Lifetime Data Anal (2017) 23:400–425
DOI 10.1007/s10985-016-9365-0

Partitioned log-rank tests for the overall homogeneity of hazard rate functions

Yukun Liu[1] · **Guosheng Yin**[2]

Received: 17 July 2015 / Accepted: 12 March 2016 / Published online: 19 March 2016

Abstract In survival analysis, it is routine to test equality of two survival curves, which is often conducted by using the log-rank test. Although it is optimal under the proportional hazards assumption, the log-rank test is known to have little power when the survival or hazard functions cross. To test the overall homogeneity of hazard rate functions, we propose a group of partitioned log-rank tests. By partitioning the time axis and taking the supremum of the sum of two partitioned log-rank statistics over different partitioning points, the proposed test gains enormous power for cases with crossing hazards. On the other hand, when the hazards are indeed proportional, our test still maintains high power close to that of the optimal log-rank test. Extensive simulation studies are conducted to compare the proposed test with existing methods, and three real data examples are used to illustrate the commonality of crossing hazards and the advantages of the partitioned log-rank tests.

Keywords Censored data · Hazard function · Log rank test · Survival difference · Survival function · Weighted tests

Electronic supplementary material The online version of this article (doi:10.1007/s10985-016-9365-0) contains supplementary material, which is available to authorized users.

✉ Guosheng Yin
gyin@hku.hk

[1] Department of Statistics and Actuarial Science, School of Statistics, East China Normal University, Shanghai, China

[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

1 Introduction

One of the most fundamental goals in survival analysis is to compare survival or hazard functions between a treatment group and a control group based on data subject to right censoring (Kleinbaum and Klein 2012). Due to its great importance in evaluating treatment effects, this problem has attracted considerable attention and a large number of testing procedures have been proposed in the literature (see for example, Fleming and Harrington 1991, Chapter 7). The most widely-used approach is a class of weighted log-rank tests, which include the usual log-rank test with unit weights (Mantel 1966; Cox 1972) and Wilcoxon tests (Gehan 1965; Breslow 1970; Peto and Peto 1972; Prentice 1978) as special cases. The log-rank test is generally optimal when the two hazard functions are proportional to each other over time, while the Wilcoxon tests have been found to be more powerful in detecting differences that are early during the follow-up time (Lee et al. 1975; Prentice and Marek 1979). Klein and Moeschberger (2003) and Hosmer and Lemeshow (1999) have more thorough discussions on these weighted testing procedures. Yin and Zeng (2005) proposed a pair chart approach to detecting survival differences, which is shown to be particularly powerful for early differences.

However, the aforementioned tests might have little power if the underlying hazard rates cross each other. This is mainly due to the cancellation of the Mantel–Haenzel test statistics before and after the crossing point. One such example is the clinical trial study for evaluating the effect of zinc nasal spray in curing common cold (Belongia et al. 2001). Statistical analysis in Belongia et al. (2001) did not find any significant treatment effect regarding cold duration, while further investigation suggested a transient reduction of symptom severity in the early stage of the medical treatment. Liu et al. (2007) found that the life-table estimates of the two hazard rates of cold resolution for the treatment and control groups cross each other. In general, it is frequently encountered in applications that the hazard rates cross each other, or that differences between two hazard rates may be apparent at one point in time but fail to exist elsewhere (Fleming et al. 1980). Examples include radiation and chemotherapy treatments for cancer and surgery (Qiu and Sheng 2008).

These various features in comparison of two survival curves have motivated development of many testing procedures that can handle the crossing hazard rates problem. These methods can be roughly classified into three groups: (i) Omnibus tests such as the modified Kolmogorov–Smirnov test (Fleming et al. 1980), the Renyi-type test (Gill 1980) and Liu et al. (2007)'s test; (ii) Weighted log-rank tests whose weights change signs before and after a potential crossing point (i.e., Mantel and Stablein 1988; Moreau et al. 1992; Qiu and Sheng 2008); and (iii) Methods based on explicitly modeling the crossing structure of the hazard rates (Anderson and Senthilselvan 1982; Breslow et al. 1984; Liu et al. 2007; Bagdonavičius et al. 2012). Comparing the above three groups of approaches, we expect the second and third groups to be more powerful in testing differences between two crossing hazard rates, because they are designed specifically for testing the alternatives of crossing hazard rates, instead of some more general alternatives that are aimed at by the first group of methods (Liu et al. 2007). However, since they are designed specifically to test crossing hazard rates, the later two groups of methods for handling the crossing hazard rates problem may lose power when two hazard rates are different but not crossing, such as parallel for the additive hazards case (Lin and Ying 1994; Qiu and Sheng 2008). On the basis of forming stochastic processes indexed by weight functions, Kosorok and Lin (1999) studied function-indexed weighted log-rank statistics for testing a wide array of stochastic ordering alternatives. Eng and Kosorok (2005) further developed a sample size formula based on the supremum log-rank statistic for planning time-to-event clinical trials where a wide variety of stochastic ordering alternatives are expected. Yang and Prentice (2010) utilized adaptive weights in the log-rank tests based on the short-term and long-term hazard ratios from the model of Yang and Prentice (2005) to detect proportional and nonproportional alternatives.

In this paper, we propose a group of new methods for comparing hazard rates by combining weighted log-rank tests and a partitioning method. The overall homogeneity of two hazard rates is equivalent to that they are both homogeneous before and after any given time point. We may apply any weighted log-rank tests to examine the homogeneity before and after a given time point. Summing up the two weighted log-rank tests naturally leads to an overall homogeneity test for the two hazard rates. To avoid the subjectiveness in choosing the cutting time point, we take the supremum of the summation of the two test statistics (before and after the partitioning point) over all time points as the proposed test for the overall homogeneity of two hazard rates. A bootstrap method is proposed to determine the distribution of the supremum test statistic under the null hypothesis.

Compared with the method in Qiu and Sheng (2008), the proposed partitioned log-rank tests have several major advantages: (i) Our tests are one-stage procedures and thus are easy to use, while in comparison, Qiu and Sheng (2008)'s method is a two-stage procedure and involves complicated weights. (ii) We can easily extend the proposed approach to testing the homogeneity of more than two hazard rates, which is not straightforward under Qiu and Sheng (2008)'s method. (iii) The construction of time partitions is very flexible, and any pair of weighted log-rank tests for the homogeneity before and after a cutting point produce a new test. Our simulation results indicate that with the commonly-used weighting schemes, the resulting new tests outperform in terms of power their competitors in almost all cases.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the weighted log-rank test and further propose the partitioned log-rank test. The asymptotic properties of the proposed test and a bootstrap procedure are presented in Sect. 3. We extend the proposed test to compare multiple groups in Sect. 4. In Sect. 5, we conduct extensive simulation studies to examine the finite sample performance of the partitioned log-rank test, and we illustrate the new testing procedure with real data examples in Sect. 6. Section 7 concludes with some remarks. All proofs are postponed in the Appendix.

2 Partitioned log-rank test

We begin with comparing two hazard rates, although our proposed tests apply generally to comparison of multiple groups. Suppose that we have two groups of survival data $\{(T_{kj}, C_{kj}) : \ j = 1, \ldots, n_k\}$, where T_{kj} $(j = 1, 2, \ldots, n_k)$ denote the death times of patients in group k $(k = 0, 1)$ and C_{kj}'s are the accompanying censoring times.

We may rewrite the observed survival data as $\{(X_{kj}, \delta_{kj}) : \ j = 1, \ldots, n_k\}$, where $X_{kj} = \min\{T_{kj}, C_{kj}\}$ and $\delta_{kj} = 1(T_{kj} \le C_{kj})$. Here $1(s < t)$ denotes an indicator function, equal to 1 if $s < t$ and 0 otherwise. Let $S_k(t)$ and $h_k(t)$ be the survival and hazard rate function of group k, respectively. Our goal is to test the equality of the hazard functions of the two groups, i.e., $H_0 : \ h_0(t) = h_1(t)$ (or $S_0(t) = S_1(t)$) for all $t \in [0, \ \infty)$, and the alternative is that there exists some difference between $h_0(t)$ and $h_1(t)$ at certain time t, i.e., $H_1 : \ h_0(t) \neq h_1(t)$ (or $S_0(t) \neq S_1(t)$) for some t.

Suppose that in total there are m distinct death times in the pooled sample of the two groups of data, i.e., $t_1 < t_2 < \cdots < t_m$, and at the death time t_i, there are d_{ki} events and r_{ki} individuals at risk in group k ($k = 0, 1$). Let $d_i = d_{0i} + d_{1i}$ and $r_i = r_{0i} + r_{1i}$. The commonly-used weighted log-rank test is defined as

$$T_w = \left\{ \sum_{i=1}^{m} w_i \left(d_{1i} - \frac{d_i r_{1i}}{r_i} \right) \right\}^2 \Big/ \sum_{i=1}^{m} \frac{w_i^2 r_{0i} r_{1i} d_i (r_i - d_i)}{r_i^2 (r_i - 1)},$$

where w_i's are prespecified weights. It includes many well-known tests as special cases. With $w_i = 1$, T_w is actually the original log-rank test (Mantel 1966; Cox 1972); and if $w_i = r_i$, T_w reduces to the Gehan–Wilcoxon test (Gehan 1965; Breslow 1970). Let $\hat{S}(t)$ be the Kaplan–Meier estimator of the common survival function $S(t)$ based on (d_i, r_i) ($i = 1, \ldots, m$), assuming that the two survival functions are equal. The test of Peto and Peto (1972) corresponds to T_w with $w_i = n\hat{S}(t_i)$ and $n = n_0 + n_1$. There are also many other weighting schemes, such as $w_i = \{\hat{S}(t_i)\}^\alpha$ (Fleming and Harrington 1981) and $w_i = r_i^{1/2}$ (Tarone and Ware 1977). See Jones and Crowley (1989) for more discussions on various weighted log-rank tests.

If the two hazard rate functions cross each other, then T_w may lose substantial power because there is some cancellation in $\sum_{i=1}^{m} w_i (d_{1i} - d_i r_{1i}/r_i)$. We propose a group of new tests by partitioning the time axis to circumvent this problem. Testing $H_0 : \ h_0(t) = h_1(t)$ for all $t \in [0, \ \infty)$ is equivalent to testing both $H_{01} : \ h_0(t) = h_1(t)$ for all $t \in [0, \ u)$ and $H_{02} : \ h_0(t) = h_1(t)$ for all $t \in [u, \ \infty)$ simultaneously for all $u \in [0, \ \infty)$, where u is a partitioning time point. We can apply any homogeneity test for two hazard rates to both H_{01} and H_{02}, and the summation of the respective homogeneity tests for H_{01} and H_{02} leads to a natural test for the overall homogeneity H_0.

In particular, we choose the weighted log-rank test as the basis for testing each sub-hypothesis of H_{01} and H_{02}. Define the lower and upper components of the test by the partitioning time point t,

$$D_l(t) = \sum_{i=1}^{m} w_i \left(d_{1i} - \frac{d_i r_{1i}}{r_i} \right) 1(t_i < t), \quad D_u(t) = \sum_{i=1}^{m} w_i \left(d_{1i} - \frac{d_i r_{1i}}{r_i} \right) 1(t_i \ge t)$$

with the accompanying variance estimates,

$$V_l(t) = \sum_{i=1}^{m} \frac{w_i^2 r_{0i} r_{1i} d_i (r_i - d_i)}{r_i^2 (r_i - 1)} 1(t_i < t), \ \ V_u(t) = \sum_{i=1}^{m} \frac{w_i^2 r_{0i} r_{1i} d_i (r_i - d_i)}{r_i^2 (r_i - 1)} 1(t_i \ge t).$$

Throughout the paper, $0/0$ is defined to be zero. Given a time point t, a natural test for H_0 is based on the test statistic

$$\tilde{T}(t) = \frac{D_l^2(t)}{V_l(t)} + \frac{D_u^2(t)}{V_u(t)}.$$

When $t = t_1$ or t_m, $\tilde{T}(t)$ reduces to the usual weighted log-rank test.

Compared with the usual weighted log-rank tests, the test based on $\tilde{T}(t)$ loses little power if the two underlying hazard rates do not cross, while it gains much power if the two underlying hazard rates cross only once and the crossing point happens to be at time t. However, this test is not easy to use since it is difficult to know the exact crossing point in practice. Furthermore, its testing capability is questionable if the two underlying hazard rates cross twice or multiple times. To avoid the arbitrariness of t and possible power loss, we propose to test H_0 by a supremum statistic,

$$T = \sup_{t \in [0,\infty]} \tilde{T}(t) = \max_{1 \le i \le m} \tilde{T}(t_i),$$

which we call a partitioned log-rank test. Similar to the usual log-rank type tests, different weight choices result in different performances for the partitioned log-rank test.

3 Asymptotics

3.1 Two-sample comparison

To understand the theoretical performance of the proposed partitioned log-rank test, we investigate its large-sample properties under the null hypothesis and a series of local alternatives, respectively. Suppose that $\{(T_{kj}, C_{kj}) : j = 1, \ldots, n_k\}$ are independent and identically distributed copies of (T_k, C_k) and that T_k and C_k are independent ($k = 0, 1$). Denote $n = n_0 + n_1$. For group k, let $\pi_k(t) = S_k(t)L_k(t)$, where $S_k(t)$ and $L_k(t)$ are the respective survivor functions of T_k and C_k.

It is convenient to use the notation of counting process when studying the limiting distributions of $\tilde{T}(t)$ and T. Denote $N_k(t) = \sum_{j=1}^{n_k} 1(X_{kj} \le t, \delta_{kj} = 1)$ and $Y_k(t) = \sum_{j=1}^{n_k} 1(X_{kj} \ge t)$ as the event and at-risk processes of group k, respectively. Let $Y.(t) = Y_0(t) + Y_1(t)$, $N.(t) = N_0(t) + N_1(t)$, and $W(t) = w_{\tilde{i}}$, where $\tilde{i} = \max\{i : t_i \le t\}$ and $\{t_i : i = 1, \ldots, m\}$ are the distinct failure times in the pooled sample. Define

$$Z(t) = n^{-1/2} \int_{(0,t]} W(s) \left\{ \frac{Y_1(t)}{Y.(t)} dN_0(s) - \frac{Y_0(t)}{Y.(t)} dN_1(s) \right\},$$

$$\hat{\sigma}(t) = n^{-1} \int_{(0,t]} W^2(s) \frac{Y_0(s)Y_1(s)}{Y.(s)} \left\{ 1 - \frac{\Delta N.(s) - 1}{Y.(s) - 1} \right\} \frac{dN.(s)}{Y.(s)},$$

and $\Delta N.(s) = N.(s) - N.(s-)$. It follows that $D_l(t) = n^{1/2} Z(t)$, $D_u(t) = n^{1/2}\{Z(\infty) - Z(t)\}$, $V_l(t) = n\hat{\sigma}(t)$ and $V_u(t) = n\{\hat{\sigma}(\infty) - \hat{\sigma}(t)\}$.

3.2 Limiting distribution under H_0

Suppose that $n_k/n = \rho_k + o(1)$ with $\rho_k \in (0, 1)$. Let $\Lambda_c(t)$ be the common cumulative hazard function of the two groups under the null hypothesis $H_0 : S_0(t) = S_1(t)$. Define

$$\sigma(t) = \int_{(0,t]} W_0^2(s) \frac{\rho_0\rho_1\pi_0(s)\pi_1(s)}{\pi.(s)} \{1 - \Delta\Lambda_c(s)\} d\Lambda_c(s),$$

where $\pi.(s) = \rho_0\pi_0(s) + \rho_1\pi_1(s)$, $\Delta\Lambda_c(s) = \Lambda_c(s) - \Lambda_c(s-)$ is the jump of $\Lambda(s)$ at s and $d\Lambda_c(s) = \Lambda_c(s + ds) - \Lambda_c(s)$ is the differential of $\Lambda_c(s)$.

The limiting distribution of the proposed partitioned log-rank test under the null hypothesis is given as follows.

Theorem 1 *Suppose that Assumptions 2, 3 and 4 in the Appendix hold for $p = 1$, and that the failure time T_k is independent of the censoring time C_k for each k. If the null hypothesis $H_0 : S_0(t) = S_1(t)$ is true, then as n goes to infinity, it holds that*

(a) $\hat{\sigma}(t)$ converges in probability to $\sigma(t)$ for any $t \in [0, \infty)$,

(b) $\tilde{T}(t) \xrightarrow{d} \chi_2^2$, the chi-squared distribution with two degrees of freedom, for each $t \in \Omega = \{t : 0 < \sigma(t) < \sigma(\infty)\}$, where $\xrightarrow{d}$ means convergence in distribution, and

(c) $T \xrightarrow{d} \sup_{0<t<1} \left[\{B(t)\}^2/t + \{B(1) - B(t)\}^2/(1 - t)\right]$, where $B(t)$ is a Brownian motion.

Let $\hat{S}(t)$ be the Kaplan–Meier estimator in the pooled sample and $\hat{\pi}(s) = Y(s)/n$. If $f(s)$ is a nonnegative bounded continuous function with bounded variation on [0, 1], then $W(s) = f(\hat{S}(s-))$ and $W(s) = f(\hat{\pi}(s))$ satisfy Assumptions 2, 3 and 4. See the proof of Theorem 7.2.1 of Fleming and Harrington (1991). Theorem 1 holds for the log-rank, Gehan–Wilcoxon, and Peto–Peto tests, because their weight functions are respectively $f(s) = 1$, $f(\hat{\pi}(s))$ and $f(\hat{S}(s-))$ with $f(s) = s$, and both $f(s) = 1$ and $f(s) = s$ are bounded variation functions on [0, 1].

Due to its complicated form, the limiting distribution of T under H_0, i.e., the distribution of $\sup_{0<t<1} \left[\{B(t)\}^2/t + \{B(1) - B(t)\}^2/(1 - t)\right]$, is not easy to use to calculate the p value of T. Nevertheless, the limiting distribution does not depend on any parameter of the underlying population survival function or censoring function, which implies that the proposed partitioned log-rank test T is asymptotically pivotal. We may determine the p value for the proposed test by simulations and generating data from any distribution. Unfortunately, our simulation experience indicates that the distribution of T converges so slowly that its limiting distribution is generally a poor approximation for small and moderate sample sizes. Instead, we propose a bootstrap method to determine the p value of T as follows.

1. Let T denote the partitioned log-rank test based on the two groups of original data, $\{(Z_{kj}, \delta_{kj}) : j = 1, \ldots, n_k\}$ $(k = 0, 1)$.
2. Denote the pooled data set by $\{(Z_j, \delta_j) : j = 1, \ldots, n\}$. Let $\{(Z_{kj}^*, \delta_{kj}^*) : j = 1, \ldots, n_k\}$ $(k = 0, 1)$ be two random bootstrap samples with sample sizes n_0 and n_1 sampled from the pooled data set with replacement. Denote the partitioned log-rank test based on these two samples by T^*.

3. Repeating step 2 for a large number of times, say, B times, we obtain B partitioned log-rank test statistics, denoted by $T_1^*, \ldots, T_B^*$. We have two ways to conduct hypothesis testing based on the partitioned log-rank test given the significance level α.
   (a) The p value of the partitioned log-rank test is defined as the proportion of T_i^*'s greater than T. We reject the null hypothesis if the p value is less than α.
   (b) Let c_α be the $B \times \alpha$ largest value of T_i^*'s, and we reject the null hypothesis if $T > c_\alpha$.

3.3 Power under local alternatives

To investigate the testing capability of the proposed partitioned log-rank test, we consider a series of local alternatives, where we allow $S_k(t)$, the survivor function of the failure time in group k, to vary as the sample size n increases. Let $\Lambda_k(t) = -\int_{(0,t]} dS_k(s)/\{1 - S_k(s)\}$.

Assumption 1 For both $k = 0$ and 1, the following conditions hold as $n \to \infty$.

(i) $\sup_{0 \le t \le \infty} |S_k(t) - S_c(t)| \to 0$ for a survival function $S_c(t)$ with respect to which each $S_k(t)$ is absolutely continuous.
(ii) Let $\Lambda_c(t)$ be the cumulative hazard function of $S_c(t)$. $\sqrt{n}\,\{d\Lambda_k(t)/d\Lambda_c(t) - 1\} \to \gamma_k(t)$ uniformly on each closed subinterval of $\{t : S_c(t+) > 0\}$, where $\gamma_k(t)$ is a real-valued function.
(iii) $\sup_{0 \le t \le \infty} |Y_k(t)/n - \rho_k \pi_k(t)| \to 0$, where we redefine $\pi_k(t) = S_c(t) L_k(t)$ under (i),
(iv) $\int_{(0,t]} W_0(s)\pi_k(s)|\gamma_k(s)|d\Lambda_c(s) < \infty$ for $t \in [0, \infty)$.

Assumption 1, adopted from Theorem 7.4.1 of Fleming and Harrington (1991), defines a series of local alternatives. The next theorem gives the local power of the proposed test under such local alternatives.

Theorem 2 *Assume the conditions in Theorem 1 with Assumption 1 in place of the null hypothesis H_0. As n goes to infinity, we have*

(a) *$\hat{\sigma}(t)$ converges in probability to $\sigma(t)$ for any $t \in [0, \infty)$, and*
(b) *$T \xrightarrow{d} \sup_{t \in \Omega} A(t)$, where Ω is defined in Theorem 1 and*

$$A(t) = \frac{\Big[B\{\sigma(t)\} + R(t)\Big]^2}{\sigma(t)} + \frac{\Big[B\{\sigma(\infty)\} - B\{\sigma(t)\} + R(\infty) - R(t)\Big]^2}{\sigma(\infty) - \sigma(t)}$$

with $B(t)$ a Brownian motion and

$$R(t) = \int_0^t W_0(s) \frac{\rho_0 \rho_1 \pi_0(s) \pi_1(s)}{\pi.(s)} \{\gamma_0(s) - \gamma_1(s)\} d\Lambda_c(s).$$

Under the conditions of Theorem 1, the limiting distribution of T is a proper distribution, and theoretically the critical value is a bounded constant given a significance level. Roughly speaking, Theorem 2 implies that the proposed partitioned log-rank test can detect alternatives departing from H_0 at a rate $n^{-1/2}$ if there exists an open interval G such that $R(t) \neq 0$ for all $t \in G$. In comparison, the usual weighted log-rank test can detect alternatives at the same rate only if $R(\infty) \neq 0$. See, for example, Theorem 7.4.1 of Fleming and Harrington (1991). Clearly the set of detectable alternatives under the usual weighted log-rank test is only a proper subset of that of the proposed partitioned log-rank test. We consequently expect that whenever the weighted log-rank test have desirable power, the proposed partitioned log-rank test would also have desirable power, while the opposite is not true, particularly when the underlying two hazard rate functions cross once or multiple times.

4 Extension to multiple groups

The proposed testing procedure can be easily extended to compare the hazard rates of multiple groups, which is an obvious advantage over Qiu and Sheng (2008)'s two-stage test. Suppose that we compare $p+1$ $(p > 1)$ groups of survival data, and define the same notation as in Sect. 3, while allowing k to range from 0 to p. The problem of interest is to test $H_0 : S_0(t) = S_1(t) = \cdots = S_p(t)$ versus H_1 : at least one of the equalities fail to hold for some t.

We still use $t_1 < t_2 < \cdots < t_m$ to denote the distinct death times across all the $p+1$ samples. Let $d_i = \sum_{k=0}^{p} d_{ki}$ and $r_i = \sum_{k=0}^{p} r_{ki}$. Define

$$D_l(t) = \sum_{i=1}^{m} w_i \left(d_{1i} - \frac{r_{1i} d_i}{r_i},\ d_{2i} - \frac{r_{2i} d_i}{r_i},\ \ldots, d_{pi} - \frac{r_{pi} d_i}{r_i} \right)^{\mathrm{T}} 1(t_i < t),$$

$$D_u(t) = \sum_{i=1}^{m} w_i \left(d_{1i} - \frac{r_{1i} d_i}{r_i},\ d_{2i} - \frac{r_{2i} d_i}{r_i},\ \ldots, d_{pi} - \frac{r_{pi} d_i}{r_i} \right)^{\mathrm{T}} 1(t_i \geq t),$$

where "T" denotes the vector transpose. Let $V_i = (v_{i,jk})_{1 \leq j,k \leq p}$ be a $p \times p$ matrix with $v_{i,jk} = r_{ji}\{r_i 1(j=k) - r_{ki}\} d_i (r_i - d_i)/\{r_i^2 (r_i - 1)\}$. The lower and upper variance–covariance matrices accompanying the partition time point t are respectively given by

$$V_l(t) = \sum_{i=1}^{m} V_i 1(t_i < t) \quad \text{and} \quad V_u(t) = \sum_{i=1}^{m} V_i 1(t_i \geq t).$$

As a result, the proposed partitioned log-rank test in the multiple group case is

$$T = \sup_{t \in [0,\infty]} \tilde{T}(t),$$

where $\tilde{T}(t) = \{D_l(t)\}^{\mathrm{T}} \{V_l(t)\}^{-1} D_l(t) + \{D_u(t)\}^{\mathrm{T}} \{V_u(t)\}^{-1} D_u(t)$.

To extend Theorem 1 from $p = 1$ to $p \geq 1$, we define more notation as follows. Let $Y.(t) = \sum_{k=0}^{p} Y_k(t)$, $N.(t) = \sum_{k=0}^{p} N_k(t)$, $Y(t) = (Y_0(t), \ldots, Y_p(t))^{\mathrm{T}}$ and

$N(t) = (N_0(t), \ldots, N_p(t))^{\mathrm{T}}$. Furthermore, let

$$Z(t) = n^{-1/2} \int_{(0,t]} W(s) \left\{ I_{p+1} - Y(s) J^{\mathrm{T}} / Y(s) \right\} dN.(s),$$
$$\widehat{\Sigma}(t) = n^{-1} \int_{(0,t]} W^2(s) \left\{ Y_*(s) - Y(s) Y^{\mathrm{T}}(s) / Y.(s) \right\} \left\{ 1 - \frac{\Delta N.(s) - 1}{Y.(s) - 1} \right\} \frac{dN.(s)}{Y.(s)},$$

where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix and J a $(p+1)$-variate vector with all elements being one and $Y_*(t) = \mathrm{diag}\{Y(t)\}$. Hereafter we use Y_* to denote the diagonal matrix $\mathrm{diag}\{Y\}$ for any vector Y. It can be shown that $D_l(t) = n^{1/2} Z_{[-1]}(t)$, $D_u(t) = n^{1/2}\{Z_{[-1]}(\infty) - Z_{[-1]}(t)\}$, $V_l(t) = n\widehat{\Sigma}_{[-1,-1]}(t)$ and $V_u(t) = n\{\widehat{\Sigma}_{[-1,-1]}(\infty) - \widehat{\Sigma}_{[-1,-1]}(t)\}$, where and in what follows for a generic vector Z, $Z_{[-1]}$ denotes Z with its first row removed, and for a generic matrix Σ, $\Sigma_{[-1,-1]}$ denotes Σ with both its first row and first column removed.

Theorem 3 *Suppose Assumptions 2, 3 and 4 in the Appendix hold, that the failure time T_k is independent of the censoring time C_k for each k. Under $H_0 : S_0(t) = S_1(t) = \cdots = S_p(t)$, as n goes to infinity, it holds that*

(a) *$\widehat{\Sigma}_{[-1,-1]}(t)$ converges in probability to $\Sigma_{[-1,-1]}(t)$ for $\Sigma(t)$ given in (2) in the Appendix and any $t \in [0, \infty)$,*
(b) *$\tilde{T}(t) \xrightarrow{d} \chi^2_{2p}$ for each $t \in \Omega$, a set consisting of t values such that both $\Sigma_{[-1,-1]}(t)$ and $\Sigma_{[-1,-1]}(\infty) - \Sigma_{[-1,-1]}(t)$ are nonsingular, and*
(c) *$T \xrightarrow{d} \sup_{t\in\Omega} A_1(t)$, where the stochastic process $A_1(t)$ is defined in (3).*

Similar to the case with $p = 1$, the limiting distribution of the partitioned log-rank test in Theorem 3 is also too complicated to use in practice. We again adopt the bootstrap method (with slight modification) to calculate the p value of the partitioned log-rank test for general p.

In parallel to Theorem 2, the next theorem explores the local power of the partitioned log-rank test test for general p.

Theorem 4 *Assume the conditions in Theorem 2 hold for $k = 0, 1, \ldots, p$. As n goes to infinity, we have*

(a) *$\widehat{\Sigma}_{[-1,-1]}(t)$ converges in probability to $\Sigma_{[-1,-1]}(t)$ for any $t \in [0, \infty)$, and*
(b) *$T \xrightarrow{d} \sup_{t\in\Omega} A_2(t)$, where the stochastic process $A_2(t)$ is defined in (5).*

Although both the partitioned and weighted log-rank tests can detect alternatives departing from H_0 at a rate of $n^{-1/2}$, the sets of their detectable alternatives are different. The partitioned log-rank tests can detect the alternatives in Assumption 1 if there exists an open interval G such that $R(t) \neq 0$ for all $t \in G$, where $R(t)$ is defined in (4). However, the weighted log-rank test can detect the alternatives only if $R(\infty) \neq 0$. The conclusion for multiple group comparison remains the same that the partitioned log-rank test generally performs better than the weighted log-rank test.

5 Simulation studies

To evaluate the finite-sample performance of the proposed tests, we conduct extensive simulations and make thorough comparisons with existing methods. Let SUP-LR, SUP-GW and SUP-PP denote the partitioned log-rank tests corresponding to $w_i = 1, r_i$ and $n\hat{S}(t_i)$. We compare these three new tests with the usual log-rank test (LR for short), the Gehan–Wilcoxon test (GW), Peto and Peto (1972)'s test (PP) and Qiu and Sheng (2008)'s test (QS). The significance level is set to be 5 %, and the type I error and power for each configuration are calculated based on 2000 replications. The p values of the LR, GW and PP tests are determined by their limiting chi-squared distributions. We employ the proposed bootstrap procedure with 1000 bootstrap samples to calculate the p values for both the proposed tests and the QS test.

We consider four simulation examples: All except the third example compare two groups of survival times ($p = 1$), and the third compares three groups ($p = 2$). In each scenario, we chose the sample sizes across the $p + 1$ groups of data to be equal, $n_k = 50$ or 100, and we examined both censoring and non-censoring cases. In the case of censoring, the censoring times for Examples 1, 2 and 3 were generated from the uniform distribution Unif(0, 2), which led to censoring rates between 24 and 54 % for different scenarios. The censoring times in Examples 4 were generated from Unif(1, 3) or Unif(5, 7), which led to censoring rates between 10 and 55% for different scenarios.

Example 1 (Qiu and Sheng 2008) Consider four hazard rate functions $h_0(t) = 1$, $h_1(t) = 2$, $h_2(t) = 0.3 + t$ and $h_4(t) = 1.2 + 0.6t$. We generate data from the following four cases: (a) $\{h_0, h_0\}$, (b) $\{h_0, h_1\}$, (c) $\{h_0, h_2\}$ and (d) $\{h_0, h_3\}$ for two-sample comparison.

Example 1 is adopted from Qiu and Sheng (2008) and in this example both censoring and non-censoring cases are considered. Among the four different pairs of hazard rates, the two hazard rate functions under comparison coincide in case (a), parallel and also proportional in case (b), cross only one time in case (c) and neither parallel nor cross in case (d). As shown in Fig. 1, the plots represent four common patterns between two hazard rate functions that cross at most one time.

The simulated rejection probabilities in percentage are tabulated in Table 1 for the four scenarios of Example 1. Case (a) is the null hypothesis and thus the rejection probability corresponds to the type I error rate, while the rest are the power of the tests. We observe that all the methods control the type I error rates around the nominal level, which implies that the proposed bootstrap method performs well in calculating the p values for the proposed tests and the QS test. When the null hypothesis does not hold in cases (b), (c) and (d), the proposed tests and the QS test behave similarly. When the two underlying hazard rate functions cross and no censoring is accommodated such as in case (c), the new tests and the QS test have remarkable power gain over the weighted log-rank tests (LR, GW and PP). In cases (b) and (d), all the test results are comparable, although the LR test has slight advantages. As expected, the power of all the tests increases as the sample size increases.

We also examine the performance of the partitioned log-rank test calibrated by its limiting distribution, which, denoted by Q, is defined in result (c) of Theorem 1.

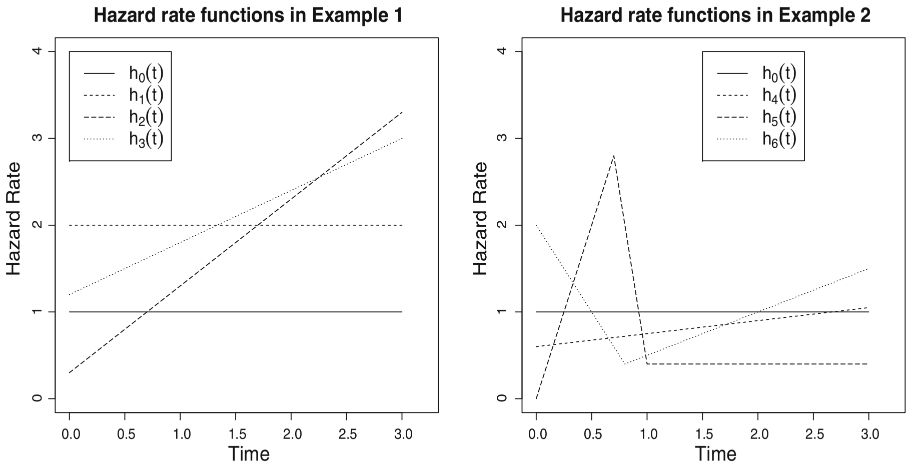

Fig. 1 Display of the hazard rate functions in Examples 1 and 2

Specifically, we take the upper 5 % quantile of Q as the critical value of the partitioned log-rank test. Since the upper 5 % quantile of Q has no closed form, we use a numerical procedure to approximate it.

(1) Generate $M = 10,000$ observations $x_1, \ldots, x_M$ from the standard normal distribution $N(0, 1)$;
(2) For $i = 1, \ldots, M - 1$, calculate $B_i = M^{-1/2} \sum_{j=1}^{i} x_j$ and $y_i = B_i^2/(i/M) + (B_M - B_i)^2/\{1 - (i/M)\}$;
(3) Calculate $t = \max_{1 \le i \le M-1} y_i$;
(4) Repeat steps (1), (2) and (3) $N = 10,000$ times. Denote the resulting t values as $t_1, \ldots, t_N$. The upper 5 % quantile of Q can be approximated by the upper 5 % quantile of $\{t_1, t_2, \ldots, t_N\}$.

We use SUP-LR*, SUP-GW* and SUP-PP* to denote the partitioned log-rank tests LR, GW and PP calibrated by the limiting distribution Q. Their simulated sizes and

Table 1 Simulated sizes (%) in column (a) and powers (%) in columns (b), (c) and (d) of the tests under comparison for Example 1 with two groups of survival data

n_k	Log-rank methods	Non-censoring				Censoring			
		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
50	LR	5.70	91.65	4.00	61.05	4.90	79.70	16.35	31.10
	GW	4.65	83.85	19.65	34.10	4.40	70.00	37.25	18.70
	PP	4.60	83.75	20.15	33.95	4.05	73.00	31.50	22.25
	QS^1	4.60	86.40	58.05	52.00	4.75	70.25	33.65	26.35
	SUP-LR	5.15	83.05	59.60	57.85	5.40	66.25	35.85	25.70
	SUP-GW	5.05	83.25	53.30	54.20	5.00	67.40	33.70	24.35
	SUP-PP	5.05	83.15	52.50	54.10	4.90	66.90	34.50	25.20
	SUP-LR*	1.83	71.56	38.81	40.72	1.04	45.30	13.70	11.15
	SUP-GW*	1.54	72.14	31.64	35.81	0.83	46.76	12.29	9.96
	SUP-PP*	1.50	71.99	30.93	35.44	0.94	46.73	12.52	10.66
100	LR	4.45	99.75	3.95	90.65	4.30	98.15	27.40	54.70
	GW	4.00	98.75	36.50	59.25	4.15	94.20	62.70	31.70
	PP	4.00	98.70	37.05	59.10	4.65	95.60	51.95	38.00
	QS^1	4.35	99.35	89.50	86.25	5.25	96.05	57.75	47.05
	SUP-LR	5.35	98.70	91.45	88.10	5.20	92.55	63.40	46.00
	SUP-GW	4.40	98.70	86.55	84.95	5.35	92.40	61.00	44.00
	SUP-PP	4.30	98.80	86.30	84.90	5.20	92.50	62.05	44.80
	SUP-LR*	2.05	97.29	82.03	80.19	1.32	84.46	41.39	28.57
	SUP-GW*	2.02	97.57	76.52	76.05	1.41	85.88	40.42	26.81
	SUP-PP*	2.04	97.56	76.12	75.82	1.41	85.68	40.72	28.38

LR the usual log-rank test, *GW* the Gehan–Wilcoxon test, *PP* the Peto–Peto test, *QS*^{1} Qiu and Sheng (2008)'s test designed for at most one crossing point, *SUP-LR*, *SUP-GW*, *SUP-PP* are the corresponding partitioned log-rank tests based on bootstrap, *SUP-LR**, *SUP-GW**, and *SUP-PP** are the partitioned log-rank tests calibrated by their limiting distributions

Table 2 Power (%) of the tests under comparison for Example 2 with two groups of survival data subject to censoring

Log-rank methods	$n_k = 50$				$n_k = 100$			
	(e)	(f)	(g)	(h)	(e)	(f)	(g)	(h)
LR	6.90	11.70	46.05	53.60	9.40	17.45	75.80	82.20
GW	5.75	26.00	28.65	72.25	5.60	46.50	47.65	95.55
PP	4.65	21.90	36.45	67.60	4.70	38.15	61.90	93.05
QS^2	48.70	20.70	64.35	53.20	83.95	38.15	92.30	85.05
SUP-LR	44.20	32.30	64.00	65.15	79.10	55.10	92.35	92.45
SUP-GW	51.35	30.90	69.45	64.90	84.20	53.90	94.45	93.40
SUP-PP	49.05	31.30	67.85	64.30	83.30	54.40	93.80	92.85

LR the usual log-rank test, *GW* the Gehan–Wilcoxon test, *PP* the Peto–Peto test, *QS*^{2} Qiu and Sheng (2008)'s test designed for two crossing points, and *SUP-LR*, *SUP-GW* and *SUP-PP* are the corresponding partitioned log-rank tests

powers based on 10,000 repetitions are also reported in Table 1. It can be seen that the test sizes are substantially less than the significance level, which leads to severely lower powers than those of the bootstrap calibrated SUP-LR, SUP-GWand SUP-PP. As the sample size increases, the limiting distribution Q provides a more reasonable approximation for the finite-sample distribution of the partitioned log-rank test. Hence, we recommend using the bootstrap method to calibrate the partitioned log-rank test.

Example 2 Define another three hazard rate functions, $h_4(t) = 0.6 + 0.15t$,

$$h_5(t) = \begin{cases} 4t, & t \in [0,\ 0.7] \\ 8.4 - 8t, & t \in (0.7,\ 1] \\ 0.4, & t \in (1,\ \infty) \end{cases} \quad \text{and} \quad h_6(t) = \begin{cases} 2 - 2t, & t \in [0,\ 0.8] \\ 0.5t, & t \in (0.8,\ \infty) \end{cases}.$$

Consider comparisons of the following four pairs of hazard functions: (e) $\{h_0, h_5\}$, (f) $\{h_0, h_6\}$, (g) $\{h_4, h_5\}$ and (h) $\{h_4, h_6\}$, where $h_0(t) = 1$ is defined in Example 1.

For both Examples 2 and 3, we focus on cases subject to censoring. As shown in Fig. 1, the four pairs of hazard rates in Example 2 are designed to have exactly two crosses. All the cases of (e)–(h) are constructed under the alternatives of the hypothesis tests, and the simulated power values are presented in Table 2. In case (e), the weighted log-rank tests completely fail to detect the survival difference between the two groups, because the hazard rate functions under comparison cross twice that may result in much cancellation in the numerator of the test statistic, $\sum_i w_i(d_{i1} - r_{i1}d_i/r_i)$. In comparison, the QS^2 test succeeds in overcoming such a weakness of the weighted log-rank tests and attains desirable power, since it is specially designed to detect such an alternative where there are two crossing points in the two hazard rate functions. On the other hand, all the proposed tests are very competitive with the QS^2 test for both sample sizes $n_k = 50$ and 100. In cases (f) and (h), the proposed tests have about 10 % power gain over the QS^2 test and even the GW and PP tests are also more powerful, for which a possible explanation is that the two crossing points ($t = 2$ and 1.714) are

Table 3 Simulated sizes (%) in column (a) and powers (%) in columns (b), (c) and (d) of the tests under comparison for Example 3 with three groups of survival data subject to censoring.

Log-rank methods	$n_k = 50$				$n_k = 100$			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
LR	5.30	40.05	24.70	27.35	5.57	65.90	44.85	48.62
GW	5.80	28.55	32.70	36.90	5.62	49.52	60.53	67.25
PP	5.85	33.80	29.60	36.00	5.50	57.27	55.25	65.27
SUP-LR	6.00	61.85	37.65	59.55	6.45	91.62	69.03	90.92
SUP-GW	7.00	68.85	37.40	67.25	5.55	94.72	66.35	93.30
SUP-PP	6.65	66.95	37.85	65.05	5.60	93.80	68.23	92.95

LR the usual log-rank test, *GW* the Gehan–Wilcoxon test, *PP* the Peto–Peto test, *SUP-LR*, *SUP-GW* and *SUP-PP* are the corresponding partitioned log-rank tests

much larger and there are very few observations around these points. If we ignore the observations larger than $t = 1.7$, the two hazard rate functions in both cases (f) and (h) cross only one time. This may downplay the power of the QS^2 test, as it is specially designed for situations with two crosses.

Example 3 Given the hazard rate functions defined in Examples 1 and 2, consider the following four combinations of hazard rate functions: (A) $\{h_0, h_0, h_0\}$, (B) $\{h_0, h_4, h_5\}$, (C) $\{h_0, h_2, h_4\}$ and (D) $\{h_0, h_2, h_6\}$.

Example 3 is designed to compare three groups of survival data, for which the proposed tests are readily applicable while the QS test is not. As a result, for the data generated from Example 3, we only compare the proposed tests with the three weighted log-rank tests.

The null hypothesis $H_0 : h_0(t) = h_1(t) = h_2(t)$ holds in case (A) and is violated in the other three cases. In particular, the hazard rate functions in each of cases (B), (C) and (D) cross at least once. From Table 3, we observe that the type I error rates of the proposed tests in case (A) are slightly inflated while still acceptable. With a larger sample size, the sizes of our tests would be close to the nominal level. For power comparison, the proposed tests are uniformly more powerful than the weighted log-rank tests in the three cases of (B), (C) and (D). For example, in cases (B) and (D), the proposed tests have around 30 % gain in power over the weighted log-rank tests for both $n_k = 50$ and 100. In several extreme cases, the power of our partitioned log-rank tests even double that of the counterparts.

Example 4 From each of the Weibull, Gamma and Lognormal distribution families, we choose three distributions in a way that the hazards of any two of them would cross. To be specific, let the density functions of a Weibull distribution (WB), a Gamma distribution (GM) and a Lognormal distribution (LN) be

$$\begin{aligned} f_{\mathrm{WB}}(x; \alpha, \beta) &= \alpha\beta x^{\alpha-1} \exp(-\beta x^{\alpha}), \\ f_{\mathrm{GM}}(x; \alpha, \beta) &= \beta^{\alpha} x^{\alpha-1} \exp(-\beta x)/\Gamma(\alpha), \\ f_{\mathrm{LN}}(x; \alpha, \beta) &= x^{-1}(2\pi\beta^2)^{-1/2} \exp[-\{\ln(x) - \alpha\}^2/(2\beta^2)], \end{aligned}$$

Table 4 Parameters of the distributions in Example 4 and censoring rates

Distribution parameters	Weibull			Gamma			Lognormal		
	WB1	WB2	WB3	GM1	GM2	GM3	LN1	LN2	LN3
α	0.5	1	3	1	2	4	0	0.1	0
β	0.2633	0.1	0.0021	0.5	0.6667	1	0.5	1	2
Censoring rate (%)	42.0	33.6	10.0	24.3	40.0	55.7	10.7	20.3	30.6

respectively, where $\Gamma(\cdot)$ is the Gamma function. If censoring is considered, we set the censoring distribution accompanying the Weibull and Gamma distributions to be Unif$(1, 3)$, and that accompanying the Lognormal distribution to be Unif$(5, 7)$. The distribution parameters and censoring rates are presented in Table 4; and their hazard functions and survival functions are displayed in Fig. 2. Under two-sample comparisons, we generate data from nine cases: (1) WB2 and WB2, (2) WB2 and WB1, (3) WB2 and WB3, (4) GM2 and GM2, (5) GM2 and GM1, (6) GM2 and GM3, (7) LN2 and LN2, (8) LN2 and LN1, (9) LN2 and LN3.

The simulation results based on 2000 repetitions at the 5 % significance level are tabulated in Table 5. The two hazard functions cross in all the nine cases, and the partitioned log-rank tests always produce the largest power. As expected, the log-rank tests including LR, GW and PP have little power in cases (2), (3), (8) and (9). Their unexpected high power values in cases (5) and (6) may be due to the fact that the involved hazard functions cross at a later time. The crossing points are around 5, and the populations GM2 and GM3 have at least 80 % probability between 0 and 5, as shown by the survival functions of the Gamma distributions in Fig. 2.

Overall, the proposed partitioned log-rank tests are able to maintain the type I error rates for testing the homogeneity of hazard or survivor functions. Whether the hazard functions under comparison cross or not, our methods are well adapted to all kinds of alternatives and thus can always attain desirable testing power. Both the proposed tests and the weighted log-rank tests are applicable to the general cases of comparing multiple groups of survival data. The proposed tests display comparable testing capability wherever the weighted log-rank tests perform well. When the hazard rate functions under study cross at least once, the weighted log-rank tests tend to lose much power and even possibly fail to detect the differences. Remarkably, the proposed tests work well regardless of whether the hazard rate functions cross once or twice, or two- or multi-group comparisons. The QS test is specially developed for two-group comparison. and is thus not applicable for general multi-group comparison. With two groups, the proposed tests tend to have very close performance to the QS test when the two hazard rate functions cross at most once, while the proposed tests often outperform the QS test when there are two crosses.

As pointed out by an anonymous referee, it would be instructive to quantify the difference of the two survival curves in the cases where the null hypothesis is violated, so that it is generally known which scenario is more difficult and which is easier to detect. We take the Integral of the Squared Difference of two Survival functions (ISDS for short) as a measure of the survival difference of two populations. The ISDS values

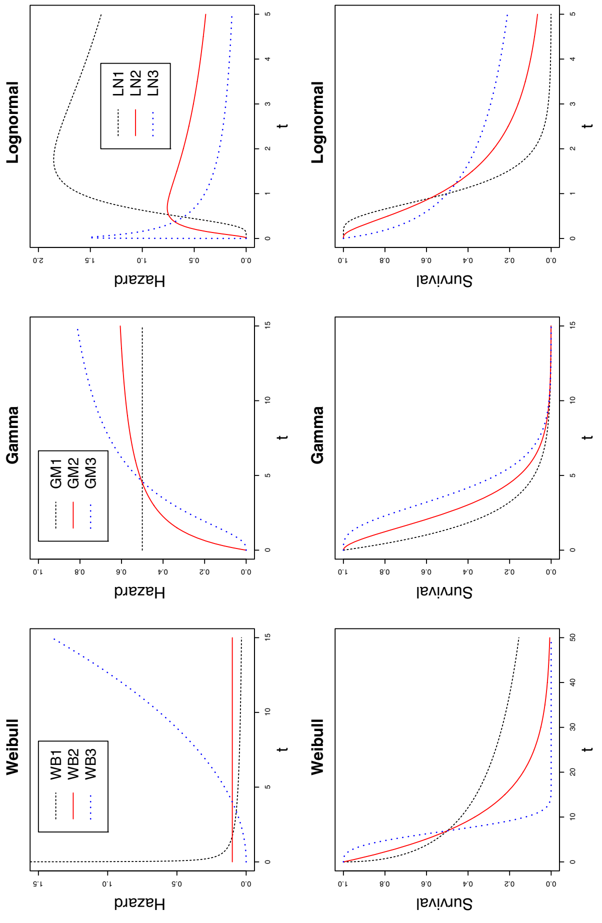

Fig. 2 Hazard rate functions and survival functions of three Weibull distributions (*left panel*), three Gamma distributions (*middle panel*) and three Lognormal distributions (*right panel*)

Table 5 Rejection probabilities (%) of the tests under comparison with data generated from Example 4

Methods	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$(n_0, n_1) = (50, 50)$, Non-censoring									
LR	5.35	47.05	68.75	5.20	62.65	52.95	5.90	59.50	26.35
GW	4.55	6.45	7.65	5.65	87.60	81.80	4.70	11.15	6.50
PP	4.60	6.25	7.80	5.70	87.70	82.05	4.70	10.85	6.70
QS^1	4.90	92.75	100.00	5.90	74.55	72.00	5.70	91.90	91.95
SUP-LR	7.05	97.70	100.00	5.85	78.00	75.35	6.05	97.00	94.50
SUP-GW	5.50	95.00	100.00	6.20	78.60	76.05	6.40	93.70	91.65
SUP-PP	5.40	94.85	100.00	6.15	78.70	75.85	6.40	93.55	91.55
$(n_0, n_1) = (50, 50)$, Censoring									
LR	4.10	9.35	25.20	5.45	66.95	59.90	5.20	32.70	4.60
GW	4.80	16.60	48.15	5.90	87.90	82.50	4.60	6.65	14.60
PP	4.55	15.35	41.30	5.85	88.05	82.55	4.35	7.60	12.20
QS^1	6.10	46.35	81.35	5.95	75.10	71.70	6.75	85.40	80.80
SUP-LR	5.80	47.90	85.20	5.90	78.65	76.55	6.50	89.60	81.50
SUP-GW	5.60	47.40	84.40	6.20	78.70	76.85	6.85	85.40	78.95
SUP-PP	5.65	47.65	85.15	6.20	78.65	76.40	6.75	87.05	80.55
$(n_0, n_1) = (100, 100)$, Non-censoring									
LR	5.30	79.50	94.60	6.15	89.30	79.80	6.00	89.65	49.05
GW	5.45	7.65	8.75	4.95	99.15	98.30	5.60	15.35	8.70
PP	5.55	7.60	8.80	4.95	99.15	98.30	5.55	15.05	9.35
QS^1	6.55	99.95	100.00	6.05	96.95	96.65	6.40	99.90	99.95
SUP-LR	6.45	100.00	100.00	6.45	98.40	98.30	6.35	99.95	100.00
SUP-GW	6.35	99.95	100.00	6.65	98.60	98.05	6.40	99.90	99.90
SUP-PP	5.95	99.95	100.00	6.50	98.60	98.10	6.55	99.90	99.85
$(n_0, n_1) = (100, 100)$, Censoring									
LR	5.20	14.80	47.10	5.85	92.00	88.00	5.45	59.60	5.40
GW	5.25	30.75	78.30	4.85	99.20	98.40	5.45	6.75	25.60
PP	5.35	28.60	69.40	4.85	99.20	98.40	5.65	8.55	20.90
QS^1	6.55	77.30	98.70	5.85	97.20	96.85	6.05	99.00	98.75
SUP-LR	6.60	81.40	99.60	6.30	98.85	98.45	7.10	99.75	98.95
SUP-GW	6.85	80.85	99.40	6.40	98.75	98.50	6.50	99.05	98.50
SUP-PP	6.90	81.30	99.60	6.50	98.75	98.45	6.60	99.40	98.70
ISDS	0	1.9715	0.8097	0	0.1811	0.1809	0	0.0890	0.3017

Columns (1), (4) and (7) are sizes and the rest are powers

LR the usual log-rank test, *GW* the Gehan–Wilcoxon test, *PP* the Peto–Peto test, QS^1 Qiu and Sheng (2008)'s test designed for one crossing point, *SUP-LR*, *SUP-GW*, *SUP-PP* are the corresponding partitioned log-rank tests, *ISDS* the Integral of the Squared Difference of two Survival functions

of cases (a)–(h) in Examples 1 and 2 are 0, 0.0833, 0.0213, 0.0353, 0.0235, 0.0120, 0.0462, and 0.0390, respectively; and those for cases (2), (3), (5), (6), (8) and (9) in Example 4 are 1.9715, 0.8097, 0.1811, 0.1809, 0.0890, and 0.3017, respectively. It can be seen from Tables 1, 2 and 5 that roughly speaking, the larger the ISDS, the larger the power of the partitioned log-rank test. Hence, ISDS does shed light on the difficulty of detecting the inequality of two survival functions: the case with a smaller ISDS value is more difficult to detect. On the other hand, censoring may interplay between ISDS and power. For example, the partitioned log-rank tests have similar power in cases (c) and (d) with no censoring, while the power in case (c) is larger than that in case (d) when there is censoring. In addition, the tests under consideration are based on hazard differences, while ISDS measures the difference of two survival functions. As a result, there is no deterministic relationship between ISDS and power, which can be complicated in the presence of censoring. For example, the partitioned log-rank tests give larger power values in case (3) than in case (2), although case (3) (ISDS $= 0.8097$) has a smaller ISDS value than case (2) (ISDS $= 1.9715$).

6 Real data analysis

As illustrations, we apply the proposed partitioned log-rank tests to three real data sets: the rats data, the kidney data, and the gastric cancer data, which are detailed below. For each data, the Kaplan–Meier survival curves and the smoothed hazard rate function estimates are displayed in Fig. 3. For comparison, Table 6 shows the test statistics and the accompanying p values of the weighted log-rank tests, the proposed tests and the QS test.

6.1 Rats data

The rats data set (Mantel et al. 1977) was obtained from a study on the tumorigenesis of an experimental drug. This study involved a total of 300 rats, with half male and half female. Following Qiu and Sheng (2008), we focused on the female half. The 150 female rats were divided into 50 litters of size 3. For each litter, one rat was randomly selected and treated with the experimental drug, and the other two were administered a placebo. There are 29 censored observations for the times to tumor in the treatment group, and 81 censored observations in the control group.

Although the two Kaplan–Meier survival curves in Fig. 3 appear to cross at around 60 weeks, their hazard rate functions do not have an obvious crossing point. It is not surprising to observe in Table 6 that all the tests produce highly significant results. In particular, the proposed tests result in the smallest p values, the QS and LR tests follow the second, and the two weighted log-rank tests yield the largest p values. According to the definition of the QS p-value, if the LR p value is smaller than $1-(1-0.05)^{1/2}$, the QS p value is defined to be the same as the LR p value.

6.2 Kidney data

Nahman et al. (1992) reported a study to assess times to the first exit-site infection (in months) in patients with renal insufficiency. The study involved 43 patients who utilized a surgically placed catheter (group 1) and 76 patients who utilized a percutaneous placement of their catheter (group 2). Censoring was mainly caused by catheter

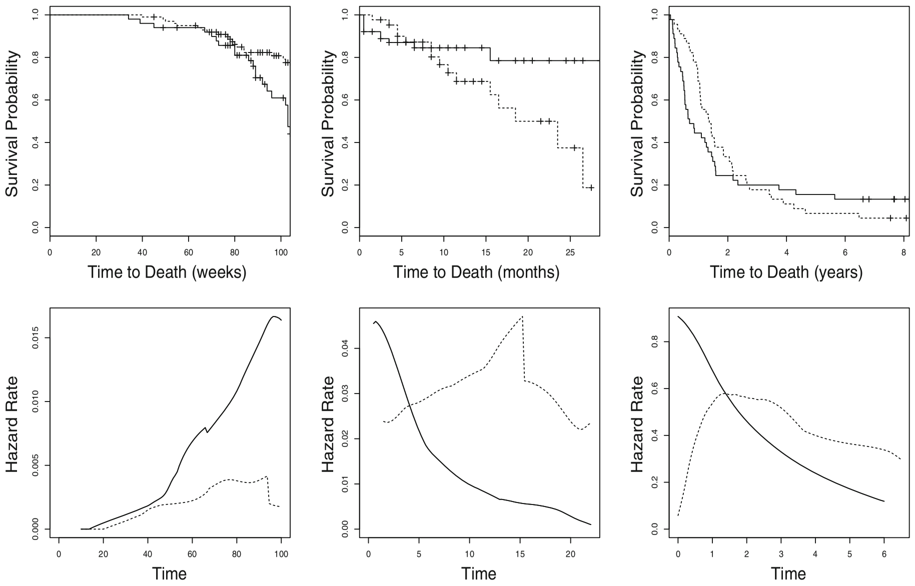

Fig. 3 Kaplan–Meier survival curves and the accompanying hazard rate functions (after smoothing) of the three data-sets: the rats, kidney, and gastric cancer data sets (from *left* to *right*)

Table 6 Test statistics and accompanying p values of the tests under comparison applied to three real data sets

Data	Test	LR	GW	PP	SUP-LR	SUP-GW	SUP-PP	QS
Rats	Statistic	8.5945	4.9284	6.9018	20.5099	20.6351	20.3868	11.8105
	p value	0.0034	0.0264	0.0086	0.0000	0.0000	0.0000	0.0034
Kidney	Statistic	2.5295	0.0021	1.3618	10.2389	9.0278	9.9972	7.7821
	p value	0.1117	0.9636	0.2432	0.0050	0.0080	0.0080	0.0260
Gastric	Statistic	0.2252	3.9637	4.0939	17.3028	15.3378	15.3065	17.1130
	p value	0.6351	0.0465	0.0430	0.0030	0.0060	0.0058	0.0257

LR the usual log-rank test, *GW* the Gehan–Wilcoxon test, *PP* the Peto–Peto test, *SUP-LR*, *SUP-GW*, *SUP-PP* are the corresponding partitioned log-rank tests, *QS* Qiu and Sheng (2008)'s test

failure, and as a result, there were 27 censored observations in group 1 and 65 censored observations in group 2. The data set can be found in Klein and Moeschberger (2003), and has been analyzed by Lin and Wang (2004) and Qiu and Sheng (2008).

As shown in Fig. 3, the survival curves and the hazard rate functions are very different between the two groups and both cross at one time point. The survival curves cross at roughly 6 months, while the hazard rate functions cross at roughly 4 months. All the weighted log-rank tests yield nonsignificant results and, in particular, the GW test (with a p value of 0.9636), completely fails to detect the survival difference. The partitioned log-rank tests (with p values of between 0.0050 and 0.0080) and the QS test (with a p value of 0.0260) achieve very desirable results, while the former have much smaller p values than the latter. As a further comparison, the test statistic in Lin and Wang (2004) is 2.2516, and the accompanying p value is 0.0242 with a two-sided test. As a conclusion, our approaches produce the most powerful test for this data set, which can be also confirmed by the enormous differences between the two survival curves in Fig. 3.

6.3 Gastric cancer data

The third example was a study reported by the Gastrointestinal Tumor Study Group (1982) to compare chemotherapy with combined chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer. A total of 90 patients were admitted into this clinical trial and they were randomly divided into two groups. Each treatment group had 45 patients, with two observations in the chemotherapy group and six in the combination therapy group censored. The data set is available in the R package `YPmodel`, and the improved log-rank test in Yang and Prentice (2010) gives a p value of 0.0304.

Similar to the kidney data, both the survival curves and the hazard rate functions of the gastric cancer data cross at one time point. The crossing points are roughly 2.8 and 1.6 years, respectively. The survival curves seems to be close to each other with a switch leading to roughly equal areas between the two curves, while the hazard rate functions are very different, especially at the early follow-up times. All the tests except for the LR test produce significant results, as with equal weights the LR test cancels out the differences before and after the crossing point. Once again the partitioned log-rank tests produce the smallest p values between 0.0030 and 0.0060, which are much smaller than those of the Yang and Prentice (2010) test, the QS test, the GW and PP tests. This indicates that the proposed tests usually provide the strongest evidence for the heterogeneity of the survival or hazard rate functions under comparison.

7 Discussion

We have proposed the partitioned log-rank test for the homogeneity of two or multiple hazard rates by partitioning the weighted log-rank test at a certain time point. We investigate the asymptotic properties of the new test under the null hypothesis and a series of local alternatives. Despite a concise form, the limiting distribution of the proposed test is not reliable to use. Instead, we propose to calculate the p value via a bootstrap method. Our simulation study indicates that the proposed tests are as powerful as the usual weighted log-rank tests if the hazard rate functions do not cross, and are much more powerful than the latter otherwise. In addition, the proposed tests outperform Qiu and Sheng (2008)'s two-stage procedure if the hazard rate functions have two crosses, although they are generally as powerful if the hazard rate functions have at most one cross.

In the definition of $\tilde{T}(u)$, survival times are partitioned into two exclusive intervals, while we may partition them into more exclusive intervals based on multiple cutting points and define new tests accordingly. However, based on our simulation experience, the power gain is very limited if we partition the time points into more than two intervals, which however complicates the computation immensely.

Although introduced in the context of right censoring, the partitioned log rank test also applies to survival data subject to other censoring schemes such as left censoring and interval censoring (Klein and Moeschberger 2003; Sun 2006; Chen et al. 2012). Through partitioning, the log rank test tailored for left or interval censoring can be first conducted in each time segment and then combined for overall inference.

Acknowledgements We would like to thank the Associate Editor and two anonymous referees for their careful reading and many insightful suggestions, which strengthened the work immensely. Liu's research was supported by grants from the National Natural Science Foundation of China (11371142), the Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019); and Yin's research by a grant (17125814) from the Research Grants Council of Hong Kong.

Appendix: Assumptions and Proofs

Assumptions and Lemma

We impose the following three assumptions, which correspond respectively to condition (1) of Corollary 7.2.1, and conditions (2) and (3) of Theorem 6.2.1 in Fleming and Harrington (1991). Let $\xi = \sup\{t : \prod_{k=0}^{p} \pi_k(t) > 0\}$.

Assumption 2 Assume that $W(s)$ converges in probability to $W_0(s)$ uniformly on $[0, t]$ for any $t \in [0, \xi]$, where $W_0(s)$ is a nonnegative, left-continuous function with right-hand limits such that $W_0(s) < \infty$ for any $s \leq \xi$, $W_0(s) = 0$ for $s > \xi$, and $W_0(s+) \equiv \lim_{u \downarrow s} W_0(u)$ has bounded variation on each closed subinterval of $[0, \xi]$.

Assumption 3 When $\xi \notin \{t : \prod_{k=0}^{p} \pi_k(t) > 0\}$, it holds for $k = 0, 1, \ldots, p$ that

(a) $\int_{(0,t]} W_0^2(s)\pi_k(s)\{1 - \Delta\Lambda_k(s)\}d\Lambda_k(s) < \infty$, and

(b) $\lim\limits_{t \uparrow \xi} \limsup\limits_{n \to \infty} P\left\{\int_{(t,\xi]} n^{-1} W^2(s) Y_k(s) d\Lambda_k(s) > \epsilon\right\} = 0$ for any $\epsilon > 0$.

Assumption 4 When $\xi < \infty$, it holds for $k = 0, 1, \ldots, p$ that

$$\lim_{n \to \infty} P\left\{\int_{\xi}^{\infty} n^{-1} W^2(s) Y_k(s) d\Lambda_k(s) > \epsilon\right\} = 0$$

for any $\epsilon > 0$.

Under the above assumptions, we present a lemma below which plays a fundamental role in the proofs of all theorems in this paper. For group k ($k = 0, 1, \ldots, p$), recall that $S_k(t)$ and $L_k(t)$ denote the respective survival functions of T_k and C_k, $\Lambda_k(t) = -\int_{(0,t]} dS_k(s)/\{1 - S_k(s)\}$, and $\pi.(t) = \sum_{k=0}^{p} \rho_k \pi_k(t)$, where $\rho_k \equiv \lim_{n \to \infty} n_k/n \in (0, 1)$ and $\pi_k(t) = S_k(t)L_k(t)$. Let $M(t) = N(t) - \int_{(0,t]} Y_*(s) d\Lambda(s)$ with $\Lambda(t) = (\Lambda_0(t), \ldots, \Lambda_p(t))^{\mathrm{T}}$ and

$$\tilde{Z}(t) = n^{-1/2} \int_{(0,t]} W(s) \left\{I_{p+1} - Y(s) J^{\mathrm{T}} / Y.(s)\right\} dM(s). \tag{1}$$

Define $\widetilde{\Sigma}(t) \equiv (\tilde{\sigma}_{kl}(t))_{0 \leq k,l \leq p}$ as

$$\widetilde{\Sigma}(t) = \int_{(0,t]} \left[\left\{I - \frac{\pi(s) J^{\mathrm{T}}}{\pi.(s)}\right\} W_0^2(s) \pi_*(s) \{I - \Delta\Lambda_*(s)\} d\Lambda_*(s) \left\{I - \frac{J\pi^{\mathrm{T}}(s)}{\pi.(s)}\right\}\right],$$

where $\pi(s) = (\rho_0\pi_0(s), \ldots, \rho_p\pi_p(s))^{\mathrm{T}}$.

Lemma 1 *Let $Q = (Q_0, Q_1, \ldots, Q_p)^{\mathrm{T}}$ be a $(p+1)$-variate Gaussian process. Suppose that all the components Q have independent increments, $Q_k(0) = 0$ almost surely, for any $0 \leq s \leq t$, $\mathbb{E}\{Q_k(t)\} = 0$ and $\mathbb{E}\{Q_k(t)Q_l(s)\} = \tilde{\sigma}_{kl}(s \wedge t)$, where $\tilde{\sigma}_{kl}(t)$'s are continuous functions. Under Assumptions 2–4, as n goes to infinity, $\tilde{Z}$ defined in (1) converges weakly to Q in $(\mathcal{D}[0, \infty])^{p+1}$, where $\mathcal{D}[0, \infty]$ is the space of functions on $[0, \infty]$ that are right-continuous with finite left-hand limits.*

Proof of Lemma 1 Along the lines of the proof of Theorem 6.2.1 in Fleming and Harrington (1991), the lemma can be proved by showing that under Assumptions 2, 3 and 4, the components of $\tilde{Z}(t)$ satisfy both (3.17) and (3.18) of Theorem 5.3.5 in Fleming and Harrington (1991). □

Proofs of Theorems 1, 2, 3 and 4

We first prove Theorems 3 and 4. Theorems 1 and 2 follows immediately as they are the special cases of Theorems 3 and 4 with $p = 1$.

Proof of Theorem 3. When the $p + 1$ survivor functions are all equal, let Λ_c be the common cumulative hazard function. Then $\widetilde{\Sigma}(t)$ reduces to

$$\begin{aligned}\Sigma(t) &\equiv (\sigma_{kl}(t))_{0\le k,l\le p}\\ &= \int_{(0,t]} W_0^2(s)\left\{\pi_*(s) - \frac{\pi(s)\pi^{\mathrm{T}}(s)}{\pi.(s)}\right\}\{1-\Delta\Lambda_c(s)\}\,d\Lambda_c(s). \qquad (2)\end{aligned}$$

Thus, part (a) follows from Lemma 7.2.1 in Fleming and Harrington (1991).

It follows from Lemma 1 that under H_0, as $n \to \infty$, the multivariate stochastic process $Z_{[-1]}$ converges weakly to $Q_{[-1]}$ which has independent increments and variance–covariance matrix $\Sigma_{[-1,-1]}(t)$. This implies that under H_0, as $n \to \infty$,

$$\begin{aligned}n^{-1/2}D_l(t) &= Z_{[-1]}(t) \text{ converges in distribution to } N(0, \Sigma_{[-1,-1]}(t)),\\ n^{-1/2}D_u(t) &= Z_{[-1]}(\infty) - Z_{[-1]}(t) \text{ converges in distribution to } N(0, \Sigma_{[-1,-1]}(\infty)\\ &\quad -\Sigma_{[-1,-1]}(t)),\end{aligned}$$

and that $n^{-1/2}D_l(t)$ and $n^{-1/2}D_u(t)$ are asymptotically independent. Therefore, part (b) holds.

Accordingly, the stochastic process $\tilde{T}$ converges weakly to $A_1(t)$ with

$$\begin{aligned}A_1(t) &= Q^{\mathrm{T}}_{[-1]}(t)\Sigma^{-1}_{[-1,-1]}(t)Q_{[-1]}(t)\\ &\quad+\Big\{Q_{[-1]}(\infty) - Q_{[-1]}(t)\Big\}^{\mathrm{T}}\Big\{\Sigma_{[-1,-1]}(\infty) - \Sigma_{[-1,-1]}(t)\Big\}^{-1}\{Q_{[-1]}(\infty)\\ &\quad-Q_{[-1]}(t)\}, \qquad (3)\end{aligned}$$

which implies part (c). □

Proof of Theorem 4. Under Assumptions 1 and 2, $W(t)$, $Y_k(t)/n$, $Y.(t)/n$ converge in probability to $W_0(s)$, $\rho_k\pi.(s)$ and $\pi.(s)$ uniformly on $[0, \xi)$, respectively. Under (iv) of Assumption 1, each element of $\Sigma(t)$ is finite, which means $\Sigma(t)$ is well defined. Consequently, $\widehat{\Sigma}(t)$ converges in probability to $\Sigma(t)$, which implies part (a).

To prove part (b), we recall that $Z(t) = \tilde{Z}(t) + \tilde{R}(t)$, where $\tilde{Z}$ is defined in (1) and

$$\tilde{R}(t) = n^{-1/2}\int_{(0,t]} W(s)\Big\{Y_*(s) - Y(s)Y^{\mathrm{T}}(s)/Y.(s)\Big\}\,d\Lambda_c(s).$$

Clearly, $\tilde{Z}$ satisfies the conditions of Lemma 1 and converges weakly to Q, which is defined in Lemma 1. Meanwhile, by the same arguments as those in proving (a),

$$\tilde{R}(t) = \int_{(0,t]} \frac{W(s)}{n}\left\{Y_*(s) - \frac{Y(s)Y^{\mathrm{T}}(s)}{Y.(s)}\right\}\cdot n^{1/2}\left\{\frac{d\Lambda(s)}{d\Lambda_c(s)} - J\right\}d\Lambda_c(s)$$

converges in probability to

$$R(t) = \int_{(0,t]} W_0(s) \left\{ \pi_*(s) - \frac{\pi(s)\pi^{\mathrm{T}}(s)}{\pi.(s)} \right\} \gamma(s) d\Lambda_c(s), \tag{4}$$

where $\gamma(s) = (\gamma_0(s), \ldots, \gamma_p(s))^{\mathrm{T}}$. This in conjunction with the weak convergence of $\tilde{Z}$ implies that Z converges weakly to $Q + R$. Combining the weak convergence of $Z_{[-1]}$ and $\hat{\Sigma}_{[-1,-1]}$, we conclude that $\tilde{T}$ converges weakly to $A_2(t)$ over $t \in \Omega$, where

$$\begin{aligned} A_2(t) = {} & (Q + R)^{\mathrm{T}}_{[-1]}(t) \Sigma^{-1}_{[-1,-1]}(t) (Q + R)_{[-1]}(t) \\ & + \{(Q + R)_{[-1]}(\infty) - (Q + R)_{[-1]}(t)\}^{\mathrm{T}} \{\Sigma_{[-1,-1]}(\infty) - \Sigma_{[-1,-1]}(t)\}^{-1} \\ & \times \{(Q + R)_{[-1]}(\infty) - (Q + R)_{[-1]}(t)\}. \end{aligned} \tag{5}$$

Consequently, T converges in distribution to $\sup_{t \in \Omega} A_2(t)$. □

Proof of Theorem 1. This theorem is a special case of Theorem 3 with $p = 1$. It can be verified that $\Sigma_{[-1,-1]}(t) = \sigma(t)$ when $p = 1$, therefore results (a) and (b) follow immediately.

We need only prove result (c). When $p = 1$, part (c) of Theorem 3 implies that T converges in distribution to the supremum of

$$\frac{\{Q_1(t)\}^2}{\sigma(t)} + \frac{\{Q_1(\infty) - Q_1(t)\}^2}{\sigma(\infty) - \sigma(t)}, \tag{6}$$

where Q_1 is a Gaussian process with independent increments and variance σ.

Let $\{B(t) : t \in [0, \infty)\}$ denote a Brownian motion, then the supremum of (6) has the same distribution as

$$\begin{aligned} & \sup_{t \in \Omega} \left\{ \frac{[B\{\sigma(t)\}]^2}{\sigma(t)} + \frac{[B\{\sigma(\infty)\} - B\{\sigma(t)\}]^2}{\sigma(\infty) - \sigma(t)} \right\} \\ & = \sup_{0 < s < \sigma(\infty)} \left\{ \frac{\{B(s)\}^2}{s} + \frac{[B\{\sigma(\infty)\} - B(s)]^2}{\sigma(\infty) - s} \right\} \\ & = \sup_{0 < t < 1} \left[\frac{\{B'(t)\}^2}{t} + \frac{\{B'(1) - B'(t)\}^2}{1 - t} \right] \end{aligned}$$

with $B'(t) = B\{\sigma(\infty)s\}/\{\sigma(\infty)\}^{1/2}$. Since $B'(t)$ is still a Brownian motion, this leads to result (c). □

Proof of Theorem 2. This theorem is a special case of Theorem 4 with $p = 1$. It can be verified that $\Sigma_{[-1,-1]}(t)$, $R_{[-1]}(t)$ and $A_2(s)$ reduce respectively to the $\sigma(t)$, $R(t)$ and $A(s)$ defined in Theorem 2. This completes the proof. □

References

Anderson JA, Senthilselvan A (1982) A two-step regression model for hazard functions. Appl Stat 31:44–51

Bagdonavičius V, Levuliene R, Nikulin M (2012) Modelling and testing of presence of hazard rates crossing under censoring. Commun Stat-Simul Comput 41:980–991

Belongia EA, Berg R, Liu K (2001) A randomized trial of zinc nasal spray for the treatment of upper respiratory illness in adults. Am J Med 111:103–108

Breslow NE (1970) A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57:579–594

Breslow NE, Edler L, Berger P (1984) A two-step censored-data rank test for acceleration. Biometrics 40:1049–62

Chen D, Sun J, Peace KE (2012) Interval-censored time-to-event data: methods and applications. Chapman & Hall/CRC, Boca Raton

Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc Ser B 34:187–220

Eng KH, Kosorok MR (2005) A sample size formula for the supremum log-rank statistic. Biometrics 61:86–91

Fleming TR, Harrington DP (1981) A class of hypothesis tests for one and two samples censored survival data. Commun Stat Theory Methods 10:763–794

Fleming TR, Harrington DP (1991) Counting processes survival analysis. Wiley, New York

Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP (1980) Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. Biometrics 36:607–625

Gehan E (1965) A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52:203–23

Gill RD (1980) Censoring and stochastic integrals, vol 124., Mathematical centre tractsMathematical Centrum, Amsterdam

Hosmer DW, Lemeshow S (1999) Applied survival analysis: regression modeling of time to event data. Wiley, New York

Jones MP, Crowley J (1989) A general class of nonparametric tests for survival analysis. Biometrics 45:157–170

Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data. Springer, New York

Kleinbaum DG, Klein M (2012) Survival analysis: a self-learning text, 3rd edn. Springer, New York

Kosorok MR, Lin C-Y (1999) The versatility of function-indexed weighted log-rank statistics. J Am Stat Assoc 94:320–332

Lee ET, Desu MM, Gehan EA (1975) A Monte Carlo study of the power of some two-sample tests. Biometrika 62:425–432

Lin DY, Ying Z (1994) Semiparametric analysis of the additive risk model. Biometrika 81:61–71

Liu K, Qiu P, Sheng J (2007) Comparing two crossing hazard rates by Cox proportional hazards modeling. Stat Med 26:375–391

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

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50:163–170

Mantel N, Stablein DM (1988) The crossing hazard function problem. Statistician 37:59–64

Mantel N, Bohidar NR, Ciminera JL (1977) Mantel–Haenszel analysis of litter-matched time-to-response data, with modifications for recovery of interlitter information. Cancer Res 37:3863–3868

Moreau T, Maccario J, Lellouch J, Huber C (1992) Weighted log rank statistics for comparing two distributions. Biometrika 79:195–198

Nahman NS, Middendorf DF, Bay WH, McElligott R, Powell S, Anderson J (1992) Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. J Am Soc Nephrol 3:103–107

Peto R, Peto J (1972) Asymptotically efficient rank-invariant test procedures (with discussion). J R Stat Soc Ser A 135:185–207

Prentice RL (1978) Linear rank tests with right censored data. Biometrika 65:167–79

Prentice RL, Marek P (1979) A qualitative discrepancy between censored data rank tests. Biometrics 35:861–867

Qiu P, Sheng J (2008) A two-stage procedure for comparing hazard rate functions. J R Stat Soc Ser B 70:191–208

Schein PD, Bruckner HW, Douglass HO, Mayer R et al, Gastrointestinal Tumor Study Group (1982) A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Cancer 49:1771–1777

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Tarone RE, Ware J (1977) On distribution-free tests for equality of survival distributions. Biometrika 64:156–160

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

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

Yin G, Zeng D (2005) Pair chart test for an early survival difference. Lifetime Data Anal 11:117–129