1 = 2, \lambda_2 = 2$; $t \geq 0.5$, $\lambda_1 = 4, \lambda_2 = 0.4$ and (e) early and late occurring survival differences, $t < 0.2$, $\lambda_1 = 3, \lambda_2 = 0.75$; $0.2 \leq t < 0.4$, $\lambda_1 = 0.75, \lambda_2 = 3$; $0.4 \leq t < 0.6$, $\lambda_1 = 1, \lambda_2 = 1$; $t \geq 0.6$, $\lambda_1 = 3, \lambda_2 = 0.75$.

In each case, the censoring distributions are uniform and the censoring percentages in the sample are $0\%, 20\%, 40\%$ and 60%. Empirical size and power of the tests were evaluated considering $\alpha = 0.05$ for sample sizes $n_1 = n_2 = 20, 50, 70, 100$ and 150. For each sample size, 2000 replications were performed for each configuration of survival and censoring distributions and in each one of them 500 bootstrap repetitions were performed.

The simulations were conducted using R (https://www.r-project.org//R Core Team, Vienna, Austria), specifically, we simulate the data with the `rpwexp` function of the `nhpsim` package [18], which simulates the exponential distribution by parts and allows to specify a distribution where the risk rate changes with time, `logrank.test` and `logrank.maxtest` functions of the `nph` package [19] were used to calculate the statistics $WLR^{0,0}$, $WLR^{0,1}$, $WLR^{1,0}$, $WLR^{1,1}$, WLR_{max4}, WLR_{max3} and the `survtest` function of the `SurvBin` package [20] to calculate the statistics $WKM^{0,0}$, $WKM^{0,1}$, $WKM^{1,0}$, $WKM^{1,1}$ and we implement a function to calculate the maximum weighted Kaplan-Meier test statistics, WKM_{max4} and WKM_{max3}. Finally, we calculate the D statistic with the `rmst2` function of the `survRM2` package [21].

As can be seen in Table 1, the simulated type I error of the thirteen test statistics for all sample sizes are close to the nominal, two-sided 5% significance level. Therefore, we can assess their power performance.

The power of the thirteen test statistics for each simulation scenario is shown in Table 2. Figure

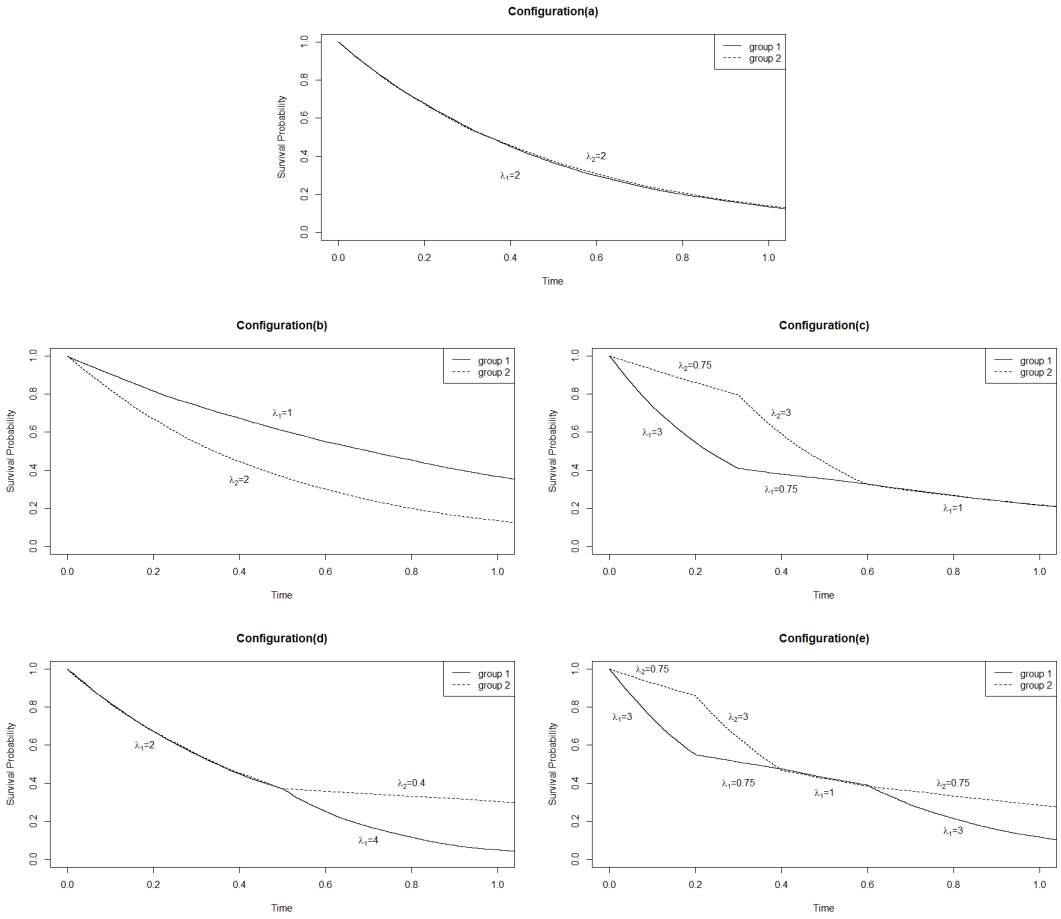

Figure 1. Survival functions considered in the size and power simulations.

Table 1. Simulated type I error.

$n = n_1 = n_2$	20				50				70				100				150			
censoring %	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%
Average number of treatment events	20	17.047	14.951	13.375	50	42.544	37.414	33.559	70	59.6735	52.5075	47.2100	100	85.173	74.849	67.163	150	127.5745	112.0740	100.4480
Average number of control events	20	18.2915	16.8435	15.5985	50	45.7340	42.1155	39.0165	70	63.9560	58.9985	54.7105	100	91.4575	84.3035	78.1045	150	137.117	126.2645	117.0625
Configuration (a) - null case																				
$WLR^{0,0}$	0.0475	0.0450	0.0435	0.0420	0.0525	0.0560	0.0515	0.0540	0.0515	0.0515	0.0495	0.0475	0.0425	0.0445	0.0500	0.0510	0.0470	0.0470	0.0485	0.0485
$WKM^{0,0}$	0.0495	0.0485	0.0495	0.0435	0.0505	0.0555	0.0545	0.0485	0.0535	0.0530	0.0515	0.0480	0.0435	0.0470	0.0490	0.0470	0.0480	0.0495	0.0475	0.0510
$WLR^{0,1}$	0.0525	0.0480	0.0490	0.0435	0.0515	0.0535	0.0535	0.0460	0.0505	0.0525	0.0485	0.0530	0.0435	0.0475	0.0515	0.0485	0.0440	0.0455	0.0430	0.0480
$WKM^{0,1}$	0.0475	0.0475	0.0525	0.0465	0.0520	0.0495	0.0510	0.0540	0.0485	0.0485	0.0505	0.0515	0.0445	0.0490	0.0485	0.0525	0.0460	0.0510	0.0470	0.0490
$WLR^{1,0}$	0.0450	0.0465	0.0475	0.0485	0.0470	0.0575	0.0575	0.0530	0.0435	0.0505	0.0470	0.0465	0.0500	0.0550	0.0490	0.0475	0.0485	0.0510	0.0520	0.0515
$WKM^{1,0}$	0.0460	0.0465	0.0445	0.0500	0.0520	0.0560	0.0575	0.0535	0.0455	0.0480	0.0475	0.0480	0.0530	0.0520	0.0505	0.0485	0.0495	0.0485	0.0510	0.0495
$WLR^{1,1}$	0.0520	0.0450	0.0480	0.0440	0.0560	0.0530	0.0565	0.0545	0.0485	0.0495	0.0450	0.0455	0.0480	0.0470	0.0490	0.0450	0.0445	0.0465	0.0480	0.0540
$WKM^{1,1}$	0.0475	0.0485	0.0510	0.0455	0.0545	0.0565	0.0555	0.0510	0.0505	0.0530	0.0550	0.0495	0.0460	0.0440	0.0435	0.0435	0.0465	0.0480	0.0460	0.0510
WLR_{max4}	0.0510	0.0505	0.0480	0.0450	0.0530	0.0525	0.0550	0.0495	0.0510	0.0530	0.0500	0.0540	0.0445	0.0400	0.0375	0.0450	0.0440	0.0510	0.0455	0.0515
WKM_{max4}	0.0490	0.0505	0.0530	0.0465	0.0500	0.0515	0.0505	0.0510	0.0515	0.0540	0.0495	0.0485	0.0450	0.0475	0.0440	0.0480	0.0435	0.0490	0.0480	0.0455
WLR_{max3}	0.0495	0.0485	0.0445	0.0435	0.0530	0.0530	0.0515	0.0490	0.0490	0.0530	0.0495	0.0545	0.0440	0.0385	0.0395	0.0430	0.0445	0.0480	0.0435	0.0495
WKM_{max3}	0.0490	0.0500	0.0525	0.0465	0.0490	0.0505	0.0485	0.0510	0.0500	0.0525	0.0500	0.0485	0.0455	0.0470	0.0450	0.0475	0.0460	0.0480	0.0475	0.0450
D	0.0495	0.0495	0.0490	0.0465	0.0505	0.0540	0.0480	0.0535	0.0535	0.0520	0.0550	0.0510	0.0435	0.0465	0.0520	0.0535	0.0480	0.0485	0.0505	0.0505

Table 2. Simulated power.

$n = n_1 = n_2$	20				50				70				100				150			
censoring %	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%
Configuration (b) - proportional hazards																				
$WLR^{0,0}$	0.5350	0.4770	0.4325	0.4080	0.9135	0.8720	0.8300	0.7895	0.9730	0.9565	0.9330	0.9140	0.9990	0.9980	0.9920	0.9840	1.0000	0.9995	0.9990	0.9985
$WKM^{0,0}$	0.5495	0.4930	0.4310	0.3980	0.9170	0.8655	0.8295	0.7860	0.9735	0.9570	0.9330	0.9100	0.9995	0.9985	0.9925	0.9855	1.0000	0.9995	0.9990	0.9980
$WLR^{0,1}$	0.4390	0.4000	0.3465	0.3260	0.8135	0.7615	0.7160	0.6680	0.9230	0.8970	0.8425	0.8015	0.9900	0.9740	0.9545	0.9225	0.9990	0.9975	0.9920	0.9885
$WKM^{0,1}$	0.5360	0.4810	0.4145	0.3680	0.9060	0.8560	0.8145	0.7700	0.9725	0.9555	0.9275	0.8920	0.9990	0.9955	0.9890	0.9760	1.0000	0.9995	0.9985	0.9970
$WLR^{1,0}$	0.4450	0.4170	0.3845	0.3605	0.8280	0.7915	0.7620	0.7285	0.9285	0.9135	0.8855	0.8540	0.9900	0.9790	0.9720	0.9575	0.9995	0.9990	0.9980	0.9960
$WKM^{1,0}$	0.4160	0.3765	0.3605	0.3325	0.8100	0.7750	0.7355	0.7155	0.9185	0.9000	0.8705	0.8410	0.9835	0.9755	0.9665	0.9530	0.9990	0.9980	0.9980	0.9950
$WLR^{1,1}$	0.4805	0.4370	0.3940	0.3695	0.8600	0.8180	0.7730	0.7330	0.9505	0.9270	0.9020	0.8660	0.9960	0.9880	0.9760	0.9615	1.0000	0.9990	0.9980	0.9955
$WKM^{1,1}$	0.4695	0.4350	0.3965	0.3670	0.8585	0.8270	0.7905	0.7530	0.9490	0.9340	0.9085	0.8860	0.9970	0.9920	0.9835	0.9765	0.9995	0.9995	0.9980	0.9975
WLR_{max4}	0.4930	0.4400	0.3890	0.3640	0.8730	0.8455	0.7965	0.7680	0.9645	0.9430	0.9170	0.8935	0.9970	0.9925	0.9875	0.9750	1.0000	0.9990	0.9990	0.9985
WKM_{max4}	0.5365	0.4810	0.4155	0.3725	0.9035	0.8580	0.8160	0.7745	0.9705	0.9540	0.9295	0.8985	0.9995	0.9960	0.9905	0.9810	1.0000	0.9995	0.9990	0.9990
WLR_{max3}	0.4940	0.4465	0.3900	0.3670	0.8765	0.8465	0.7945	0.7645	0.9655	0.9445	0.9220	0.8935	0.9975	0.9935	0.9860	0.9775	1.0000	0.9990	0.9985	0.9990
WKM_{max3}	0.5360	0.4825	0.4160	0.3705	0.9070	0.8585	0.8155	0.7740	0.9720	0.9530	0.9305	0.8990	0.9995	0.9960	0.9900	0.9810	1.0000	0.9995	0.9990	0.9990
D	0.5495	0.4800	0.4015	0.3445	0.9170	0.8445	0.7765	0.7095	0.9735	0.9445	0.8985	0.8380	0.9995	0.9955	0.9750	0.9410	1.0000	0.9995	0.9965	0.9860
Configuration (c) - early survival differences																				
$WLR^{0,0}$	0.1195	0.1260	0.1530	0.1595	0.2320	0.2545	0.3020	0.3285	0.2625	0.3145	0.3675	0.4225	0.3545	0.4120	0.4720	0.5390	0.4730	0.5565	0.6385	0.7040
$WKM^{0,0}$	0.0800	0.0955	0.1185	0.1380	0.1150	0.1490	0.1940	0.2415	0.1455	0.1890	0.2455	0.3075	0.1785	0.2380	0.3095	0.3840	0.2275	0.3305	0.4245	0.5130
$WLR^{0,1}$	0.0600	0.0575	0.0555	0.0570	0.0635	0.0640	0.0615	0.0635	0.0840	0.0785	0.0745	0.0685	0.0990	0.1000	0.0950	0.0885	0.1300	0.1315	0.1215	0.1045
$WKM^{0,1}$	0.0590	0.0645	0.0825	0.0825	0.0655	0.0780	0.1020	0.1225	0.0730	0.0940	0.1115	0.1300	0.0845	0.1060	0.1260	0.1545	0.0890	0.1265	0.1620	0.1995
$WLR^{1,0}$	0.3265	0.3410	0.3490	0.3620	0.6615	0.6765	0.6920	0.7135	0.8100	0.8165	0.8350	0.8515	0.9035	0.9205	0.9290	0.9355	0.9810	0.9860	0.9885	0.9890
$WKM^{1,0}$	0.2740	0.2930	0.3225	0.3500	0.5285	0.5720	0.6155	0.6585	0.6600	0.7115	0.7500	0.7870	0.7940	0.8385	0.8710	0.8970	0.9230	0.9495	0.9635	0.9740
$WLR^{1,1}$	0.0780	0.0810	0.0950	0.0945	0.0775	0.0910	0.1040	0.1155	0.0795	0.0930	0.1060	0.1260	0.0820	0.0945	0.1045	0.1335	0.0920	0.1085	0.1310	0.1620
$WKM^{1,1}$	0.1590	0.1790	0.2070	0.2260	0.2735	0.2930	0.3410	0.3790	0.3195	0.3715	0.4260	0.4770	0.3970	0.4525	0.5095	0.5650	0.5345	0.6030	0.6575	0.7150
WLR_{max4}	0.2270	0.2355	0.2485	0.2560	0.5560	0.5745	0.5950	0.6210	0.7320	0.7565	0.7640	0.7885	0.8655	0.8930	0.9030	0.9160	0.9785	0.9830	0.9830	0.9855
WKM_{max4}	0.1075	0.1250	0.1600	0.1770	0.3335	0.3870	0.4490	0.4925	0.4865	0.5570	0.6260	0.6625	0.6565	0.7255	0.7810	0.8245	0.8550	0.8975	0.9320	0.9485
WLR_{max3}	0.2385	0.2460	0.2565	0.2635	0.5730	0.5900	0.6120	0.6350	0.7475	0.7730	0.7810	0.8020	0.8790	0.8955	0.9120	0.9215	0.9800	0.9840	0.9860	0.9895
WKM_{max3}	0.1065	0.1255	0.1615	0.1775	0.3360	0.3895	0.4525	0.4940	0.4925	0.5630	0.6285	0.6660	0.6605	0.7310	0.7860	0.8270	0.8605	0.8990	0.9345	0.9505
D	0.0800	0.0765	0.0895	0.0865	0.1150	0.1145	0.1185	0.1175	0.1455	0.1340	0.1400	0.1285	0.1785	0.1635	0.1590	0.1655	0.2275	0.2165	0.2055	0.1900

Table 2. Simulated power (continuation).

$n = n_1 = n_2$	20				50				70				100				150			
censoring %	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%	0%	20%	40%	60%
Configuration (d) - late survival differences																				
$WLR^{0,0}$	0.2855	0.2230	0.1745	0.1405	0.6840	0.5615	0.4430	0.3580	0.8330	0.7365	0.6025	0.4985	0.9495	0.8850	0.7810	0.6790	0.9950	0.9760	0.9180	0.8330
$WKM^{0,0}$	0.4870	0.3790	0.2865	0.2235	0.9415	0.8455	0.7115	0.5805	0.9930	0.9315	0.8600	0.7485	0.9980	0.9840	0.9400	0.8785	1.0000	0.9960	0.9790	0.9465
$WLR^{0,1}$	0.5870	0.4850	0.3985	0.3160	0.9735	0.9330	0.8605	0.7725	0.9985	0.9835	0.9580	0.9155	1	0.9985	0.9940	0.9820	1.0000	1.0000	0.9995	0.9990
$WKM^{0,1}$	0.5515	0.4525	0.3680	0.2920	0.9720	0.9125	0.8070	0.7065	0.9980	0.9690	0.9205	0.8620	0.9990	0.9935	0.9770	0.9485	1.0000	0.9990	0.9945	0.9870
$WLR^{1,0}$	0.0940	0.0825	0.0705	0.0640	0.1610	0.1400	0.1200	0.1055	0.1905	0.1505	0.1280	0.1125	0.2760	0.2235	0.1870	0.1590	0.3755	0.2945	0.2345	0.1900
$WKM^{1,0}$	0.0620	0.0550	0.0535	0.0505	0.1105	0.0935	0.0875	0.0755	0.1165	0.1040	0.0870	0.0865	0.1730	0.1385	0.1240	0.1025	0.2350	0.1875	0.1475	0.1280
$WLR^{1,1}$	0.2765	0.2265	0.1760	0.1450	0.5920	0.4925	0.4220	0.3470	0.7355	0.6370	0.5375	0.4695	0.8810	0.8040	0.7140	0.6180	0.9660	0.9245	0.8520	0.7640
$WKM^{1,1}$	0.0870	0.0740	0.0655	0.0600	0.1920	0.1670	0.1420	0.1205	0.2530	0.2070	0.1765	0.1575	0.3925	0.3255	0.2680	0.2220	0.5440	0.4545	0.3780	0.3085
WLR_{max4}	0.5105	0.4010	0.3095	0.2405	0.9490	0.8870	0.7835	0.6795	0.9920	0.9700	0.9245	0.8595	0.9990	0.9975	0.9880	0.9625	1.0000	1.0000	0.9995	0.9965
WKM_{max4}	0.5490	0.4235	0.3345	0.2620	0.9695	0.8945	0.7775	0.6630	0.9980	0.9625	0.9075	0.8335	0.9990	0.9925	0.9695	0.9355	1.0000	0.9990	0.9900	0.9815
WLR_{max3}	0.5235	0.4135	0.3185	0.2490	0.9545	0.8960	0.7960	0.6895	0.9930	0.9720	0.9320	0.8730	0.9995	0.9975	0.9900	0.9675	1.0000	1.0000	0.9995	0.9965
WKM_{max3}	0.5490	0.4240	0.3380	0.2640	0.9700	0.8975	0.7815	0.6660	0.9980	0.9635	0.9095	0.8355	0.9990	0.9925	0.9705	0.9370	1.0000	0.9990	0.9900	0.9825
D	0.4870	0.3930	0.3425	0.3010	0.9415	0.8585	0.7590	0.6945	0.9930	0.9300	0.8660	0.8140	0.9980	0.9505	0.8985	0.8610	1.0000	0.9545	0.9060	0.8800
Configuration (e) - early and late occurring survival differences																				
$WLR^{0,0}$	0.3290	0.2745	0.2500	0.2155	0.7680	0.6745	0.5900	0.5320	0.8985	0.8260	0.7620	0.7010	0.9765	0.9435	0.8975	0.8535	0.9985	0.9940	0.9775	0.9580
$WKM^{0,0}$	0.3335	0.2750	0.2390	0.2065	0.8005	0.7000	0.5970	0.5280	0.9210	0.8420	0.7685	0.6965	0.9855	0.9515	0.9070	0.8415	0.9995	0.9925	0.9820	0.9530
$WLR^{0,1}$	0.2500	0.1820	0.1425	0.1150	0.7085	0.5575	0.4015	0.3030	0.8685	0.7385	0.5865	0.4420	0.9640	0.8945	0.7660	0.6105	0.9960	0.9735	0.9055	0.8095
$WKM^{0,1}$	0.3250	0.2625	0.2280	0.1905	0.8180	0.7035	0.5790	0.4855	0.9320	0.8410	0.7525	0.6515	0.9895	0.9525	0.8995	0.8110	0.9995	0.9930	0.9740	0.9375
$WLR^{1,0}$	0.2795	0.2625	0.2560	0.2465	0.5955	0.5770	0.5535	0.5310	0.7505	0.7250	0.7095	0.7005	0.8830	0.8705	0.8510	0.8360	0.9725	0.9675	0.9610	0.9520
$WKM^{1,0}$	0.2075	0.1975	0.1975	0.2015	0.4680	0.4645	0.4555	0.4475	0.6320	0.6300	0.6170	0.6130	0.7855	0.7790	0.7705	0.7640	0.9255	0.9200	0.9125	0.9110
$WLR^{1,1}$	0.1340	0.1140	0.1035	0.0895	0.2770	0.2260	0.1855	0.1505	0.3780	0.3060	0.2325	0.1960	0.5140	0.4190	0.3260	0.2710	0.6505	0.5365	0.4350	0.3590
$WKM^{1,1}$	0.1480	0.1400	0.1375	0.1345	0.3565	0.3330	0.3010	0.2910	0.4975	0.4680	0.4385	0.4160	0.6545	0.6135	0.5845	0.5405	0.8105	0.7830	0.7450	0.7190
WLR_{max4}	0.2765	0.2290	0.2120	0.1840	0.7120	0.6140	0.5230	0.4740	0.8780	0.7885	0.7110	0.6530	0.9640	0.9140	0.8590	0.8045	0.9975	0.9870	0.9605	0.9405
WKM_{max4}	0.3195	0.2655	0.2370	0.2005	0.8005	0.6800	0.5790	0.4965	0.9200	0.8305	0.7445	0.6755	0.9855	0.9450	0.8950	0.8220	0.9995	0.9910	0.9725	0.9480
WLR_{max3}	0.2830	0.2375	0.2195	0.1915	0.7245	0.6300	0.5405	0.4880	0.8865	0.7980	0.7210	0.6695	0.9710	0.9225	0.8710	0.8215	0.9975	0.9885	0.9670	0.9445
WKM_{max3}	0.3200	0.2670	0.2375	0.2010	0.8040	0.6820	0.5825	0.4985	0.9205	0.8325	0.7460	0.6805	0.9860	0.9465	0.8970	0.8260	0.9995	0.9915	0.9730	0.9495
D	0.3335	0.2785	0.2385	0.2010	0.8005	0.7085	0.6135	0.5280	0.9210	0.8325	0.7590	0.6775	0.9855	0.9275	0.8780	0.7985	0.9995	0.9760	0.9430	0.9070

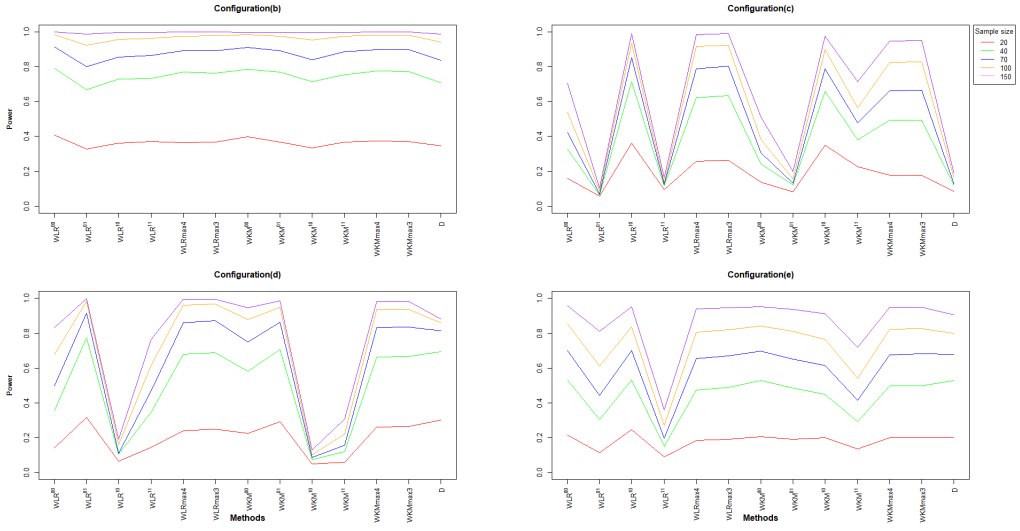

Figure 2. Simulated power of the statistic tests for 60% censoring.

2 displays the powers under 60% censoring. Under proportional hazard configuration, the log-rank test ($WLR^{0,0}$) and Pepe-Flemings's weighted Kaplan-Meier test ($WKM^{0,0}$) have maximum power as expected. However, the power of WKM_{max3} and WKM_{max4} are really close. The D statistic is also competitive but under non-censoring. Therefore, these two combination tests have almost the same sensitivity as $WLR^{0,0}$, i.e. the optimal test against proportional hazards alternative.
For early survival differences, as expected, $WLR^{1,0}$ and $WKM^{1,0}$ have greater power than the other statistics. Even though, WLR_{max3} fares poorer than them, it is the best among the rest of considered test statistics. In the case of the late survival differences, $WLR^{0,1}$ and $WKM^{0,1}$ have greater power than the others. Nevertheless, the difference in power between these and WKM_{max3} is small. Finally, in the case of early and late survival differences, there is not a clear more powerful test statistic but WKM_{max3} is consistently a good choice for all censoring and sample sizes.
In summary, assuming ignorance of the likely treatment effect, except for early survival differences, the WKM_{max3} test statistic is preferable to the other maximum combination test statistics. For early survival differences, the WLR_{max3} test statistic exhibits a higher power than WKM_{max3} but the latter is higher than the well-known log-rank test. In fact, [14] showed a minimal power gain power for WKM and RSMT to the LR test when there is a delayed effect but WKM_{max3} has better performance than WKM and RMST.

4. Application to a real data set

We ilustrate the performance of the thirteen test statistics with a real data set, chosen because they appear to show late difference. The data is available in the package `KMsurv` [22] with the name `bmt`. The `bmt` data come from a study on the survival of bone marrow transplant patients with different types leukemia. The complete data frame has 22 variables and to carry out the study we worked with the following: disease free survival time, that is, the time either relapse or death (t2), censoring indicator, 1: relapse or death, 0: censored survival time (d3), disease group, acute lymphoblastic leukemia (ALL) consisting of 38 patients and low risk acute myeloid leukemia (AML) consisting of 54 patients.

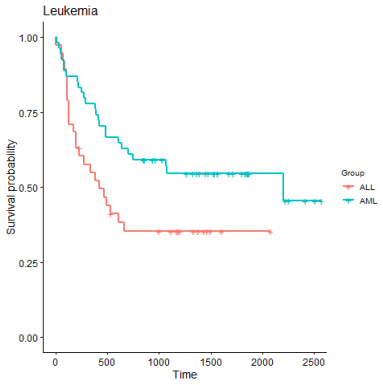

Figure 3. Survival functions for bone marrow transplant patients.

Table 3. Test statistics and their correponding p-values.

	statistic	p-value
$WLR^{0,0}$	2.1748	0.026
$WLR^{0,1}$	1.6935	0.094
$WLR^{1,0}$	2.2032	0.024
$WLR^{1,1}$	2.0612	0.042
WLR_{max4}	2.2032	0.054
WLR_{max3}	2.2032	0.054
$WKM^{0,0}$	2.3419	0.016
$WKM^{0,1}$	2.3127	0.024
$WKM^{1,0}$	2.3516	0.016
$WKM^{1,1}$	2.3589	0.018
WKM_{max4}	2.3589	0.020
WKM_{max3}	2.3516	0.020
D	415.9541	0.080

Kaplan-Meier curves are shown in Figure 3 and exhibit a late difference in survival. The p-values associated to the proposed test statistics are summarized in Table 3, from which it can be seen that most of the test statistics reject the null hypothesis $H_0 : S_1 = S_2$ at the 0.05% level. However, D, $WLR^{0,1}$, WLR_{max3} and WLR_{max4} yield a nonsignificant p-value. Although, WLR_{max3} and WLR_{max4} p-values are in the borderline. Therefore, WKM_{max3} and WKM_{max4} yield stronger evidence against the null hypothesis of no difference between the two groups than WLR_{max3} and WLR_{max4}. Therefore, one can miss a potential regulatory submission opportunity if the primary analysis is carried out based on the well-known RMST or MaxCombo tests (WLR_{max4} or WLR_{max3}).

5. Discussion

In practice, the log-rank test is the most used to test the equality of two survival distributions in the presence of censoring and it is known to have optimum power to detect a difference in survival distributions when proportional hazards holds. However, non-proportional hazard ratio between the control and experimental arms occurs fairly often in clinical trials. For example, the proportional hazards (PH) assumption often does not hold for the primary time-to-event (TTE) efficacy endpoint in trials of novel immuno-oncology drugs. For some immuno-oncology drugs, a delayed treatment effect is observed, where the survival curves for the immuno-oncology drug and standard of care are not separated initially but start separating after some period of time. In other oncology trials, the separation between the survival curves occurs early on but then the distance between the curves diminishes over time. There may also be cases where the survival curves cross each other, i.e., the PH assumption is violated. With a preconceived type of treatment effect, then the weighted log rank and weighted Kaplan-Meier test statistics are good alternatives to the log-rank test. However, if the direction of the violation from non-PH is not known, then a combination of the previous test statistics should be used. There are a number of papers studying these combinations based on weighted log rank and weighted Kaplan-Meier test statistics separately. Recently, [13] and [14] used Monte Carlo simulation to study the performance of nine different statistics based mainly on weighted log-rank tests. They conclude that a modified version of WLR_{max3} and WLR_{max4}, respectively are preferred in the absence of prior knowledge regarding the PH or non-PH patterns. [15] concluded that MaxCombo test shows clear advantages when a large treatment benefit emerges later on the trial. They introduced a design approach for confirmatory trials with WLR_{max4} and developed a stepwise and iterative approach for calculating sample size when the final analysis is based on this test statistic. Our conclusion is that WLR_{max3} is better than the MaxCombo test, WLR_{max4} and WKM_{max3} is preferable to WLR_{max3} in most of the cases for small and moderate sample sizes. Performance of the KM based methods is known that depend on the lenght of study period (t_c) and censoring pattern. We have minimum of the largest observed time in each of the two groups as a choice of t_c. There are other issues such as the interpretation of the WLR test statistics result in terms of the clinical effect. [23] proposed a Cox-model based time-varying treatment effect estimate to complement the WLR test statistics. However, investigation of these topics is beyond the scope of the present paper.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The second author was supported by grant PID2019-106861RB-I00 funded by MCIN/AEI/10.13039/501100011033 and grant A-SEJ-154-UGR20 funded by FEDER/Junta de Andalucía/Department of Economic Transformation, Industry, Knowledge and Universities and the first author was supported by grant PID2019-104681RB-I00 funded by MCIN/AEI/10.13039/501100011033.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. S. Herbst, P. Baas, D. W. Kim, E. Felip, J. L. Pérez-Gracia, J. Y. Han, et al., Pembrolizumab Versus Docetaxel for Previously Treated, PDL1-Positive, Advanced Non-Small-Cell Lung Cancer (KEYNOTE-010): A Randomised Controlled Trial, *The Lancet*, **387** (2016), 1540—1550. https://doi.org/10.1016/S0140-6736(15)01281-7
2. M. S. Pepe, T. R. Fleming, Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data, *Biometrics*, **45** (1989), 497–507. http://dx.doi.org/10.2307/2531492
3. B. Huang, P. F. Kuan, Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point, *Pharmaceut. Statist.*, **17** (2018), 202–213. http://dx.doi.org/10.1002/pst.1846
4. S. G. Self, An adaptive weighted log-rank test with application to cancer prevention and screening trials, *Biometrics*, **47** (1991), 975–986. http://dx.doi.org/10.2307/2532653
5. T. R. Fleming, D. P. Harrington, A class of hypothesis tests for one and two samples of censored survival data, *Commun. Statist. Theory Methods*, **13** (1981), 2469–2486. http://dx.doi.org/10.1080/03610928108828073
6. D. P. Harrington, T. R. Fleming, A class of rank test procedures for censored survival data, *Biometrika*, **69** (1982), 553–566. http://dx.doi.org/10.1093/biomet/69.3.553
7. T. R. Fleming, D. P. Harrington, *Counting Processes and Survival Analysis*, New York: Wiley, (1991). http://dx.doi.org/10.1002/9781118150672
8. J. W. Lee, Some versatile tests based on the simultaneous use of weighted log-rank statistics, *Biometrics*, **52** (1996), 721–725. http://dx.doi.org/10.2307/2532911
9. S. H. Lee, On the versatility of the combination of the weighted log-rank statistics, *Comput. Statist. Data Anal.*, **51** (2007), 6557–6564. http://dx.doi.org/10.1016/j.csda.2007.03.006
10. T. G. Karrison, Versatile tests for comparing survival curves based on weighted log-rank statistics, *Stat. J.*, **16** (2016), 678–690. http://dx.doi.org/10.1177/1536867X1601600308
11. Y. Shen, J. Cai, Maximum of the Weighted Kaplan-Meier tests with application to cancer prevention and screening trials, *Biometrics*, **57** (2001), 837–843. http://dx.doi.org/10.1111/j.0006-341X.2001.00837.x

12. S. H. Lee, Maximum of the weighted Kaplan-Meier tests for the two-sample censored data, *J. Statist. Comput. Simul.*, **81** (2011), 1017–1026. http://dx.doi.org/10.1080/00949651003627753

13. P. Royston , M. K. B. Parmar, A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome, *Trials*, **21** (2020), 315. http://dx.doi.org/10.1186/s13063-020-4153-2

14. R. S. Lin, J. Lin, S. Roychoudhury, K. M. Anderson, T. Hu, B. Huang, et al., Alternative Analysis Methods for Time to Event Endpoints under Non-proportional Hazards: A Comparative Analysis, *Statist. Biopharm. Res.*, **12** (2020), 187–198. https://doi.org/10.1080/19466315.2019.1697738

15. S. Roychoudhury, K. M. Anderson, J. Ye, P. Mukhopadhyay, Robust Design and Analysis of Clinical Trials with Nonproportional Hazards: A Straw Man Guidance From a Cross- Pharma Working Group, *Statist. Biopharm. Res.*, **15** (2021), 280–294. http://dx.doi.org/10.1080/19466315.2021.1874507

16. E. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53** (1958), 457–481. http://dx.doi.org/10.1080/01621459.1958.10501452

17. T. R. Fleming, D. P. Harrington, M. O'Sullivan, Supremum versions of the log-rank and generalized Wilcoxon statistics, *J. Am. Stat. Assoc.*, **82** (1987), 312–320. http://dx.doi.org/10.1080/01621459.1987.10478435

18. W. Yang, W. Haiyan, A. Keaven, R. Satrajit, H. Tianle, L. Hongliu, R package nphsim: Non proportional hazards sample size and simulation, (2017). Available from: `https://github.com/keaven/nphsim`

19. R. Ristl, N. Ballarini, R package nph: Planning and Analysing Survival Studies under Non-Proportional Hazards, (2020). Available from: `https://CRAN.R-project.org/package=nph`

20. M. Bofill, G. Gómez, R package SurvBin: Two-sample statistics for binary and time-to-event outcomes, (2020). Available from: `https://github.com/MartaBofillRoig/SurvBin`

21. H. Uno, L. Tian, M. Horiguchi, A. Cronin, C. Battioui, J. Bell, R package survRM2: Comparing Restricted Mean Survival Time, (2020). Available from: `https://CRAN.R-project.org/package=survRM2`

22. J. P Klein, M. L. Moeschberger, J. Yan, R package KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis, (2012). Available from: `https://CRAN.R-project.org/package=KMsurv`

23. R. S. Lin, L. F. León, Estimation of treatment effects in weighted log-rank tests, *Contempor. Clin. Trials Commun.*, **8** (2017), 147–155. http://dx.doi.org/10.1016/j.conctc.2017.09.004

Supplementary

R-code for the p-value calculation

```
library(KMsurv)
data(bmt)
B=500
```

```
bmt2=datos[(bmt$group==1)|(bmt$group==2),]
df=data.frame(time=bmt2$t2, cens=1-bmt2$d3,
event=ifelse(bmt2$d3==1, TRUE, FALSE), group=bmt2$group)

n=nrow(df)
ng1=as.numeric(table(df$group)[1])
ng2=as.numeric(table(df$group)[2])

max_wlr=logrank_maxtest(df$time, df$event, df$group,
rho = c(0,0,1,1), gamma = c(0,1,0,1))
Test_wlr00=max_wlr$tests$z[1]

est_wlr00=vector()

for (b in 1:B){
    index=sample(1:n, replace=FALSE)
    dfindex=df[index,]
    g1=dfindex[1:ng1,]
    g1$group=1
    g2=dfindex[(ng1+1):(ng1+ng2),]
    g2$group=2

    combinedB=rbind(g1,g2)

    max_wlr=logrank_maxtest(combinedB$time, combinedB$event,
    combinedB$group,  rho = c(0,0,1,1), gamma = c(0,1,0,1))
    est_wlr00[b]=max_wlr$tests$z[1]
}

for(j in 1:B){
   if(abs(est_wlr00[j])>abs(Test_wlr00)) n_wlr=n_wlr+1
}

prop_wlr=n_wlr/ B
```