



## More powerful logrank permutation tests for two-sample survival data

Marc Ditzhaus, Sarah Friedrich

### Angaben zur Veröffentlichung / Publication details:

Ditzhaus, Marc, and Sarah Friedrich. 2020. "More powerful logrank permutation tests for two-sample survival data." *Journal of Statistical Computation and Simulation* 90 (12): 2209–27. <https://doi.org/10.1080/00949655.2020.1773463>.

# More powerful logrank permutation tests for two-sample survival data

Marc Ditzhaus <sup>a</sup> and Sarah Friedrich <sup>b</sup>

<sup>a</sup>Department of Statistics, TU Dortmund University, Dortmund, Germany; <sup>b</sup>Department of Medical Statistics, University Medical Centre Goettingen, Göttingen, Germany

## ABSTRACT

Weighted logrank tests are a popular tool for analysing right-censored survival data from two independent samples. Each of these tests is optimal against a certain hazard alternative, for example, the classical logrank test for proportional hazards. But which weight function should be used in practical applications? We address this question by a flexible combination idea leading to a testing procedure with broader power. Besides the test's asymptotic exactness and consistency, its power behaviour under local alternatives is derived. All theoretical properties can be transferred to a permutation version of the test, which is even finitely exact under exchangeability and showed a better finite sample performance in our simulation study. The procedure is illustrated in a real data example.

## KEYWORDS

Local alternatives; right censoring; two-sample survival model; weighted logrank test

## 1. Introduction

Deciding whether there is a difference between two treatments is only one example of the variety of two-sample problems. Within the right censoring survival set-up the classical logrank test, first proposed by [1,2], is very popular in practice. It is well known that the logrank test is optimal for proportional hazard alternatives but may lead to wrong decisions when the relationship of the hazards is time-dependent. Adding a weight function, we obtain optimal tests for other kinds of alternatives. These so-called weighted logrank tests are well studied in the literature; see [3–9]. However, no weighted logrank test is a so-called omnibus test, i.e. a consistent test for all alternatives. Depending on the pre-chosen weight, the corresponding logrank test is consistent for specific alternatives, details can be found in Section 2. This is in line with the result of [10] that any test has only reasonable power for a finite-dimensional subspace of the nonparametric two-sample alternative. A lot of effort was made to obtain tests with a good performance for a huge class of alternatives. Fleming et al. [11] suggested a supremum version of the logrank test with the purpose of power robustification. The funnel test of [12] had the same aim, to lose a little power for some alternatives and to gain a substantial power amount for other alternatives

in reverse. Lai and Ying [13] proposed to estimate the weight function. Since they use kernel estimators, a great amount of data is needed for a suitable performance and, hence, it is not usable for various applications. Adaptive weights were discussed by [14,15]. Jones and Crowley [16,17] generalized many previous tests to a huge class of nonparametric single-covariate tests. Several researchers followed the idea to combine different weighted logrank tests. For instance, [18–20] took the maximum. Bathke et al. [21] considered the censored empirical likelihood with constraints corresponding to different weights. The supremum of function-indexed weighted logrank tests was studied by [22].

Finally, we like to focus on the paper of [23], which motivated the present paper. Adapting the concept of broader power functions by [24,25] to the right-censored survival set-up, they first choose a vector of weighted logrank statistics. Roughly speaking, this vector is then adaptively projected onto a space corresponding to the closed hazard alternative. In this way, they ensure asymptotic optimality against the given alternatives of interest. A permutation version of their test solves the problem of the test statistic's unknown limit distribution. While their procedure is theoretically optimal (in some sense), it has the following disadvantages, which may explain why the method is not used in practice: (1) Due to the projection terminology the paper is quite hard to read and understand. (2) Their permutation approach requires to view the critical value as part of the test statistic, which is a rather counter-intuitive approach. (3) Their method is not implemented in any statistical software. In this paper, we present a solution for all these points. (1) We only use the typical survival notation and our statistic is a simple quadratic form. (2) We explain how to appropriately choose the weights for the logrank statistic such that the asymptotic results are not affected, but the corresponding permutation test becomes far more intuitive. (3) Our novel method is implemented in an R package called *mdir.logrank*, which is available on CRAN. A simulation study promises a good finite sample performance of our permutation test under the null and a good power behaviour under various alternatives.

The paper is structured as follows: In Section 2, we define the set-up and introduce the notation used throughout the paper. Section 3 introduces our test statistic and its asymptotic properties. In order to improve its small sample behaviour, we introduce a permutation version of our test in Section 4. Section 5 contains extensive simulation studies analysing type-I error rates and power behaviour of the proposed tests. A real-world example is analysed in Section 6, where we also give a short description of our R package. We conclude with a discussion in Section 7. All proofs are deferred to the Appendix.

## 2. Two-sample survival set-up

We consider the standard two-sample survival set-up given by survival times  $T_{j,i} \sim F_j$  and censoring times  $C_{j,i} \sim G_j$  ( $j = 1, 2$ ;  $i = 1, \dots, n_j$ ) with continuous distribution functions  $F_j, G_j$  on the positive line. As usual, all random variables  $T_{1,1}, C_{1,1}, \dots, T_{2,n_2}, C_{2,n_2}$  are assumed to be independent. Let  $n = n_1 + n_2$  be the pooled sample size, which is supposed to go to infinity in our asymptotic consideration. All limits  $\rightarrow$  are meant as  $n \rightarrow \infty$  if not stated otherwise. We are interested in the survival times' distributions  $F_1, F_2$ , but only the possibly censored survival times  $X_{j,i} = \min(T_{j,i}, C_{j,i})$  and their censoring status  $\delta_{j,i} = \mathbf{1}\{X_{j,i} = T_{j,i}\}$  ( $j = 1, 2$ ;  $i = 1, \dots, n_j$ ) are observable.

Throughout, we adopt the counting process notation of [3]. Let  $N_{j,i}(t) = \mathbf{1}\{X_{j,i} \leq t, \delta_{j,i} = 1\}$  and  $Y_{j,i}(t) = \mathbf{1}\{X_{j,i} \geq t\}$  ( $t \geq 0$ ). Then  $N_j(t) = \sum_{i=1}^{n_j} N_{j,i}(t)$  counts the number of events in group  $j$  up to  $t$  and  $Y_j(t) = \sum_{i=1}^{n_j} Y_{j,i}(t)$  equals the number of individuals in group  $j$  at risk at time  $t$ . Analogously, the pooled versions  $N = N_1 + N_2$  and  $Y = Y_1 + Y_2$  can be interpreted. Using these processes we can introduce the famous Kaplan–Meier and Nelson–Aalen estimators. Andersen et al. [3] proved that both estimators obey a central limit theorem, or, in other words, they are asymptotically normal. The Nelson–Aalen estimator  $\hat{A}_j$  given by

$$\hat{A}_j(t) = \int_{[0,t]} \frac{\mathbf{1}\{Y_j > 0\}}{Y_j} dN_j \quad (t \geq 0; j = 1, 2)$$

is the canonical nonparametric estimator of the (group specific) cumulative hazard function  $A_j(t) = -\log(1 - F_j(t)) = \int_0^t (1 - F_j)^{-1} dF_j$ . Similarly, for the pooled sample we introduce  $\hat{A}(t) = \int_0^t \mathbf{1}\{Y > 0\}/Y dN$  ( $t \geq 0$ ). In the following, we need the Kaplan–Meier estimator  $\hat{F}$  (only) for the pooled sample. It is

$$1 - \hat{F}(t) = \prod_{(j,i): X_{j,i} \leq t} \left(1 - \frac{\delta_{j,i}}{Y(X_{j,i})}\right) = \prod_{(j,i): X_{j,i} \leq t} \left(1 - \frac{\Delta N(X_{j,i})}{Y(X_{j,i})}\right) \quad (t \geq 0),$$

where  $\Delta f(t) = f(t) - f(t-)$  denotes the jump height in  $t$  for  $f: \mathbb{R} \rightarrow \mathbb{R}$ .

In the subsequent sections, we study the two-sample testing problem

$$H_=: F_1 = F_2 \quad \text{versus} \quad K_=: F_1 \neq F_2. \quad (1)$$

Weighted logrank tests are well known and often applied in practice for this testing problem. An introduction to these tests in their general form can be found in the books of [3,5]. First, choose a weight function  $w \in \mathcal{W} = \{w: [0, 1] \rightarrow \mathbb{R} \text{ continuous and of bounded variation}\}$ . Then, the corresponding weighted logrank statistic is

$$T_n(w) = \left(\frac{n}{n_1 n_2}\right)^{1/2} \int_{[0,\infty)} w(\hat{F}(t-)) \frac{Y_1(t) Y_2(t)}{Y(t)} [d\hat{A}_1(t) - d\hat{A}_2(t)].$$

By [7],  $T_n(w)$  is asymptotically normal and its asymptotic variance can be estimated by

$$\hat{\sigma}_n^2(w) = \frac{n}{n_1 n_2} \int_{[0,\infty)} w(\hat{F}(t-))^2 \frac{Y_1(t) Y_2(t)}{Y(t)} d\hat{A}(t). \quad (2)$$

Tests based on  $T_n(w)$  or studentized versions based on  $T_n(w)/\hat{\sigma}_n(w)$  are not omnibus tests for (1). But they have good properties for specific semiparametric hazard alternatives depending on the pre-chosen weight function  $w$ . Among others,  $T_n(w)$  is consistent for alternatives of the form

$$K_w: A_2(t) = \int_{[0,t]} 1 + \vartheta w \circ F_1 dA_1, \quad (3)$$

where we consider all constants  $\vartheta \neq 0$  leading to a non-negative integrand  $1 + \vartheta w \circ F_1$  over the whole line. For example, the classical logrank test with weight  $w \equiv 1$

is consistent against the proportional hazard alternative  $K_{\text{prop}} : A_2(t) = (1 + \vartheta)A_1(t)$ ,  $0 \neq \vartheta \in (-1, \infty)$ , and even optimal for so-called local alternatives  $K_{\text{loc}} : A_2(t) = (1 + n^{-1/2}\vartheta)A_1(t)$ , see [7]. Choosing  $w_{\text{prop}} \equiv 1$  we weight all time points equally. Instead of this, we can also give more weight to departures of the null  $A_1 = A_2$  at early times by setting  $w_{\text{early}}(u) = u(1 - u)^3$  or at central times, which are close to the median  $F_1(1/2)$ , by  $w_{\text{cent}}(u) = u(1 - u)$  ( $u \in [0, 1]$ ). All these are examples for stochastic ordered alternatives, i.e. we have  $F_1 \leq F_2$  or  $F_1 \geq F_2$ , depending on the sign of  $\vartheta$ . Even the local increments  $A_2(t, t + \varepsilon]$  are ordered since all  $w$  are strictly positive. An example without the latter property is the crossing hazard weight  $w_{\text{cross}}(u) = 1 - 2u$  with a change of sign at  $u = 1/2$ . Since  $w_{\text{prop}}$  and  $w_{\text{cross}}$  are orthogonal in  $L^2(0, 1)$ , i.e.  $\int_0^1 w_{\text{prop}}(x)w_{\text{cross}}(x) dx = 0$ , it is not surprising that the classical logrank test has no asymptotic power for the crossing hazard alternative  $K_w$  with  $w = w_{\text{cross}}$ , and vice versa. Our paper's aim is to combine the good properties of  $T_n(w)$  for different weight functions  $w$  to obtain a powerful test for various hazard alternatives simultaneously.

### 3. Our test and its asymptotic properties

For the asymptotic set-up, we need two (very common) assumptions. First assume that no group size vanishes:  $0 < \liminf_{n \rightarrow \infty} n_1/n \leq \limsup_{n \rightarrow \infty} n_1/n < 1$ . Let  $\tau = \inf\{u > 0 : [1 - G_1(u)][1 - G_2(u)][1 - F_1(u)][1 - F_2(u)] = 0\}$ , where the convention  $\inf \emptyset = \infty$  is used. To observe not only censored data it is convenient to suppose  $F_1(\tau) > 0$  or  $F_2(\tau) > 0$  in the case of  $\tau < \infty$ .

The basic idea of our test is to first choose an arbitrary amount of hazard directions/weights  $w_1, \dots, w_m \in \mathcal{W}$  ( $m \in \mathbb{N}$ ) with  $\mathcal{W} = \{w : [0, 1] \rightarrow \mathbb{R} \text{ continuous and of bounded variation}\}$  as above and to consider the vector  $T_n = [T_n(w_1), \dots, T_n(w_m)]^T$  of the corresponding weighted logrank tests. In the spirit of (2), let the empirical covariance matrix  $\widehat{\Sigma}_n$  of  $T_n$  be given by its entries

$$(\widehat{\Sigma}_n)_{r,s} = \frac{n}{n_1 n_2} \int_{[0, \infty)} w_s(\widehat{F}(t-)) w_r(\widehat{F}(t-)) \frac{Y_1(t) Y_2(t)}{Y(t)} d\widehat{A}(t) \quad (r, s = 1, \dots, m).$$

The studentized version of the statistic  $T_n$  is the quadratic form  $S_n = T_n^T \widehat{\Sigma}_n^- T_n$ , where  $A^-$  denotes the Moore–Penrose inverse of the matrix  $A$ . We suggest to use  $S_n$  for testing (1). For our asymptotic results, we restrict our considerations to linear independent weights in the following sense.

**Assumption 3.1:** Suppose for all  $\varepsilon \in (0, 1)$  that  $w_1, \dots, w_m$  are linearly independent on  $[0, \varepsilon]$ , i.e.  $\sum_{i=1}^m \beta_i w_i(x) = 0$  for all  $x \in [0, \varepsilon]$  implies  $\beta_1 = \dots = \beta_m = 0$ .

Many typical hazard weights are polynomial, for example, the ones we introduced in Section 1. For these weights the linear independence on  $[0, 1]$  is equivalent to the one on  $[0, \varepsilon]$ . Consequently, it is easy to check whether the pre-chosen weights fulfil Assumption 3.1.

**Theorem 3.1 (Convergence under the null):** *Let Assumption 3.1 be fulfilled. Then  $S_n$  converges in distribution to a  $\chi_m^2$ -distributed random variable under the null hypothesis.*

Regarding Theorem 3.1, we define our test by  $\phi_{n,\alpha} = \mathbf{1}\{S_n > \chi_{m,\alpha}^2\}$  [ $\alpha \in (0, 1)$ ], where  $\chi_{m,\alpha}^2$  is the  $(1 - \alpha)$ -quantile of the  $\chi_m^2$ -distribution. Under Assumption 3.1,  $\phi_{n,\alpha}$  is asymptotically exact, i.e.  $E_{H=}(\phi_{n,\alpha}) \rightarrow \alpha$ . We want to point out that Assumption 3.1 is not needed to obtain distributional convergence under the null, see [23]. But the degree of freedom  $k$  of the limiting  $\chi_k^2$ -distribution depends in general on the unknown asymptotic set-up and may be less than  $m$  if Assumption 3.1 does not hold. For this case, Brendel et al. [23] suggested to estimate  $k$  by its consistent estimator  $\kappa = \text{rank}(\widehat{\Sigma}_n)$  and use the data depended critical value  $\widehat{c}_\alpha(\widehat{\Sigma}_n) = \chi_{\kappa,\alpha}^2$ . Since this critical value is data-dependent, they write it as part of the test statistic  $S_n^{\text{BJMP}} = S_n - \widehat{c}_\alpha(\widehat{\Sigma}_n)$ .

Theorem 3.1 implies that the classical statistic  $T_n(w)^2/\widehat{\sigma}_n^2(w)$  converges in distribution to a  $\chi_1^2$ -distributed random variable. The weighted logrank test  $\phi_{n,\alpha}(\widetilde{w}) = \mathbf{1}\{T_n(\widetilde{w})^2/\widehat{\sigma}_n^2(\widetilde{w}) > \chi_{1,\alpha}^2\}$  of asymptotic exact size  $\alpha \in (0, 1)$  is consistent for alternatives of the shape (3) with  $w\widetilde{w} \geq 0$  and  $\int w(x)\widetilde{w}(x) dx > 0$ . This can be concluded, for instance, from the subsequent Theorem 3.3. For  $\widetilde{w} = w_i$ , this consistency can be transferred to our  $\phi_{n,\alpha}$  and, consequently, we indeed combine the strength of each single weighted logrank test.

**Theorem 3.2 (Consistency):** *Consider a fixed alternative  $K$ . If for some  $i = 1, \dots, m$  that  $\phi_{n,\alpha}(w_i)$  is consistent for testing  $H=$  versus  $K$ , i.e. the error of second kind  $E_K[1 - \phi_{n,\alpha}(w_i)]$  tends to 0 for all  $\alpha \in (0, 1)$ , then  $\phi_{n,\alpha}$  is consistent as well.*

Consequently, our test  $\phi_{n,\alpha}$  is consistent for alternatives (3) with  $w$  coming from the linear subspace  $\mathcal{W}_m = \{\sum_{i=1}^m \beta_i w_i : \beta = (\beta_1, \dots, \beta_m) \in \mathbb{R}^m, \beta \neq 0\}$  of  $\mathcal{W}$  or, more generally, with  $w$  such that  $ww_i \geq 0$  and  $\int w(x)w_i(x) dx > 0$  for some  $i = 1, \dots, m$ . Having this in mind the weights should be chosen.

In the introduction, we already mentioned local alternatives, which are small perturbations of the null assumption  $F_1 = F_2$ , or equivalently  $A_1 = A_2$ . Let  $F_0$  be a continuous (baseline) distribution and  $A_0$  be the corresponding (baseline) cumulative hazard function. From now on, the survival distributions of both groups depend on the sample size  $n$  and we write  $F_{j,n}$  as well as  $A_{j,n}$  instead of  $F_j$  and  $A_j$ , respectively. Let  $A_{1,n}$  and  $A_{2,n}$  be perturbations of the baseline  $A_0$  in (opposite) hazard directions  $w$  and  $-w$ . To be more specific, let

$$A_{j,n}(t) = \int_{[0,t]} 1 + c_{j,n}w \circ F_0 dA_0 \quad (t \geq 0), \quad c_{j,n} = \frac{(-1)^{j+1}}{n_j} \left( \frac{n_1 n_2}{n} \right)^{1/2} \quad (4)$$

for some  $w \in \mathcal{W}$  and sufficiently large  $n$  such that the integrand is non-negative over the whole line. Clearly, the two regression coefficients  $c_{j,n} = O(n^{-1/2})$  are asymptotically of rate  $n^{-1/2}$ . These coefficients are often used for two-sample rank tests. We denote by  $E_{n,w}$  the expectation under (4) and by  $E_{n,0}$  the expectation under the null  $F_{1,n} = F_{2,n} = F_0$ .

**Theorem 3.3 (Power under local alternatives):** *Suppose that  $n_1/n \rightarrow \eta \in (0, 1)$ , Assumption 3.1 and (4) for a hazard direction  $w$  hold. Define  $\psi = [(1 - G_1)(1 - G_2)]/[\eta(1 - G_1) + (1 - \eta)(1 - G_2)]$ . Under (4),  $S_n$  converges in distribution to a  $\chi_m^2(\lambda_w)$ -distributed random variable with non-centrality parameter  $\lambda_w = a_w^T \Sigma^- a_w$ , where  $a_w = (\int w \circ F_0 w_i \circ F_0 \psi dF_0)_{i \leq m}$  and the entries of  $\Sigma$  are  $\Sigma_{r,s} = \int w_r \circ F_0 w_s \circ F_0 \psi dF_0$  ( $1 \leq r, s \leq m$ ).*

From the well-known properties of non-central  $\chi^2$ -distributions, we obtain from Theorems 3.1 and 3.3 that our test is asymptotically unbiased under local alternatives, i.e.  $E_{n,w}(\phi_{n,\alpha}) \rightarrow P(Z_w > \chi_{m,\alpha}^2) \geq \alpha$  with  $Z_w \sim \chi_m^2(\lambda_w)$  for  $\lambda_w$  from Theorem 3.3. In the proof of Theorem 3.3, we show that the limiting covariance  $\Sigma$  is invertible. That is why  $a_w \neq 0$  implies  $\lambda_w = a_w^T \Sigma^{-1} a_w > 0$  and  $E_{n,w}(\phi_{n,\alpha}) \rightarrow P(Z_w > \chi_{m,\alpha}^2) > \alpha$ . Clearly,  $w \in \mathcal{W}_m$  lead to  $a \neq 0$  and, hence, our test has nontrivial power for local alternatives in hazard direction  $w$  coming from the linear subspace  $\mathcal{W}_m$ . For this kind of alternatives the test is even admissible, a certain kind of optimality which says that there is no test which achieves better asymptotic power for all hazard alternatives  $w \in \mathcal{W}_m$  simultaneously.

**Theorem 3.4 (Admissibility):** *Suppose that Assumption 3.1 holds. Then there is no test sequence  $\varphi_n$  ( $n \in \mathbb{N}$ ) of asymptotic size  $\alpha$ , i.e.  $\limsup_{n \rightarrow \infty} E_{H=}(\varphi_n) \leq \alpha$ , such that the limit  $\liminf_{n \rightarrow \infty} [E_{n,w}(\varphi_n) - E_{n,w}(\phi_{n,\alpha})]$  is non-negative for all  $w \in \mathcal{W}_m$  and positive for at least one  $w \in \mathcal{W}_m$ .*

#### 4. Permutation test

It is well known that permutation tests are finitely exact under exchangeability, which is present in our setting under the restricted null hypothesis  $H_{\text{res}} : \{F_1 = F_2, G_1 = G_2\}$ . The test's exactness, at least asymptotically, can be preserved even beyond exchangeability by using proper studentized test statistics as the Wald-type statistic applied here. This idea was originally proposed for two-sample settings [26–29] and was later extended to more general situations [30–36].

Denote by  $X_{(1)} \leq \dots \leq X_{(n)}$  the order statistics of the pooled sample. Let  $c_{(k)} \in \{c_{1,n}, c_{2,n}\}$  and  $\delta_{(k)} \in \{0, 1\}$  ( $1 \leq k \leq n$ ) be the group and the censoring status corresponding to  $X_{(k)}$ , i.e. if  $X_{(k)} = X_{j,i}$ , then  $c_{(k)} = c_{j,n}$  and  $\delta_{(k)} = \delta_{j,i}$ . The counting processes  $N_j, Y_j$  used for the test statistic jump only at the order statistics. Their value at these points can be expressed by the components of  $c^{(n)} = (c_{(1)}, \dots, c_{(n)})$  and  $\delta^{(n)} = (\delta_{(1)}, \dots, \delta_{(n)})$ , for example,  $N_j(X_{(k)}) = \sum_{i=1}^k \delta_{(i)} \mathbf{1}\{c_{(i)} = c_{j,n}\}$ . Consequently, the test statistic only depends on  $(c^{(n)}, \delta^{(n)})$ . That is why we write  $S_n(c^{(n)}, \delta^{(n)})$  instead of  $S_n$  throughout this section. The basic idea of our permutation test is to keep  $\delta^{(n)}$  fixed and to permute  $c^{(n)}$  only, i.e. to randomly permute the group membership of the individuals. For the case  $m = 1$ , i.e.  $S_n = T_n(w)^2 / \hat{\sigma}_n^2(w)$ , this permutation idea was already used by [26] and [37]. In simulations of [26] and [38], the resulting permutation test had a good finite sample performance, even in the case of unequal censoring  $G_1 \neq G_2$ .

Let  $c_n^\pi$  be a uniformly distributed permutation of  $c^{(n)}$  and be independent of  $\delta^{(n)}$ . Denote by  $c_{n,\alpha}^*(\delta)$  ( $\alpha \in (0, 1)$ ,  $\delta \in \{0, 1\}^n$ ) the  $(1 - \alpha)$ -quantile of the permutation statistic  $S_n(c_n^\pi, \delta)$ . Then our permutation test is given by  $\phi_{n,\alpha}^* = \mathbf{1}\{S_n(c^{(n)}, \delta^{(n)}) > c_{n,\alpha}^*(\delta^{(n)})\}$ . This test shares all the asymptotic properties of the unconditional test verified in the previous section.

**Theorem 4.1:** *Suppose that Assumption 3.1 is fulfilled and fix  $\alpha \in (0, 1)$ . Then  $\phi_{n,\alpha}^*$  is asymptotically exact under  $H=$  and  $\phi_{n,\alpha}^*$  is consistent for fixed alternative  $K$  whenever  $\phi_{n,\alpha}$  is consistent for  $K$ . Under local alternatives (4)  $\phi_{n,\alpha}^*$  and  $\phi_{n,\alpha}$  are asymptotically equivalent, i.e.  $E_{n,w}(|\phi_{n,\alpha} - \phi_{n,\alpha}^*|) \rightarrow 0$ , and, hence, they have the same asymptotic power under local alternatives. In particular,  $\phi_{n,\alpha}^*$  is asymptotically admissible, compare to Theorem 3.4.*

## 5. Simulations

### 5.1. Type-I error

To analyse the behaviour of the proposed test statistic for small sample sizes, we performed a simulation study implementing various situations. All simulations were conducted with the R computing environment, version 3.2.3 [39] using 10,000 simulation and 1000 permutation run.

First, we considered the behaviour of different tests under the null hypothesis  $H_0 : F_1 = F_2$ . Survival times were generated following an exponential  $\text{Exp}(1)$  distribution. Censoring times were simulated to follow the same distribution as the survival times, but with varying parameters to reflect different proportions of censoring: No censoring, equal censoring in both groups, where the parameters were chosen such that on average 15% of individuals were censored, and unequal censoring distributions reflecting 10% and 20% censoring (on average) in the first and second group, respectively. Sample sizes were chosen to construct balanced as well as unbalanced designs, namely  $(n_1, n_2) = (25, 25)$ ,  $(n_1, n_2) = (15, 35)$ ,  $(n_1, n_2) = (50, 50)$ ,  $(n_1, n_2) = (30, 70)$ ,  $(n_1, n_2) = (100, 100)$  and  $(n_1, n_2) = (150, 50)$ . For all scenarios, we compared the performance of our test with and without permutation based on weights of the form

$$w^{(r,g)}(u) = u^r (1 - u)^g \quad (r, g \in \mathbb{N}_0), \quad w_{\text{cross}}(u) = 1 - 2u, \quad (5)$$

including the famous weights  $w^{(0,0)}$  (*proportional hazards*),  $w^{(1,1)}$  (*central hazards*) and  $w_{\text{cross}}$  (*crossing hazards*). But also mid-early, early, mid-late and late hazards are included in this class of hazard weights. We distinguished between testing based on two or four hazard directions  $w_i$ , namely proportional and crossing hazards

$$w_1(u) = 1, \quad w_2(u) = 1 - 2u$$

as well as additionally central and early hazards

$$w_3(u) = u(1 - u), \quad w_4(u) = u(1 - u)^3.$$

We compared our results to the two-sided permutation test proposed by Brendel et al. [23], which we denote as BJMP-test in the following. Recall that the data-dependent critical value is part of their test statistic  $S_n^{\text{BJMP}} = S_n - \widehat{c}_\alpha(\widehat{\Sigma}_n)$ , see the paragraph after Theorem 3.1 for details. As the statistic  $S_n$ , this critical value need to be recalculated for each permutation iteration, a somehow counter-intuitive and unusual step that is not required by our approach. The resulting type-I error rates are displayed in Table 1. As we can see from the tables, the permutation version of the test always keeps the nominal level of 5% better than the corresponding  $\chi^2$ -approximation. The results based on our permutation test and the BJMP-test are very similar. Testing based on two or four directions, in contrast, does not change the type-I error much for neither of the tests.

In order to investigate the liberal behaviour of the  $\chi^2$ -approximation in more detail, we considered growing sample sizes  $n = 50, 100, 150, \dots, 500$ . We simulated balanced ( $n_1 = n_2 = n/2$ ) as well as unbalanced ( $n_1 = 0.3n, n_2 = 0.7n$ ) designs with equal or unequal censoring as described above. The results of the type-I error rates for the  $\chi^2$ -approximation based on four hazard alternatives are displayed in Figure 1. It can be seen

**Table 1.** Type-I error rates in % (nominal level 5%) for exponentially distributed censoring and survival times, testing based on two (2 dir) or four (4 dir) hazard directions with and without permutation procedure as well as the BJMP-Test with permutation, respectively.

$(n_1, n_2)$	Censoring	Permutation test		$\chi^2$ -Approximation		BJMP-test	
		4 dir	2 dir	4 dir	2 dir	4 dir	2 dir
(25, 25)	None	5.13	4.81	6.67	6.14	5.06	4.83
	Equal	4.98	4.93	6.17	6.24	5.00	4.92
	Unequal	5.25	4.88	6.47	6.02	5.36	4.99
(15, 35)	None	5.45	4.73	7.90	6.53	5.56	4.73
	Equal	5.16	5.37	7.87	7.03	5.11	5.37
	Unequal	4.59	4.78	6.89	6.47	4.47	4.90
(50, 50)	None	4.88	4.70	5.92	5.45	4.92	4.75
	Equal	5.27	5.48	6.30	6.19	5.14	5.56
	Unequal	4.93	4.99	5.80	5.79	4.86	4.96
(30, 70)	None	5.52	5.72	7.36	6.82	5.49	5.74
	Equal	5.25	5.11	6.68	6.07	5.08	5.00
	Unequal	5.06	5.19	6.76	6.29	5.00	5.19
(100, 100)	None	5.04	5.27	5.84	5.73	5.18	5.30
	Equal	4.61	5.11	5.47	5.41	4.75	5.11
	Unequal	4.90	5.13	5.54	5.58	4.80	5.22
(150, 50)	None	5.07	5.46	6.27	6.16	5.22	5.49
	Equal	5.14	5.27	6.31	5.71	5.11	5.18
	Unequal	5.02	5.26	6.28	5.83	5.00	5.03

Equal censoring corresponds to an average of 15% censored individuals, while the unequal censoring distribution reflects an average of 10% and 20% in the first and second group.

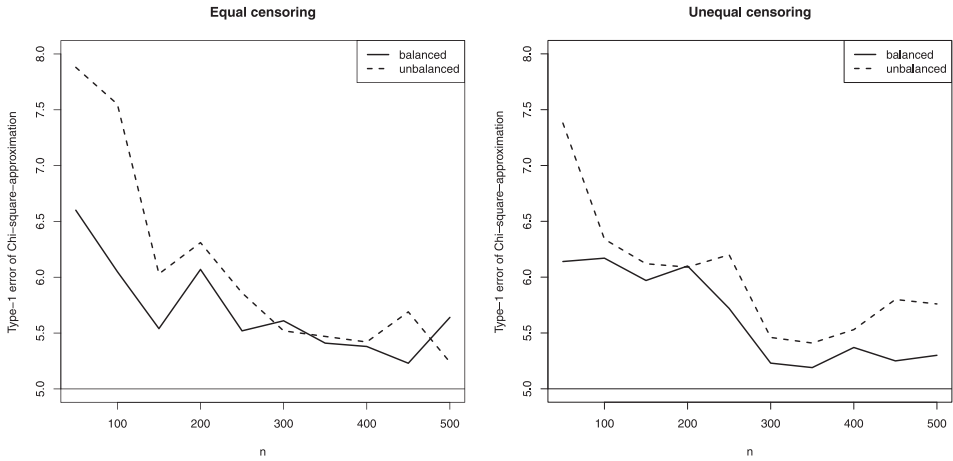
that the type-I error decreases towards the nominal 5% level with growing sample size as expected and reaches an acceptable level for about  $n = 300$  individuals in the situations considered.

## 5.2. Power behaviour against various alternatives

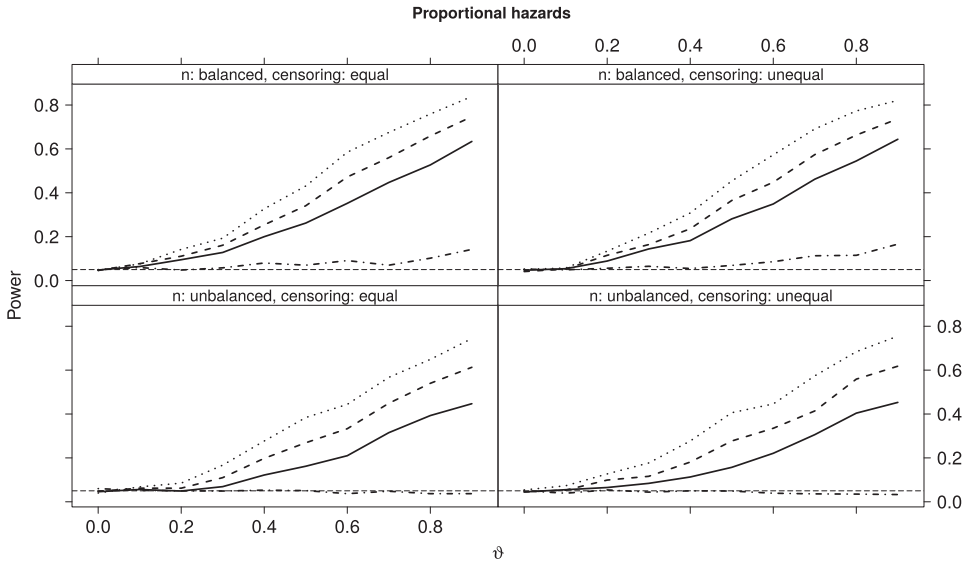
In a second simulation study, we considered the power behaviour of the test under various alternatives using 1000 simulation and 1000 permutation runs. Since we found the  $\chi^2$ -approximation to be slightly liberal in all considered scenarios, we excluded it from the power comparisons. We also excluded the BJMP-test in the plots below to increase readability, since the results were again almost indistinguishable. We again considered the exponential distribution, i.e. survival times in the first group were simulated to follow an  $\text{Exp}(1)$  distribution. The simulated data for the second group was generated according to

$$A_2(t) = \int_{[0,t]} 1 + \vartheta w_i \circ F_1 \, dA_1$$

with different weight functions  $w_i$  ( $i = 1, \dots, 4$ ) as above. Realizations of the distribution belonging to  $A_2$  were generated using an acceptance-rejection procedure. The parameter  $\vartheta$  was chosen to range from  $\vartheta = 0$  (corresponding to the null hypothesis) to  $\vartheta = 0.9$  in the case of proportional and crossing hazards, to  $\vartheta = 4.5$  for central hazards and early hazards. Censoring times were simulated as above to create equal as well as unequal censoring distributions. Sample sizes were  $(n_1, n_2) = (50, 50)$  and  $(n_1, n_2) = (30, 70)$ . For each alternative

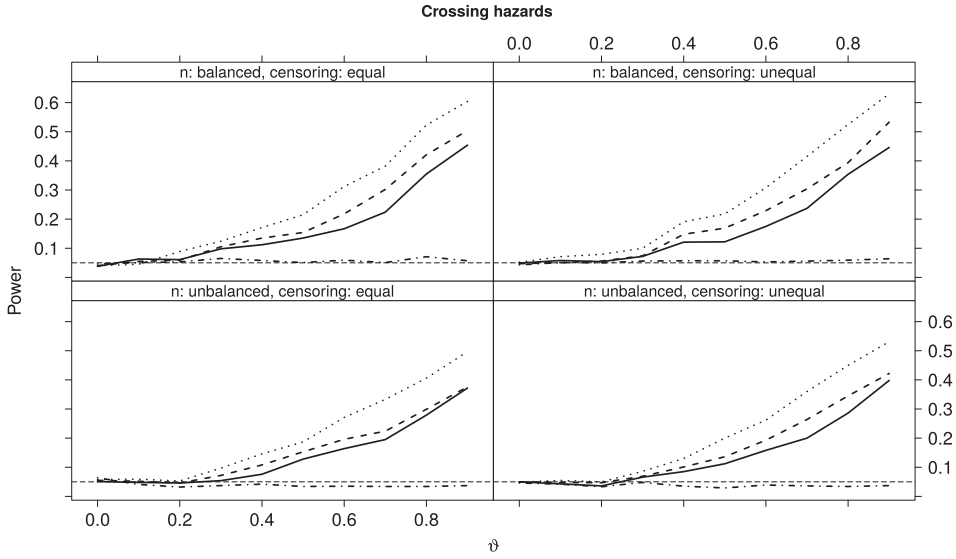


**Figure 1.** Type-I error of the  $\chi^2$ -approximation based on four hazard alternatives for growing sample sizes. The left plot corresponds to equal censoring in both groups (15%), whereas the right plot shows unequal censoring (10% and 20%, respectively). In the balanced scenario (solid line) it is  $n_1 = n_2 = n/2$ , while for the unbalanced design (dashed line)  $n_1 = 0.3n$  and  $n_2 = 0.7n$ .

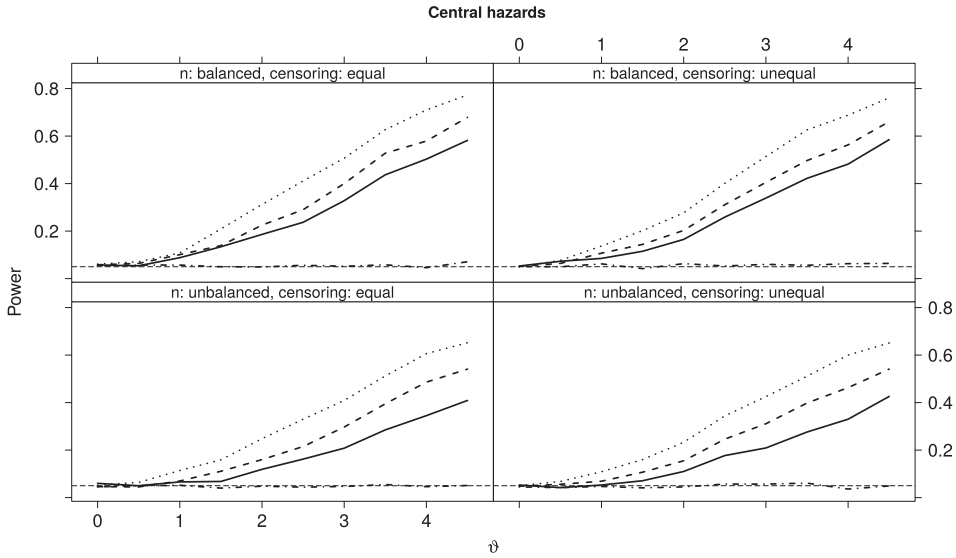


**Figure 2.** Power simulation results ( $\alpha = 5\%$ ) of the permutation test  $\phi_{n,\alpha}^*$  based on four (solid) and two (dashed) directions, the proportional hazards (logrank) test (dotted) and the crossing hazards test (dot-dash). Sample sizes are  $(n_1, n_2) = (50, 50)$  (balanced) and  $(n_1, n_2) = (30, 70)$  (unbalanced).

based on a weight function  $w_i$ , we considered our permutation test based on the two or four hazard directions  $w_i$  stated above (i.e. proportional and crossing hazards for the two directions and additionally central and early hazards for the four directions). Note that this choice is independent of the true underlying alternative). Moreover, we included the optimal test based on  $T_n(w_i)/\widehat{\sigma}_n(w_i)$  and one of the other one-directional tests based on  $T_n(w_j)/\widehat{\sigma}_n(w_j)$



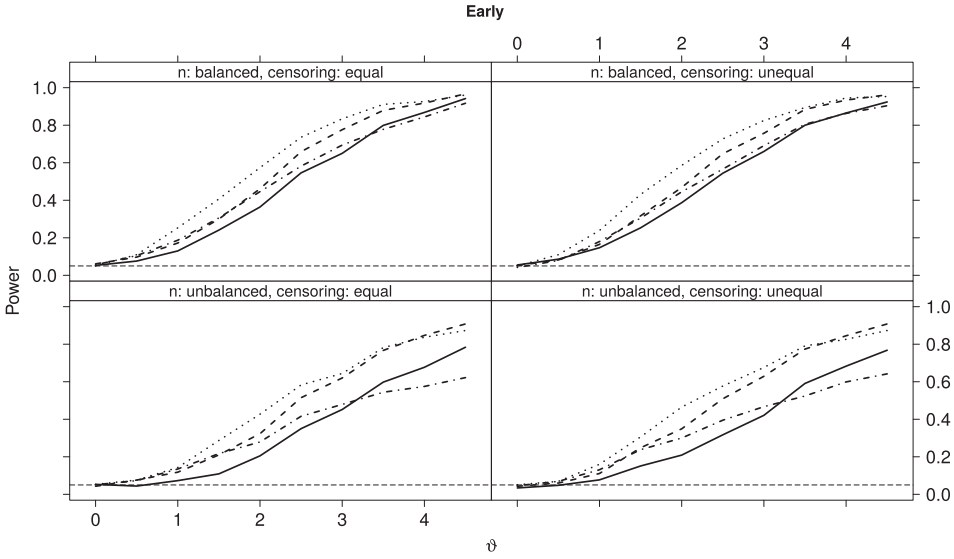
**Figure 3.** Power simulation results ( $\alpha = 5\%$ ) of the permutation test  $\phi_{n,\alpha}^*$  based on four (solid) and two (dashed) directions, the crossing hazards test (dotted) and the proportional hazards test (dot-dash). Sample sizes are  $(n_1, n_2) = (50, 50)$  (balanced) and  $(n_1, n_2) = (30, 70)$  (unbalanced).



**Figure 4.** Power simulation results ( $\alpha = 5\%$ ) of the permutation test  $\phi_{n,\alpha}^*$  based on four (solid) and two (dashed) directions, the central hazards test (dotted) and the crossing hazards test (dot-dash). Sample sizes are  $(n_1, n_2) = (50, 50)$  (balanced) and  $(n_1, n_2) = (30, 70)$  (unbalanced).

for some  $j \neq i$ . In the scenario with early hazards below (Figure 5), we considered a more extreme choice of early hazard alternatives corresponding to  $\tilde{w}_4(u) = (1 - u)^5$ .

Figures 2–5 show that choosing the wrong weight function can lead to a substantial loss in power, as already known in the literature. Moreover, both permutation tests follow the



**Figure 5.** Power simulation results ( $\alpha = 5\%$ ) of the permutation test  $\phi_{n,\alpha}^*$  based on four (solid) and two (dashed) directions, the early hazards test (dotted) and the crossing hazards test (dot-dash). Sample sizes are  $(n_1, n_2) = (50, 50)$  (balanced) and  $(n_1, n_2) = (30, 70)$  (unbalanced).

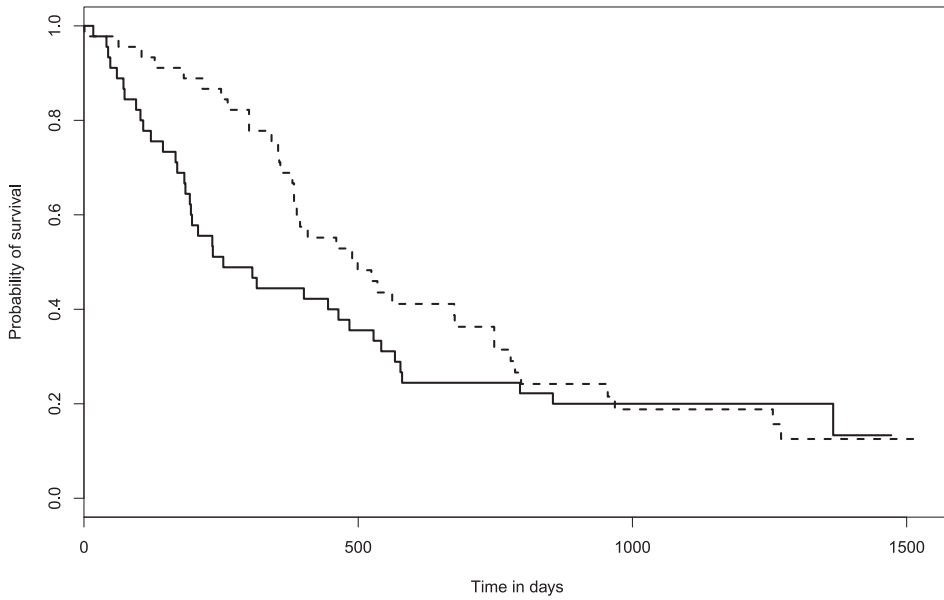
power curve of the optimal test. Furthermore, there is no notable difference between equal and unequal censoring proportions, while unbalanced designs tend to result in slightly lower power than balanced designs. Since the classical logrank test is consistent for early, central and late hazard alternatives it is not surprising that the two-direction test has reasonable power in all scenarios. In Figure 5, the power line of the four-direction test intersect one of the two-direction test and is even significantly higher for large  $\vartheta$ . This is an interesting phenomenon indicating two competing effects. On the one hand, we want to choose the true/best direction, but on the other hand, we should not choose too many weights since we would broaden the power into too many directions. In Figure 5, we see that only for a high weight effect size, the benefit of choosing the right direction can compensate the negative effect of choosing too many weights. In all other scenarios, the two-direction test has higher power than the four-direction test. Due to these observations, we advice to use the two-direction test unless specific alternatives are more relevant or interesting for the underlying statistical analysis.

## 6. Analysing a real data example using the R package *mdir.logrank*

In order to make these tests available to users, we have implemented them in an R package *mdir.logrank*, which is available on CRAN.

The package also contains the function *mdir.onesided* for the one-sided wild bootstrap test of [40] testing the stochastic ordered alternative  $K : F_1 \geq F_2, F_1 \neq F_2$ . For details on the commands and options of the package, see the corresponding manual [41].

Consider the gastrointestinal tumour study from [42], which is available in the *coin* package [43] in R. This study compared the effect of chemotherapy alone versus a combination of radiation and chemotherapy in treatment of gastrointestinal cancer. Of the



**Figure 6.** Kaplan–Meier curves for the patients receiving chemotherapy alone (dashed) and those receiving a combination of chemotherapy and radiation (solid).

**Table 2.** P-values for the single-direction crossing, proportional, early and central hazard tests as well as the multiple-direction tests based on the first two or all four hazard directions in the gastrointestinal cancer study.

	Crossing	Proportional	Early	Central	Two directions	Four directions
Permutation	0.001	0.256	0.005	0.742	0.007	0.017
$\chi^2$ -approximation	0.002	0.255	0.005	0.748	0.007	0.018

90 patients in the study, 45 were randomized to each of the two treatment groups. The Kaplan–Meier curves for the two groups are displayed in Figure 6.

In order to test whether the difference seen between the curves is statistically significant or not, we use our proposed test and its permutation version based on proportional and crossing hazards as well as additionally based on early ( $w^{(1,5)}$ ) and central hazards.

We compare the results to the corresponding single-direction tests. The resulting  $p$ -values are displayed in Table 2. The BJMP-test, which is not implemented in the package, yielded the same conclusions (results not shown). As we can see from the table, the single-direction crossing as well as early hazard tests detect significant differences between the two groups at 5% level, a finding shared by the two- and four-direction tests, while the proportional and the central hazards test do not lead to significant results. This result illustrates the problem when using the classical (single-direction) weighted logrank test since we do not know the right direction in advance. Moreover, the result confirms the advantage of combining different weights and, hence, we advice to use one of our new multiple-direction tests. Similar to the simulation study, we find that the test based on two hazard directions has higher power than the one based on four directions, in particular, the former would even reject the null at 1% level.

## 7. Discussion

The main difference between our approach and the one of [23] is the additional Assumption 3.1. The linear subset  $\mathcal{W}_m$  of  $\mathcal{W}$  plays an important role, see Theorems 3.2 and 3.4 as well as the comments to them. Concerning this set, it is not an actual restriction to consider only linearly independent weights. As already mentioned, the typical (polynomial) weights fulfil Assumption 3.1 if and only if they are linearly independent. Users of our R package *mdir.logrank* do not have to check the linear independence of the weights in advance since we implemented an automatic check. If the pre-chosen weights are linearly dependent, then a subset consisting of linearly independent weights will be selected automatically. Consequently, considering additionally Assumption 3.1 is not an actual restriction or disadvantage. In fact, we benefit from this assumption since no additional estimation step for the degree of freedom of the limiting  $\chi^2$ -distribution under the null is needed. Due to the latter, the permutation approach becomes much more intuitive, since we do not need to write the critical values as part of the test statistic as in [23]. Instead, our test is defined in the usual way, where we compare a test statistic to a critical value (either obtained based on asymptotics or obtained by the permutation procedure).

The one-sided permutation test of [23] for stochastic ordered alternatives  $K : F_1 \geq F_2, F_1 \neq F_2$  is computationally demanding. Under our Assumption 3.1 a wild bootstrap version of their test shares all asymptotic properties of the original permutation approach but reduces the computation time significantly by a factor of about 1,000, see [40]. The proofs of [40] show that the wild bootstrap approach can be used for our two-sided testing problem as well. Here, the computation times of both resampling versions are nearly indistinguishable. That is why we prefer the permutation test due to its finite exactness under the restricted null hypothesis  $H_{\text{res}} : \{F_1 = F_2, G_1 = G_2\}$ .

## Acknowledgments

The authors thank Markus Pauly for his inspirational suggestions.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Marc Ditzhaus was supported by the Deutsche Forschungsgemeinschaft [grant number PA-2409 5-1].

## ORCID

Marc Ditzhaus  <http://orcid.org/0000-0001-9235-1905>

Sarah Friedrich  <http://orcid.org/0000-0003-0291-4378>

## References

- [1] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep.* 1966;50:163–170.
- [2] Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc A Stat Soc.* 1972;135:185–206.

- [3] Andersen P, Borgan O, Gill RD, et al. Statistical models based on counting processes. New York (NY): Springer; 1993.
- [4] Bagdonavicius V, Kruopis J, Nikulin M. Non-parametric tests for censored data. Hoboken (NJ): Wiley; 2011.
- [5] Fleming T, Harrington D. Counting processes and survival analysis. New York (NY): Wiley; 1991.
- [6] Harrington D, Fleming T. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69(3):553–566.
- [7] Gill RD. Censoring and stochastic integrals. Amsterdam: Mathematisch Centrum; 1980. (Mathematical Centre Tracts; 124).
- [8] Klein J, Moeschberger M. Survival analysis: techniques for censored and truncated data. New York (NY): Springer; 1997.
- [9] Tarone R, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika*. 1977;64(1):156–160.
- [10] Janssen A. Global power functions of goodness of fit tests. *Ann Stat*. 2000;28(1):239–253.
- [11] Fleming T, Harrington D, O'Sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. *J Am Stat Assoc*. 1987;82(397):312–320.
- [12] Ehm W, Mammen E, Müller DW. Power robustification of approximately linear tests. *J Am Stat Assoc*. 1995;90(431):1025–1033.
- [13] Lai T, Ying Z. Rank regression methods for left-truncated and right-censored data. *Ann Stat*. 1991;19(2):531–556.
- [14] Yang S, Hsu L, Zhao L. Combining asymptotically normal tests: case studies in comparison of two groups. *J Stat Plan Inference*. 2005;133(1):139–158.
- [15] Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. *Biometrics*. 2010;66:30–38.
- [16] Jones M, Crowley J. A general class of nonparametric tests for survival analysis. *Biometrics*. 1989;45(1):157–170.
- [17] Jones M, Crowley J. Asymptotic properties of a general class of nonparametric tests for survival analysis. *Ann Stat*. 1990;18(3):1203–1220.
- [18] Bajorski P. Max-type rank tests in the two-sample problem. *Polska Akademia Nauk Instytut Matematyczny Zastosowania Matematyki Applicationes Mathematicae*. 1992;21(3):371–385.
- [19] Tarone R. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics*. 1981;37:79–85.
- [20] Garés V, Andrieu S, Dupuy JE, et al. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Stat Med*. 2015;34:541–557.
- [21] Bathke A, Kim MO, Zhou M. Combined multiple testing by censored empirical likelihood. *J Stat Plan Inference*. 2009;139(3):814–827.
- [22] Kosorok M, Lin CY. The versatility of function-indexed weighted log-rank statistics. *J Am Stat Assoc*. 1999;94(445):320–332.
- [23] Brendel M, Janssen A, Mayer CD, et al. Weighted logrank permutation tests for randomly right censored life science data. *Scand Stat Theory Appl*. 2014;41(3):742–761.
- [24] Behnen K, Neuhaus G. Galton's test as a linear rank test with estimated scores and its local asymptotic efficiency. *Ann Stat*. 1983;11(2):588–599.
- [25] Behnen K, Neuhaus G. Rank tests with estimated scores and their application. Stuttgart: B. G. Teubner; 1989.
- [26] Neuhaus G. Conditional rank tests for the two-sample problem under random censorship. *Ann Stat*. 1993;21(4):1760–1779.
- [27] Janssen A. Studentized permutation tests for non-iid hypotheses and the generalized Behrens-fisher problem. *Stat Probab Lett*. 1997;36(1):9–21.
- [28] Janssen A, Pauls T. How do bootstrap and permutation tests work? *Ann Stat*. 2003;31(3):768–806.
- [29] Pauly M. Weighted resampling of martingale difference arrays with applications. *Electron J Stat*. 2011;5:41–52.

- [30] Chung E, Romano JP. Exact and asymptotically robust permutation tests. *Ann Stat.* **2013**;41(2):484–507.
- [31] Pauly M, Brunner E, Konietzschke F. Asymptotic permutation tests in general factorial designs. *J R Stat Soc B Stat Methodol.* **2015**;77(2):461–473.
- [32] Friedrich S, Brunner E, Pauly M. Permuting longitudinal data in spite of the dependencies. *J Multivar Anal.* **2017**;153:255–265.
- [33] Smaga Ł. Diagonal and unscaled Wald-type tests in general factorial designs. *Electron J Stat.* **2017**;11:2613–2646.
- [34] Umlauft M, Konietzschke F, Pauly M. Rank-based permutation approaches for non-parametric factorial designs. *British J Math Stat Psychol.* **2017**;70:368–390.
- [35] Ditzhaus M, Fried R, Pauly M. Qanova: Quantile-based permutation methods for general factorial designs; 2019. ArXiv e-prints.
- [36] Harrar S, Ronchi F, Salmaso L. A comparison of recent nonparametric methods for testing effects in two-by-two factorial designs. *J Appl Stat.* **2019**;46:1649–1670.
- [37] Janssen A, Mayer CD. Conditional Studentized survival tests for randomly censored models. *Scand Stat Theory Appl.* **2001**;28(2):283–293.
- [38] Heller G, Venkatraman E. Resampling procedures to compare two survival distributions in the presence of right-censored data. *Biometrics.* **1996**;52(4):1204–1213.
- [39] R Core Team. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; **2018**. Available From: <http://www.R-project.org>.
- [40] Ditzhaus M, Pauly M. Wild bootstrap logrank tests with broader power functions for testing superiority. *Comput Stat Data Anal.* **2019**;136:1–11.
- [41] Ditzhaus M, Friedrich S. mdir.logrank: multiple-Direction Logrank Test; 2018. R package version 0.0.4. Available From: <https://CRAN.R-project.org/package=mdir.logrank>.
- [42] Stablein DM, Carter JWH, Novak JW. Analysis of survival data with nonproportional hazard functions. *Control Clin Trials.* **1981**;2:149–159.
- [43] Hothorn T, Hornik K, van de Wiel M, et al. A Lego system for conditional inference. *Am Stat.* **1996**;60:257–263.
- [44] Neuhaus G. A method of constructing rank tests in survival analysis. *J Stat Plan Inference.* **2000**;91(2):481–497.
- [45] Janssen A. Local asymptotic normality for randomly censored models with applications to rank tests. *Stat Neerl.* **1989**;43(2):109–125.
- [46] Efron B, Johnstone IM. Fisher’s information in terms of the hazard rate. *Ann Stat.* **1990**;18(1):38–62.
- [47] Ritov Y, Wellner JA. Censoring, martingales, and the cox model. *Contemp Math.* **1988**;80:191–219.
- [48] Strasser H. Mathematical theory of statistics. Vol. 7. Berlin: de Gruyter; **1985**. (Statistical Experiments and Asymptotic Decision Theory).
- [49] Anderson T. An introduction to multivariate statistical analysis. 3rd ed. Hoboken (NJ): Wiley; **2003**.

## Appendix. Proofs

### A.1 Some notes on [23]

In the subsequent proofs of our theorems, we often refer to [23]. To avoid a misunderstanding, we want to comment on three aspects regarding their results and notation. First, we want to point out that [23] interpreted the statistics as certain orthogonal projections. They expressed their test statistic as  $\|\Pi_{V_r}(\hat{\gamma}_n)\|_{\hat{\mu}_n}^2$ , which equals our  $S_n$  according to their Theorem 1. Second, our  $w_i$  corresponds to their  $\tilde{w}_i$  and their  $w_i$  in Theorem 9.1 equals  $w_i \circ F_1$  here. The third aspect concerns the definition of the test statistic. Introduce  $m_n = \min\{\max\{X_{j,i} : i = 1, \dots, n_j\} : j = 1, 2\}$ , the smallest group maximum. Fix  $\omega \in \mathcal{W}$ , Brendel et al. [23] replaced  $w(\hat{F}(t-))$  by  $w(\hat{F}(t-))\mathbf{1}\{t < m_n\}$  ( $t \geq 0$ )

in the integrands of  $T_n(w)$  and  $\widehat{\Sigma}_{r,s}$ . Let  $T_n^*(w)$  be the corresponding weighted logrank statistic, i.e.

$$T_n^*(w) = \left( \frac{n}{n_1 n_2} \right)^{1/2} \int_{[0, m_n]} w(\widehat{F}(t-)) \frac{Y_1(t)Y_2(t)}{Y(t)} [d\widehat{A}_1(t) - d\widehat{A}_2(t)].$$

All observations lying in  $(m_n, \infty)$  belong to the same group, and, hence, the integrand equals 0 on  $(m_n, \infty)$ . Consequently, only the set  $\{m_n\}$  is excluded from the integration area compared to  $T_n(w)$ . Since  $w$  is bounded we can assume  $|w| \leq K \in (0, \infty)$ . It is easy to check

$$|T_n(w) - T_n^*(w)| \leq \left( \frac{n}{n_1 n_2} \right)^{1/2} K \Delta N(m_n) \leq K \left( \frac{n}{n_1 n_2} \right)^{1/2} \rightarrow 0.$$

A comparable convergence can be shown for the entries  $\widehat{\Sigma}_{r,s}$  of  $\widehat{\Sigma}$ . Finally, the asymptotic results of [23] remain valid when we omit the additional indicator function, as we did in our definitions.

## A.2 Proof of Theorem 3.1

Considering appropriate subsequences, we can assume that  $n_1/n \rightarrow \eta \in (0, 1)$ . By Theorem 9.1 in the supplement of [23],  $T_n$  converges in distribution to  $Z \sim N(0, \Sigma)$  and  $\widehat{\Sigma}_n$  converges in probability to  $\Sigma$ , where the entries of  $\Sigma$  are

$$\Sigma_{r,s} = \int_{[0, \infty)} w_i \circ F_1 w_j \circ F_1 \psi \, dF_1 \quad (1 \leq r, s \leq m)$$

and  $\psi = [(1 - G_1)(1 - G_2)]/[\eta(1 - G_1) + (1 - \eta)(1 - G_2)]$ . Below we will verify  $\text{kern}(\Sigma) = \{0\}$ , i.e.  $\Sigma$  has full rank and is invertible. In this case, it is well known that the convergence of the Moore–Penrose inverse follows, i.e.  $\widehat{\Sigma}_n^- \rightarrow \Sigma^-$  in probability. By the continuous mapping theorem  $S_n$  converges in distribution to a  $\chi_m^2$ -distributed random variable. Observe that this convergence does not depend on  $\eta$  and the subsequence chosen at the beginning of the proof.

Let  $\beta = (\beta_1, \dots, \beta_m)^T \in \text{kern}(\Sigma)$ . Then

$$0 = \beta^T \Sigma \beta = \int_{[0, \infty)} \left( \sum_{i=1}^m \beta_i w_i \circ F_1 \right)^2 \psi \, dF_1.$$

Since  $\psi$  is positive on  $[0, \tau)$  and  $F_1$  as well as  $w_1, \dots, w_m$  are continuous functions we can deduce  $\sum_{i=1}^m \beta_i w_i(x) = 0$  for all  $x \in [0, F_1(\tau))$ . From Assumption 3.1,  $\beta_1 = \dots = \beta_m = 0$  follows.

## A.3 Proof of Theorem 3.2

Brendel et al. [23] showed, see the proof of their Theorem 2, that  $S_n \geq T_n(w_i)^2 / \widehat{\sigma}_n^2(w_i)$  for all  $1 \leq i \leq m$ . Since consistency of  $\phi_{n,\alpha}(w_i)$  implies  $P(T_n(w_i)^2 / \widehat{\sigma}_n^2(w_i) > \chi_{1,\alpha}^2) \rightarrow 1$  under  $K$  for all  $\alpha \in (0, 1)$  we can deduce that  $S_n$  converges in probability to  $\infty$  under the alternative  $K$ . Finally, the consistency of  $\phi_{n,\alpha}$  follows.

## A.4 Proof of Theorem 3.3

Following the argumentation of [23] for the proof of their Theorem 9.1 in the supplement we obtain from Theorem 7.4.1 of [5] and the Cramér–Wold device that  $T_n$  converges in distribution to a multivariate normal distributed  $Z \sim N(a, \Sigma)$  and  $\widehat{\Sigma}_n \rightarrow \Sigma$  in probability. The covariance matrix  $\Sigma$  coincides with the one introduced in the proof of Theorem 3.1 when replacing  $F_1$  by  $F_0$ . In particular,  $\Sigma$  is invertible and (strict) positive definite. By the continuous mapping theorem  $S_n$  converges in distribution to a  $\chi_m^2(\lambda)$ -distributed random variable with non-centrality parameter  $\lambda = a^T \Sigma^{-1} a$ .

## A.5 Proof of Theorem 3.4

Considering appropriate subsequences, we can suppose that  $n_1/n \rightarrow \eta \in (0, 1)$ . Let  $Q_{n,\beta}$  ( $\beta = (\beta_1, \dots, \beta_m)^T \in \mathbb{R}^m$ ;  $n \in \mathbb{N}$ ) be the common distribution of  $(X_{1,1}, \delta_{1,1}, \dots, X_{2,n_2}, \delta_{2,n_2})$  under the local alternative (4) in direction  $w = \sum_{i=1}^m \beta_i w_i$ . In particular,  $Q_{n,0}$  denotes the corresponding distribution under the null. Let  $\psi$  and  $\Sigma$  be defined as in Theorem 3.3.

**Lemma A.1:** *For every  $\beta \in \mathbb{R}^m$ , the log likelihood ratio can be expressed by*

$$\log \frac{dQ_{n,\beta}}{dQ_{n,0}} = \beta^T T_n - \frac{1}{2} \beta^T \Sigma \beta + R_n,$$

where  $R_n$  converges in  $Q_{n,0}$ -probability to 0.

**Proof:** Fix  $\beta = (\beta_1, \dots, \beta_m)^T \in \mathbb{R}^m$  and let  $w = \sum_{i=1}^m \beta_i w_i$ . Let  $\{P_\theta^* : \theta \in \Theta\}$ ,  $\Theta = (-\theta_0, \theta_0) \subset \mathbb{R}$ , be a parametrized family with cumulative hazard measures  $A_\theta^*$  given by

$$A_\theta^*(t) = \int_{[0,t]} 1 + \theta w \circ F_0 dA_0 \quad (\theta \in \Theta, t \geq 0),$$

where  $\theta_0 > 0$  is chosen such that the integrand is always positive. Plugging in  $\theta = c_{j,n}$  gives us  $A_{j,n}$  from (4) ( $j = 1, 2$ ). Let  $Q_{\theta,j}^*$  ( $j = 1, 2; \theta \in \Theta$ ) be the distribution of  $(\min(T, C), \mathbf{1}\{T \leq C\})$  for independent  $T \sim P_\theta^*$  and  $C \sim G_j$ . Obviously,  $Q_{n,\beta} = (Q_{c_{1,n},1}^*)^{n_1} \otimes (Q_{c_{2,n},2}^*)^{n_2}$ . As already stated by [23], see the top of their page 6, the family  $\theta \mapsto Q_{\theta,j}^*$  is  $L_2$ -differentiable with derivative  $L$ , say. Let  $M_j = N_j - \int Y_j dA_0$  ( $j = 1, 2$ ). Following the argumentation of [44], see also [45], we obtain

$$\log \frac{dQ_{n,\beta}}{dQ_{n,0}} = Z_n - \frac{1}{2} \sigma^2 + R_n^*, \quad Z_n = \int R(L) \left( \frac{dM_1}{n_1} - \frac{dM_2}{n_2} \right),$$

where  $R_n^*$  converges in  $Q_{n,0}$ -probability to 0,  $Z_n$  converges in distribution to  $Z \sim N(0, \sigma^2)$  for some  $\sigma \geq 0$  under  $Q_{n,0}$  and  $R$  is the operator studied by [46] and [47]. In our situation  $R(L) = w \circ F_0$ . It is easy to check that  $T_n(w)$  coincides with  $Z_n$  when we replace  $w \circ F_0$  and  $A_0$  by  $t \mapsto w \circ \widehat{F}(t-)$  and  $\widehat{A}$ , respectively. Using the standard counting process techniques, for example Theorem 4.2.1 of [7], we can conclude that  $T_n(w) - Z_n$  converges in  $Q_{n,0}$ -probability to 0. Hence,

$$\log \frac{dQ_{n,\beta}}{dQ_{n,0}} = T_n(w) - \frac{1}{2} \sigma^2 + R_n,$$

where  $R_n$  tends in  $Q_{n,0}$ -probability to 0 and  $T_n(w)$  converges in distribution to  $Z \sim N(0, \sigma^2)$  under  $Q_{n,0}$ . From the proof of Theorem 3.1, setting  $m = 1$  and  $w_1 = w$  there, we get  $\sigma^2 = \int (w \circ F_0)^2 \psi dF_0$ . Finally, observe that  $T_n(w) = \beta^T T_n$  and  $\sigma^2 = \beta^T \Sigma \beta$ . ■

Recall from the proof of Theorem 3.3 that  $T_n$  converges in distribution to  $Z \sim N(\Sigma\beta, \Sigma) = Q_\beta$  under  $Q_{n,\beta}$  for all  $\beta \in \mathbb{R}^m$  and that  $\Sigma$  is invertible. Combining these and Lemma A.1 yields that  $dQ_{n,\beta}/dQ_{n,0}$  converges in distribution under  $Q_{n,0}$  to  $dQ_\beta/dQ_0(Z)$  with  $Z \sim Q_0$ . In terms of statistical experiments, see Sections 60 and 80 of [48], the experiment sequence  $\{Q_{n,\beta} : \beta \in \mathbb{R}^m\}$  fulfils Le Cam's local asymptotic normality, in short LAN, and converges weakly to the Gaussian shift model  $\{Q_\beta : \beta \in \mathbb{R}^m\}$ .

**Remark A.1:** By Le Cam's first lemma, see Theorem 61.3 of [48],  $Q_{n,\beta}$  and  $Q_{n,0}$  are mutually contiguous for all  $\beta \in \mathbb{R}^m$ , i.e. convergence in  $Q_{n,0}$ -probability implies convergence in  $Q_{n,\beta}$ -probability, and vice versa.

From the distributional convergence of  $T_n$  mentioned above, we obtain for all  $\beta \in \mathbb{R}_m$

$$E_{n,\beta}(\phi_{n,\alpha}) = \int \mathbf{1}\{x^T \Sigma_n^- x > \chi_{m,\alpha}^2\} dQ_{n,\beta}^{T_n}(x) \rightarrow \int \mathbf{1}\{x^T \Sigma^- x > \chi_{m,\alpha}^2\} dQ_\beta(x),$$

where  $Q_{n,\beta}^{T_n}$  is the image measure of  $Q_{n,\beta}$  under the map  $T_n$ . Since  $x \mapsto x^T \Sigma^- x$  is convex we can deduce from Stein's Theorem, see Theorem 5.6.5 of [49], that  $x \mapsto \phi_\alpha^*(x) = \mathbf{1}\{x^T \Sigma^- x > \chi_{m,\alpha}^2\}$  ( $x \in \mathbb{R}^m$ ) is an admissible test in the Gaussian shift model  $\{Q_\beta : \beta \in \mathbb{R}^m\}$  for testing the null  $H : \beta = 0$  versus the alternative  $K : \beta \neq 0$ . This means that there is no test  $\varphi$  of size  $\alpha$  such that  $\int \varphi - \phi_\alpha^* dQ_\beta$  is non-negative for all  $\beta \neq 0$  and positive for at least one  $\beta$ . Now, suppose contrary to the claim of Theorem 3.4 that there is a test sequence  $\varphi_n$  ( $n \in \mathbb{N}$ ) with the mentioned properties. By Theorem 62.3 of [48], which goes back to Le Cam, there is a test  $\varphi$  for the limiting model  $\{Q_\beta : \beta \in \mathbb{R}^m\}$  such that along an appropriate subsequence  $E_{n,\beta}(\varphi_n) \rightarrow \int \varphi dQ_\beta$  for all  $\beta \in \mathbb{R}^m$ . Under our contradiction assumption, we obtain  $\int \varphi dQ_0 \leq \alpha$ ,  $\int \varphi dQ_\beta \geq \int \phi_\alpha^* dQ_\beta$  for all  $0 \neq \beta \in \mathbb{R}^m$  and  $\int \varphi dQ_\beta > \int \phi_\alpha^* dQ_\beta$  for at least one  $\beta \neq 0$ . But, clearly, this contradicts the admissibility of  $\phi_\alpha^*$ .

## A.6 Proof of Theorem 4.1

As we explain at the proof's end all statements follow from the subsequent lemma.

**Lemma A.2:** *Let  $F_1, F_2, G_1, G_2$  be fixed and independent of  $n$ . Suppose that  $S_n(c^{(n)}, \delta^{(n)})$  converges in distribution to a random variable  $Z$  on  $[0, \infty]$ . Moreover, assume that the distribution function  $t \mapsto P(Z \leq t)$  ( $t \in [0, \infty]$ ) of  $Z$  is continuous on  $[0, \infty)$ . Then the unconditional test  $\phi_{n,\alpha}$  and the permutation test  $\phi_{n,\alpha}^*$  are asymptotically equivalent, i.e.  $E(|\phi_{n,\alpha} - \phi_{n,\alpha}^*|) \rightarrow 0$  for all  $\alpha \in (0, 1)$ .*

**Proof:** Considering appropriate subsequences, we can suppose  $n_1/n \rightarrow \eta \in (0, 1)$ . By Lemma 1 of [28], it is sufficient to verify

$$\sup_{x \geq 0} \left| P(S_n(c^{(n)}, \delta^{(n)}) \leq x \mid \delta^{(n)}) - \chi_m^2([0, x]) \right| \rightarrow 0$$

in probability. Recall that  $\xi_n \rightarrow \xi$  in probability if and only if every subsequence has a subsequence such that along this sub-subsequence  $\xi_n$  converges to  $\xi$  with probability one. Define

$$H(x) = 1 - \eta[1 - F_1(x)][1 - G_1(x)] - [1 - \eta][1 - F_2(x)][1 - G_2(x)] \quad (x \geq 0),$$

$$H^1(x) = \eta \int_{[0,x]} (1 - G_1) dF_1 + (1 - \eta) \int_{[0,x]} (1 - G_2) dF_2 \quad (x \geq 0),$$

$$F^*(x) = 1 - \exp \left( -\eta \int_{[0,x]} \frac{1 - G_1}{1 - H} dF_1 - (1 - \eta) \int_{[0,x]} \frac{1 - G_2}{1 - H} dF_2 \right) \quad (x \geq 0).$$

Let  $B$  be an  $m \times m$ -matrix with entries  $B_{r,s} = \int w_r(F^*) w_s(F^*) dH^1$  ( $1 \leq r, s \leq m$ ). Following the proof's argumentation of [23] for their Theorem 4, in particular using their Lemmas 10.1 and 10.2, we can deduce: for every subsequence there is a subsequence such that along this sub-subsequence  $S_n(c_n^\tau, \delta^{(n)}(\omega))$  converges in distribution to  $\tilde{Z} \sim \chi_{\text{rank}(B)}^2$  for almost all  $\omega$ , i.e. for all  $\omega \in E$  with  $P(E) = 1$ . Consequently, it remains to show  $\text{rank}(B) = m$ , or equivalently  $\text{kern}(B) = \{0\}$ .

Let  $\beta = (\beta_1, \dots, \beta_m)^T \in \text{kern}(B)$ . First, observe that for every  $0 < x < y$  we have  $F^*(y) - F^*(x) > 0$  if and only if  $H^1(y) - H^1(x) > 0$ . Thus, we obtain from  $0 = \beta^T B \beta = \int (\sum_{i=1}^m \beta_i w_i(F^*))^2 dH^1$  and the continuity of  $F^*$  as well as of  $w_1, \dots, w_m$  that  $\sum_{i=1}^m \beta_i w_i(x) = 0$  for all  $x \in [0, F^*(\infty)]$ , where  $F^*(\infty) = \lim_{u \rightarrow \infty} F^*(u)$ . Since  $F_1(\tau) > 0$  or  $F_2(\tau) > 0$  we can conclude  $F^*(\infty) \geq F^*(\tau) > 0$  and, hence,  $\beta = 0$  follows from the linear independence of  $w_1, \dots, w_m$  on  $[0, F^*(\infty)]$ . ■

First, suppose that  $\phi_{n,\alpha}$  is consistent for a fixed alternative  $K$ , i.e.  $E(\phi_{n,\alpha}) \rightarrow 1$  for all  $\alpha \in (0, 1)$ . Then  $S_n$  converges to  $Z \equiv \infty$  in probability under  $K$ . Applying Lemma A.2 yields that  $\phi_{n,\alpha}^*$  is consistent for  $K$  as well. From Theorem 3.1 and Lemma A.2, we can conclude that  $\phi_{n,\alpha}^*$  is asymptotically

exact. To be more specific, we obtain  $E_{n,0}(|\phi_{n,\alpha} - \phi_{n,\alpha}^*|) \rightarrow 0$  for all  $\alpha \in (0, 1)$ . From Remark A.1, setting  $m = 1$  and  $w_1 = w$  there, we get  $E_{n,w}(|\phi_{n,\alpha} - \phi_{n,\alpha}^*|) \rightarrow 0$  for all  $\alpha \in (0, 1)$  and every  $w \in \mathcal{W}$ . Combining this and Theorem 3.4 proves the last statement of Theorem 4.1, the admissibility of  $\phi_{n,\alpha}^*$ .