Statistics
in Medicine

Research Article

Received 2 September 2013, Accepted 25 June 2015 Published online 20 July 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6591

A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves

Hajime Uno,[a*†] Lu Tian,[b] Brian Claggett[c] and L. J. Wei[d]

With censored event time observations, the logrank test is the most popular tool for testing the equality of two underlying survival distributions. Although this test is asymptotically distribution free, it may not be powerful when the proportional hazards assumption is violated. Various other novel testing procedures have been proposed, which generally are derived by assuming a class of specific alternative hypotheses with respect to the hazard functions. The test considered by Pepe and Fleming (1989) is based on a linear combination of weighted differences of the two Kaplan–Meier curves over time and is a natural tool to assess the difference of two survival functions directly. In this article, we take a similar approach but choose weights that are proportional to the observed standardized difference of the estimated survival curves at each time point. The new proposal automatically makes weighting adjustments empirically. The new test statistic is aimed at a one-sided general alternative hypothesis and is distributed with a short right tail under the null hypothesis but with a heavy tail under the alternative. The results from extensive numerical studies demonstrate that the new procedure performs well under various general alternatives with a caution of a minor inflation of the type I error rate when the sample size is small or the number of observed events is small. The survival data from a recent cancer comparative study are utilized for illustrating the implementation of the process.

Keywords: logrank test; perturbation resampling method; proportional hazards; robust tests

1. Introduction

In summarizing the comparisons of two survival distributions with censored event time observations, it is conventional to provide a plot of two Kaplan–Meier (KM) curves along with a *p*-value from the two-sample logrank test. Note that the logrank test statistic reflects the difference of two underlying hazard functions, not of the KM curves directly [1]. Asymptotically, the logrank test is valid nonparametrically but may perform rather poorly when the proportional hazards (PH) assumption does not hold [2,3]. As an example, in a randomized clinical trial (E4A03) recently conducted by the Eastern Cooperative Oncology Group, low-dose and high-dose dexamethasone for treating newly diagnosed multiple myeloma patients were compared with respect to the patient's overall survival [4]. Of a total of 445 enrolled patients, 222 were assigned to the low-dose group and 223 to the high-dose group. Figure 1 presents the KM curves of overall survival, based on the data collected from November 2009, for the two dose groups. The two curves are markedly separated before 30 months of follow-up but then appear to be connected with each other at the end of the study. Visually, it appears that the low dose does have a short-term survival advantage over the high dose. However, the two-sided *p*-value from the logrank test is 0.46, and the *p*-value from the Peto–Prentice–Wilcoxon test is 0.28. Neither test gives strong evidence that the low-dose group is better than the high-dose group under the intent-to-treat principle. In this example, the PH

[a]*Department of Biostatistics and Computational Biology and Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, U.S.A.*
[b]*Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, U.S.A*
[c]*Brigham and Women's Hospital, Division of Cardiovascular Medicine, Harvard Medical School, Boston, MA 02115, U.S.A.*
[d]*Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.*
[*]*Correspondence to: Hajime Uno, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, U.S.A.*
[†]*E-mail: huno@jimmy.harvard.edu*

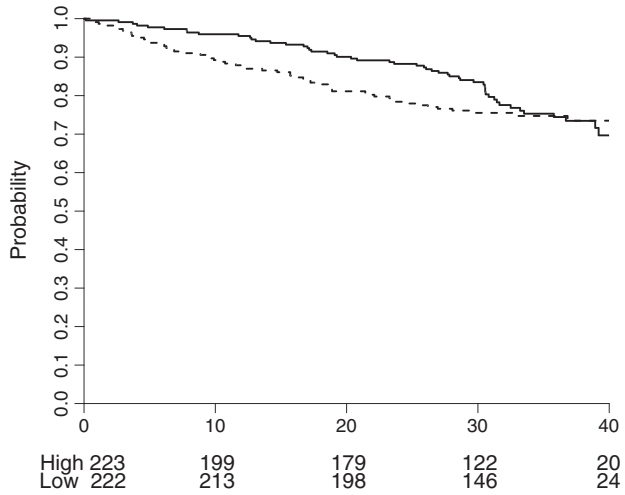

Figure 1. Overall survival curves for low-dose arm (solid line) and high-dose arm (dashed line) with the E4A03 data.

assumption is likely violated, and the logrank test may not be powerful for detecting the difference of the survival functions.

To avoid a 'fishing expedition' in which tests of equality of the survival curves are selected *ad hoc*, a pre-specified test needs to be well described in the study protocol or the statistical analysis plan before unblinding the data. Unfortunately, in general, little or no information is available regarding the profile of the potential difference between two survival curves at the design stage. To this end, various flexible test procedures have been proposed. For example, the $G^{\rho,\gamma}$ tests, which are constructed assuming a class of survival distributions indexed by ρ and γ under the alternative hypothesis [5] and a linear combination or the maximum of several test statistics [5–12], and other novel robust procedures [13, 14] have been extensively studied. The pros and cons of those procedures were discussed in a recent paper by Yang and Prentice [15]. Note that the aforementioned novel alternatives to the logrank test are more or less built with respect to a family of pre-specified survival functions. The test recently proposed by Yang and Prentice [15] is a weighted logrank test whose weights are obtained by fitting the data to the model proposed by Yang and Prentice [16], which includes the PH and the proportional odds models as special cases.

Note that a class of novel tests based on the weighted KM (WKM) statistics proposed by Pepe and Fleming [1, 17] did not obtain much attention in practice. Specifically, let $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ be the KM estimators for the two groups to be compared. A WKM test statistic is

$$WKM = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \int_0^{\tau} \hat{W}(t)\hat{D}(t)dt, \tag{1}$$

where $\hat{D}(t) = \hat{S}_2(t) - \hat{S}_1(t)$, $\tau = \sup\left[t : \min\left\{\hat{K}_1(t), \hat{K}_2(t)\right\} > 0\right]$, $\hat{K}_i(\cdot)$ denotes the left-continuous version of the KM estimator for the censoring survival function for the ith group, n_i is the sample size in group i, $(i = 1, 2)$, and $\hat{W}(\cdot)$ is the data-dependent weight function. Because we conventionally present the KM curves as in Figure 1 to show the temporal profile of the group difference, it seems natural to provide a test that directly compares two survival functions, rather than their hazard functions. For the class of test statistics in (1), Pepe and Fleming [1] considered two weighting schemes:

$$\frac{\hat{K}_1(t)\hat{K}_2(t)}{\hat{q}_1\hat{K}_1(t) + \hat{q}_2\hat{K}_2(t)} \tag{2}$$

and

$$\left\{\frac{\hat{K}_1(t)\hat{K}_2(t)}{\hat{q}_1\hat{K}_1(t) + \hat{q}_2\hat{K}_2(t)}\right\}^{1/2}, \tag{3}$$

where $\hat{q}_i$ is the proportion of subjects assigned to group i. Note that their weighting schemes only depend on the censoring distributions.

In this article, we also consider tests similar to (1) for testing against a one-sided alternative hypothesis; that is, one survival curve is greater than the other for a time interval of the follow-up. However, in choosing $\hat{W}(\cdot)$, instead of considering only the observed censoring distribution, we propose to put more weight at the time points t such that the difference $\hat{D}(t)$ is 'large'. One possible choice is to let $\hat{W}(t) = \hat{D}(t)$; the resulting test statistic, however, is like a 'chi-squared' statistic and tends to have a rather long right tail under the null hypothesis. It follows that this omnibus test may not be powerful for certain alternatives. In the next section, we propose a simple weighting scheme whose weight at time t depends on $\hat{D}(t)$. Under the null, the distribution of the test statistic has a relatively short tail, but under a general one-sided alternative, the observed test statistic tends to be large and likely to reject the null hypothesis. The new test statistic is not derived from any pre-specified class of distributions like most existing test procedures in the literature. Instead, it automatically chooses the weights adaptively based on the size of $\hat{D}(\cdot)$ or a function thereof, to effectively differentiate the null and alternative hypotheses. For the aforementioned cancer study survival data, the resulting one-sided p-value of the proposed test is 0.005. The details of implementing the new test are given in Section 2. We also conducted extensive numerical studies to assess the performance of the proposed procedure.

2. Combining weighted differences of the Kaplan–Meier curves

Let $S_1(\cdot)$ and $S_2(\cdot)$ be two survival functions for failure times T_1 and T_2, respectively. Let C_1 and C_2 be the corresponding censoring times. Also, let $\{(T_{1j}, C_{1j}),\ j = 1, \ldots, n_1\}$ and $\{(T_{2j}, C_{2j}),\ j = 1, \ldots, n_2\}$ be independent random copies from (T_1, C_1) and (T_2, C_2), respectively. Because of censoring, one can only observe $\left\{(X_{ij}, \Delta_{ij});\ i = 1, 2,\ j = 1, \ldots, n_i\right\}$, where $X_{ij} = \min(T_{ij}, C_{ij})$ and Δ_{ij} is equal to 1 if $T_{ij} \leqslant C_{ij}$ and 0 otherwise. Let $D(\cdot) = S_2(\cdot) - S_1(\cdot)$. Let $[0, \zeta]$ be a given time interval, and we assume that $\Pr(X_i > \zeta) > 0,\ i = 1, 2$. We are interested in testing the null hypothesis that $D(t) = 0$, for $t \in [0, \zeta]$, against a general one-sided alternative, that is, $D(\cdot) \geqslant 0$ with at least one $t \in [0, \zeta]$, such that $D(t) > 0$. Now, let $\hat{D}(\cdot)$ be equal to $\hat{S}_2(\cdot) - \hat{S}_1(\cdot)$, $\hat{\sigma}(\cdot)$ be its standard error estimate, and $Z(\cdot) = \hat{D}(\cdot)/\hat{\sigma}(\cdot)$, which is distributed approximately $N(0, 1)$ under the null hypothesis.

Instead of utilizing $Z(t)$ as a test statistic at a specific time point t for testing the null hypothesis, we consider a test statistic that is a weighted integration of standardized differences between two survival curves over $[0, \zeta]$. For example, one potential class of test statistics is

$$V = \int_0^{\zeta} \hat{W}(t) Z(t) dt, \tag{4}$$

where $\hat{W}(\cdot)$ is a data-dependent weight function. Note that (4) is slightly different from (1). We replace $\hat{D}(\cdot)$ by $Z(\cdot)$ because we are primarily interested in hypothesis testing. Note that we define $Z(t) = 0$ for $\hat{D}(t) = \hat{\sigma}(t) = 0$, that is, $\hat{S}_1(t) = \hat{S}_2(t) = 1$ where no events have been observed by t.

Heuristically, a test based on V would perform well if $\hat{W}(t)$ is proportional to $E\,(Z(t))$ under alternatives. That is, $\hat{W}(t)$ is large for a large observed $Z(t)$. A natural choice is to let $\hat{W}(t) = Z(t)$. However, as discussed in Section 1, the distribution of this 'chi-squared-like' statistic may have a rather long right tail under the null hypothesis, and the test may not perform well for specific alternatives. On the other hand, when $\hat{W}(\cdot)$ is constant over time, the distribution of such a 'normal-like' statistic is centered around zero and has a short tail, under the null, but this test may only be powerful when $Z(t)$ is approximately constant over $[0, \zeta]$. Therefore, the question is how to choose the weight function such that the distribution of the resulting statistic has a short right tail under the null, but the observed V is large under the alternative. A possible solution is to expand on the idea proposed by Xu *et al.* [18] for combining a small number of dependent test statistics for linkage or association across multiple phenotypic traits. Specifically, let $c \in [0, \eta]$, where η is a constant, say 4. Let $\hat{W}_c(t) = \max\{Z(t), c\}$ and

$$V_1(c) = \int_0^{\zeta} \hat{W}_c(t) Z(t) dt. \tag{5}$$

For a fixed c, say 1.65, under the null hypothesis, because $Z(t)$ is approximately $N(0, 1)$, $\hat{W}_c(t) \sim 1.65$ for most $t \in [0, \zeta]$. It follows that the distribution of $V_1(c)$, which is similar to a linear combination of dependent standard normal random variables, would not have a long right tail. On the other hand, under an alternative hypothesis, for a large observed $Z(t)$, ($\geqslant$1.65), $\hat{W}(t) = Z(t)$ and the resulting observed $V_1(c)$

would be large. On the other hand, the choice of $c = 1.65$ may not work well for cases in which $D(\cdot)$ is positive for a large portion of time points but the observed $Z(\cdot)'$s are less than 1.65 because of, for example, the low observed mortality rates. Therefore, it is not obvious how to select such a threshold value c a priori.

Here, we propose a simple, automatic way to choose c adaptively to construct a test statistic based on $\{V_1(c), 0 \leqslant c \leqslant \eta\}$. First, suppose that we can generate a good approximation to the null distribution of the process $V_1(c)$ indexed by $c \in [0, \eta]$. Let $v_1(c)$ be the observed value of $V_1(c)$, and its p-value $p(c)$ can be obtained via the approximation to the null distribution of $V_1(c)$. Let $p_b = \min\{p(c) : c \in [0, \eta]\}$, the most significant $p(c)$ in $c \in [0, \eta]$. A small p_b would support the alternative hypothesis. The question is how to choose the threshold value for claiming a 'statistical significance' based on p_b. That is, one needs to obtain the null distribution of $P_b = \min\{P(c) : c \in [0, \eta]\}$, the random counterpart of p_b, where $P(c) = S_{V_1(c)}\left(V_1(c)\right)$ and $S_{V_1(c)}(v)$ is the survival function of $V_1(c)$. Using the standard martingale theory, we have shown that $Z(\cdot)$ converges weakly to a limiting Gaussian process $G(\cdot)$ [19]. In Appendix A, we show that $V_1(c)$ and $P(c)$, as processes in c, converge weakly to $\psi(c) = \int_0^\zeta \max\{G(t), c\}G(t)dt$ and $U(c) = S_{\psi(c)}(\psi(c))$, respectively, where $S_{\psi(c)}(v)$ is the survival function of $\psi(c)$.

To empirically approximate the limiting distribution of this process under the null, one may utilize a perturbation resampling method, which has been applied successfully to various problems in survival analysis [20,21]. Specifically, the distribution of the process $\left\{\hat{S}_i(t) - S_i(t)\right\}, i = 1, 2$ can be approximated by that of

$$Q_i(t) = -\hat{S}_i(t)\sum_{j=1}^{n_i}\left[\left\{\sum_{k=1}^{n_i} I(x_{ik} \geqslant x_{ij})\right\}^{-1} \delta_{ij} I(x_{ij} \leqslant t)\xi_{ij}\right],$$

where $\left(x_{ij}, \delta_{ij}\right)$ is the observed value of $\left(X_{ij}, \Delta_{ij}\right)$, $I(\cdot)$ is the indicator function, and $\left\{\xi_{ij},\ i = 1, 2,\ j = 1, \ldots, n_i\right\}$ is a random sample from a distribution with mean 0 and variance 1, for example, the standard normal distribution. In practice, the null distribution of $V_1(c)$ can be approximated by generating M sets of $\left\{\xi_{ij}\right\}$. For each realized set $\left\{\xi_{ij}\right\}$, we compute

$$V_1^*(c) = \int_0^\zeta W_c^*(t)Z^*(t)dt, \tag{6}$$

where $Z^*(\cdot) = \left\{Q_2(\cdot) - Q_1(\cdot)\right\}/\hat{\sigma}(\cdot)$ and $W_c^*(\cdot) = \max\{Z^*(\cdot), c\}$. The set $\mathcal{D}$ of M realizations $\left\{V_1^*(c),\ c \in [0, \eta]\right\}$ serves as a reference set for the proposed test. For each of the M sets, we compute (6) and obtain the corresponding $P(c)$, using the reference set $\mathcal{D}$, denoted by $P^*(c)$. The null distribution of P_b can be estimated using the M realizations of $P_b^* = \min\{P^*(c) : 0 \leqslant c \leqslant \eta\}$ based on $\left\{V_1^*(c),\ c \in [0, \eta]\right\}$. The bona fide p-value of the proposed test is then given by $\Pr\left(P_b^* < p_b\right)$. In Appendix B, we show that conditional on the observed data, under the null, the limiting distributions of $V_1^*(c)$ and $P^*(c)$ are $\psi(c)$ and $U(c)$, respectively, and thus, the null distribution of P_b can be approximated well by $\min\{P^*(c) : 0 \leqslant c \leqslant \eta\}$.

Another potential class of test statistics is

$$V_2(c) = \int_0^\zeta \hat{W}_c(t)Z(t)d\bar{N}(t), \tag{7}$$

where $\bar{N}(t) = (n_1 + n_2)^{-1}\sum_{i=1}^2\sum_{j=1}^{n_i} I(X_{ij} \leqslant t)\Delta_{ij}$. Note that the null distribution of P_b can also be approximated via the aforementioned procedure. The role of $\bar{N}(\cdot)$ as the integrator serves as another weighting function for $Z(\cdot)$; that is, the weight is heavy for the time intervals with large numbers of observed events. The parameter η in the proposed test can be any positive constant. Empirically, we find that the choice $\eta = 4$ works well because it is unlikely $Z(\cdot)$ would be beyond 4 under the null (see Section 3 for details).

3. The E4A03 example and numerical comparison studies

First, we apply the proposed tests to the survival data from E4A03 for comparing the low-dose and high-dose groups discussed in Section 1. We calculate $V_1(c)$ and $V_2(c)$ over the range [0, 40] months at each value of $c = 0,\ 0.1,\ 0.2, \ldots, 4$, where we use Greenwood's formula for estimating variances of $\hat{S}_1(\cdot)$

and $\hat{S}_2(\cdot)$ to calculate the standard error of $\hat{D}(\cdot)$. The observed p_b values are 0.0044 and 0.0018 based on $V_1(c)$ and $V_2(c)$, respectively, both of which are obtained at $c = 0$ in this example. To construct the null reference sets for these two tests, we generate $M = 5000$ realized samples $\{\xi_{ij}\}$ from $N(0, 1)$ to obtain the null distributions of P_b. The resulting one-sided bona fide p-values, $\Pr(P_b < p_b)$, are 0.0048 and 0.0020, respectively. For comparison, we also analyze the data with the two WKM tests, (2) and (3), proposed by Pepe and Fleming [1]; the corresponding p-values are 0.007 and 0.014, respectively. The test proposed by Yang and Prentice [15] gives $p = 0.138$. Recall that the Peto–Prentice–Wilcoxon test and logrank test yield $p = 0.142$ and $p = 0.233$, respectively.

We conduct an extensive simulation study to examine the performance of the new tests. First, we assess the size and power of the tests under a similar setting to the E4A03 trial. The pattern illustrated in Figure 2(a), with no difference between the two survival functions, is considered for evaluating the empirical I error rate. The curve in Figure 2(a) is the survival function of a Weibull distribution derived as the best approximation of the low-dose group data from E4A03 using the maximum likelihood method. Figure 2(b) shows survival functions of the Weibull distributions that approximate the low-dose

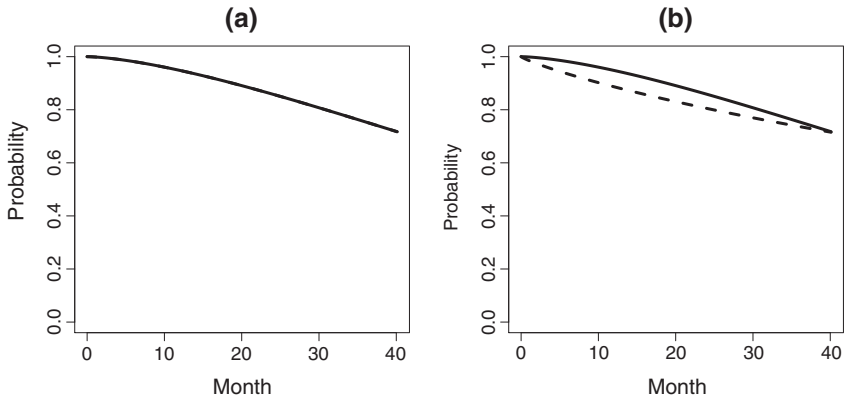

Figure 2. Comparison of survival functions considered in simulation studies.

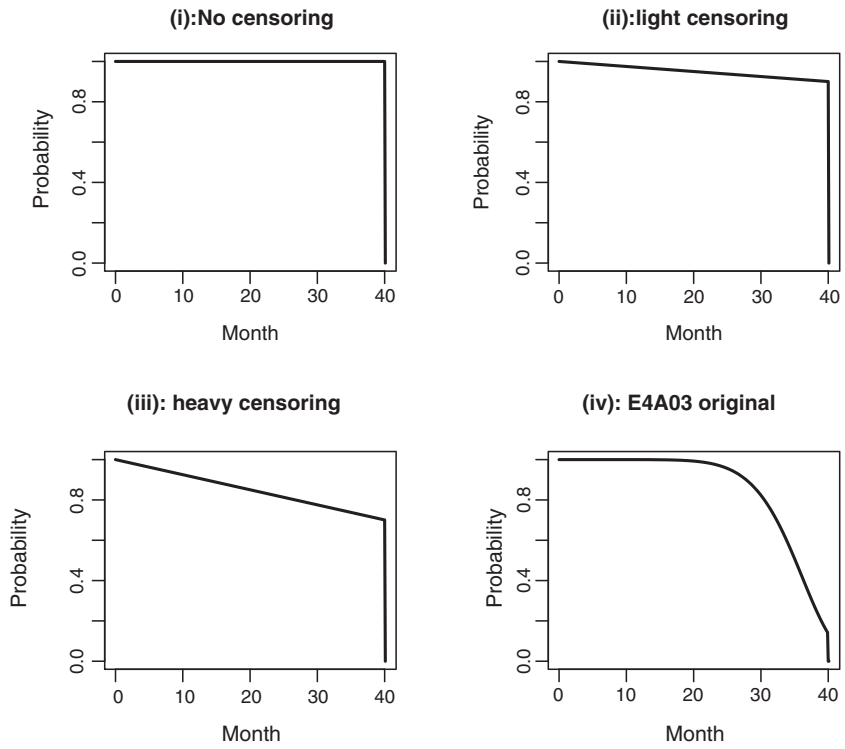

Figure 3. Survival functions of the underlying censoring distributions considered in the simulations.

(solid line) and high-dose (dashed line) groups data from the E4A03 trial, respectively. For the underlying censoring time distributions, we consider four scenarios: (i) no censoring, (ii) light censoring, (iii) heavy censoring, and (iv) the observed censoring patterns from the E4A03 trial, obtained by fitting the data with a Weibull distribution model with the low-dose and high-dose groups combined (Figure 3). For all censoring configurations, we consider an administrative censoring at 40 months. We graphically present those censoring patterns in Figure D.2.

We generate 5000 independent samples, with sample size $n = 300$ (per group), from the distributions described earlier (Figures 2 and D.2). Note that for each subject, we generate a survival time and an underlying censoring time and compute the observable time (i.e., the minimum of the survival time and the censoring time) and the censoring indicator. Similar to the analysis for E4A03 presented earlier, we let $[0, \zeta] = [0, 40]$ (months) and $M = 5000$, and $\{\xi_{ij}\}$ is from the standard normal distribution and c is evaluated in increments of 0.1 up to a maximum of $\eta = 4$. For comparators, we include the one-sided logrank, Peto–Prentice–Wilcoxon, Yang–Prentice, and Pepe–Fleming tests. Because the results of Pepe–Fleming tests with weights (2) and (3) are similar, we present the results with weight (2) only.

Table I shows the results of this numerical study. Except for the Yang–Prentice test, the empirical I error rates are nearly identical to their nominal level of 0.05. The new tests appear to be consistently more powerful than their comparators. For these cancer study data, the logrank test performs rather poorly with respect to power, as expected.

We also examined other scenarios to evaluate the performance of the new proposals. In one of the numerical studies, we consider the following patterns of differences between two survival functions

Table I. Size and power of LR, PP, and YP tests, PF test based on (2), and the new tests based on (5) [V_1] and (7) [V_2].

Size of tests Survival distributions: 2(a)				
	Censoring			
	(i)	(ii)	(iii)	(iv)
Number of events*	169.9	161.4	144.3	145.9
Test				
LR	0.051	0.049	0.048	0.051
PP	0.051	0.048	0.047	0.051
YP	0.060	0.057	0.058	0.062
PF	0.050	0.049	0.048	0.047
V_1	0.047	0.048	0.045	0.051
V_2	0.051	0.049	0.046	0.049
Power of tests Survival distributions: 2(b)				
	Censoring			
	(i)	(ii)	(iii)	(iv)
Number of events*	169.2	159.5	140.0	139.2
Test				
LR	0.107	0.121	0.162	0.226
PP	0.173	0.191	0.245	0.300
YP	0.195	0.214	0.273	0.328
PF	0.628	0.622	0.628	0.726
V_1	0.842	0.829	0.822	0.846
V_2	0.834	0.827	0.839	0.867

*Average number of the total observed events across 5000 simulation realizations.
2(a): no difference; 2(b): difference observed in E4A03 (Figure 2).
LR, logrank; PP, Peto–Prentice–Wilcoxon; YP, Yang–Prentice; PF, Pepe–Fleming.

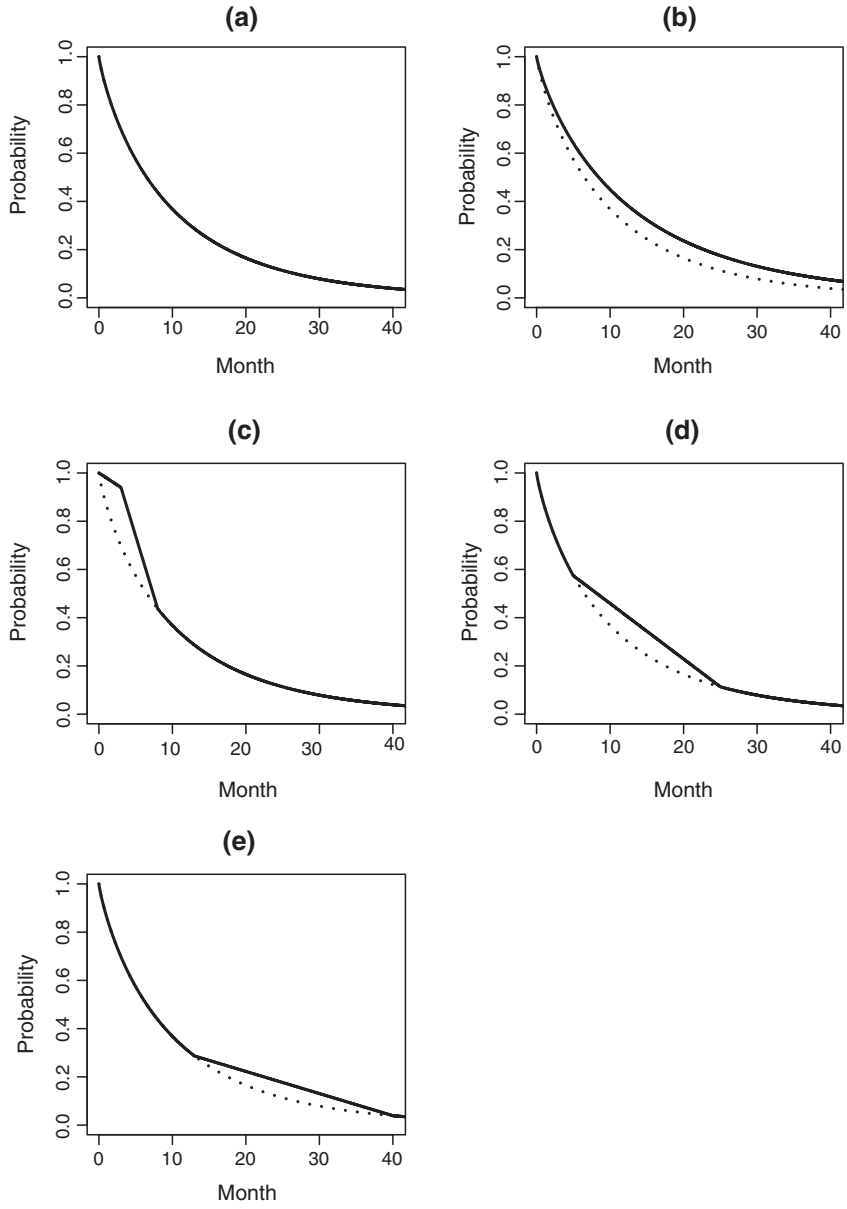

Figure 4. Comparison of survival functions considered in simulation studies.

illustrated in Figure 4. Specifically, five different patterns are examined: (a) no difference, (b) PH difference (with true hazard ratio of 0.8), (c) early difference, (d) middle difference, and (e) late difference. For this numerical study, the sample size is $n = 200$ per arm. Other configurations, such as the underlying censoring distributions (Figure D.2) and the number of iterations, are the same as previously described.

Table II presents the results of the study. The results in pattern 4(a) show that the empirical significance levels of almost all tests are close to their nominal value of 0.05, The Yang–Prentice test is still liberal. The empirical size of the test V_1 is 0.062, which is also slightly higher than 0.05 when the censoring is rather heavy (censoring pattern (iv)). Further simulation suggests that this increased I error is associated with instability in the tails of the KM curves near ζ. This issue can be addressed via a simple modification of V_1 and V_2, described in the Section 4, which allows I error rates to be maintained without loss of power. Under the PH alternative (pattern 4(b)), the test V_2 is comparable with other tests that are not designed for this specific alternative. However, the test V_1 performs as well as the logrank test in this scenario. For the early difference alternative (pattern 4(c)), the Peto–Prentice–Wilcoxon test gives a higher power than the logrank as we expected, but the new tests are even more powerful than the Wilcoxon. For pattern 4(d), our tests are more powerful than the other tests. For the late difference (pattern 4(e)), the Peto–Prentice–Wilcoxon test is worst, and V_1 is the best among all the test procedures considered. Note

Table II. Size and power of LR, PP, and YP tests, PF test based on (2), and the new tests based on (5) [V_1] and (7) [V_2], to detect the difference between two survival curves, based on 5000 of iterations, with sample size 200 per arm.

Censoring	(i) No censoring					(ii) Light censoring				
Survival distributions	4(a)	4(b)	4(c)	4(d)	4(e)	4(a)	4(b)	4(c)	4(d)	4(e)
Number of events*	384.4	377.2	384.4	384.4	384.4	375.8	368.1	375.2	374.8	375.2
Test	Size	Power				Size	Power			
LR	0.050	0.703	0.293	0.397	0.199	0.048	0.690	0.310	0.397	0.199
PP	0.051	0.610	0.843	0.385	0.104	0.052	0.605	0.846	0.370	0.103
YP	0.063	0.732	0.661	0.442	0.245	0.061	0.720	0.682	0.441	0.240
PF	0.051	0.697	0.305	0.524	0.267	0.052	0.685	0.316	0.519	0.258
V_1	0.055	0.707	0.954	0.614	0.387	0.055	0.698	0.955	0.607	0.372
V_2	0.053	0.627	1.000	0.554	0.163	0.054	0.607	1.000	0.534	0.151
Censoring	(iii) Heavy censoring					(iv) Observed in E4A03				
Survival distributions	4(a)	4(b)	4(c)	4(d)	4(e)	4(a)	4(b)	4(c)	4(d)	4(e)
Number of events*	358.3	349.7	356.7	355.6	356.7	375.4	366.3	375.4	374.5	369.5
Test	Size	Power				Size	Power			
LR	0.052	0.667	0.319	0.399	0.189	0.049	0.690	0.298	0.413	0.252
PP	0.053	0.590	0.860	0.350	0.099	0.053	0.612	0.840	0.378	0.105
YP	0.065	0.697	0.717	0.443	0.237	0.062	0.719	0.673	0.454	0.316
PF	0.056	0.664	0.333	0.513	0.234	0.051	0.680	0.327	0.562	0.251
V_1	0.058	0.678	0.954	0.601	0.354	0.062	0.715	0.968	0.646	0.389
V_2	0.055	0.594	0.999	0.512	0.134	0.054	0.613	1.000	0.557	0.142

*Average number of the total observed events across 5000 simulation realizations.
4(a) no difference, 4(b) proportional hazards, 4(c) early difference, 4(d) difference in middle, and 4(e) late difference (Figure 4).
LR, logrank; PP, Peto–Prentice–Wilcoxon; YP, Yang–Prentice; PF, Pepe–Fleming.

that V_2 is not as powerful under this setting because most of the failures are observed prior to the time at which the survival curves separate.

We also observe how the distribution of the selected value of c corresponding to p_b, the smallest p-values, varies across simulation scenarios. For each simulation scenario, we obtain the selected optimal c for each of the generated 5000 independent samples, and draw the histogram. Figure 5(a)–(e) shows the results from simulation scenarios corresponding to Figure 4(a)–(e) with Figure 3(i) no censoring, for the test based on V_1. The selected values of c rarely reach 4 in our simulations, which suggests that $[0, 4]$ is a reasonable search range for c, in practice. This figure further suggests that the choice of c is indeed data dependent and cannot be pre-specified.

Because we are particularly interested in the performance of the newly proposed tests under the PH alternatives, we conducted another set of power comparisons for several scenarios (Figure 6 shows the patterns of the paired survival functions considered). With censoring distributions as described for the previous cases (Figure D.2), Table III shows the results of this specific set of simulations. It appears that the logrank test indeed outperforms our tests under the PH assumption, but the gain of power is modest (at most about 10%).

In summary, while the test based on V_2 performs particularly well when the survival curves separate early, the test based on V_1 appears to be more generally useful, demonstrating power broadly comparable with that of the logrank test under the PH alternative and *exceeding* the power of all comparator tests under all other scenarios, including those proposed by Pepe and Fleming [1]. Such robust performance characteristics are likely indicative of the fact that the proposed tests are not derived to detect a specific departure from the null hypothesis. Unlike other procedures, the new proposals are not derived under the assumption of a specific type of alternative hypotheses.

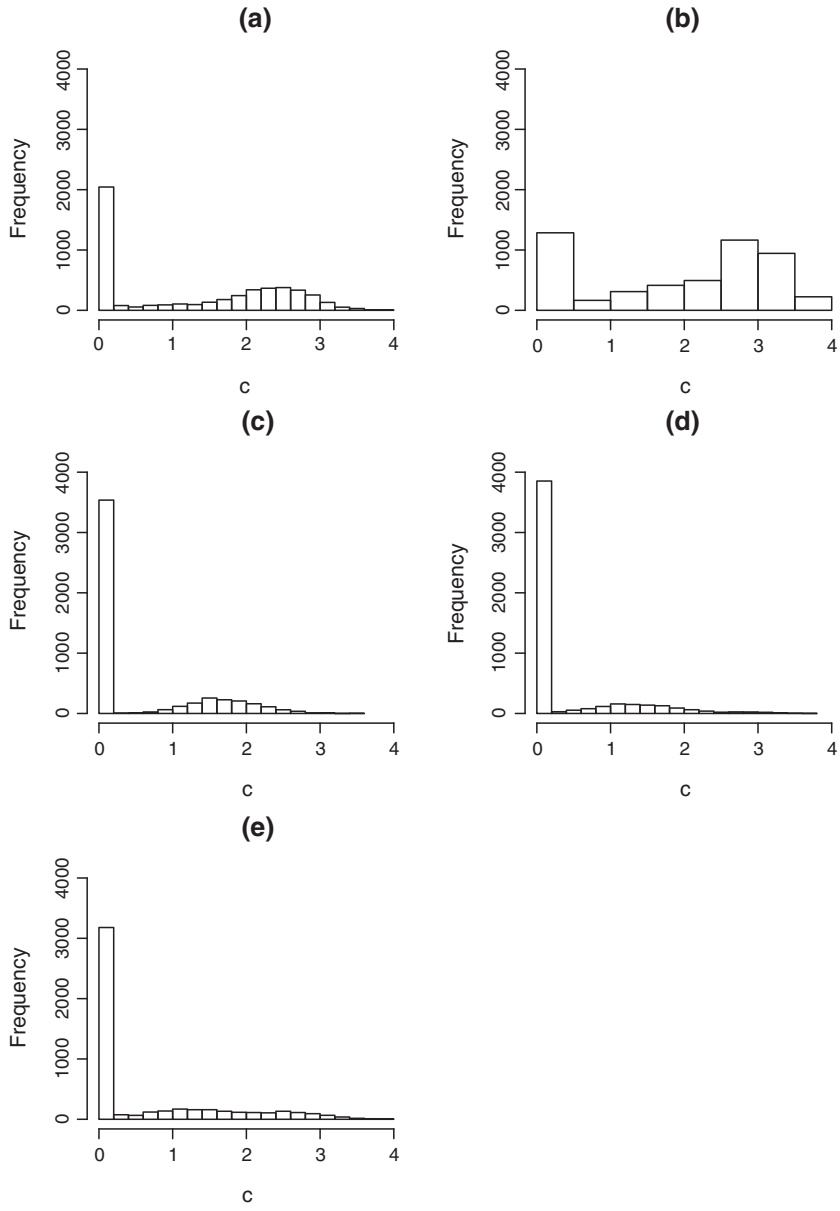

Figure 5. Frequency of the selected value of c for $V_1(c)$ in simulation studies.

4. Remarks

When the censoring is heavy and ζ is a pre-specified time point, the KM estimate may be unstable in the neighborhood of ζ, and the test based on V_1 may be slightly liberal (Table II(iv)) even when the sample size is moderately large and the event rate is not that low. One simple remedy is to modify the upper integral limit for V_1 and V_2 to be the minimum among ζ and the largest observed failure time from each of the two groups. With this modification, we find that the I error rate is preserved well, and there is negligible loss of power for tests with V_1 and V_2. We summarize the results of this finding in Table C.1 of Appendix C for the setting identical to that for Table II(iv). Note that for other scenarios considered in Table II, the operating characteristic profiles for the modified versions of V_1 and V_2 are identical to those reported in Table II because the KM curves are quite stable around ζ. Note that if ζ is not pre-specified, we may let ζ be ∞ for these modified tests.

In general, the results from an extensive numerical study indicate that the type I error rates of the proposed tests are well preserved with a moderate number of observed events in a comparative trial. The type I error rate of test V_1 may be slightly off from its nominal value when the sample size or the number

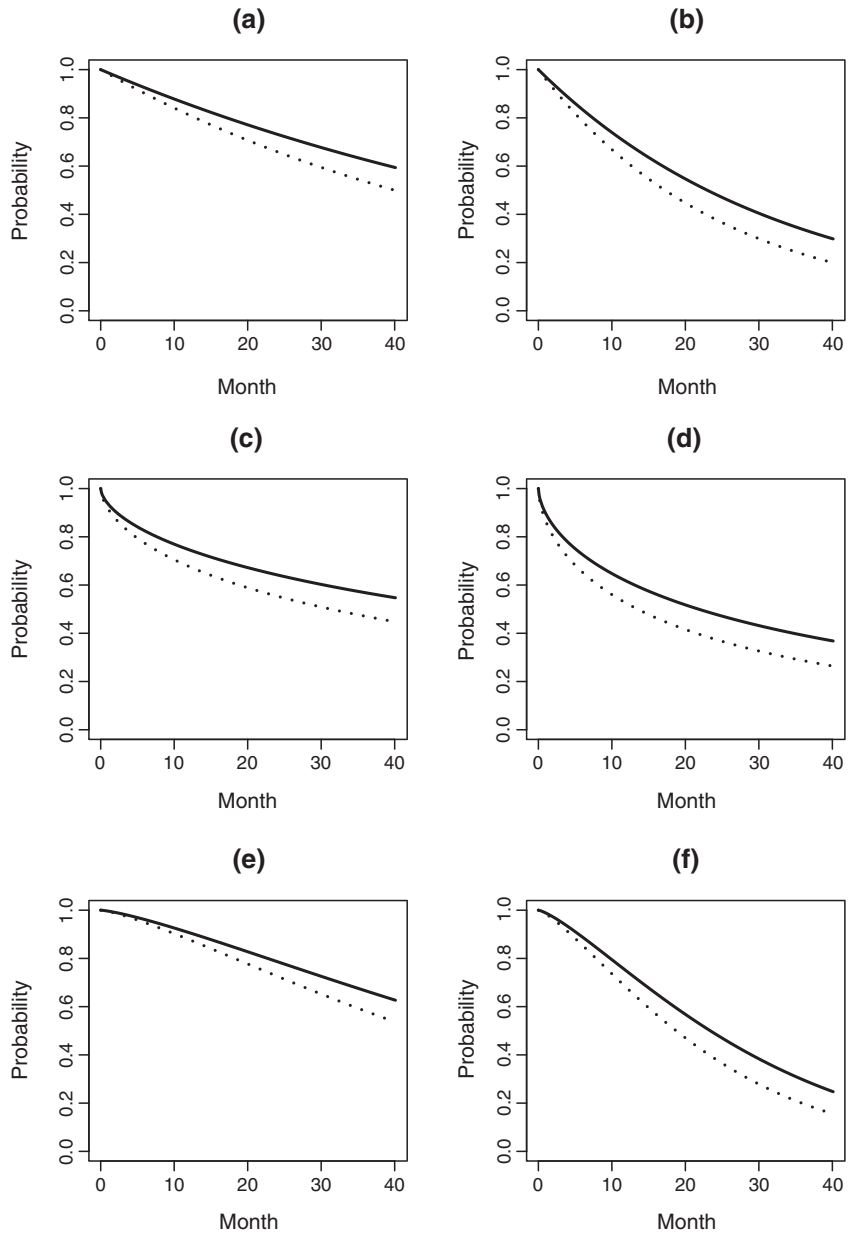

Figure 6. Comparison of survival functions considered in simulation studies under proportional hazards alternatives.

of observed events is very small. For instance, when the number of events is less than 25 or the size of the risk set at the truncation time is less than 30, the empirical type I error rate of the new test based on V_1 would be off by 0.02 to 0.03 with a nominal value of 0.05. Otherwise, for all cases considered in the numerical study, the empirical error rate is practically identical to its nominal value.

The perturbation resampling procedure is utilized to construct the null distributions of V_1 and V_2. The null distributions may also be generated via the standard permutation method. With an extensive study, we find that the permutation method controls the type I error rate at the nominal level with moderate sample sizes. For small sample size, Latta [22] showed that the permutation tests may not work well. On the other hand, for large sample size, the permutation tests are valid even when the censoring distributions are not equal between two groups to be compared [23, 24]. In Appendix D, we present the results from our extensive simulation study regarding the performance of the permutation tests based on V_1 and V_2.

Unlike the estimates of the hazard function or the cumulative version thereof, the KM plots are informative and easily interpretable in describing the temporal profile between-group differences. It seems natural to have a companion two-sample test for a global statistical assessment of such differences. The

Table III. Power of LR, PP, and YP tests, PF test based on (2), and the new tests based on (5) [V_1] and (7) [V_2], to detect the proportional hazards difference between two survival curves (Figure 6), based on 5000 of iterations, with sample size 300 per arm.

Censoring	(i) No censoring						(ii) Light censoring					
Survival distributions	6(a)	6(b)	6(c)	6(d)	6(e)	6(f)	6(a)	6(b)	6(c)	6(d)	6(e)	6(f)
Number of events*	271.7	450.1	301.2	410.1	250.5	478.5	259.7	432.9	291.8	398.6	237.6	457.5
Test	Power						Power					
LR	0.762	0.920	0.800	0.893	0.733	0.932	0.761	0.913	0.797	0.890	0.728	0.921
PP	0.757	0.895	0.791	0.873	0.730	0.904	0.750	0.885	0.787	0.869	0.718	0.893
YP	0.788	0.930	0.828	0.906	0.760	0.940	0.782	0.925	0.823	0.903	0.753	0.932
PF	0.700	0.888	0.763	0.873	0.650	0.902	0.687	0.886	0.762	0.877	0.634	0.897
V_1	0.695	0.895	0.766	0.879	0.637	0.911	0.685	0.891	0.768	0.879	0.620	0.899
V_2	0.667	0.862	0.712	0.829	0.638	0.884	0.652	0.856	0.705	0.826	0.616	0.877
Censoring	(iii) Heavy censoring						(iv) Observed in E4A03					
Survival distributions	6(a)	6(b)	6(c)	6(d)	6(e)	6(f)	6(a)	6(b)	6(c)	6(d)	6(e)	6(f)
Number of events*	235.3	398.2	272.9	375.8	211.5	415.5	241.9	415.6	281.7	388.8	214.6	435.1
Test	Power						Power					
LR	0.719	0.893	0.779	0.875	0.682	0.901	0.726	0.902	0.781	0.886	0.681	0.914
PP	0.706	0.863	0.765	0.852	0.670	0.871	0.720	0.876	0.769	0.864	0.674	0.885
YP	0.748	0.905	0.803	0.887	0.706	0.912	0.754	0.913	0.803	0.899	0.700	0.923
PF	0.657	0.871	0.750	0.864	0.600	0.877	0.652	0.873	0.747	0.865	0.590	0.878
V_1	0.661	0.877	0.757	0.871	0.594	0.882	0.689	0.893	0.776	0.886	0.615	0.904
V_2	0.619	0.833	0.678	0.812	0.578	0.853	0.615	0.841	0.679	0.815	0.567	0.856

*Average number of the total observed events across 5000 simulation realizations.
LR, logrank; PP, Peto–Prentice–Wilcoxon; YP, Yang–Prentice; PF, Pepe–Fleming.

logrank test statistic is sensitive for testing the equality of two hazard functions, especially under the PH model, but not necessarily for comparing two survival functions directly. The new proposal, on the other hand, is shown to be a useful tool to serve this purpose.

In this paper, we present the new proposals as one-sided testing procedures, which may be more appropriate than a two-sided, omnibus test for demonstrating that a new treatment tends to prolong the patient's survival time. On the other hand, our test statistics can be easily generalized to deal with a general alternative, which includes the scenario with crossing survival functions. For example, such an omnibus test can be obtained by replacing $Z(t)$ in $V_1(c)$ or $V_2(c)$ with $|Z(t)|$. It is straightforward to show that these generalized tests are consistent for a general alternative hypothesis. For other one-sided tests considered in this paper, it is common practice to consider the absolute value of the one-sided test statistic as a two-tailed test procedure. As kindly suggested by a referee, we conduced a simulation study to examine the performance of our proposed tests, for example, when the survival functions cross during the follow-up period (Figure 7). Again, we used the patterns of censoring distributions in Figure 3 for our study. We report the power comparisons in Table IV. Our tests are more powerful than almost all others, with the exception of the Yang and Prentice test [15]. Note that if we are interested in testing against such a general alternative hypothesis, tests such as the Kolmogorov–Smirnov type or Cramér–von Mises type of tests [25] may be considered.

The proposed method can be generalized trivially to handle the stratified case. Specifically, for each stratum, we perform our test and obtain its p-value. To combine these test results across strata, one may combine the stratum-specific p-values or their transformations thereof to assess the relative merit of the two groups to be compared. For example, a common approach is to utilize Fisher's method by summing the logarithms of p-values [26]. Under the null hypothesis, the resulting test statistic has a chi-squared distribution. The pros and cons of using this combination technique are well studied under a general setting. An attractive alternative is to use Stouffer's linear combination of weighted stratum-

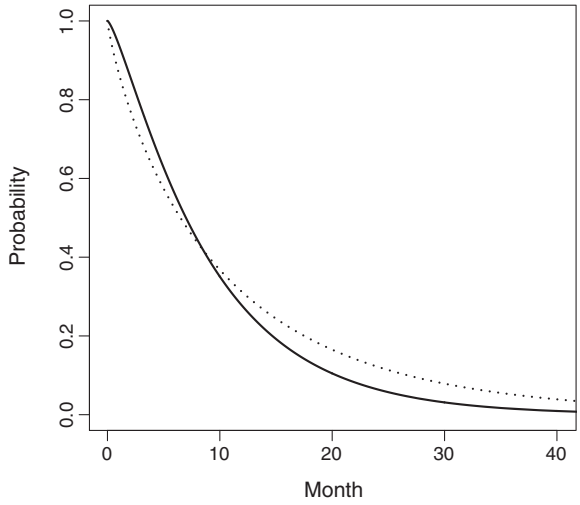

Figure 7. A cross-survival scenario considered in the simulation for two-sided tests.

Table IV. Power of two-sided tests (LR, PP, and YP tests, PF test based on (2), and the new tests based on (5) [V_1] and (7) [V_2]) with crossing survival curves (Figure 7), based on 5000 of iterations, with sample size 200 per arm.

	Censoring			
	(i)	(ii)	(iii)	(iv)
Number of events*	390.3	381.6	363.9	383.2
Test				
LR	0.199	0.184	0.152	0.182
PP	0.090	0.097	0.106	0.095
YP	0.554	0.534	0.490	0.527
PF	0.168	0.158	0.132	0.136
V_1	0.744	0.728	0.691	0.733
V_2	0.705	0.685	0.670	0.677

*Average number of the total observed events across 5000 simulation realizations.

LR, logrank; PP, Peto–Prentice–Wilcoxon; YP, Yang–Prentice; PF, Pepe–Fleming.

specific z-scores based on their p-values [27]. We may use equal weight for such a procedure or, for the current case, the weights being proportional to the stratum size or the number of observed events in the stratum.

Because the weighted logrank test is constructed by combining 2×2 tables at the observed event times, in principle, one may utilize the same idea to choose the weight associated with each table. However, the operating characteristics of the resulting test procedure are not clear. Further research is needed along this line. The new tests automatically and empirically adjust the weighting functions, which are not required to be pre-specified in the protocol or statistical analysis plan and are not restricted to be powerful only for specific alternatives.

Although hypothesis testing provides statistical evidence of treatment difference, the estimation of the treatment difference is also important for assessing the magnitude of the difference. The standard companion quantification procedure to the logrank test is the hazard ratio estimation. However, when the PH assumption is violated, the resulting estimate is difficult to interpret, as it is not simply an average of the true hazard ratio over time [28, 29]. An alternative, model-free approach is to use the restricted mean event time as the parameter of interest, which can be estimated by the area under the KM curve. Inferences about the difference or ratio of two restricted mean survival times can be made

[30–34]. Further research is warranted to connect the proposed testing procedure with these closely related estimators.

An R package (survTest2) to implement the proposed method is available on request from the author (email: huno@jimmy.harvard.edu).

Appendix A: Weak convergence of $V_1(c)$ and $P(c)$

Note that under the null hypothesis, $Z(t)$ converges weakly to $G(t)$ as a process. Assuming that $c_2 - c_1 = o(1)$, we have

$$\left|V_1(c_2) - V_1(c_1)\right| \leqslant |c_2 - c_1| O\left(\int_0^{\zeta} |Z(t)|dt\right) = o_p(1). \tag{A.1}$$

Coupled with the continuous mapping theorem, (A.1) implies that $V_1(c)$ converges weakly to $\psi(c) = \int_0^{\zeta} \max\{G(t), c\}G(t)dt$.

Next, we show that under H_0, $P(c) \to U(c), c \in [0, \eta]$ in distribution as $n \to \infty$, where the limiting process $U(c) = S_{\psi(c)}\{\psi(c)\}$ and $S_{\psi(c)}(v) = \mathrm{pr}\{\psi(c) \geqslant v\}$, the survival function of $\psi(c)$. Here, we consider the survival function of $V_1(c)$, $S_{V_1(c)}(v) = \mathrm{pr}\{V_1(c) \geqslant v\}$ as well. To simplify the notation in the following argument, we use $S_c(v)$ and $\hat{S}_c(v)$ for $S_{\psi(c)}(v)$ and $S_{V_1(c)}(v)$, respectively. Because of the weak convergence of $V_1(c)$, for any $c_1, c_2, \ldots, c_K \in [0, \eta]$, we have

$$\sup_{1 \leqslant k \leqslant K} \sup_{v \in (-\infty, \infty)} |\hat{S}_{c_k}(v) - S_{c_k}(v)| = o(1)$$

as $n \to \infty$. Thus,

$$\sup_{1 \leqslant k \leqslant K} |P(c_k) - S_{c_k}\{V_1(c_k)\}| = \sup_{1 \leqslant k \leqslant K} |\hat{S}_{c_k}\{V_1(c_k)\} - S_{c_k}\{V_1(c_k)\}| = o(1), \quad \text{a.s,}$$

which implies that

$$\begin{aligned}
&\left| \mathrm{pr}\left\{P(c_1) \geqslant p_1, \ldots, P(c_K) \geqslant p_K\right\} - \mathrm{pr}\left\{U(c_1) \geqslant p_1, \ldots, U(c_K) \geqslant p_K\right\} \right| \\
\leqslant &\left| \mathrm{pr}\left\{P(c_1) \geqslant p_1, \ldots, P(c_K) \geqslant p_K\right\} - \mathrm{pr}\left[S_{c_1}\{V_1(c_1)\} \geqslant p_1, \ldots, S_{c_K}\{V_1(c_K)\} \geqslant p_K\right] \right| \\
&+ \left| \mathrm{pr}\left\{V_1(c_1) \leqslant S_{c_1}^{-1}(p_1), \ldots, V_1(c_K) \leqslant S_{c_K}^{-1}(p_K)\right\} - \mathrm{pr}\left\{\psi(c_1) \leqslant S_{c_1}^{-1}(p_1), \ldots, \psi(c_K) \leqslant S_{c_K}^{-1}(p_K)\right\} \right| \\
= &\, o(1) \quad \text{for} \quad 0 \leqslant p_1, \ldots, p_K \leqslant 1.
\end{aligned}$$

Thus, for any $c_1, c_2, \ldots, c_K \in [0, \eta]$, the joint distribution of $\{P(c_1), \ldots, P(c_K)\}$ converges to that of $\{U(c_1), \ldots, U(c_K)\}$. In addition, (A.1) implies that for $c_2 - c_1 = o_p(1)$,

$$\begin{aligned}
|P(c_2) - P(c_1)| &= |S_{c_2}\{V_1(c_2)\} - S_{c_1}\{V_1(c_1)\}| + o_p(1) \\
&\leqslant |S_{c_2}\{V_1(c_2)\} - S_{c_1}\{V_1(c_2)\}| + |S_{c_1}\{V_1(c_2)\} - S_{c_1}\{V_1(c_1)\}| + o_p(1) = o_p(1).
\end{aligned}$$

Thus, under the null, $P(c)$ converges weakly to the process $U(c)$ indexed by c.

Appendix B: Approximation of the null distribution of P_b by P_b^*

Let $\mathcal{O}$ be the observed data. Let $\hat{S}_c^*(v) = \mathrm{pr}\{V_1^*(c) \geqslant v|\mathcal{O}\}$. Note that under the null, $Z^*(t)|\mathcal{O}$ converges weakly to $G(t)$ as a process almost surely. Coupled with the continuous mapping theorem, $V_1^*(c)|\mathcal{O}$ converges weakly to $\psi(c)$ almost surely. Thus, using the same argument in the Appendix A, we can show that for any $c_1, c_2, \ldots, c_K \in [0, \eta]$,

$$\sup_{1 \leqslant k \leqslant K} \sup_{v \in (-\infty, \infty)} |\hat{S}_{c_k}^*(v) - S_{c_k}(v)| = o(1)$$

almost surely as $n \to \infty$, which implies $\{P^*(c_1), \ldots, P^*(c_K)\} \mid \mathcal{O} \to \{U(c_1), \ldots, U(c_K)\}$. Also, $|P^*(c_2) - P^*(c_1)| = o_p(1)$, for $|c_1 - c_2| = o(1)$ is derived in the same way. Therefore, $P^*(c) = \hat{S}_c^*\{V_1^*(c)\}$, conditional on the observed data, converges weakly to the process $U(c), c \in [0, \eta]$ almost surely. Therefore, the null distribution of $P_b = \min\{P(c) : 0 \leqslant c \leqslant \eta\}$ can be approximated by $P_b^* = \min\{P^*(c) : 0 \leqslant c \leqslant \eta\}$.

Appendix C: Simulation results with a modified version of proposed tests that corresponds to Table II(iv)

Under the same setting as was used to generate Table II(iv), we conducted further simulation studies using the modified version of our proposed tests illustrated in Section 4. Table C.1 shows the results of the simulation studies. The I error is well preserved to the nominal level, and power profiles are also well preserved compared with those presented in Table II(iv).

Table C.1. Contrast of size and power of V_1 and V_2 between their original versions (Table II(iv)) and the modified versions, based on 5000 of iterations, with sample size 200 per arm.

Censoring	(iv) Observed in E4A03				
Survival distributions	4(a)	4(b)	4(c)	4(d)	4(e)
Test	Size	Power			
V_1 (modified)	0.050	0.685	0.971	0.658	0.381
V_1 (Table II(iv))	0.062	0.715	0.968	0.646	0.389
V_2 (modified)	0.054	0.609	1.000	0.557	0.143
V_2 (Table II(iv))	0.054	0.613	1.000	0.557	0.142

Figure 4(a) no difference, (b) proportional hazards, (c) early difference, (d) difference in middle, and (e) late difference.

Appendix D: Simulation results for permutation tests

Following the suggestion by a reviewer, we have conducted another set of numerical studies to examine the behavior of the permutation version of our test procedures under a null setting (Figure D.1). For censoring, in addition to the four censoring patterns shown in Figure 3 (i.e., (i) no censoring, (ii) light censoring, (iii) heavy censoring, and (iv) observed in E4A03), we considered a case when the censoring distributions are different between two groups (v). The distributions of censoring time are shown in Figure D.2. In these settings, the permutation and the perturbation versions of V_1 and V_2 preserve I error rate well even when the censoring distributions are different (Table D.1).

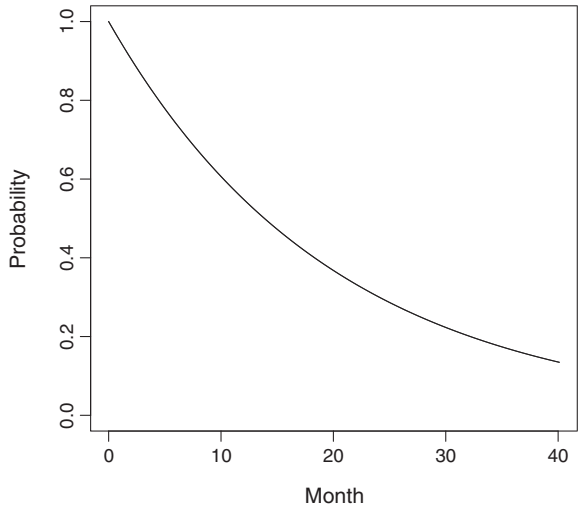

Figure D.1. Survival time distribution considered in the simulation for permutation tests.

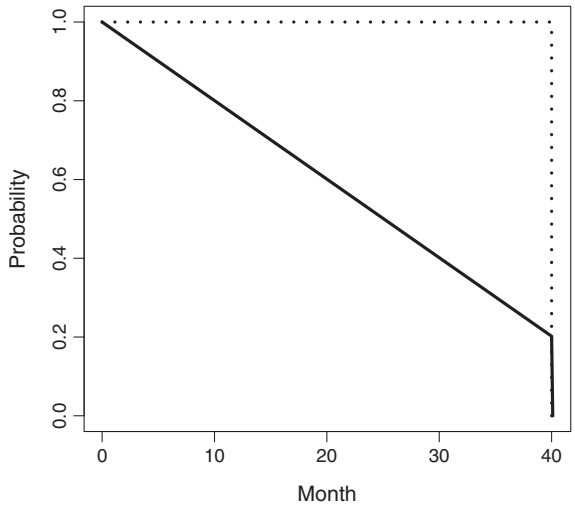

Figure D.2. Censoring pattern (v) considered in the simulation for permutation tests under unequal censoring time distributions. Solid and dotted lines indicate survival functions of underlying censoring time distributions of two groups.

Table D.1. Size of LR, PP, and YP tests, PF test based on (2), and the new tests based on (5) [V_1] and (7) [V_2] (both perturbation and permutation versions), with sample size 500 per arm.

Size of tests Survival distributions: 7					
	Censoring				
	(i)	(ii)	(iii)	(iv)	(v)
Number of events*	864.3	834.8	775.8	814.9	745.9
Test					
LR	0.051	0.052	0.053	0.050	0.054
PP	0.051	0.054	0.055	0.050	0.053
YP	0.059	0.061	0.061	0.057	0.061
PF	0.050	0.054	0.052	0.052	0.054
V_1 (perturbation)	0.050	0.054	0.053	0.053	0.055
V_2 (perturbation)	0.049	0.052	0.053	0.049	0.054
V_1 (permutation)	0.048	0.052	0.053	0.051	0.051
V_2 (permutation)	0.049	0.051	0.052	0.050	0.052

*Average number of the total observed events across 5000 simulation realizations.
LR, logrank; PP, Peto–Prentice–Wilcoxon; YP, Yang–Prentice; PF, Pepe–Fleming.

Acknowledgements

The authors are grateful to the reviewers for insightful comments/suggestions on the manuscript. They also thank the Eastern Cooperative Oncology Group for permission to use the data. This work was supported in part by grants R01-HL089778, R01-GM079330, R01-AI024643, UM1AI068616 and UM1AI068634 from the NIH.

References

1. Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics* 1989; **45**:497–507.
2. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977; **64**(1):156–160.

3. Lagakos S, Schoenfeld D. Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* 1984; **40**(4):1037–1048.
4. Rajkumar SV, Jacobus S, Callander NS, Fonseca R, Vesole DH, Williams ME, Abonour R, Siegel DS, Katz M, Greipp PR, for the Eastern Cooperative Oncology Group. Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial. *The Lancet Oncology* 2010; **11**(1):29–37.
5. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
6. Tarone RE. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics* 1981; **37**:79–85.
7. Fleming TR, Harrington DP. Evaluation of censored survival data test procedures based on single and multiple statistics. In *Topics in Applied Statistics*. Marcel Dekker: New York, 1984; 97–123.
8. Gastwirth JL. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* 1985; **80**(390):380–384.
9. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 1990; **77**(4):853–864.
10. Self SG. An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics* 1991; **47**(3):975–986.
11. Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* 1996; **52**:721–725.
12. Kosorok MR, Lin CY. The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association* 1999; **94**(445):320–332.
13. Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* 1991; **19**(2):531–556.
14. Pecková M, Fleming TR. Adaptive test for testing the difference in survival distributions. *Lifetime Data Analysis* 2003; **9**:223–238.
15. Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. *Biometrics* 2010; **66**(1):30–38.
16. Yang S, Prentice R. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika* 2005; **92**(1):1–17.
17. Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: large sample and optimality considerations. *Journal of the Royal Statistical Society Series B* 1991; **53**(2):341–352.
18. Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* 2003; **4**(2):223–229.
19. Gill R. Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics* 1983; **11**:49–58.
20. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**(1):73–81.
21. Parzen MI, Wei LJ, Ying Z. Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics* 1997; **24**(3):309–314.
22. Latta R. A Monte Carlo study of some two-sample rank tests with censored data. *Journal of the American Statistical Association* 1981; **76**(375):713–719.
23. Neuhaus G. Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics* 1993; **21**(4):1760–1779.
24. Heimann G, Neuhaus G. Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics* 1998; **54**:168–184.
25. Schumacher M. Two-sample tests of the Cramér–von Mises and Kolmogorov–Smirnov type for randomly censored data. *International Statistical Review* 1984; **52**:263–281.
26. Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 1925.
27. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM, Jr. *The American soldier*, Adjustment during Army Life, vol. 1. Princeton University Press: Princeton, 1949.
28. Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika* 1986; **73**(2):363–369.
29. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics* 2000; **1**(4):423–439.
30. Murray S, Tsiatis AA. Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics* 1999; **55**(4):1085–1092.
31. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 2011; **30**(19):2409–2421.
32. Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, Wei LJ. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials* 2012; **9**(5):570–577.
33. Tian L, Zhao L, Wei L. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 2014; **15**(2):222–233.
34. Yang S. Semiparametric inference on the absolute risk reduction and the restricted mean survival difference. *Lifetime Data Analysis* 2013; **19**(2):219–241.