Annals of the Institute of Statistical Mathematics (2021) 73:1037–1061
https://doi.org/10.1007/s10463-020-00777-w

A permutation test for the two-sample right-censored model

Grzegorz Wyłupek[1]

Received: 24 March 2020 / Revised: 19 September 2020 / Accepted: 26 October 2020
Published online: 12 January 2021

Abstract
The paper presents a novel approach to solve a classical two-sample problem with right-censored data. As a result, an efficient procedure for verifying equality of the two survival curves is developed. It generalizes, in a natural manner, a well-known standard, that is, the log-rank test. Under the null hypothesis, the new test statistic has an asymptotic Chi-square distribution with one degree of freedom, while the corresponding test is consistent for a wide range of the alternatives. On the other hand, to control the actual Type I error rate when sample sizes are finite, permutation approach is employed for the inference. An extensive simulation study shows that the new test procedure improves upon classical solutions and popular recent developments in the field. An analysis of the real datasets is included. A routine, written in R, is attached as Supplementary Material.

Keywords Incomplete observations · Laguerre polynomials · Permutation test · Survival analysis · Two-sample test · Weighted log-rank test

1 Introduction

We consider the two-sample censorship model with $n = n_1 + n_2$ independent observations made of n_l individuals from the lth population, $l = 1, 2$. The ith subject in the lth sample has nonnegative, independent, latent survival and censoring times X^0_{li} and U_{li} with the corresponding continuous distribution function F_l and G_l, respectively, $i = 1, \dots, n_l$, $l = 1, 2$. The observable random variables are $X_{li} = \min\{X^0_{li}, U_{li}\}$ together with their censoring statuses $\Delta_{li} = \mathbb{1}(X^0_{li} \le U_{li})$, where

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-020-00777-w) contains supplementary material, which is available to authorized users.

✉ Grzegorz Wyłupek
wylupek@math.uni.wroc.pl

[1] Institute of Mathematics, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland

$\mathbb{1}(\cdot)$ is the indicator of the set $\cdot$. As a result, what we have at our disposal is the set of incomplete observations

$$(X_{li}, \Delta_{li}) = (\min\{X_{li}^o, U_{li}\}, \mathbb{1}(X_{li}^o \leq U_{li})), \quad i = 1, \dots, n_l, \quad l = 1, 2.$$

On their basis, we will test

$$\begin{aligned} \mathcal{H} : F_1(x) = F_2(x), &\quad \text{for all} \quad x \geq 0, \\ \mathcal{A} : F_1(x) \neq F_2(x), &\quad \text{for some} \quad x \geq 0, \end{aligned}$$

in the presence of infinite-dimensional nuisance parameters G_1 and G_2.

The problem is very important and has an enormous significance in practice. Reliability, social sciences, and medicine are the disciplines, where the censored data are commonplace. Therefore, a really good solution is highly desirable. Notwithstanding, there are a lot of solutions of the above problem; many of them have serious drawback, i.e., they are sensitive to some specific discrepancies from the null model. For instance, the popular log-rank test (Mantel 1966) is asymptotically optimal for the Lehmann's alternatives of the form $F_1(x) = 1 - [1 - F_2(x)]^\theta$, $\theta > 0, \theta \neq 1, x \geq 0$, when $G_1 = G_2$. So, it means that, if F_1 and F_2 are the cumulative distribution functions of an exponential distribution, a proportional hazard difference occurs under the alternative. The generalized Wilcoxon test (Gehan 1965; Peto and Peto 1972) inherits the weaknesses of its counterpart for complete data. The reason is that it is the most sensitive to one direction, which is an early hazard difference, while the other ones can be barely detectable, cf. Sect. 4.2.2 and Appendix G in Supplementary Material. In late seventies, Prentice (1978) provides a general source of a construction of the linear rank statistics sensitive mainly to one direction. It covers, among others, the log-rank and Wilcoxon-type tests. Those and many other solutions belong to a wider class of statistics, called the weighted log-rank statistics (WLR for brevity), while a difference between them simply relies on a choice of a weight function. Efron (1967), Tarone and Ware (1977), Gill (1980), Harrington and Fleming (1982), and Fleming et al. (1987) are just a few of them. For an overview of such solutions, see, for instance, Letón and Zuluaga (2005). Furthermore, the tests based on the Kolmogorov–Smirnov statistic (Fleming et al. 1980; Schumacher 1984), the Cramér–von Mises statistic (Koziol 1978; Schumacher 1984), as well as the weighted Kaplan–Meier statistic, WKM for short, (Pepe and Fleming 1989, 1991; Lee et al. 2008), although consistent (in the case of the WKM test under stochastically ordered alternatives), for small and moderate sample sizes, are able to detect only certain specific deviations from the null hypothesis. For the evidence in the case of Renyi-type Kolmogorov–Smirnov test, see Sect. 4.2.2 and Appendix G in Supplementary Material. Therefore, there were many attempts to increase the range of their sensitivity. Lu et al. (1994) introduced a bootstrap version of the test based on the horizontal shift function related to the *Q*–*Q* plot, while Li et al. (1996) proposed a Chi-square test and a Kolmogorov-type test based on the vertical shift function related to the *P*–*P* plot employing the bootstrap procedure of Efron (1981) as well. Lee (1996) proposed a linear combination of the four weighted log-rank statistics sensitive to the early, middle, late, and crossing hazard differences, as well as a maximum type test statistic based on them. Lin and Kosorok (1999) introduced a huge class of function-indexed test statistics and, also, defined a versatile procedure based on them. Chi and Tsai (2001) developed some versatile tests being a function of the log-rank and WKM statistics. Wu and Gilbert (2002) proposed a weighted log-rank tests optimal for detection early and/or late survival differences. Also, on their basis, they defined two kinds of versatile test procedures. Lin and Wang (2004) defined a test statistic, which is a standardized sum of squares of differences between the observed and conditionally expected, given the past, number of events in the first group at each failure time point. Lee (2007) suggested for testing a combination of the two selected WLR statistics. Qiu and Sheng (2008) proposed a two-stage procedure for comparing two hazard rate functions. In the first stage, the log-rank test is used, while in the second stage, a subtle procedure focused on detection crossing hazard rates is employed. Kraus (2009) worked out an adaptive test related to a system of Legendre polynomials providing a generalization of the log-rank and Wilcoxon test simultaneously. Martínez-Camblor (2010) defined the test, which is based on the idea of adaptation of the likelihood ratio statistic proposed by Zhang and Wu (2007) for uncensored data. His approach simply relies on the replacement of the nonparametric estimator of the unknown cumulative distribution functions by the Kaplan and Meier (1958) estimate. Yang and Prentice (2010) developed several new generalizations of the log-rank test including an adaptively weighted log-rank test based on the Yang and Prentice (2005) model. Darilay and Naranjo (2011) proposed a pretest for using log-rank or Wilcoxon solution. Chang et al. (2012) proposed a combination of the WLR and WKM statistics with a jackknife selection of the dominating solution. Chauvel and O'Quigley (2014) describe a class of tests based on O'Quigley (2003) approach and discuss several representants of them. Koziol and Jia (2014) introduced a weighted (Lin and Wang 2004) statistic for crossing hazards. By contrast, Brendel et al. (2014) investigated an adaptive projection-type test sensitive to the three directions corresponding to the proportional, crossing, and central hazards. Their procedure is patterned after the Behnen and Neuhaus (1983) solution for complete data. Callegaro and Spiessens (2017) defined new tests sensitive, in particular, to nonproportional hazard difference. Garès et al. (2017) investigated the Fleming–Harrington test for late hazard difference, see Fleming and Harrington (1991, p. 257), for the general definition of the $G^{p,q}$ class of tests. Hsieh and Chen (2017) introduced two strategies for testing equality of survival functions based on several, known from the literature, test procedures. Liu and Yin (2017) proposed the partitioned log-rank test for the homogeneity of two hazard rates by partitioning the weighted log-rank statistic at a certain time point. Next, they defined a versatile procedure based on the supremum of such partitioned log-rank statistics over all distinct death times in the pooled sample. Arboretti et al. (2018) recommended nonparametric combination tests under the constraint that the observations are exchangeable. Also, cf. Arboretti et al. (2010), as well as chapter 9 in Pesarin and Salmaso (2010), for earlier contributions.

In spite of the rich arsenal of the two-sample tests for censored data, it seems that there is still some space for the introduction of a new solution which improves upon their finite sample properties. Therefore, in this paper, we develop an efficient generalization of the log-rank test, which is sensitive to an arbitrary large number of directions determine by the efficient score functions. It leads to a consistent test for a wide range of deviations from the null model.

The paper is organized as follows. Section 2 reveals the details of a construction of a new test statistic. The asymptotic outcomes and some finite sample theoretical properties of the proposed test are gathered in Sect. 3. Section 4 and Appendix G demonstrate the results of the conducted simulation study, while Sect. 5 presents the real data examples. Section 6 concludes with some discussion. A derivation of the efficient score functions, an additional interpretation of the related WLR statistic, and the proofs are deferred to Appendices A, B, and C, respectively. Appendix D presents a description of the algorithm enabling estimation of the power function of a test. A detailed description of the competitive solutions is given in Appendix E. An `R` code permitting to calculate, among others, values of the new test statistic and p values of the related test is included in Supplementary Material. A description of the `R` code is placed in Appendix F. All the appendices are part of Supplementary Material.

2 A new test statistic

In this section, we construct a new data-driven test in the problem $(\mathcal{H}, \mathcal{A})$. First, we reparametrize $\mathcal{H}$ in term of the weighted difference of the hazard functions $a(\cdot)[\lambda_1(\cdot) - \lambda_2(\cdot)]$ and expand that function in a Fourier series in a system of orthonormal functions. Next, we estimate the Fourier coefficients in the expansion. Then, we justify a selection of the system. After that, we built a quadratic form, W_d, being a sum of squares of the standardized empirical (estimated) Fourier coefficients. Finally, the selection rule T finishes the job selecting the number of the summands from the data at hand.

2.1 Reparametrization of $\mathcal{H}$

Assume, in this subsection, that the respective derivatives f_1,f_2 of F_1, F_2 exist. Then,

$$\lambda_l(y) = \frac{f_l(y)}{1 - F_l(y)}, \quad y \in [0, \infty), \quad l = 1, 2, \tag{1}$$

are the unknown hazard functions, while the null hypothesis is equivalent to

$$\lambda_1(y) = \lambda_2(y), \quad \text{for all} \quad y \geq 0. \tag{2}$$

Let $n = n_1 + n_2$ and $\eta = \lim_{n\to\infty}(n_1/n)$, $\eta \in (0, 1)$. Let $f(y) = \eta f_1(y) + (1 - \eta)f_2(y)$, $y \geq 0$, be the pooled density function. Set $\pi_l(x) = [1 - F_l(x)][1 - G_l(x)] = P(X_{l1} > x)$, $l = 1, 2, \pi(x) = \eta\pi_1(x) + (1 - \eta)\pi_2(x)$, and $\tau = \inf\{x : \pi_1(x)\pi_2(x) = 0\}$, $x \in [0, +\infty)$. Multiplying both sides of (2) by $\pi_1(y)\pi_2(y)/[\pi(y)f(y)] =: a(y)$, $y \in [0, \tau]$, we obtain that the null hypothesis, restricted to the interval $[0, \tau]$, is equivalent to

$$a(y)[\lambda_1(y) - \lambda_2(y)] = 0, \quad \text{for all} \quad y \in [0, \tau]. \tag{3}$$

It should be emphasized, in this place, that a difference between F_1 and F_2 can only be detected on the interval $[0, \tau]$. Therefore, a restriction of the comparisons to them is actually not a limitation. Since we prefer to work on [0, 1) instead on the half-line $[0, +\infty)$, we make a transformation via the cumulative distribution function in the combined sample, i.e., $F = \eta F_1 + (1 - \eta)F_2$.

This yields

$$a(F^{-1}(t))[\lambda_1(F^{-1}(t)) - \lambda_2(F^{-1}(t))] = 0, \quad \text{for all} \quad t \in [0, F(\tau)], \tag{4}$$

where F^{-1} is the inverse function. Now, we expand the left-hand side of the above equation in a system $\{\ell_j^0\}_{j=1}^{\infty}$ of orthonormal functions in $L^2([0, 1), dt)$. As a result, under the null model, all the Fourier coefficients in the expansion, i.e.,

$$\begin{aligned}&\int_0^{F(\tau)} \ell_j^0(t)\, a(F^{-1}(t))[\lambda_1(F^{-1}(t)) - \lambda_2(F^{-1}(t))]\mathrm{d}t \\ &\quad = \int_0^{\tau} \ell_j^0(F(y))\, \frac{\pi_1(y)\,\pi_2(y)}{\pi(y)} \mathrm{d}[\Lambda_1(y) - \Lambda_2(y)],\end{aligned}$$

$j = 1, 2, 3, \dots$, vanish.

2.2 Empirical Fourier coefficients

Recall that, we only know the incomplete observations

$$(X_{li}, \Delta_{li}) = (\min\{X_{li}^o, U_{li}\}, \mathbb{1}(X_{li}^o \le U_{li})), \quad i = 1, \dots, n_l, \quad l = 1, 2.$$

For the notational convenience, we will also write that set skipping the first subscript

$$(X_i, \Delta_i) = (\min\{X_i^o, U_i\}, \mathbb{1}(X_i^o \le U_i)), \quad i = 1, \dots, n, \quad n = n_1 + n_2,$$

having in mind that the first n_1 observations belong to the first group, while the remaining ones are a part of the second group. On their basis, we define the following counting processes

$$\begin{aligned}N_l(x) &= \sum_{i=1}^{n_l} N_{li}(x) = \sum_{i=1}^{n_l} \Delta_{li}\, \mathbb{1}(X_{li} \le x), \\ Y_l(x) &= \sum_{i=1}^{n_l} Y_{li}(x) = \sum_{i=1}^{n_l} \mathbb{1}(X_{li} \ge x), \quad l = 1, 2, \ x \in [0, +\infty).\end{aligned} \tag{5}$$

The process N_l counts the observed deaths in group l through time x. Therefore, it is called the death process. The process Y_l is the number of subjects still at risk at x in the lth group. In the literature, it is known as the at-risk process. We also put $N(x) = N_1(x) + N_2(x)$, $Y(x) = Y_1(x) + Y_2(x)$, $x \in [0, +\infty)$. Set $\hat{\tau} = \inf\{x : Y_1(x)Y_2(x) = 0\}$, $\hat{F}(x) = 1 - \prod_{X_i<x}\{[Y(X_i) - 1]/[Y(X_i)]\}^{\Delta_i}$, i.e.,

the left-continuous version of the Kaplan and Meier (1958) estimate. Since $\int_0^y [\mathrm{d}N_l(x)/Y_l(x)]$ is the estimate of $\Lambda_l(y)$, $l = 1, 2$, a natural, from a counting processes viewpoint, estimator of the jth Fourier coefficient is the rescaled weighted log-rank statistic

$$\sqrt{\frac{n}{n_1 n_2}}\,\mathcal{L}_j = \frac{n}{n_1 n_2}\int_0^{\hat{\tau}} \ell_j^0(\hat{F}(x))\frac{Y_1(x)Y_2(x)}{Y(x)}\left(\frac{\mathrm{d}N_1(x)}{Y_1(x)} - \frac{\mathrm{d}N_2(x)}{Y_2(x)}\right). \tag{6}$$

On the other hand, $\mathcal{L}_j$ is the rescaled, by the factor $\sqrt{n_1 n_2/n}$, jth empirical (estimated) Fourier coefficient in the expansion of the function $a(F^{-1}(\cdot))[\lambda_1(F^{-1}(\cdot)) - \lambda_2(F^{-1}(\cdot))]$ in the system $\{\ell_j^0\}_{j=1}^{\infty}$.

2.3 A selection of the system

There are a lot of systems of orthonormal functions in $L^2([0,1), dt)$. The Legendre polynomials, the Haar functions, and the trigonometric functions are just a few of them. In this subsection, we define a system of transformed Laguerre polynomials and justify their usefulness in survival analysis settings.

Let $\{L_j\}_{j=0}^{\infty}$ be the system of Laguerre polynomials on $[0, +\infty)$. The first four polynomials have the following form

$$L_0(x) = 1, \quad L_1(x) = -x + 1, \quad L_2(x) = (1/2)(x^2 - 4x + 2),$$
$$L_3(x) = (1/6)(-x^3 + 9x^2 - 18x + 6),$$

while the general description provides that

$$L_j(x) = \frac{e^x}{j!}\frac{\mathrm{d}^j}{\mathrm{d}x^j}(x^j e^{-x}) = \sum_{k=0}^{j}\frac{(-1)^k}{k!}\binom{j}{k}x^k, \quad j = 0, 1, 2, \dots, \quad x \in [0, +\infty). \tag{7}$$

The system is complete in $L^2([0,\infty), e^{-x}dx)$ and L_js satisfy

$$\int_0^{\infty} L_i(x)L_j(x)e^{-x}\mathrm{d}x = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{8}$$

Therefore, the related system of functions on [0, 1) given by

$$\ell_j(t) = L_j(-\log(1-t)), \quad j = 0, 1, 2, \dots, \quad t \in [0, 1), \tag{9}$$

where $-\log(1-t)$ is the quantile function of the exponential distribution with the parameter 1, is orthonormal and complete in $L^2([0,1), dt)$.

We will build the test statistic on the basis of the weighted log-rank statistics related to the projected functions from the basis $\{\ell_j\}_{j=0}^{\infty}$. The goal of usage of a projection is a possible elimination of the infinite-dimensional nuisance parameters. To be specific, we calculate the projection of a score function ℓ_j onto the space spanned by the nuisance parameters G_1 and G_2 in the parametrization of Neuhaus (2000, p. 487), formula

(3.5), and exploit their residuum. Actually, we utile Neuhaus' recipe for the so-called efficient score function ℓ_j^0, formula (3.33), p. 491, which leads to the formula

$$\ell_j^0(t) = \ell_j(t) - \frac{1}{1-t}\int_t^1 \ell_j(s)\mathrm{d}s, \quad j = 0, 1, 2, \dots, \tag{10}$$

in our notation. A prompt calculation yields

$$\ell_0^0(t) = 0 \quad \text{and} \quad \ell_j^0(t) = \ell_{j-1}(t), \quad j = 1, 2, 3, \dots, \quad t \in [0, 1). \tag{11}$$

The details are given in Supplementary Material, Appendix A. The system of the efficient score functions $\{\ell_j^0\}_{j=1}^\infty$ is orthonormal and complete because it is equivalent to the system of the score functions $\{\ell_j\}_{j=0}^\infty$. It looks like there is no other system of orthonormal functions having such a property and, therefore, there is no other one better suited for the employment in our problem. Furthermore, the work of Prentice (1978) provides the interpretation of the components in (10). Namely, the minuend is the score function corresponding to the uncensored observations, while the subtrahend is the score function related to the censored data.

2.4 An auxiliary statistic

In this subsection, we introduce the standardized weighted log-rank statistics related to the efficient score functions from the system $\{\ell_j^0\}_{j=1}^\infty$ and define the auxiliary statistic being the sum of their squares.

The weighted log-rank statistic with the weight function corresponding to the jth efficient score function, in the above sense, has the form

$$\mathcal{L}_j = \sqrt{\frac{n}{n_1 n_2}} \int_0^{\hat{\tau}} \ell_j^0(\hat{F}(x)) \frac{Y_1(x)Y_2(x)}{Y(x)} \left(\frac{\mathrm{d}N_1(x)}{Y_1(x)} - \frac{\mathrm{d}N_2(x)}{Y_2(x)} \right), \quad j = 1, 2, 3, \dots. \tag{12}$$

An instantaneous integration provides

$$\begin{aligned}\mathcal{L}_j &= \sqrt{\frac{n}{n_1 n_2}} \left\{ \sum_{i=1}^{n_1} w_j(X_i) \frac{Y_1(X_i)Y_2(X_i)}{Y(X_i)} \frac{\Delta_i}{Y_1(X_i)} - \sum_{i=n_1+1}^{n} w_j(X_i) \frac{Y_1(X_i)Y_2(X_i)}{Y(X_i)} \frac{\Delta_i}{Y_2(X_i)} \right\} \\ &= \sqrt{\frac{n_1 n_2}{n}} \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} w_j(X_i) \frac{Y_2(X_i)}{n_2} \frac{n}{Y(X_i)} \Delta_i - \frac{1}{n_2} \sum_{i=n_1+1}^{n} w_j(X_i) \frac{Y_1(X_i)}{n_1} \frac{n}{Y(X_i)} \Delta_i \right\},\end{aligned}$$

where $w_j(X_i) = \ell_j^0(\hat{F}(X_i))\, \mathbb{1}\big(Y_1(X_i)Y_2(X_i) > 0\big)$, $i = 1, \dots, n$. Under the null hypothesis, $\mathcal{L}_j$ has an asymptotic normal distribution $N(0, \sigma_j^2)$, see Appendix C in Supplementary Material, with

$$\sigma_j^2 = \int_0^\tau \left[\ell_j^0(F(x))\right]^2 \frac{\pi_1(x)\,\pi_2(x)}{\pi(x)}\, \mathrm{d}\Lambda(x), \tag{13}$$

where Λ is the pooled cumulative hazard function under $\mathcal{H}$. Its consistent estimator, under the null model,

$$\hat{\sigma}_j^2 = \frac{n}{n_1 n_2} \int_0^{\hat{\tau}} \left[\ell_j^0(\hat{F}(x))\right]^2 \frac{Y_1(x)Y_2(x)}{Y(x)} \frac{\mathrm{d}N(x)}{Y(x)} = \frac{n}{n_1 n_2} \sum_{i=1}^{n} \left[w_j(X_i)\right]^2 \frac{Y_1(X_i)Y_2(X_i)}{Y(X_i)} \frac{\Delta_i}{Y(X_i)},$$

cf. formula (3.3.12), p. 47, Gill (1980), allows us to define the components

$$C_j = \frac{\mathcal{L}_j}{\hat{\sigma}_j}, \quad j = 1, 2, 3, \dots, \tag{14}$$

which, under the null hypothesis, have the asymptotic $N(0,\ 1)$ distribution. Obviously, they are correlated, while the asymptotic correlation of the ith and jth component is $\sigma_{ij}/(\sigma_i\, \sigma_j)$, where

$$\sigma_{ij} = \int_0^{\tau} \ell_i^0(F(x))\, \ell_j^0(F(x))\, \frac{\pi_1(x)\, \pi_2(x)}{\pi(x)}\, \mathrm{d}\Lambda(x), \quad i,j = 1, 2, 3, \dots, i \neq j. \tag{15}$$

It should be noted that, when at least one sample is heavy censored, a value of $\mathcal{L}_j$ can be 0 and a value of $\hat{\sigma}_j$ can be 0 as well. As a result, the symbol 0/0 occurs, which simply means that $C_j = 0$.

Note that the component C_1 is the standardized log-rank statistic and the corresponding two-sided asymptotically level α test has the form $\{C_1 < \Phi^{-1}(\alpha/2)\} \cup \{C_1 > \Phi^{-1}(1-\alpha/2)\}$, where Φ is the cumulative distribution function of the $N(0,\ 1)$ distribution. Therefore, its natural generalization can be written as the quadratic form

$$W_d = \sum_{j=1}^{d} C_j^2, \tag{16}$$

which measures deviations from $\mathcal{H}$ in the first d standardized Fourier coefficients (directions). Even though the components are (asymptotically) correlated, for simplicity and also keeping in mind their interpretation, we do not utile their correlations to define the statistic W_d. Now, the choice of the parameter d is a delicate question. The reason is that it has a great influence on the finite sample behavior of the resulting test. Therefore, we will select the number of summands in W_d in a data-driven way.

2.5 A new selection rule and the corresponding test

It is known that, in many cases, when the data are complete, a simplified Schwarz (BIC) selection rule

$$S = \min\left\{ d : 1 \le d \le d(n),\ W_d - d \log n \ge W_j - j \log n, j = 1, \dots, d(n) \right\}, \tag{17}$$

where $d(n)$ is a nondecreasing sequence of natural numbers, is a good device to choose d. For the evidence, see, for instance, Janic-Wróblewska and Ledwina (2000) or Wyłupek (2010), where the two-sample problem for complete data was considered. Therefore, we will use it.

However, under the alternative, when the source of deviations from the null model comes from the further directions, the penalty $d \log n$ for the dimension d can be too heavy, and, a better choice relies on the usage of the Akaike's penalty and the corresponding (AIC) selection rule

$$A = \min\big\{\, d \,:\, 1 \le d \le d(n),\ W_d - 2d \ge W_j - 2j,\ j = 1, \dots, d(n) \big\}. \tag{18}$$

Since, in practice, we do not know what kind of deviations from the null model occurs, if any, a compromise will be provided as a data-driven combination of those two rules. First, note that under the null hypothesis C_js should take small values, which find a reflection in a value of the indicator

$$J(n, c) = \mathbb{1}\Big(\max_{1 \le j \le d(n)} |C_j| \le \sqrt{c \log n} \Big), \quad c > 0. \tag{19}$$

The above will serve as a switch between the rule S, desirable under the null model assuring relatively small critical values, and the rule A, welcome under the alternative leading to the selection of higher dimensions, which results in more frequent rejections of the null hypothesis. The idea behind comes from Inglot and Ledwina (2006). The random penalty

$$\Pi_d = \begin{cases} d \log n, & \text{if } J(n, c) = 1, \\ 2d, & \text{if } J(n, c) = 0, \end{cases} \tag{20}$$

leads us to the rule

$$T = \min\big\{\, d \,:\, 1 \le d \le d(n),\ W_d - \Pi_d \ge W_j - \Pi_j,\ j = 1, \dots, d(n) \big\}, \tag{21}$$

which under a proper selection of the parameter c inherits good properties of the rules S and A. Usually, the choice of the parameter c c.a. 2 leads to a sensible behavior of the corresponding test. Finally, we will reject the null hypothesis for large values of the statistic W_T.

Basic asymptotics of the new test procedure based on W_T, as well as finite sample properties of their permutation counterpart are presented in the next section.

3 Theoretical results

3.1 Asymptotic outcomes

Let $\tau_L = \sup\{\, x : L(x) < 1 \,\}$ for any cumulative distribution function L. Put $H(x) = 1 - \eta [1 - F_1(x)][1 - G_1(x)] - (1 - \eta)[1 - F_2(x)][1 - G_2(x)]$, $x \in [0, \infty)$, and $H^{-1}(t) = \inf \{\, x : H(x) \ge t \,\}$, $t \in [0, 1)$. Define $t_{\max} = 1 - \eta[1 - F_1(\tau_{G_2})][1 - G_1(\tau_{G_2})] - (1 - \eta)[1 - F_2(\tau_{G_1})][1 - G_2(\tau_{G_1})]$. Obviously, $t_{\max} = 1$, if and only if, either $\max\{\tau_{F_1}, \tau_{F_2}\} \le \min\{\tau_{G_1}, \tau_{G_2}\}$ or $\tau_{G_1} = \tau_{G_2}$.

Theorem 1 *Assume that $d(n) = d$ is fixed and does not depend on n, while c is an arbitrary positive constant. If the null hypothesis $\mathcal{H}$ is true and $\tau \in (0, H^{-1}(t_{\max}))$, we have $C_j \xrightarrow{\mathcal{D}} N(0,1)$, $j = 1,2,3,\ldots, T \xrightarrow{\mathcal{P}} 1$, and $W_T \xrightarrow{\mathcal{D}} \chi_1^2$.*

Theorem 2 *Assume that $d(n) = d$ is fixed and does not depend on n, while c is an arbitrary positive constant. If $\tau \in (0, H^{-1}(t_{\max}))$, and*

$$\int_0^{\tau} \ell_j^0(F(x)) \frac{\pi_1(x)\,\pi_2(x)}{\pi(x)} \mathrm{d}[\Lambda_1(x) - \Lambda_2(x)] \neq 0, \qquad \text{for some} \qquad j \in \{1,\ldots,d\},$$

then $\lim_{n\to\infty} P(T \geq j) = 1$, while the test rejecting $\mathcal{H}$ for large values of W_T is consistent.

Assumption on τ is natural from the point of view of the statistical practice and is dictated by unboundedness of the Laguerre polynomials. Under finite sample sizes, to apply the test procedure, we need to set $d(n)$. Therefore, the fixed $d(n)$ is also sufficient in practical applications. An alternation of the parameter $d(n)$ in Theorem 1 does not change the asymptotic outcomes concerning the rule T and the statistic W_T. Since $d(n)$ is fixed, Theorem 2 asserts the consistency of the proposed test for an arbitrary large class of the alternatives set in advance. Our experience prompts that a selection $d(n) = d = 12$ allows one to detect any desirable departure from the null model, see Sect. 4.2.2 and Appendix G in Supplementary Material. Theorem 1 also implies that the test $\Phi_{n,\alpha} = \mathbb{1}(W_T > q_{\chi_1^2}(1-\alpha))$, where $q_{\chi_1^2}(1-\alpha)$ is the $(1-\alpha)$-quantile of the Chi-square distribution with one degree of freedom, is asymptotically distribution-free. However, under $\mathcal{H}$, the convergence of the test statistic W_T to the limiting χ_1^2 distribution is slow. Therefore, it is of no practical usage. Furthermore, under finite sample sizes, infinite-dimensional nuisance parameters G_1, G_2, do not make the corresponding test distribution-free any more. Therefore, to control the error of the first kind in this case, we treat our solution as a permutation test. See the algorithm in Appendix D in Supplementary Material. Such approach leads to an exact test under the restricted null hypothesis $\mathcal{H}_0$, i.e., $\mathcal{H}_0 : F_1 = F_2, G_1 = G_2$, and provides reasonable control of the power function, under $\mathcal{H}$, when $G_1 \neq G_2$. A formal justification of those facts is given below.

3.2 Permutation approach

Define the vectors of the order statistics $\boldsymbol{X}_{()} = (X_{1:n}, \ldots, X_{n:n})$ together with the vector of their concomitants $\boldsymbol{\Delta}_{()} = (\Delta_{1:n}, \ldots, \Delta_{n:n})$. Moreover, we introduce the vector of anti-ranks of $X_1, \ldots, X_n$, that is, $\boldsymbol{D} = (D_1, \ldots, D_n)$, where $X_{i:n} = X_{D_i}$. Now, we will implicitly start to consider the statistic W_T as a function of $(\boldsymbol{D};(\boldsymbol{X}_{()}, \boldsymbol{\Delta}_{()}))$. Given $(\boldsymbol{X}_{()}, \boldsymbol{\Delta}_{()}) = (\boldsymbol{x}, \boldsymbol{\delta})$, let $q_{W_T}(1-\alpha, \boldsymbol{x}, \boldsymbol{\delta})$ be the permutation $(1-\alpha)$-quantile of the $W_T(\boldsymbol{D};(\boldsymbol{x},\boldsymbol{\delta}))$ statistic. Then, $\Phi_{n,\alpha,\mathrm{perm}} = \mathbb{1}(W_T > q_{W_T}(1-\alpha, \boldsymbol{x}, \boldsymbol{\delta}))$ is the corresponding permutation test.

Lemma 1 *Under $\mathcal{H}_0$ (the restricted $\mathcal{H}$), the vector $\boldsymbol{D}$ is independent from $(\boldsymbol{X}_{()}, \boldsymbol{\Delta}_{()})$, and has a uniform distribution on the set of all permutations of the vector $(1,\ldots,n)$.*

Lemma 2

(i) *Under $\mathcal{H}_0$, the test $\Phi_{n,\alpha,\mathrm{perm}}$ is distribution-free.*
(ii) *Under $\mathcal{H}$, the tests $\Phi_{n,\alpha}$ and $\Phi_{n,\alpha,\mathrm{perm}}$ are asymptotically equivalent.*
(iii) *Under $\mathcal{H}$, the test $\Phi_{n,\alpha,\mathrm{perm}}$ is asymptotically distribution-free.*

Lemma 1 is well known and explains the above notation. The construction of $\Phi_{n,\alpha,\mathrm{perm}}$ implies (i) in Lemma 2. That statement means that the permutation test is an exact level α test under finite sample sizes when the restricted null hypothesis $\mathcal{H}_0 : F_1 = F_2, G_1 = G_2$ holds. However, the statement (ii) says that under the null hypothesis $\mathcal{H} : F_1 = F_2, G_1, G_2$ – nuisance parameters, the unconditional $\Phi_{n,\alpha}$ test and the conditional $\Phi_{n,\alpha,\mathrm{perm}}$ test are asymptotically equivalent. It means that under $\mathcal{H}$, the permutation $W_T(\boldsymbol{D};(\boldsymbol{x},\boldsymbol{\delta}))$ test statistic has the asymptotic Chi-square distribution with one degree of freedom. As a result, the permutation $\Phi_{n,\alpha,\mathrm{perm}}$ test is an asymptotic level α test in the problem $(\mathcal{H}, \mathcal{A})$. Such an outcome assures that, for sufficiently large sample sizes, we are able to control the Type I error under $F_1 = F_2$, $G_1 \neq G_2$, i.e., when $\mathcal{H}$ is true, but $\mathcal{H}_0$ does not hold. The results of the simulation study presented in Sect. 4.3.1 show that, in this case, the Type I error is controlled even under small sample sizes. This is an appealing feature, which each reasonable solution of the considered testing problem should have. The proofs of the properties (ii) and (iii), as well as Theorems 1 and 2 are relegated to Supplementary Material, Appendix C.

4 Numerical experiment

Since the problem is classical and its history is very long, to provide profound comparisons, we have tried to select, on the one hand, a few classical well known solutions which are frequently applied in practice, and, on the other hand, to include into the simulation study several recent, quite new, supposedly powerful and versatile procedures. The specially selected competitors are introduced in Sect. 4.1. Next, in Sect. 4.2, we consider equal censoring distributions, i.e., the case when $G_1 = G_2$, and investigate their behavior in comparison with the proposed test checking the errors of both kinds under small and moderate sample sizes. After that, in Sect. 4.3, we analyze the behavior of the tests under comparison when the censoring distributions are distinct. In that case, we also examine the errors of the first kind and investigate the powers. All computations have been carried out in `R` under the `seed 1` and at the significance level $\alpha = 0.05$. The parameters defining the test statistic W_T are set to be $c = 2$ and $d = 12$. In the simulations, we estimated the values of the power functions of the tests using 1000 Monte Carlo runs and 1000 permutation or bootstrap runs unless the asymptotic critical values have been used. See, for instance, Table 1, hereunder.

Table 1 Empirical errors of the first kind under $\mathcal{H}_0 : F_1 = F_2 \sim Exp(1), G_1 = G_2 \sim U(0, 2)$ against $n, n_1 = n_2, d = 12, c = 2$, $\alpha = 0.05$, 1000 MC runs, 1000 permutation/bootstrap runs. Errors multiplied by 100

n	26	50	100	150	200	250	300	
Test	Type I errors under $\mathcal{H}_0$							Method
G	6.6	4.7	6.0	4.8	4.1	4.4	5.2	Permutation
M	5.9	4.8	5.2	6.2	4.5	4.3	5.3	Permutation
R	6.1	5.2	5.4	5.8	4.4	4.5	5.0	Permutation
FH	6.0	4.5	4.8	5.8	4.6	4.9	4.4	Permutation
LW	5.1	5.3	3.9	5.7	4.3	4.4	5.7	Asymptotic
QS	5.7	6.6	6.3	5.6	4.7	4.4	5.1	Bootstrap
YP	6.5	6.9	6.7	6.5	5.6	5.1	6.0	Asymptotic
LY	7.0	7.6	6.2	7.0	6.3	5.8	6.7	Bootstrap
NPC	4.9	5.1	5.2	5.7	5.2	6.2	4.6	NPC technique
W_T	5.8	5.3	6.0	5.8	4.7	4.7	5.6	Permutation

4.1 Competitive tests

We included into the simulation study the most popular: Wilcoxon (Gehan 1965) and log-rank (Mantel 1966) tests, the classical: Renyi-type Kolmogorov–Smirnov (Gill 1980), Fleming and Harrington (1991, p. 257), $[p = 0, q = 1]$ statistics, as well as the selected recent developments in the field: the Lin and Wang (2004) proposal, the Qiu and Sheng (2008) test procedure, the Yang and Prentice (2010) solution, the Liu and Yin (2017) statistic, and the Arboretti et al. (2018) Tippet nonparametric combination test. In the figures, they are denoted as G, M, R, FH, LW, QS, YP, LY, and NPC, respectively. For their exact forms and additional computational comments, see Appendix E in Supplementary Material.

4.2 Equal censoring distributions

4.2.1 'Type I error's control, $G_1 = G_2$

In the first step, we examine the errors of the first kind of the tests under comparison.

Notwithstanding, all of the tests are asymptotically level α tests, under finite sample sizes, the control of the Type I error can sometimes be a challenge. A well-known remedy can be permutation approach or a bootstrap method. In our paper, the former is apply to the G, M, R, FH, and W_T test statistics, while the latter is exploited in the QS and LY procedures. The detailed algorithm is presented in Supplementary Material, Appendix D. By contrast, the Lin and Wang (2004) and Yang and Prentice (2010) solutions are investigated as the asymptotic tests, while the p value of the NPC test of Arboretti et al. (2018) is calculated using the nonparametric combination technique described in Supplementary Material, Appendix E. The errors obtained in such a way completes Table 1. Specifically, we display Type I errors under $F_1 = F_2$, which are fixed to be an exponential distribution with parameter 1, while the censoring distributions G_1, G_2 are equal and set to be the uniform ones on [0, 2]. This leads to c.a. 43% censored observations in each sample. The sample sizes in both groups are equal, i.e., $n_1 = n_2$. We increase the total sample size, $n = n_1 + n_2$, from 50 to 300, by 50, and investigate the errors. Furthermore, we include the case $n_1 = n_2 = 13$, which corresponds to the sample sizes considered in the real data example, see Sect. 5.2, hereunder. It can be seen that the actual Type I error of the YP and LY solutions is a little bit inflated, while all the remaining tests keep the nominal level well. We also verified the cases when the underlying distribution $F_1 = F_2$ is log-normal or Weibull one. Since the outcomes were similar, we do not present them.

4.2.2 Power comparisons, $G_1 = G_2$

In the second step, we examine powers analyzing the behavior of the tests considered under nine examples covering a large set of possible deviations from the null model. Most of the cases either have been frequently analyzed in the literature or are their simple modifications. Cf. Fleming et al. (1987), Letón and Zuluaga (2005), Fleming et al. (1980), Schumacher (1984), Pepe and Fleming (1989), Lee et al. (2008), Lu et al. (1994), Li et al. (1996), Lee (1996), Chi and Tsai (2001), Wu and Gilbert (2002), Lin and Wang (2004), Lee (2007), Qiu and Sheng (2008), Kraus (2009), Martínez-Camblor (2010), Chang et al. (2012), Brendel et al. (2014), Callegaro and Spiessens (2017), Hsieh and Chen (2017), Liu and Yin (2017), and Arboretti et al. (2018), among others. They include models with proportional, early, middle, late, crossing, and subtle difference hazard rates. All the alternatives are expressed in terms of the hazard functions.

Description of the alternatives

Example 1 $\lambda_1(t) = 1$, $\lambda_2(t) = 1.5$.

Example 2 $\lambda_1(t) = 2 \cdot \mathbb{1}_{[0,0.2)}(t) + 0.75 \cdot \mathbb{1}_{[0.2,0.4)}(t) + \mathbb{1}_{[0.4,+\infty)}(t)$,
$\lambda_2(t) = 0.75 \cdot \mathbb{1}_{[0,0.2)}(t) + 2 \cdot \mathbb{1}_{[0.2,0.4)}(t) + \mathbb{1}_{[0.4,+\infty)}(t)$.

Example 3 $\lambda_1(t) = 2 \cdot \mathbb{1}_{[0,0.1)}(t) + 3 \cdot \mathbb{1}_{[0.1,0.4)}(t) + 0.75 \cdot \mathbb{1}_{[0.4,0.7)}(t) + \mathbb{1}_{[0.7,+\infty)}(t)$,
$\lambda_2(t) = 2 \cdot \mathbb{1}_{[0,0.1)}(t) + 0.75 \cdot \mathbb{1}_{[0.1,0.4)}(t) + 3 \cdot \mathbb{1}_{[0.4,0.7)}(t) + \mathbb{1}_{[0.7,+\infty)}(t)$.

Example 4 $\lambda_1(t) = 2 \cdot \mathbb{1}_{[0,0.2)}(t) + 3 \cdot \mathbb{1}_{[0.2,0.6)}(t) + 0.75 \cdot \mathbb{1}_{[0.6,0.9)}(t) + \mathbb{1}_{[0.9,+\infty)}(t)$,
$\lambda_2(t) = 2 \cdot \mathbb{1}_{[0,0.2)}(t) + 0.75 \cdot \mathbb{1}_{[0.2,0.6)}(t) + 3 \cdot \mathbb{1}_{[0.6,0.9)}(t) + \mathbb{1}_{[0.9,+\infty)}(t)$.

Example 5 $\lambda_1(t) = 2 \cdot \mathbb{1}_{[0,0.6)}(t) + 4 \cdot \mathbb{1}_{[0.6,+\infty)}(t)$, $\lambda_2(t) = 2 \cdot \mathbb{1}_{[0,0.6)}(t) + 0.4 \cdot \mathbb{1}_{[0.6,+\infty)}(t)$.

Example 6 $\lambda_1(t) = 2 \cdot \mathbb{1}_{[0,0.5)}(t) + 4 \cdot \mathbb{1}_{[0.5,+\infty)}(t)$,
$\lambda_2(t) = 2 \cdot \mathbb{1}_{[0,0.1)}(t) + 3 \cdot \mathbb{1}_{[0.1,0.4)}(t) + 0.75 \cdot \mathbb{1}_{[0.4,0.7)}(t) + \mathbb{1}_{[0.7,+\infty)}(t)$.

Example 7 $\lambda_1(t) = 1$, $\lambda_2(t) = t + 0.3$.

Example 8 $\lambda_1(t) = 1$, $\lambda_2(t) = 4t \cdot \mathbb{1}_{[0,0.7)}(t) + [8.4 - 8t] \cdot \mathbb{1}_{[0.7,1)}(t) + 0.4 \cdot \mathbb{1}_{[1,+\infty)}(t)$.

Example 9 $\lambda_1(t) = 1 + 0.6\cos(7t)$, $\lambda_2(t) = 1 - 0.6\cos(7t)$.

In each case, the censoring distributions in both samples are equal, $G_1 = G_2$, and set to be the uniform ones on [0, 2]. It leads to the so-called moderate censoring, which seems to be the most frequent in statistical practice.

The first alternative is *Proportional Hazard Model*, the second one is *Early Hazard Difference*, the third one can be labeled by *Middle/Early Hazard Difference*, the fourth *Middle Hazard Difference*, the fifth *Late Hazard Difference*. The remaining alternatives are *Crossing Hazard Difference*, whereas the last of them can be treated as a *Subtle Hazard Difference*.

To calculate powers, we fix the alternative. The sample sizes in both groups are equal, $n_1 = n_2$. We increase the total sample size, $n = n_1 + n_2$, from 50 to 300, by 50, and investigate the powers. The results are depicted in Figs. 1, 2, and 3. The significance level $\alpha = 0.05$.

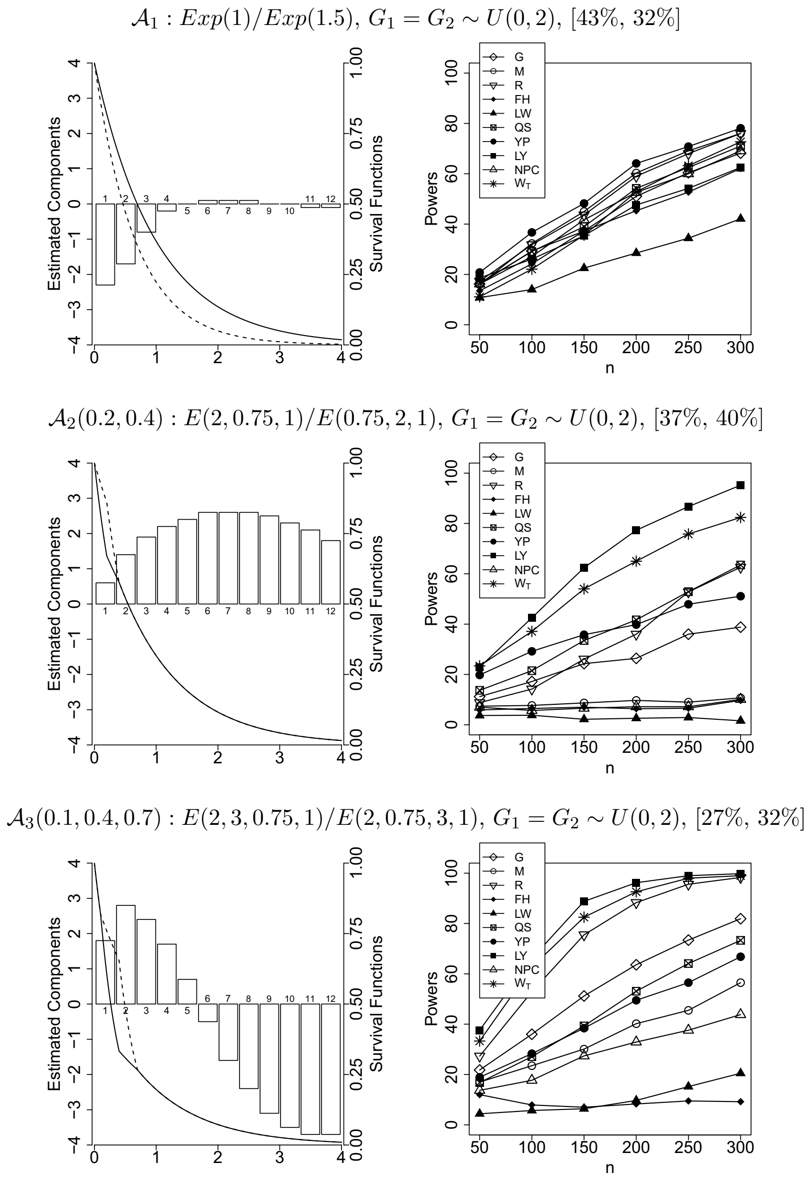
Since the density function can be expressed as the function of the hazard rate $\lambda(t)$

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(s)\mathrm{d}s\right],$$

the von Neumann accept/reject algorithm allows one to generate the observations from the above models with ease.

The notation in the figures is as follows. The first alternative is denoted as $\mathcal{A}_1$: Exp(1)/Exp(1.5), while the last one as $\mathcal{A}_9$: Cos(7, 0.6)/Cos(7, –0.6). The notation of the remaining alternatives is similar and will be explained on the basis of the second example: $\mathcal{A}_2(0.2, 0.4) : E(2, 0.75, 1)/E(0.75, 2, 1)$. The string (0.2, 0.4) means that the half-line $[0, \infty)$ has been partitioned into the intervals [0, 0.2), [0.2, 0.4), $[0.4, \infty)$, while the values of the hazard functions on the successive intervals are defined as 2, 0.75, 1, and 0.75, 2, 1, in the first and second groups, respectively, (cf. the definition of Example 2). Note that the second distribution in the alternative $\mathcal{A}_7$ is the linear hazard with the coefficients 0.3, 1, i.e., $\lambda_2(t) = t + 0.3$. A similar description is used in the case of $\mathcal{A}_8$. However, this time, the linear hazard is different on the intervals [0, 0.7), [0.7, 1), $[1, \infty)$. Cf. the definition of Example 8. In the label of each alternative, one can also find the information on the censoring distributions and the percentages of the censored observations in the first and second groups, respectively. Each figure consists of two panels. The right panel presents the powers of the tests against the total sample size n. The left panel shows the survival functions in both groups together with the average values of the first 12 components, C_js, calculated under 1000 MC runs, when $n = 200$. Such a manner of the presentation demonstrates real differences between the consecutive alternatives and enables one to investigate the deviations from the null model in terms of those objects.

Under the first alternative, the first component is dominating, which results well behavior of the Mantel (1966) test. Slightly larger powers of the Yang and Prentice (2010) procedure are caused by the larger errors of the first kind of that test under the null model, see Table 1. Also, the Renyi-type Kolmogorov–Smirnov test of Gill



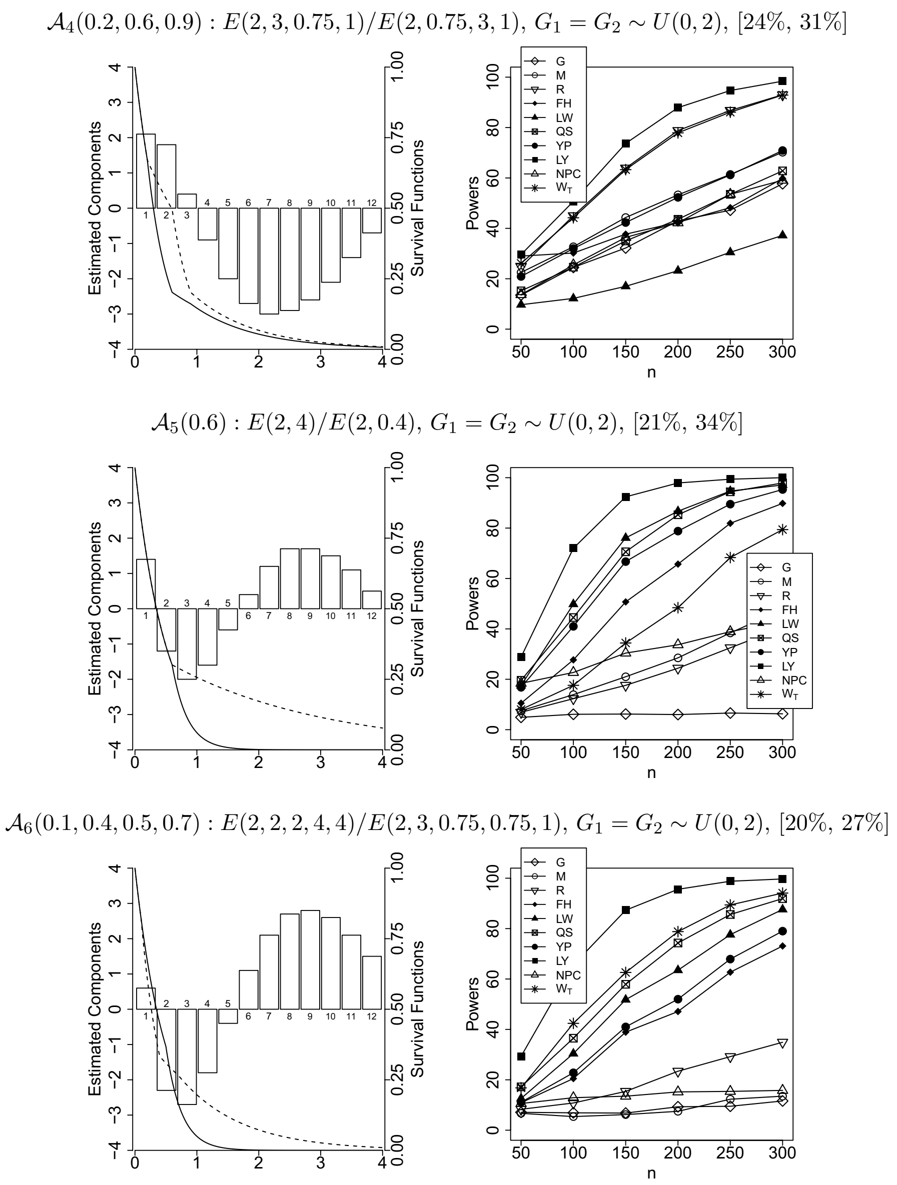
$\mathcal{A}_1 : Exp(1)/Exp(1.5),\ G_1 = G_2 \sim U(0,2),\ [43\%,\ 32\%]$

$\mathcal{A}_2(0.2, 0.4) : E(2, 0.75, 1)/E(0.75, 2, 1),\ G_1 = G_2 \sim U(0,2),\ [37\%,\ 40\%]$

$\mathcal{A}_3(0.1, 0.4, 0.7) : E(2, 3, 0.75, 1)/E(2, 0.75, 3, 1),\ G_1 = G_2 \sim U(0,2),\ [27\%,\ 32\%]$

Fig. 1 Left panel: Survival functions S_1 (—), S_2 (- -), in the first and second samples, respectively. The bars represent the average values of the components C_js, $j = 1, \dots, 12$, under $n = 200$. Right panel: Empirical powers against n; $\alpha = 0.05$; $n_1 = n_2$; $d = 12$; $c = 2$. Based on 1000 MC runs and 1000 permutation/bootstrap runs. Powers multiplied by 100

(1980) is the leading test. Except the Lin and Wang (2004) procedure, the remaining tests do not lose too much to the best ones. The second alternative is difficult to detect for the Mantel (1966) test because the first component is the smallest one.



$\mathcal{A}_4(0.2, 0.6, 0.9) : E(2, 3, 0.75, 1)/E(2, 0.75, 3, 1)$, $G_1 = G_2 \sim U(0, 2)$, [24%, 31%]

$\mathcal{A}_5(0.6) : E(2, 4)/E(2, 0.4)$, $G_1 = G_2 \sim U(0, 2)$, [21%, 34%]

$\mathcal{A}_6(0.1, 0.4, 0.5, 0.7) : E(2, 2, 2, 4, 4)/E(2, 3, 0.75, 0.75, 1)$, $G_1 = G_2 \sim U(0, 2)$, [20%, 27%]

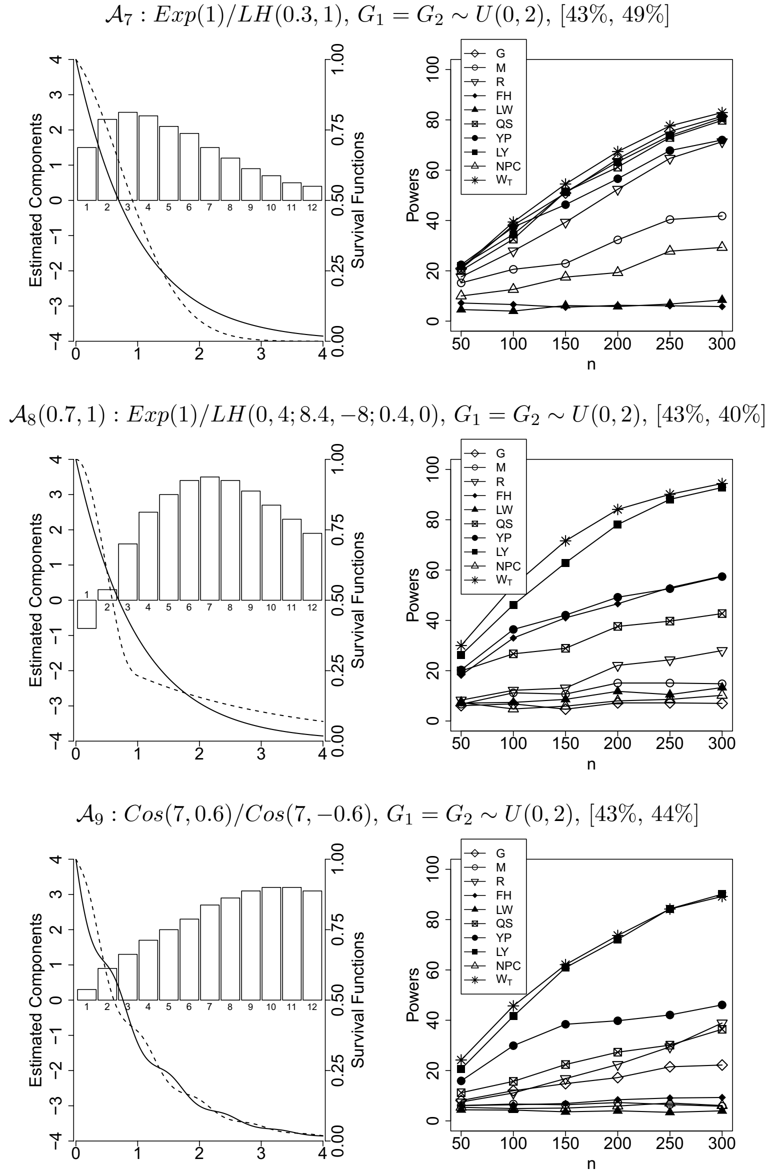
Fig. 2 Left panel: Survival functions S_1 (—), S_2 (- -), in the first and second samples, respectively. The bars represent the average values of the components C_js, $j = 1, \dots, 12$, under $n = 200$. Right panel: Empirical powers against n; $\alpha = 0.05$; $n_1 = n_2$; $d = 12$; $c = 2$. Based on 1000 MC runs and 1000 permutation/bootstrap runs. Powers multiplied by 100

The Fleming and Harrington (1991) test, the NPC test of Arboretti et al. (2018), as well as the Lin and Wang (2004) solution also behave badly. The Gehan (1965) test works better, but still weakly. The R test and the recent developments based on the

QS and YP statistics are inefficient in detection of this type of disturbance from the null model, whereas the advantage of the new solution can be clearly seen. The reason is a large number of the significant components in combination with the flexible model selection. The competitive (Liu and Yin 2017) procedure works even better than our proposal, but analyzing the behavior under the alternative, we should also have in mind what is going on under $\mathcal{H}$. The third alternative is similar to the previous one, but there, middle hazard difference also occurs. In those cases, the worst solutions, among the competitive tests, are the Fleming and Harrington (1991) and Lin and Wang (2004) tests. The Arboretti et al. (2018), Mantel (1966), Yang and Prentice (2010), Qiu and Sheng (2008), and Gehan (1965) tests have moderate powers. In this case, the Renyi-type test of Gill (1980) works very well. The new test is better, while the power of the Liu and Yin (2017) procedure is even slightly greater. Under the fourth alternative, the ordering of the best three tests is the same as in the previous case, whereas the remaining solutions lose much. In the next example, the Gehan (1965) test completely breaks down. The powers of the Renyi-type Gill (1980), Mantel (1966), and NPC (Arboretti et al. 2018) tests are small. This time the Fleming and Harrington (1991) test supposedly optimal has the moderate power. The behavior of the Liu and Yin (2017) test is the best. The Lin and Wang (2004), Qiu and Sheng (2008), as well as Yang and Prentice (2010) procedures work equally well. In this example, the new test loses the power because the components are less informative. Under the sixth alternative two the most popular classical solutions, that is, the Gehan (1965) and Mantel (1966) tests completely break down. The optimal situation for the QS solution results in their high power. Nevertheless, the new test is slightly better, while the LY test outperforms it. The next two alternatives were studied by Liu and Yin (2017). The crossing hazard difference results in very good behavior of the new test, which is the best one. The last alternative demonstrates subtle deviations from the null model. Since the first component is the smallest one, the Mantel (1966) test poorly works, just like the Fleming and Harrington (1991), Arboretti et al. (2018), and Lin and Wang (2004) solutions. The behavior of the Gehan (1965) test is slightly better, but still insufficient. The same concerns the R, QS, and YP procedures. Since the larger the number of the component the larger their magnitude, the new solution has high power and together with the LY test significantly outperforms the remaining procedures.

4.3 Different censoring distributions

In this section, we consider the same configurations of the distributions of the survival times as in Sect. 4.2; however, the censoring distributions are different. In each scenario, we set that the censoring distribution in the first group is uniform on [0, 1.5], while the censoring distribution in the second group is uniform on [0, 2.5]. In Sect. 4.3.1, we check how the solutions under comparison control the error of the first kind, while in Sect. 4.3.2, we examine their behavior under the alternatives.



$\mathcal{A}_7 : Exp(1)/LH(0.3, 1),\ G_1 = G_2 \sim U(0, 2),\ [43\%,\ 49\%]$

$\mathcal{A}_8(0.7, 1) : Exp(1)/LH(0, 4; 8.4, -8; 0.4, 0),\ G_1 = G_2 \sim U(0, 2),\ [43\%,\ 40\%]$

$\mathcal{A}_9 : Cos(7, 0.6)/Cos(7, -0.6),\ G_1 = G_2 \sim U(0, 2),\ [43\%,\ 44\%]$

Fig. 3 Left panel: Survival functions S_1 (—), S_2 (- -), in the first and second samples, respectively. The bars represent the average values of the components C_js, $j = 1, \dots, 12$, under $n = 200$. Right panel: Empirical powers against n; $\alpha = 0.05$; $n_1 = n_2$; $d = 12$; $c = 2$. Based on 1000 MC runs and 1000 permutation/bootstrap runs. Powers multiplied by 100

4.3.1 'Type I error's control, $G_1 \neq G_2$

In Table 2, we present the errors of the first kind of the investigated procedures under $F_1 = F_2$, which are fixed to be an exponential distribution with parameter 1, and different censoring distributions, specifically, $G_1 \sim U(0, 1.5)$, $G_2 \sim U(0, 2.5)$. This leads to c.a. 52% and 37% censored observations in the first and second samples, respectively. The sample sizes in both groups are equal, i.e., $n_1 = n_2$. We increase the total sample size, $n = n_1 + n_2$, from 50 to 300, by 50, and investigate the errors. Furthermore, we include the case $n_1 = n_2 = 13$, which corresponds to the sample sizes considered in the real data example, see Sect. 5.2, hereunder.

It can be seen that the actual Type I error of the Liu and Yin (2017) test significantly exceeds the significance level $\alpha = 5\%$. It makes that solution an unsafe procedure. We also verified the cases when the underlying distribution $F_1 = F_2$ is log-normal or Weibull one. The outcomes were similar. Therefore, we exclude it from further comparisons. In the case of the NPC test of Arboretti et al. (2018), the situation is even more dramatic. The Type I error grows together with the sample sizes and already exceeds 95% when $n_1 = n_2 = 100$. This is the reason why we also exclude that solution from the further comparisons.

4.3.2 Power comparisons, $G_1 \neq G_2$

We repeat the experiment from Sect. 4.2.2 changing the censoring distributions into different ones. The obtained results are depicted in Figures S1–S3 in Supplementary Material, Appendix G. Commenting very briefly on the outcomes, we can say that the ordering of the tests under comparison is similar to the case when the censoring distributions were equal. Therefore, the advantages of the new solution have been shown.

5 Real data examples

In this section, we analyze two medical real datasets.

5.1 Locally unresectable gastric cancer: chemotherapy versus chemotherapy with radiation therapy

We analyze the data frequently investigated in the literature. The dataset represents $n = 90$ individuals divided into two equal groups of $n_1 = n_2 = 45$ patients. There are two and six censored observations in the first and second samples, respectively. The problem concerns comparing chemotherapy with combined chemotherapy and radiation therapy, in the treatment for locally unresectable gastric cancer, and is based on the work of the Gastrointestinal Tumor Study Group (1982). The data is easily accessible in the R package YPmodel.

Table 2 Empirical errors of the first kind under $\mathcal{H}: F_1 = F_2 \sim Exp(1)$, $G_1 \sim U(0, 1.5)$, $G_2 \sim U(0, 2.5)$ against $n, n_1 = n_2, d = 12, c = 2$, $\alpha = 0.05$, 1000 MC runs, 1000 permutation/bootstrap runs. Errors multiplied by 100

n	26	50	100	150	200	250	300	
Test	Type I errors under $\mathcal{H}$							Method
G	5.5	4.8	5.7	4.9	4.6	4.9	5.0	Permutation
M	6.0	5.0	5.2	5.4	4.3	5.1	4.7	Permutation
R	6.1	5.1	5.5	5.5	4.7	4.5	4.7	Permutation
FH	5.4	6.3	6.8	6.0	5.5	6.8	6.2	Permutation
LW	4.7	5.1	4.8	6.5	5.9	5.0	5.2	Asymptotic
QS	4.7	4.9	6.1	5.1	4.5	3.7	4.6	Bootstrap
YP	6.6	6.3	6.7	6.1	5.1	4.9	4.8	Asymptotic
LY	8.3	6.5	8.1	8.7	10.0	9.6	10.8	Bootstrap
NPC	10.7	30.8	60.9	85.0	95.5	98.0	99.3	NPC technique
W_T	4.4	6.0	5.9	5.7	4.8	5.5	4.8	Permutation

Figure 4 presents the estimated Kaplan–Meier survival curves as well as the values of the first 12 components C_js.

It can be seen that the estimated survival curves are different and cross each other. Undeniable evidence in favor of the alternative $\mathcal{A}$, at the asymptotic nominal level, is provided by the second, third, and fourth components. Nevertheless, we calculated their permutation p values. The results for all the components considered are displayed in Table 3.

Table 4 contains the p values of the tests under consideration calculated using either permutation, bootstrap, or asymptotic approach. A selection of a method is the same as in the conducted simulation study, cf. Tables 1 and 2.

Two classical solutions, that is, the Mantel (1966) and Renyi-type (Gill 1980) tests, do not reject the null hypothesis. The Arboretti et al. (2018) procedure is also insignificant. The p values of the Gehan (1965), as well as Fleming and Harrington

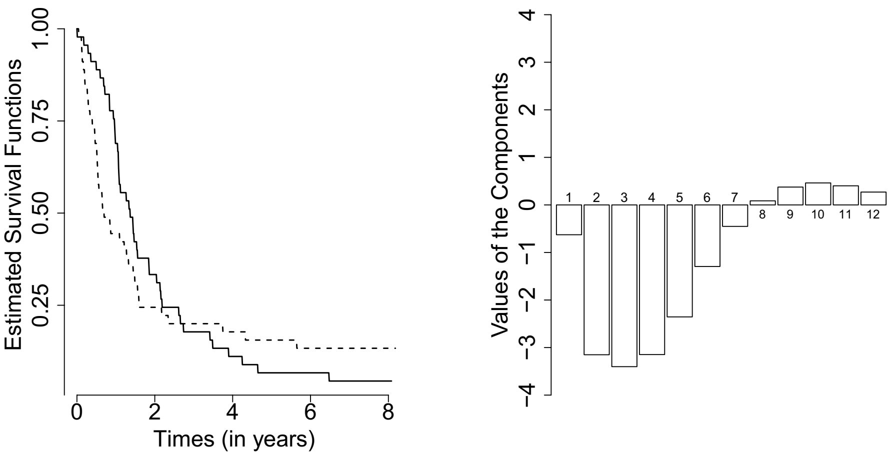


Fig. 4 Left panel: Estimated survival functions $\hat{S}_1$ (—), $\hat{S}_2$ (- -), in the first and second samples, respectively. Right panel: The bars represent the values of the components C_j, $j = 1, \ldots, 12$

Table 3 Empirical p values of the components C_j, $j = 1, \dots, 12$, $n_1 = n_2 = 45$. Based on 10,000 permutation runs

	C_j											
j	1	2	3	4	5	6	7	8	9	10	11	12
p value	0.553	0.001	0.001	0.002	0.016	0.195	0.649	0.933	0.714	0.645	0.686	0.788

Table 4 Empirical p values of the tests under consideration, $n_1 = n_2 = 45, d = 12, c = 2$. Based on 10,000 permutation/bootstrap runs

Test	G	M	R	FH	LW	QS	YP	LY	NPC	W_T
p value	0.0465	0.5530	0.2825	0.0501	0.2788	0.0257	0.0304	0.0026	0.8687	0.0114

(1991) tests are close to the nominal significance level $\alpha = 0.05$. The remaining tests, except the asymptotic (Lin and Wang 2004) solution, reject the null hypothesis. The new proposal is one of the two tests, which provide the strongest evidence against the null model with p value 0.0114.

5.2 Advanced ovarian carcinoma: cyclophosphamide versus cyclophosphamide plus adriamycin

Here, we analyze the data investigated in Edmonson et al. (1979). The patients were treated using either cyclophosphamide alone (1 g/m^2) or cyclophosphamide (500 mg/m^2) plus adriamycin (40 mg/m^2). The data are easily accessible in the `R` package `survival` with the label `ovarian`. There are $n_1 = n_2 = 13$ observations in each sample. Six observations in the first group and eight in the second one are censored.

Figure 5 presents the estimated Kaplan–Meier survival curves, as well as the values of the first 12 components C_js.

Large values of the components C_5, C_6, C_7 suggest the alternative $\mathcal{A}$ at the asymptotic nominal level smaller than 0.01. The permutation p values of all the components considered are presented in Table 5.

Table 6 contains the p values of the tests under consideration calculated using either permutation, bootstrap, or asymptotic approach. A selection of a method is the same as in the conducted simulation study, cf. Tables 1 and 2.

The classical solutions, that is, the Gehan (1965), Mantel (1966), Renyi-type (Gill 1980), as well as Fleming and Harrington (1991) tests do not reject the null hypothesis with large p values. The same concerns the Lin and Wang (2004), Qiu and Sheng (2008), and Arboretti et al. (2018) tests. The remaining competitive procedures are also insignificant at the significance level $\alpha = 0.05$. The only test which validates the hypothesis $\mathcal{H}$ in favor of the alternative $\mathcal{A}$ with p value 0.0170 is the new proposal.

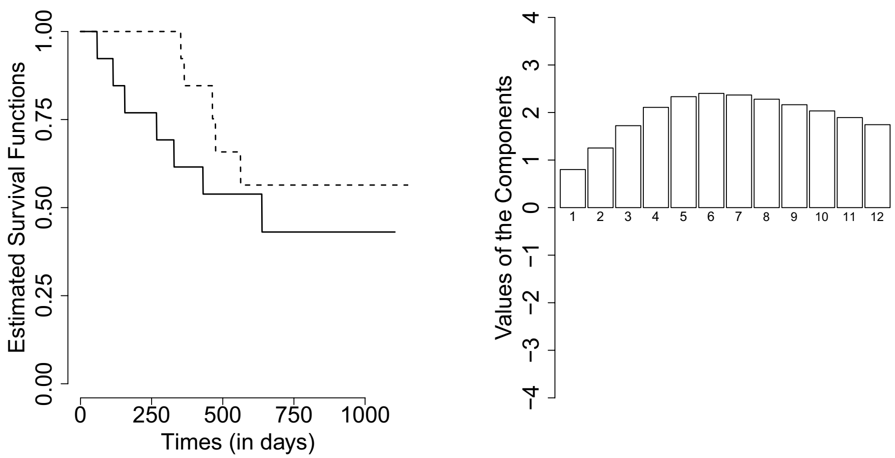


Fig. 5 Left panel: estimated survival functions $\hat{S}_1$ (—), $\hat{S}_2$ (- -), in the first and second samples, respectively. Right panel: the bars represent the values of the components C_j, $j = 1, \dots, 12$

Table 5 Empirical p values of the components C_j, $j = 1, \dots, 12$, $n_1 = n_2 = 13$. Based on 10,000 permutation runs

	C_j											
j	1	2	3	4	5	6	7	8	9	10	11	12
p value	0.455	0.229	0.090	0.032	0.013	0.007	0.009	0.013	0.022	0.033	0.048	0.074

Table 6 Empirical p values of the tests under consideration, $n_1 = n_2 = 13$, $d = 12, c = 2$. Based on 10,000 permutation/bootstrap runs

Test	G	M	R	FH	LW	QS	YP	LY	NPC	W_T
p value	0.2102	0.4548	0.6154	0.1477	0.8291	0.1362	0.0785	0.0827	0.4701	0.0170

6 Discussion

The paper has been aimed at proposing a solution, which controls the error of the first kind and is sensitive in detection of a wide range of the alternatives, generalizing the popular log-rank test, as well as providing a new source of characterization of the discrepancies from the null model.

The first issue has been addressed by a permutation idea resulting in an exact test under finite sample sizes and the restricted null hypothesis $\mathcal{H}_0$, as well as asymptotic distribution-freeness under general $\mathcal{H}$. An employment of the Laguerre polynomials together with the weighted log-rank statistics and building on their basis a kind of an efficient score statistic in combination with a proper model selection realizes the remaining goals. The relation (11) shows that usage of that system presumably leads to the one and only natural generalization of the log-rank test, where the Fourier coefficients defined corresponding to the consecutive polynomials allow one to characterize the source of discrepancies from the null model, whereas the newly proposed test procedure enables one to detect them. Such an approach leads to the flexible solution which competes well with the best tests and simultaneously is a safe procedure controlling the Type I error at the assumed significance level α.

Acknowledgements The author is grateful to the Associate Editor and Reviewer for the comments which led to the improvement in the presentation. He also thanks A. Callegaro, Y-M. Chang, V. Garès, J-J. Hsieh, and L. Salmaso for sending the copies of their papers. The research has been supported by the Grant 1407/M/IM/15 indirectly awarded by the Polish Ministry of Science and Higher Education. Calculations have been carried out in Wrocław Centre for Networking and Supercomputing (http://www.wcss.wroc.pl) under Grant No. 199. The cooperation of the Centre is gratefully acknowledged.

References

Arboretti, R., Fontana, R., Pesarin, F., Salmaso, L. (2018). Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. *Statistical Methods in Medical Research*, *27*, 3739–3769.

Arboretti, R. G., Bolzan, M., Campigotto, F., Corain, L., Salmaso, L. (2010). Combination-based permutation testing in survival analysis. *Quaderni di Statistica*, *12*, 15–38.

Behnen, K., Neuhaus, G. (1983). Galton's test as a linear rank test with estimated scores and its local asymptotic efficiency. *Annals of Statistics*, *11*, 588–599.

Brendel, M., Janssen, A., Mayer, C.-D., Pauly, M. (2014). Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, *41*, 742–761.

Callegaro, A., Spiessens, B. (2017). Testing treatment effect in randomized clinical trials with possible non-proportional hazards. *Statistics in Biopharmaceutical Research*, *9*, 204–211.

Chang, Y.-M., Chen, C.-S., Shen, P.-S. (2012). A jackknife-based versatile test for two-sample problems with right-censored data. *Journal of Applied Statistics*, *39*, 267–277.

Chauvel, C., O'Quigley, J. (2014). Tests for comparing estimated survival functions. *Biometrika*, *101*, 535–552.

Chi, Y., Tsai, M.-H. (2001). Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan–Meier statistics. *Communications in Statistics*: *Simulation and Computation*, *30*, 743–759.

Darilay, A. T., Naranjo, J. D. (2011). A pretest for using logrank or Wilcoxon in the two-sample problem. *Computational Statistics and Data Analysis*, *55*, 2400–2409.

Edmonson, J. H., Fleming, T. R., Decker, D. G., Malkasian, G. D., Jorgensen, E. O., Jefferies, J. A., Webb, M. J., Kvols, L. K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer Treatment Reports*, *63*, 241–247.

Efron, B. (1967). The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *4*, 831–853.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, *76*, 312–319.

Fleming, T. R., Harrington, D. P. (1991). *Counting processes and survival analysis*. New York: Wiley.

Fleming, T. R., Harrington, D. P., O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association*, *82*, 312–320.

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*, *36*, 607–625.

Garès, V., Andrieu, S., Dupuy, J.-F., Savy, N. (2017). On the Fleming–Harrington test for late effects in prevention randomized controlled trials. *Journal of Statistical Theory and Practice*, *11*, 418–435.

Gastrointestinal Tumor Study Group. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer*, *49*, 1771–1777.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, *52*, 203–223.

Gill, R. D. (1980). Censoring and stochastic integrals. *Mathematical Centre Tracts* 124. Amsterdam: Mathematisch Centrum. http://oai.cwi.nl/oai/asset/11499/11499A.pdf.

Harrington, D. P., Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, *69*, 553–566.

Hsieh, J.-J., Chen, H.-Y. (2017). A testing strategy for two crossing survival curves. *Communications in Statistics-Simulation and Computation*, *46*, 6685–6696.

Inglot, T., Ledwina, T. (2006). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and Its Applications*, *417*, 124–133.

Janic-Wróblewska, A., Ledwina, T. (2000). Data driven rank test for two-sample problem. *Scandinavian Journal of Statistics*, *27*, 281–297.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Koziol, J. A. (1978). A two sample Cramér–von Mises test for randomly censored data. *Biometrical Journal*, *20*, 603–608.

Koziol, J. A., Jia, Z. (2014). Weighted Lin–Wang tests for crossing hazards. *Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/2014/643457.

Kraus, D. (2009). Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. *Journal of Statistical Planning and Inference*, *139*, 3559–3569.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, *52*, 721–725.

Lee, S.-H. (2007). On the versatility of the combination of the weighted log-rank statistics. *Computational Statistics and Data Analysis*, *51*, 6557–6564.

Lee, S.-H., Lee, E.-J., Omolo, B. O. (2008). Using integrated weighted survival difference for the two-sample censored data problem. *Computational Statistics and Data Analysis*, *52*, 4410–4416.

Letón, E., Zuluaga, P. (2005). Relationships among tests for censored data. *Biometrical Journal*, *47*, 377–387.

Li, G., Tiwari, R. C., Wells, M. T. (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association*, *91*, 689–698.

Lin, Ch.-Y., Kosorok, M. R. (1999). A general class of function-indexed nonparametric tests for survival analysis. *Annals of Statistics*, *27*, 1722–1744.

Lin, X., Wang, H. (2004). A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal*, *46*, 489–496.

Liu, Y., Yin, G. (2017). Partitioned log-rank tests for the overall homogeneity of hazard rate functions. *Lifetime Data Analysis*, *23*, 400–425.

Lu, H. H. S., Wells, M. T., Tiwari, R. C. (1994). Inference for shift functions in the two-sample problem with right-censored data: With applications. *Journal of the American Statistical Association*, *89*, 1017–1026.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, *50*, 163–170.

Martínez-Camblor, P. (2010). Comparing *k*-independent and right censored samples based on the likelihood ratio. *Computational Statistics*, *25*, 363–374.

Neuhaus, G. (2000). A method of constructing rank tests in survival analysis. *Journal of Statistical Planning and Inference*, *91*, 481–497.

O'Quigley, J. (2003). Khalamadze-type graphical evaluation of the proportional hazard assumption. *Biometrika*, *90*, 577–584.

Pepe, M. S., Fleming, T. R. (1989). Weighted Kaplan–Meier statistics: A class of distance tests for censored survival data. *Biometrics*, *45*, 497–507.

Pepe, M. S., Fleming, T. R. (1991). Weighted Kaplan–Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society, Series B*, *53*, 341–352.

Pesarin, F., Salmaso, L. (2010). *Permutation tests for complex data*: *Theory, applications and software*. Chichester: Wiley.

Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, *135*, 185–206.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, *65*, 167–179.

Qiu, P., Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society, Series B*, *70*, 191–208.

Schumacher, M. (1984). Two-sample tests of Cramér–von Mises- and Kolmogorov–Smirnov-type for randomly censored data. *International Statistical Review*, *52*, 263–281.

Tarone, R. E., Ware, J. (1977). On distribution-free test for equality of survival distributions. *Biometrika*, *64*, 156–160.

Wu, L., Gilbert, P. B. (2002). Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics*, *58*, 997–1004.

Wyłupek, G. (2010). Data-driven *k*-sample tests. *Technometrics*, *52*, 107–123.

Yang, S., Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, *92*, 1–17.

Yang, S., Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, *66*, 30–38.

Zhang, J., Wu, Y. (2007). *k*-sample tests based on the likelihood ratio. *Computational Statistics and Data Analysis*, *51*, 4682–4691.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.