Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Ball divergence for the equality test of crossing survival curves

Na You[1,2] | Xueyi He[1] | Hongsheng Dai[2,3] | Xueqin Wang*[4]

[1]School of Mathematics, Sun Yat-sen University, Guangdong, China
[2]Department of Mathematical Sciences, University of Essex, Colchester, U.K.
[3]School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, U.K.
[4]School of Management, University of Science and Technology of China, Anhui, China

Correspondence
*Xueqin Wang, School of Management, University of Science and Technology of China, Anhui, China. Email: wangxq20@ustc.edu.cn

Summary

It is a very common problem to test survival equality using the right-censored time-to-event data in clinical research. Although the log-rank test is popularly used in various studies, it may become insensitive when the proportional hazards assumption is violated. As follows, there have a variety of statistical methods being proposed to identify the discrepancy between crossing survival curves or hazard functions. The omnibus tests against general alternatives are usually preferred due to their wide applicability to complicated scenarios in real applications. In this paper, we propose two novel statistics to estimate the ball divergence using the right-censored survival data, and then implement them in the equality test on survival time in two independent groups. The simulation analysis demonstrates their efficiency in identifying the survival discrepancy. Compared to the existing methods, our proposed methods present higher power in situations with complex distributions, especially when there is a scale shift between groups. Real examples illustrate its advantage in practical applications.

KEYWORDS:
Crossing survival curves; Censored data; Two-sample test; Nonparameteric; Permutation.

1 | INTRODUCTION

The comparison of survival time in two independent groups is a very common problem in medical research. For instance, to evaluate the treatment effect of some new drug on tumor patients, the relapse or distant metastasis occurrence time of the treated group with the new drug would be compared to that of the control group who accepts the standardized treatment. Moreover, to explore potential risk factors, the survival times in the stratified groups, say males and females, would be compared. Due to different individual admission times and the limited duration of the clinical trial, the observed survival data may be censored if the event has not happened before the trial ends. With these right-censored observations, many statistical methods have been proposed for the equality test of survival times in two arms. As a representation being well known, the log-rank (LR) test[1] is most widely used in different disciplines, however, it may become insensitive in distinguishing the alternative of crossing survival curves. Therefore, the development of statistical methods for testing the crossing alternatives is of great interest for a long history[2,3,4,5,6,7,8,9].

In the literature, there were two types of crossings being considered, that is the crossing of hazard rates or survival curves[7]. Accordingly, two types of methods that are based on the comparison of hazard rates or survival curves were motivated. The LR test summarizes the differences between the observed hazard rates of one group at each event time and their expectations under the null hypothesis as the test statistic. In the cases of crossing hazard rates, these differences at different event time points present opposite signs and therefore are canceled out, yielding the power loss of LR test. To overcome the obstacle, the amended test statistics utilized various weight functions to emphasize the differences between two hazard rates at some stages[10],

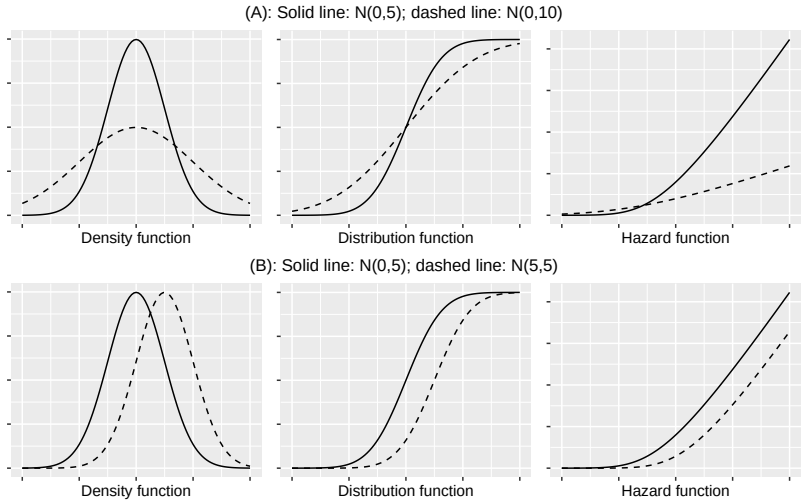

Figure 1 Illustrative examples of scale shift (A) and mean shift (B) between two probability measures. Left column: density function $f(x)$; middle column: distribution function $F(x) = \int_{-\infty}^{x} f(y)dy$; right column: hazard function $h(x) = f(x)/(1-F(x))$.

or considered the supremum or the square sum of these differences[2,5,6,11]. Taking the maximum of multiple subjects is another common strategy to set up the statistic to test equivalence on multiple dimensions[12]. In the meantime, to quantify the discrepancy between two survival curves, the difference between their Kaplan-Meier estimators was used to construct the test statistics, incorporated with some weight functions, absolute difference or the supremum to handle the crossing[13,3,14,15,16]. Among these methods, some were constructed for specific scenarios such as the crossing at early stage, while most recent ones were omnibus for general alternatives of nonidentical hazard functions or survival curves. In practice, it is uncertain to know whether there is a crossing or what kind of crossing it is, thus the omnibus tests against general alternatives are usually preferred. In order to investigate the performance of existing methods to guide the practical use, the numerical analysis was conducted by Li et al.[17] via a series of simulations and Dormuth et al.[18] via plenty of real datasets. It is shown that two-stage (TW) method[8], the adaptive Neyman smoothing (NS) test[9], and the test utilizing the combination of multiple hazard weights (mdir)[11] outperform the others with higher powers in general scenarios.

No matter whether the hazard rate functions or survival curves were considered to construct the test statistics, the equality of two probability measures on the survival time was of interest. To quantify the dissimilarity between probability measures, several innovative statistical divergences and their sample estimators were proposed in modern research, such as the energy distance[19], MMD[20], HHG[21] and Ball divergence[22]. Besides much attention they have attracted in dealing with high-dimensional data, another important feature that makes them outstanding is their nonparametric nature. The statistical inference based on these quantities is model-free with weak assumptions, therefore can deal with a vast range of situations. They were popularly adapted to the univariate survival analysis. Matabuena and Padilla[23] adapted the statistics to estimate the energy distance and MMD measure using the right-censored data. Gorfine et al.[24] extended the HHG test for comparing two or more survival distributions using right-censored data.

For two probability measures T_1 and T_2 with a mean shift, i.e., $T_2 = T_1 + \mu$ where μ is a constant, their distribution functions F_1 and F_2 satisfy $F_2(t) = F_1(t-\mu)$. Since the distribution function is non-decreasing, $F_2(t) \geq F_1(t)$ when $\mu < 0$, or $F_2(t) \leq F_1(t)$ when $\mu > 0$, thus F_1 and F_2 never cross over. For T_1 and T_2 with a scale shift which can be expressed as $T_1 = \mu + \epsilon$ and $T_2 = \mu + \sigma\epsilon$, where μ and $\sigma > 0$ are constants and ϵ is a random variable with the distribution function $F(t)$ and mean zero, $F_1(t) = F(t-\mu)$ and $F_2(t) = F((t-\mu)/\sigma)$. When $\sigma > 1$, $F_2(t) \geq F_1(t)$ for $t < \mu$, then they intersect at $t = \mu$, and $F_2(t) \leq F_1(t)$ for $t > \mu$. When $\sigma < 1$, $F_2(t) \leq F_1(t)$ for $t < \mu$, and then $F_2(t) \geq F_1(t)$ after the intersection at $t = \mu$. Therefore, the distribution functions cross due to the scale rather than mean shift between the probability measures. We present two illustrative examples in Figure 1. It was demonstrated that Ball divergence is more powerful in testing the scale shift[22]. In this paper, we propose two novel statistics to estimate the Ball divergence using the right-censored survival data, and implement them in the two-sample equality test for crossing survival curves.

The paper is structured as follows. In Section 2, we introduce the test statistics and hypothesis testing procedure. A series of simulation studies are conducted in Section 3 to investigate their performance in finite samples. In Section 4, we apply the proposed testing procedure to multiple real datasets, and a short discussion is given in Section 5.

2 | METHODS

In Banach space $(V, ||\cdot||)$, the Ball divergence of two Borel probability measures μ and ν was defined as[22],

$$D(\mu,\nu)=\iint_{V\times V}(\mu-\nu)^2\bar{B}(x,\rho(x,y))\Big(\mu(dx)\mu(dy)+\nu(dx)\nu(dy)\Big),$$

where $\bar{B}(x,\rho(x,y))$ indicates the closed ball with center x and radius $\rho(x,y)=||x-y||$. For two survival times $T_1\sim F_1$ over $(0,\tau_1]$ and $T_2\sim F_2$ over $(0,\tau_2]$, their Ball divergence is given by

$$D(F_1,F_2)=\int_0^{\tau}\int_0^{\tau}(F_1-F_2)^2\bar{B}(x,|x-y|)\Big(F_1(dx)F_1(dy)+F_2(dx)F_2(dy)\Big),$$

where $|x-y|$ is the absolute difference between $x,y\in(0,\tau]$, and $\tau=\max(\tau_1,\tau_2)$. Let $s=\min(2x-y,y)$ and $\upsilon=\max(2x-y,y)$, and $\gamma_k(x,y)=\mathbb{P}(s<T_k\le\upsilon)$ for $k=1,2$. The Ball divergence has more specific expression, i.e.,

$$D(F_1,F_2)=D_1+D_2,\quad\text{with}\begin{cases}D_1=\int_0^{\tau}\int_0^{\tau}[\gamma_1(x,y)-\gamma_2(x,y)]^2F_1(dx)F_1(dy),\\ D_2=\int_0^{\tau}\int_0^{\tau}[\gamma_1(x,y)-\gamma_2(x,y)]^2F_2(dx)F_2(dy).\end{cases}\tag{1}$$

Note that $\gamma_1(x,y)$ and $\gamma_2(x,y)$ respectively measure the probability that T_1 and T_2 belong to the time interval $(s,\upsilon]$.

Suppose there are two independent samples $T_{k,i}\sim F_k$, $i=1,2,\dots,n_k$, $k=1,2$, which may not be observed but censored by $C_{k,i}\sim G_k$, $i=1,2,\dots,n_k$, $k=1,2$, respectively. Note that $T_{k,i}$ and $C_{k,i}$ are mutually independent of each other. The observed data are then represented by $X_{k,i}=\min(T_{k,i},C_{k,i})$ and $\delta_{k,i}=I(T_{k,i}\le C_{k,i})$, $i=1,2,\dots,n_k$, $k=1,2$. Our target is to test the hypothesis

$$H_0:F_1=F_2\quad\text{versus}\quad H_a:F_1\neq F_2.\tag{2}$$

For $k=1,2$, let $H_k(t)=n_k^{-1}\sum_{i=1}^{n_k}I(X_{k,i}>t)$ and

$$\Gamma_k(x,y)=\frac{H_k(s)}{\hat{\bar{G}}_k(s)}-\frac{H_k(\upsilon)}{\hat{\bar{G}}_k(\upsilon)},$$

where $\hat{\bar{G}}_k$ is the Kaplan-Meier estimate for $\bar{G}_k=1-G_k$.

Lemma 1. For $\forall\ x,y$ such that $\bar{G}_k(\upsilon)>0$, $\Gamma_k(x,y)$ is a consistent estimator for $\gamma_k(x,y)$.

With the survival observations $(X_{k,i},\delta_{k,i})$, $i=1,2,\dots,n_k$, $k=1,2$, let

$$\mathcal{D}_k=\frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\frac{\delta_{k,i}\delta_{k,j}}{\hat{\bar{G}}_k(X_{k,i})\hat{\bar{G}}_k(X_{k,j})}\Big[\Gamma_1(X_{k,i},X_{k,j})-\Gamma_2(X_{k,i},X_{k,j})\Big]^2,\quad k=1,2,$$

and

$$\mathcal{D}=\mathcal{D}_1+\mathcal{D}_2.\tag{3}$$

Theorem 1. $\mathcal{D}$ is a consistent estimator for the Ball divergence $D(F_1,F_2)$ (1).

Besides, the Kaplan-Meier integral[25] provides a general method to estimate statistical functionals using the right-censored survival data. We order the observations in two samples and denote them as $X_{k(1)}\le X_{k(2)}\le\cdots\le X_{k(n_k)}$, with the corresponding status $\delta_{k(i)}$, $i=1,2,\dots,n_k$, $k=1,2$.

Theorem 2. Let

$$W_{k(i)}=\frac{\delta_{k(i)}}{n_k-i+1}\prod_{j=1}^{i-1}\Big(\frac{n_k-j}{n_k-j+1}\Big)^{\delta_{k(j)}},$$

and

$$\mathcal{D}_k^A = \sum_{i=1}^{n_k}\sum_{j=1}^{n_k} W_{k(i)}W_{k(j)}\left[\sum_{u=1}^{n_1} W_{1(u)}\eta(X_{k(i)}, X_{k(j)}, X_{1(u)}) - \sum_{v=1}^{n_2} W_{2(v)}\eta(X_{k(i)}, X_{k(j)}, X_{2(v)})\right]^2, \text{ for } k = 1, 2,$$

where $\eta(x, y, z) = I\big(z \in \bar{B}(x, |x-y|\big) = I\big(\min(2x-y, y) < z \le \max(2x-y, y)\big)$.

$$\mathcal{D}^A = \mathcal{D}_1^A + \mathcal{D}_2^A \tag{4}$$

provides a consistent estimator for the Ball divergence $D(F_1, F_2)$ (1), alternatively.

We call (3) and (4) as the survival Ball divergence, denoted by BD(1) and BD(2) respectively, and use them as the test statistic for the hypothesis (2). In order to quantify the significance of the observed data, it is useful to explore the distribution of test statistic under null hypothesis. However,

$$\mathcal{D}_k = \frac{1}{n_k^2}\frac{1}{n_1^2 n_2^2}\sum_{i,j=1}^{n_k}\sum_{u,u'=1}^{n_1}\sum_{v,v'=1}^{n_2} \frac{\delta_{k,i}\delta_{k,j}\Big[\xi_1(X_{k,i}, X_{k,j}, X_{1,u}) - \xi_2(X_{k,i}, X_{k,j}, X_{2,v})\Big]\Big[\xi_1(X_{k,i}, X_{k,j}, X_{1,u'}) - \xi_2(X_{k,i}, X_{k,j}, X_{2,v'})\Big]}{\hat{\bar{G}}_k(X_{k,i})\hat{\bar{G}}_k(X_{k,j})},$$

with $\xi_k(x, y, z) = I\big(z > \min(2x-y, y)\big)/\hat{\bar{G}}_k\big(\min(2x-y, y)\big) - I\big(z > \max(2x-y, y)\big)/\hat{\bar{G}}_k\big(\max(2x-y, y)\big)$, and

$$\mathcal{D}_k^A = \sum_{i,j=1}^{n_k}\sum_{u,u'=1}^{n_1}\sum_{v,v'=1}^{n_2} W_{k(i)}W_{k(j)}\Big[W_{1(u)}\eta(X_{k(i)}, X_{k(j)}, X_{1(u)}) - W_{2(v)}\eta(X_{k(i)}, X_{k(j)}, X_{2(v)})\Big]\cdot \Big[W_{1(u')}\eta(X_{k(i)}, X_{k(j)}, X_{1(u')}) - W_{2(v')}\eta(X_{k(i)}, X_{k(j)}, X_{2(v')})\Big]$$

are both Kaplan-Meier V-statistics[26], whose asymptotic distributions are complicated to be inferred in theory and also by numerical approximation. Thus, we use the permutation method[27,23] to calculate p-value for the hypothesis testing (2) using the test statistics BD(1) or BD(2). Assign the group label $K_i = 1$ to $(X_{1,i}, \delta_{1,i})$, $i = 1, 2, \dots, n_1$, and $K_{n1+j} = 2$ to $(X_{2,j}, \delta_{2,j})$, $j = 1, 2, \dots, n_2$. With a permutation of $(K_1, \dots, K_{n_1+n_2})$, denoted by $(K_1^*, \dots, K_{n_1+n_2}^*)$, the samples are divided into two groups according to $K_i^* = 1$ or 2, and the test statistic can be calculated. The permutation repeats B times, and the p-value is given by the proportion of test statistics with permuted labels that are greater or equal to the observed statistic.

3 | SIMULATIONS

A series of experiments were simulated to numerically evaluate the performance of our proposed testing procedure with BD(1) or BD(2) in a variety of scenarios. The type-I error rates under null hypotheses and the testing powers under alternative settings were investigated. For null hypotheses, two experiments, 1 and 2, were considered. In experiment 1, the survival times T_{ki}, $i = 1, 2, \dots, n_k$, for $k = 1, 2$, are both simulated from the Weibull distribution $\bar{F}_1(t) = \bar{F}_2(t) = \exp(-\lambda t^p)$ with $\lambda = 2$ and $p = 4$, while in experiment 2, they are generated from the Gompertz distribution $\bar{F}_1(t) = \bar{F}_2(t) = \exp\big(\lambda(1 - \exp(\alpha t))/\alpha\big)$ with $\lambda = 3$ and $\alpha = 1$. The survival functions in these two experiments are plotted in Figure 2, presenting an inverse S and L shaped survival curves. The testing powers under alternative hypotheses are assessed in the following experiments 3-9, where T_{1i}, $i = 1, 2, \dots, n_1$ and T_{2j}, $j = 1, 2, \dots, n_2$ are from different survival functions. As shown in Figure 2, two survival curves to be tested in experiments 3-4 are not crossing while they intersect each other in experiments 5-9. More specifically, the hazard functions of two groups in experiment 3 are proportional, while not in experiment 4. For the crossed survival curves in experiments 5-7, the intersections are set at early, median and later stages respectively, as done by Li et al.[17]. Two more complicated scenarios are implemented in experiments 8-9, where $\bar{F}_k(t)$ may be a mixture distribution. To illustrate the dispersion of two probability measures in experiments 3-9, we plot their density functions in Figure 2 as well. It is shown that, in experiments 3-4, there mainly show location shifts, meanwhile in experiments 5-9, there additionally appear scale shifts. Particularly in experiment 9, it is rather a scale shift than a location shift. The distributions and their parameters used for data generation are listed in Table 1. For each experiment, the censoring times C_{1i}, $i = 1, 2, \dots, n_1$ and C_{2j}, $j = 1, 2, \dots, n_2$ are subject to the uniform distribution $U(0, C_1)$ and $U(0, C_2)$ respectively, where C_1 and C_2 are chosen to yield no censoring, or the censoring rates of both groups at low level around 20%, median level around 40%, or high level around 60%.

First, the consistency of our proposed statistics BD(1) and BD(2) was investigated. For each experiment, with different settings of C_1 and C_2, we increase the sample sizes $n_1 = n_2 = 30$, 50 to 100. Under each scenario, the statistics BD(1) and BD(2) were calculated according to (3) and (4) with the simulated data. The simulation is repeated 500 times. The average and standard

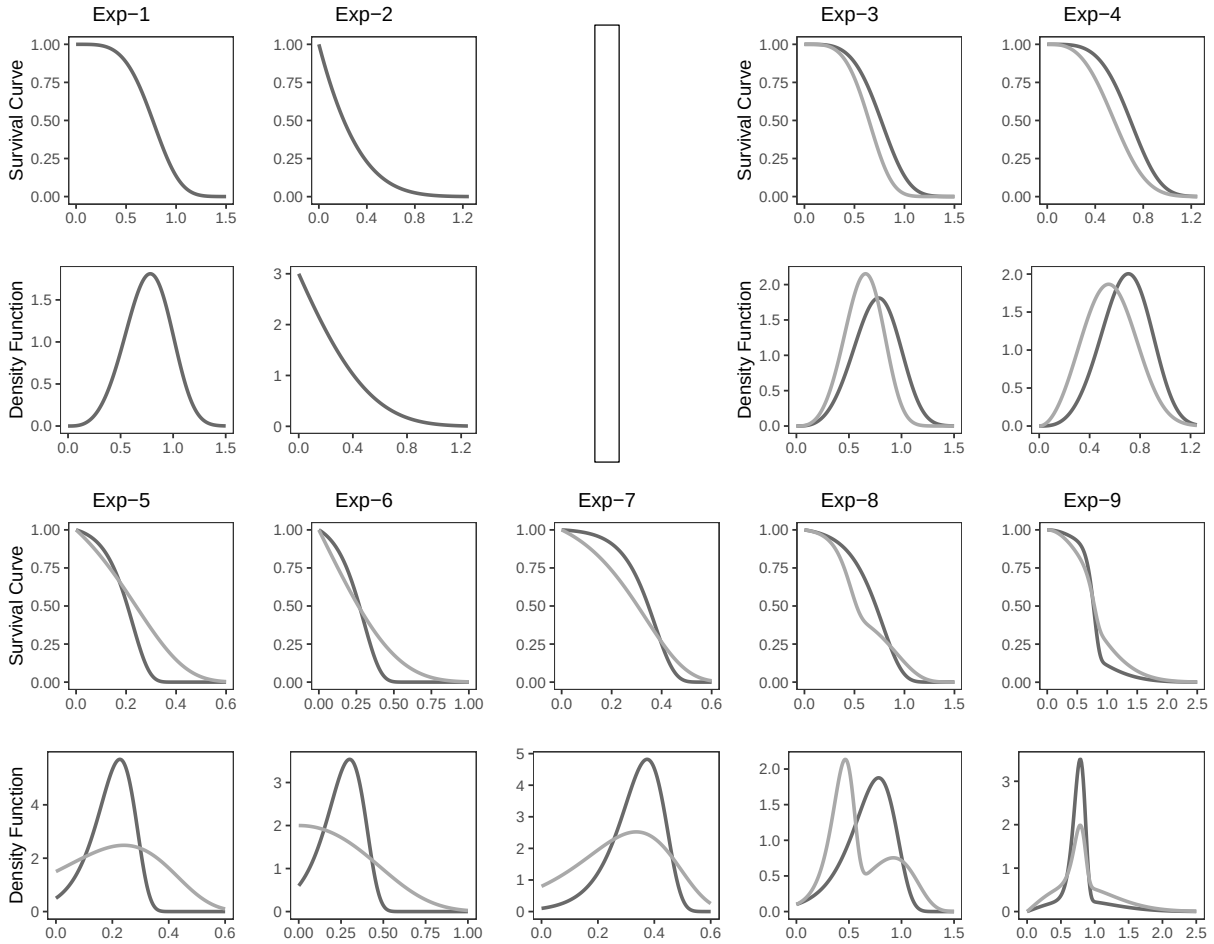

Figure 2 Survival curves and density functions to be tested in the simulation experiments 1-9.

deviation of the statistics from 500 repetitions are calculated for evaluation, being presented in Table 2. The Ball divergence is zero if and only if $F_1 = F_2$ [22], however, it is hard to be computed analytically when $F_1 \neq F_2$ due to its complex expression (1). For experiments 1-2 with null hypotheses $F_1 = F_2$, $D(F_1, F_2) = 0$, while for experiments 3-9, we used the Monte Carlo method to approximate $D(F_1, F_2)$ by its sample estimator, where two random samples of size 5000 were generated from F_1 and F_2, respectively, and the sample estimator was calculated using R package `Ball` [28]. It is shown in Table 2, as the sample size increases, both BD(1) and BD(2) approach $D(F_1, F_2)$, and their standard deviations become smaller. As mentioned previously, both D_k and D_k^A are Kaplan-Meier V-statistics, which are biased estimators [29], so that BD(1) and BD(2) present positive biases from $D(F_1, F_2)$ in most scenarios. The bias reduces as the sample size goes larger, demonstrating the consistency of the statistics as proved. It is indicated that the estimation bias of BD(2) is larger than that of BD(1) when the censoring rate is high, especially in the scenario with a smaller sample size.

Next, the performance of our testing procedure is investigated for each experiment with the sample sizes $n_1 = n_2 = 30$ or 50 and different censoring rates. Each scenario with different parameter settings is repeated 500 times. Given the significance level $\alpha = 0.05$, the proportion that the null hypothesis is rejected is calculated, which estimates the type-I error rate in experiments 1-2 and the testing power in experiments 3-9. We summarize the results from our testing procedure using BD(1) and BD(2) in Figure 3 and 4. Li et al. [17] and Dormuth et al. [18] concluded that two-stage (TS) method [8], the adaptive Neyman smoothing (NS) test [9], and the test employing the combination of multiple hazard weights (mdir) [11] outperform the other existing methods with higher powers in general scenarios. Therefore, we compare the results of our method to that of the classical LR test, TS, NS, mdir, and the newly developed ED, MMD, and HHG methods for survival data analysis. Matabuena and Padilla [23] provided the R source code to calculate ED with Euclidean distance and MMD with Gaussian and Laplacian kernels while presenting the methodologies. We used their code to apply ED and MMD methods. The results of MMD with two kernels are very similar,

Table 1 Survival distributions and their parameters used for data generation in simulations.

	Exp	Distribution	Parameter Setting	
			Group 1	Group 2
Type-I Error Rate	1	Weibull	$\lambda = 2, p = 4$	
	2	Gompertz	$\lambda = 3, \alpha = 1$	
Power	3	Weibull	$\lambda = 2, p = 4$	$\lambda = 4, p = 4$
	4	Weibull	$\lambda = 3, p = 4$	$\lambda = 4, p = 3$
	5	Gompertz	$\lambda = 0.5, \alpha = 15$	$\lambda = 1.5, \alpha = 5$
	6	Gompertz	$\lambda = 0.6, \alpha = 9$	$\lambda = 2, \alpha = 2$
	7	Gompertz	$\lambda = 0.1, \alpha = 13$	$\lambda = 0.8, \alpha = 6$
	8	Gompertz mixture	$\lambda = 0.1, \alpha = 5$	$\lambda = 0.1, \alpha = 4$ (50%) $\lambda = 0.1, \alpha = 10$ (50%)
	9	Weibull mixture	$\lambda = 10, p = 10$ (70%) $\lambda = 1, p = 2$ (30%)	$\lambda = 10, p = 20$ (30%) $\lambda = 1, p = 2$ (70%)

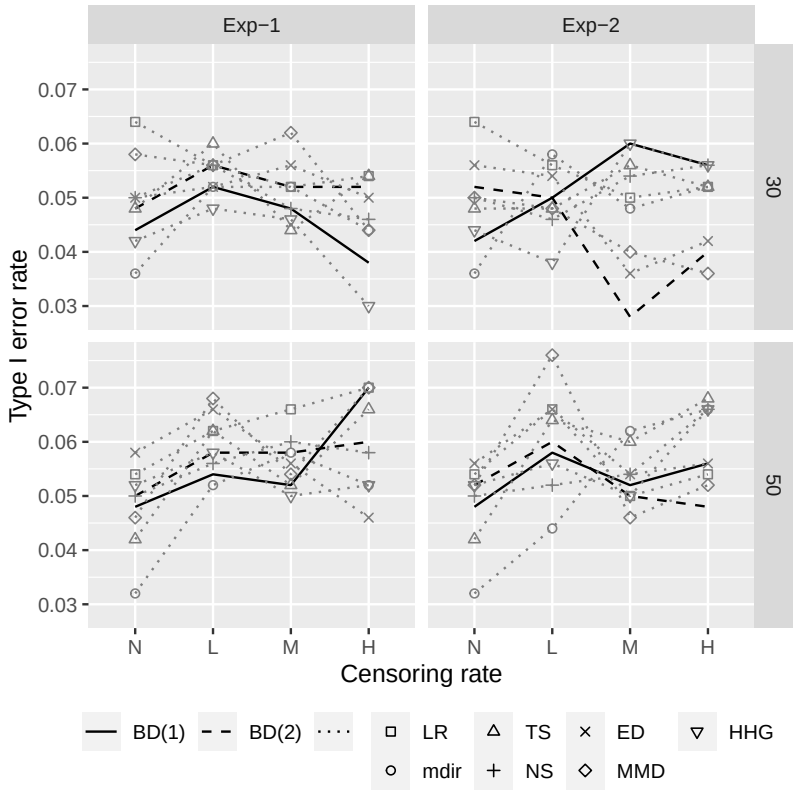

Figure 3 Type-I error rates of our proposed method using two statistics BD(1) and BD(2) in experiments 1 and 2 with the sample sizes $n_1 = n_2 = 30$ and 50 at the nominal significance level $\alpha = 0.05$, with comparison to those of Logrank (LR), two-stage (TS), Neyman smoothing (NS), mdir, energy distance (ED), MMD and HHG methods.

thus we only present those with Gaussian kernel. The other methods were accomplished using the corresponding R packages listed by Dormuth et al.[18].

It is shown by experiments 1 and 2 in Figure 3, all methods can control the type-I error rates around the nominal level of 0.05 in different scenarios with varying sample sizes and censoring rates. Figure 4 presents the testing powers of all methods for diverse alternative hypotheses. In each experiment, their powers increase as the sample size becomes larger, and decrease as the censoring rate goes higher. For experiment 3 where the hazard functions of two groups are proportional and two survival curves

Table 2 The average and standard deviation in the parentheses of the estimated survival Ball divergence BD(1) or BD(2) in 500 repetitions of experiments 1-9 with varying sample sizes and censoring levels.

Exp	$n_1 = n_2$	BD(1)			BD(2)			$D(F_1, F_2)$
		Censoring level			Censoring level			
		Low	Median	High	Low	Median	High	
1	30	.028(.02)	.04(.026)	.054(.034)	.031(.021)	.047(.031)	.082(.057)	0
	50	.018(.012)	.023(.016)	.032(.02)	.018(.013)	.026(.018)	.041(.029)	
	100	.008(.006)	.011(.008)	.016(.01)	.008(.006)	.012(.008)	.019(.014)	
2	30	.027(.018)	.031(.021)	.031(.024)	.03(.021)	.056(.038)	.088(.059)	0
	50	.017(.011)	.02(.014)	.02(.014)	.018(.012)	.033(.025)	.055(.046)	
	100	.008(.006)	.01(.007)	.011(.009)	.008(.006)	.016(.011)	.027(.021)	
3	30	.065(.039)	.075(.044)	.091(.06)	.071(.043)	.087(.055)	.152(.103)	.033
	50	.054(.03)	.058(.035)	.069(.041)	.057(.032)	.064(.038)	.104(.066)	
	100	.044(.02)	.046(.022)	.053(.028)	.045(.02)	.048(.023)	.071(.041)	
4	30	.07(.041)	.078(.046)	.089(.06)	.075(.045)	.09(.055)	.175(.112)	.041
	50	.056(.032)	.061(.035)	.068(.041)	.059(.033)	.066(.039)	.119(.07)	
	100	.047(.021)	.049(.022)	.051(.026)	.048(.021)	.051(.022)	.084(.044)	
5	30	.094(.042)	.099(.047)	.089(.047)	.1(.045)	.108(.055)	.109(.065)	.072
	50	.084(.031)	.087(.036)	.08(.035)	.087(.033)	.09(.039)	.074(.045)	
	100	.077(.023)	.077(.026)	.07(.026)	.078(.023)	.076(.027)	.054(.03)	
6	30	.079(.038)	.08(.04)	.072(.038)	.083(.04)	.089(.051)	.12(.072)	.054
	50	.067(.028)	.068(.031)	.059(.029)	.07(.029)	.066(.033)	.076(.045)	
	100	.06(.02)	.059(.022)	.049(.021)	.061(.021)	.053(.023)	.053(.028)	
7	30	.073(.039)	.081(.045)	.081(.049)	.077(.041)	.09(.051)	.146(.094)	.044
	50	.06(.029)	.064(.032)	.061(.033)	.061(.03)	.067(.034)	.095(.056)	
	100	.051(.02)	.053(.021)	.048(.023)	.052(.02)	.054(.021)	.065(.034)	
8	30	.087(.044)	.095(.051)	.088(.056)	.093(.047)	.115(.062)	.211(.12)	.057
	50	.074(.033)	.077(.036)	.067(.036)	.076(.035)	.085(.041)	.15(.084)	
	100	.066(.024)	.068(.027)	.057(.025)	.067(.024)	.071(.029)	.118(.06)	
9	30	.078(.045)	.082(.048)	.09(.057)	.083(.048)	.095(.055)	.126(.078)	.056
	50	.065(.032)	.069(.037)	.07(.041)	.067(.034)	.075(.041)	.086(.048)	
	100	.058(.023)	.059(.026)	.058(.029)	.059(.023)	.062(.027)	.066(.034)	

are non-crossing, the LR test achieves the highest power. However, for experiment 4 when the hazards are not proportional though two survival curves are still non-crossing, ED outperforms the LR test. The other methods including ours, provide comparable but lower powers than ED and LR tests in these two experiments. It is worth noting that ED with Euclidean distance degenerates to quantify the mean difference between two probability measures, so that it is sensitive to the location shift, similar to the classical LR test, whereas the others that were designed for the crossing alternatives are less sensitive than them. Although LR test and ED perform better in experiments 3 and 4, their powers greatly drop in experiments 5-9 when the survival curves are crossing, especially the LR test, meanwhile, the other methods provide higher powers. When the data of two groups are generated from the simple parametric models with both location and scale shifts as in experiments 5-7, mdir and NS show the highest powers, followed by HHG, MMD, and our method. However, when the underlying model becomes more complex such as the mixture model in experiment 8, HHG, MMD, and our method present very similar results to that of mdir and NS, and higher powers than TS, ED, and LR test. In particular, for experiment 9 where there is only a scale shift, HHG, MMD, and our method surpass all the others and result in significantly higher powers. It is shown that HHG, MMD, and our method are more sensitive to the scale shift than the other compared methods. The nonparametric nature of HHG, MMD, and our method warrants their efficiency for general alternatives. The extension for survival analysis aimed to quantify the divergence between two probability measures with the right-censored data, no matter whether it is in location or scale. However, the other typically

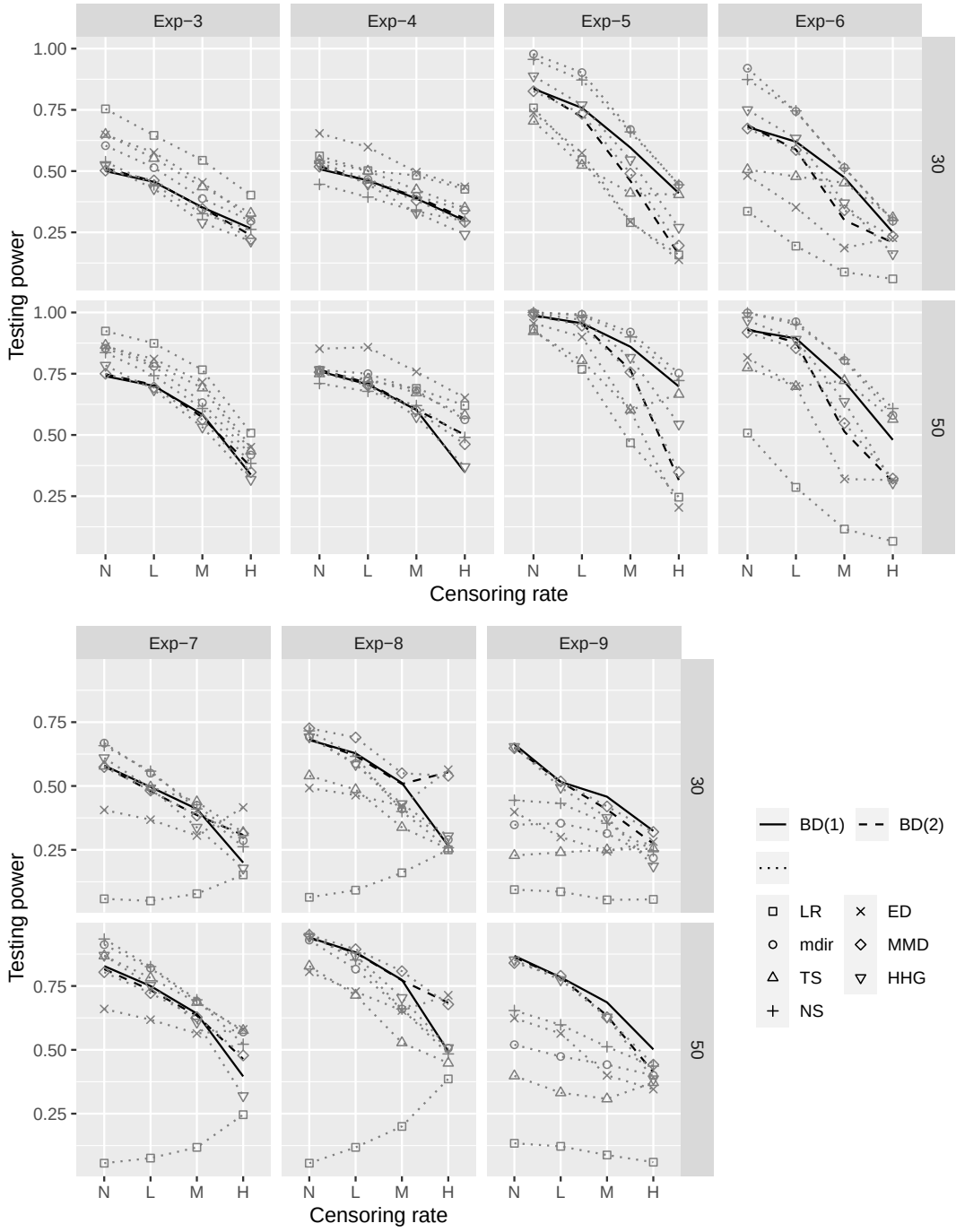

Figure 4 Testing powers of our proposed method using two statistics BD(1) and BD(2) in experiments 3-8 with the sample sizes $n_1 = n_2 = 30$ and 50 at the nominal significance level $\alpha = 0.05$, with comparison to those of Logrank (LR), two-stage (TS), Neyman smoothing (NS), mdir, energy distance (ED), MMD and HHG methods.

developed methods try to estimate the survival curves or hazard functions first, and then use them to measure the difference between the probability measures, which lose power for the estimation.

The consistency result of the Ball divergence estimator does not depend on the sample size ratio, ensuring its testing power in imbalanced samples[22]. In order to further investigate the performance of our proposed method, we set $n_1 = 30$, $n_2/n_1 = 4$ or 16 in experiments 3-9, and present the powers of all aforementioned methods in Figure 5. For experiments 3-8, the results remain in a similar pattern to those of $n_2/n_1 = 1$ in Figure 4, while for experiment 9, our method reaches the highest powers, presenting a

significant advantage over HHG and MMD. The nonparametric statistical divergences ED, MMD, and BD are all kernel-based approaches[22], and BD is associated with HHG with a proper weight[30]. The different definitions of these divergences account for their varied performance in survival analysis. In summary, for balanced samples, the kernel-based approaches perform similarly and are superior to the typical methods designed for crossing in identifying the scale shift, while for imbalanced samples, our method outperforms both the typical competitors and other kernel-based approaches.

The results of our methods using BD(1) and BD(2) are both illustrated in Figure 4 and 5. When the censoring rate is low, their powers are very close. However, as the censoring rate increases, the testing power from BD(2) drops lower than that from BD(1), which can be clearly seen in Figure 5 with a larger sample size than those in Figure 4. It is noteworthy that the observed higher power of BD(2) than BD(1) at the high level of censoring in Figure 4, even the elevated tail in experiment 8 when $n_1 = n_2 = 30$, does not necessarily indicate the superior performance of BD(2) compared to BD(1). According to the estimation results in Table 2, BD(2) may present nonnegligible positive bias at the high level of censoring, in particular with a small sample size, which makes it very likely to be rejected. Compared to BD(2), BD(1) exhibits greater resilience to the high level of censoring. Therefore, our method using BD(1) is more recommended for use in practice. Note that MMD was adapted in a similar manner as BD(2) to handle censoring, resulting in a similar performance to BD(2) in most simulation scenarios but with lower power for imbalanced samples.

4 | REAL DATASETS

The crossing survival curves are very commonly seen in real applications. Although it was alarmed early that LR test may lose power without the proportional hazards assumption, it is still popularly used in clinical research, even for crossing survival curves[31]. To convey the state-of-the-art statistical tests to the community and evaluate their performance to help the choice in practice, Dormuth et al.[18] collected 18 datasets with crossing survival curves from the most recent clinical oncology publications on PubMed. With a significance level of 0.05, mdir rejects the most number of null hypotheses of equal survival in two arms of these datasets, among 11 methods being considered by Dormuth et al.[18]. We applied the proposed and aforementioned peer methods to the analysis of these datasets. The sample sizes and the censoring rates in both groups of each dataset, and their p-values are listed in Table 3. There are 7 datasets, presented in the first 7 rows, where all methods provide nonsignificant p-values that are larger than 0.1, due to limited sample sizes, high censoring rates, or trivial differences between two groups. In the following two datasets, all methods except LR reject the null hypothesis at the significance level of 0.05, indicating the ineffectiveness of LR test in testing crossing survival curves. In the rest 9 datasets, NS and mdir reject the null hypothesis in 6 cases with the most significant p-values. Although BD(1) only reports p-values that are less than 0.05 in 3 datasets, there are two more p-values of 0.06 which are very close to the significance level. Meanwhile, BD(2) identifies the survival inequality in 5 datasets. Besides, MMD rejects the null hypothesis in 5 cases; HHG, TS, and ED all succeed in 4 ones, and LR in only one dataset with one more p-value of 0.06. In simulations, it is indicated that our method shows an advantage in complex distributions, and is especially sensitive to the scale shift. However, as stated by Dormuth et al.[18], they selected the datasets with criteria that there are only one or two crossings, so that the underlying distribution may not be much complicated. Moreover, a location shift is more expected than a scale shift in clinical studies. Our method rejected the null hypothesis in as comparably many datasets as NS and mdir, demonstrating its power in detecting the survival difference for general crossing survival curves.

When the crossing pattern becomes more complicated, there usually involves a scale shift. The R package `KMsurv` collected a series of survival data sets from the book of Klein and Moeschberger[32], one of which is from a randomized clinical trial to investigate the therapeutic effectiveness of an experimental treatment using the combination of AZT, zalcitabine and saquinavir for HIV patients, compared to the standard treatment using both AZT and zalcitabine. There included 34 patients, being assigned at 1:1 ratio into two groups for different treatments. The time to CD4 counts reaching a specified level after drug administration is recorded for each patient, with the censoring rate of 17.6% and 23.5% respectively in the standard and experimental treatment group. We present the Kaplan-Meier estimators of two survival curves under different treatments in Figure 6. It is seen that, after two crossings at the beginning, those two curves depart but approach to each other again in the median survival stage. Thereafter, they go separate and then get close again at the end. Note that it is impossible to know if there is a location or scale shift by the observation of survival curves. The R package `survPresmooth`[33] implements nonparametric presmoothed estimators of the main functions in survival analysis, including the survival, density, hazard, and cumulative hazard functions. We plot the estimated density functions of this data set on the right side of Figure 6. It is obviously seen that the patients in the two groups have distinct survival patterns. The p-values for equality testing using different methods are presented in the last row of Table 3.

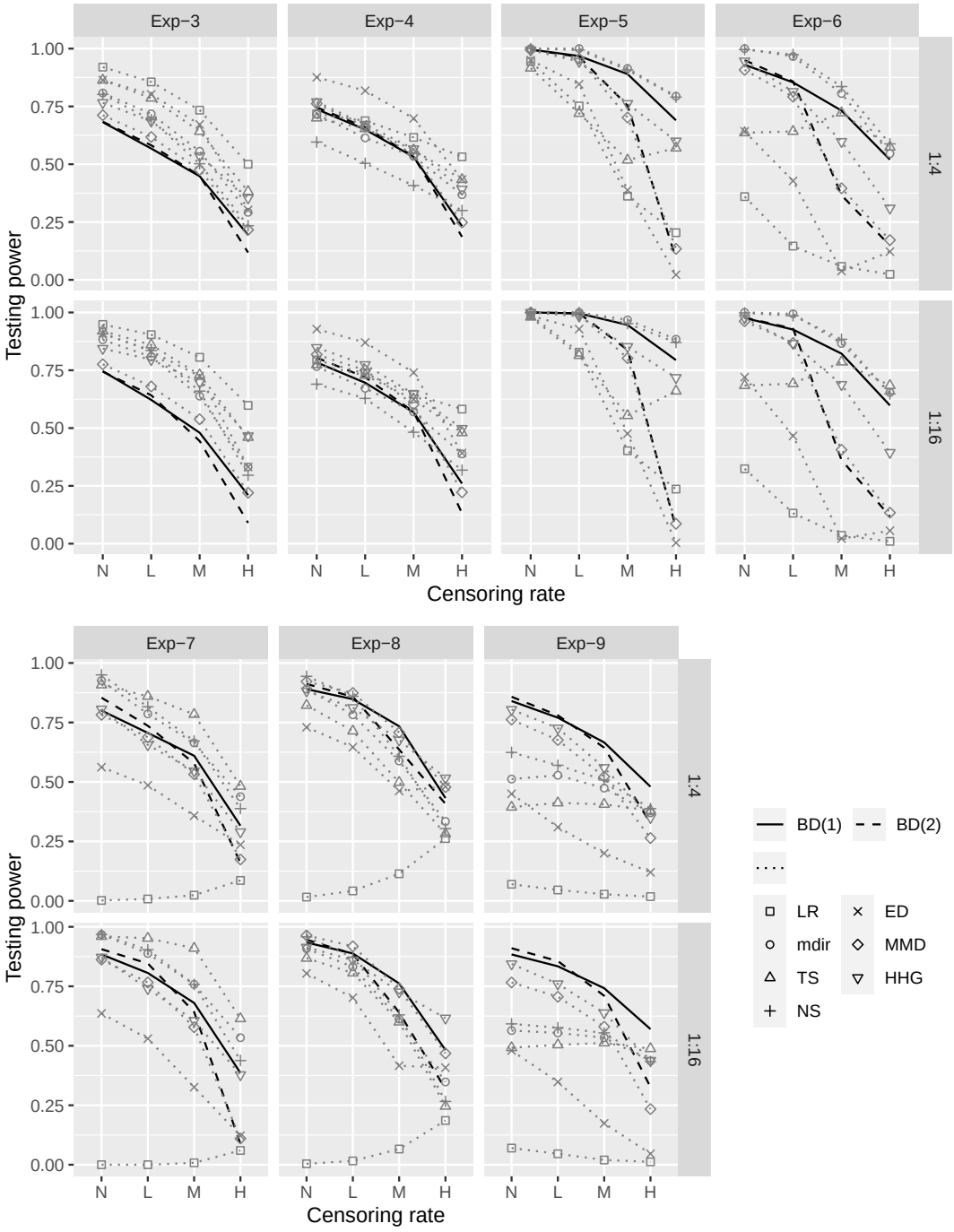

Figure 5 Testing powers of different methods in experiments 3-8 with the sample sizes $n_1 = 30$, $n_2/n_1 = 4$ and 16, at the nominal significance level $\alpha = 0.05$, with comparison to those of Logrank (LR), two-stage (TW), Neyman smoothing (NS), mdir, energy distance (ED), MMD and HHG methods.

Only MMD and our methods reject the null hypothesis at the significance level of 0.05, besides ED reports a p-value of 0.06. It is illustrated that these nonparametric statistical methods are more powerful in identifying the survival difference between complex underlying distributions.

Note that due to the disease or population heterogeneity, it is very common in reality that the survival distribution under the same treatment has multiple modes and presents as a mixture, where a scale shift usually presents between groups. Our method and MMD show an advantage over NS and mdir in identifying this change in survival. However, whether or not there is a scale

Table 3 The p-values from our proposed method and compared methods in real data analysis, where n_k is the sample size and R_k indicates the censoring rate of the kth group, for $k = 1, 2$.

Dataset	n_1	R_1	n_2	R_2	LR	ED	TS	NS	mdir	HHG	MMD	BD(1)	BD(2)
Cortes[1]	19	.16	36	.19	.19	.83	.84	.24	.44	.43	.93	.50	.80
Godfrey[2]	192	.95	190	.96	.49	.54	.90	.50	.77	.28	.39	.71	.59
Golan[3]	62	.52	92	.57	.61	.89	.23	.61	.62	.59	.90	.76	.97
Hammel[4]	57	.60	89	.63	.22	.28	.25	.26	.14	.23	.28	.50	.14
Kotani[5]	66	.17	60	.43	.14	.50	.51	.17	.27	.45	.55	.54	.32
Mukai[6]	14	.36	32	.56	.16	.77	.54	.21	.36	.18	.68	.21	.79
Toxopeus[7]	173	.61	208	.53	.91	.16	.15	.90	.10	.35	.12	.32	.13
Bellmunt[8]	272	.17	270	.19	.49	.00	.03	.00	.00	.00	.00	.00	.00
Fradet[9]	272	.09	270	.14	.40	.00	.03	.00	.00	.00	.00	.00	.00
Bang[10]	48	.27	50	.18	.37	*.02*	*.05*	*.03*	*.05*	.13	*.04*	.06	*.03*
Becker[11]	25	.00	152	.11	.09	.22	.27	*.01*	*.02*	.16	.42	.19	.28
Ferris[12]	121	.15	240	.20	.33	.08	*.04*	*.00*	*.02*	*.01*	*.01*	*.01*	*.01*
Jones[13]	37	.38	94	.34	.17	*.01*	*.04*	*.01*	*.02*	*.05*	*.00*	.12	*.01*
Jones20[14]	71	.56	69	.70	*.05*	*.00*	.33	.08	.12	.12	*.00*	*.00*	*.00*
Kreuzer[15]	37	.46	26	.50	.53	.15	.13	.08	.10	.34	.24	.55	.21
Lu[16]	139	.43	141	.38	.06	*.00*	*.04*	*.01*	*.00*	*.05*	*.03*	.06	*.04*
Malone[17]	215	.82	217	.87	.11	.25	.66	.13	.30	.09	.34	.16	.32
Motzer[18]	411	.22	410	.21	.07	.66	.29	*.00*	*.02*	*.02*	.47	*.03*	.20
Drughiv	17	.18	17	.24	.15	.06	.36	.18	.14	.11	*.05*	*.05*	*.03*

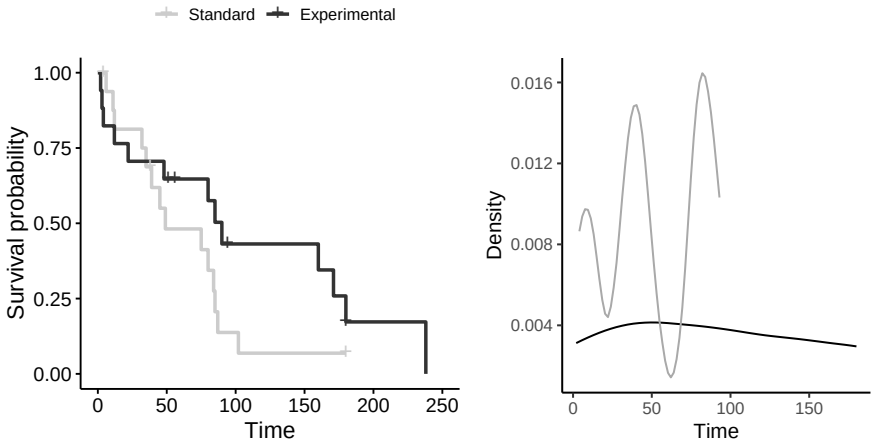

Figure 6 The Kaplan-Meier estimators of the survival curves (left) and their density estimators (right) of HIV patients in two groups with the standard or experimental treatment.

shift can not be certainly known by the inspection of survival curves or even density functions, except for the obvious patterns observed in the HIV drug data set. Actually, both location and scale shifts occur simultaneously for common cases. If only the location shift is of interest, we suggest NS and mdir, while if a scale shift is expected, then our method is recommended. We present two test statistics BD(1) and BD(2). They make the same decisions for most hypotheses, but may choose different ones such as in the 13th and 18th data sets in Table 3. When they are discordant, BD(2) gives the same results as MMD, due to the fact that both of them utilized the Kaplan-Meier integral to deal with the censoring, whereas BD(1) incorporated the inverse probability weights.

5 | DISCUSSION

In literature, for non-proportional hazard alternatives, either crossing survival curves or crossing hazard functions were conceived in the construction of test statistics. Although their null hypotheses are the same, these two alternatives are not equivalent but overlapped[7]. Moreover, it is ambiguous to infer crossings by inspection of estimated survival curves or cumulative hazard functions. These facts bring confusion to users regarding the choice of testing methods. Although the omnibus tests attempt against general alternatives, Janssen[34] indicated that there exists no test globally with high power, and any nonparametric test can only outperform in a finite-dimensional subspace. It is essentially meaningful to clarify in which situations the proposed test is more applicable to guide practical use. No matter which type of crossing was presumed in mind, those existing methods reveal the divergence between probability measures. In this paper, we utilized the estimation of Ball divergence from right-censored survival data to propose a testing method for general alternative settings. The test statistics compare the numbers of observations from different groups on a collection of intervals formulated by the sample data, thus they are rank-based and in line with LR test. Our method inherits the advantage of Ball divergence, being sensitive to the scale shift between probability measures.

The proposed test statistics are consistent estimators of Ball divergence defined by Pan et al.[22]. As the domain of survival time being R^+, the ball $B(x, |x-y|)$ reduces to the interval $(s, v]$, where $s = \min(2x-y, y)$, and $v = \max(2x-y, y)$. On the other hand,

$$\begin{aligned}[\gamma_1(x,y) - \gamma_2(x,y)]^2 &= [(1-\gamma_1(x,y)) - (1-\gamma_2(x,y))]^2 \\ &= [P(T_1 \notin \bar{B}(x, |x-y|)) - P(T_2 \notin \bar{B}(x, |x-y|))]^2 \\ &= [(P(T_1 \le s) - P(T_2 \le s)) + (P(T_1 > v) - P(T_2 > v))]^2.\end{aligned}$$

The ball divergence can also be regarded as a measure of probability distribution difference on both tails out of the interval $(s, v]$. When there is only a location shift between T_1 and T_2, $P(T_1 \le s) - P(T_2 \le s)$ and $P(T_1 > v) - P(T_2 > v)$ have reverse signs. Meanwhile, when T_1 differs from T_2 in terms of scale rather than location, they have concordant signs. The accrued divergence enhances the significance of testing statistics, and therefore improves their power performance in different scenarios. The crossing is induced by a scale shift, where the location shift usually occurs as well, so that some omnibus tests, such as NS and mdir, provide comparable results to our method in many situations, as illustrated by simulations and real data analysis. For a robust powerful test, the Cauchy combination test can be further implemented to integrate the p-values from these competitive methods[35]. Nevertheless, their extensions may be challenging. In contrast, the Ball divergence was originally developed for multivariate analysis, so our method can be naturally extended to the cases of multiple time-to-event endpoints. Moreover, the Ball divergence defines a metric on the difference between two probability measures. It can be adapted with conditional probability to control potential covariates in the comparison of survival. We leave these studies for future work.

The Ball divergence provides a framework to measure the discrepancy between two probability measures. It integrates the square of their difference over a ball Borel σ-algebra. For multi-dimensional space, within or out of a ball is a natural choice of the σ-algebra. However, for one-dimensional space such as R^+, it may be either an interval, or both left and right tails out of the interval, or even the only left or right tail, i.e., for $x, y \in R^+$, $\gamma_k(x, y)$ can be either $P(s < T_k \le v)$, $P(T_k \le s) + P(T_k > v)$, $P(T_k \le s)$, or $P(T_k > v)$. Different settings of $\gamma_k(x, y)$ correspond to varying choices of σ-algebra, leading to different definitions of the statistical divergence. Considering the right-censoring of survival data, $\gamma_k(x, y) = P(T_k > v) = P(T_k > \max(x, y))$ may be an alternative to simplify the notations and reduce the computational complexity. With the simplified version, more complicated statistical inferences can be more easily developed, such as the conditional divergence and independence test between the survival time and potential covariates. Its performance will be investigated in future work.

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and the referees for their valuable comments that improved the presentation of the paper. You's research is partially supported by National Natural Science Foundation of China (12126610), and Guangdong Basic and Applied Basic Research Foundation (2023A1515012254). Wang's research is partially supported by National Natural Science Foundation of China (12231017, 72171216, 71921001, 71991474), and the Science and Technology Program of Guangzhou, China (202002030129).

DATA AVAILABILITY STATEMENT

The 18 reconstructed datasets that support the findings of this study were kindly shared from Dr. Ina Dormuth, TU Dortmund University. Our proposed method was implemented as R package `SurvBD`, and publicly available at https://github.com/scrcss/SurvBD.

References

1. Mantel N, others . Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966; 50(3): 163–170.
2. Gill R. Censoring and stochastic integrals. *Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum* 1980.
3. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* 1989: 497–507.
4. Shen Y, Fleming TR. Weighted mean survival test statistics: a class of distance tests for censored survival data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997; 59(1): 269–280.
5. Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2004; 46(5): 489–496.
6. Lee SH. On the versatility of the combination of the weighted log-rank statistics. *Computational statistics & data analysis* 2007; 51(12): 6557–6564.
7. Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. *Statistics in medicine* 2007; 26(2): 375–391.
8. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008; 70(1): 191–208.
9. Kraus D. Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. *Journal of statistical planning and inference* 2009; 139(10): 3559–3569.
10. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. John Wiley & Sons . 2011.
11. Ditzhaus M, Friedrich S. More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation* 2020; 90(12): 2209–2227.
12. Liu A, Li Q, Liu C, Yu K, Yu KF. A rank-based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association* 2010; 105(490): 578–587.
13. Schumacher M. Two-Sample tests of Cramér–von Mises-and Kolmogorov–Smirnov-Type for randomly censored data. *International Statistical Review/Revue Internationale de Statistique* 1984: 263–281.
14. Shen Y, Cai J. Maximum of the weighted Kaplan-Meier tests with application to cancer prevention and screening trials. *Biometrics* 2001; 57(3): 837–843.
15. Lin X, Xu Q. A new method for the comparison of survival distributions. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 2010; 9(1): 67–76.
16. Liu T, Ditzhaus M, Xu J. A resampling-based test for two crossing survival curves. *Pharmaceutical Statistics* 2020; 19(4): 399–409.
17. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One* 2015; 10(1): e0116774.

18. Dormuth I, Liu T, Xu J, Yu M, Pauly M, Ditzhaus M. Which test for crossing survival curves? A user's guideline. *BMC medical research methodology* 2022; 22(1): 1–7.

19. Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference* 2013; 143(8): 1249–1272.

20. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A kernel method for the two-sample-problem. *Advances in neural information processing systems* 2006; 19: 513–520.

21. Heller R, Heller Y, Gorfine M. A consistent multivariate test of association based on ranks of distances. *Biometrika* 2013; 100(2): 503–510.

22. Pan W, Tian Y, Wang X, Zhang H. Ball divergence: nonparametric two sample test. *Annals of statistics* 2018; 46(3): 1109.

23. Matabuena M, Padilla OHM. Energy distance and kernel mean embeddings for two-sample survival testing. *arXiv preprint arXiv:1912.04160* 2019.

24. Gorfine M, Schlesinger M, Hsu L. K-sample omnibus non-proportional hazards tests based on right-censored data. *Statistical Methods in Medical Research* 2020; 29(10): 2830–2850.

25. Stute W. Kaplan–Meier integrals. *Handbook of Statistics* 2003; 23: 87–104.

26. Fernández T, Rivera N. Kaplan-Meier V-and U-statistics. *Electronic Journal of Statistics* 2020; 14(1): 1872–1916.

27. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall, New York . 1993.

28. Zhu J, Pan W, Zheng W, Wang X. Ball: An R package for detecting distribution difference and association in metric spaces. *Journal of Statistical Software* 2021; 97(6): 1–31.

29. Lee AJ. *U-statistics: Theory and Practice*. Routledge . 2019.

30. Pan W, Wang X, Zhang H, Zhu H, Zhu J. Ball covariance: A generic measure of dependence in Banach space. *Journal of the American Statistical Association* 2020; 115(529): 307-317.

31. Kristiansen I. PRM39 survival curve convergences and crossing: a threat to validity of meta-analysis?. *Value in health* 2012; 15(7): A652.

32. Klein JP, Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. Springer . 2003.

33. López-de-Ullibarri I, Jácome MA. survPresmooth: an R package for presmoothed estimation in survival analysis. *Journal of Statistical Software* 2013; 54: 1–26.

34. Janssen A. Global power functions of goodness of fit tests. *The Annals of Statistics* 2000; 28(1): 239–253.

35. Long M, Li Z, Zhang W, Li Q. The Cauchy combination test under arbitrary dependence structures. *The American Statistician* 2023; 77(2): 134–142.

36. Stute W, Wang JL. The strong law under random censorship. *The Annals of statistics* 1993: 1591–1607.

37. Stute W, Wang JL. Multi-sample U-statistics for censored data. *Scandinavian journal of statistics* 1993: 369–374.

APPENDIX

Proof of Lemma 1. Under the assumption of independent censoring, as $n_k \longrightarrow \infty$,

$$H_k(t) = \frac{1}{n_k}\sum_{i=1}^{n_k} I(X_{k,i} > t) \xrightarrow{a.s.} P(X_{k,i} > t) = P(T_{k,i} > t)P(C_{k,i} > t) = P(T_k > t)\bar{G}_k(t).$$

The Kaplan-Meier estimator $\hat{\bar{G}}_k(t) \xrightarrow{a.s.} \bar{G}_k(t)$ for the continuous distribution $G_k(t)$[36]. Therefore, for any $t > 0$ such that $\bar{G}_k(t) > 0$,

$$\frac{H_k(t)}{\hat{\bar{G}}_k(t)} \xrightarrow{a.s.} \frac{P(T_k > t)\bar{G}_k(t)}{\bar{G}_k(t)} = P(T_k > t).$$

$$\begin{aligned}\Gamma_k(x,y) = \frac{H_k(s)}{\hat{\bar{G}}_k(s)} - \frac{H_k(v)}{\hat{\bar{G}}_k(v)} &\xrightarrow{a.s.} P(T_k > s) - P(T_k > v)\\ &= P(s < T_k \le v) = \gamma_k(x,y).\end{aligned}$$

□

Proof of Theorem 1.

$$\begin{aligned}\mathcal{D}_k =& \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\frac{\delta_{k,i}\delta_{k,j}}{\hat{\bar{G}}_k(X_{k,i})\hat{\bar{G}}_k(X_{k,j})}\Big[\Gamma_1(X_{k,i},X_{k,j}) - \Gamma_2(X_{k,i},X_{k,j})\Big]^2\\ =& \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\frac{\delta_{k,i}\delta_{k,j}}{\hat{\bar{G}}_k(X_{k,i})\hat{\bar{G}}_k(X_{k,j})}\Big[\big(\Gamma_1(X_{k,i},X_{k,j}) - \gamma_1(X_{k,i},X_{k,j})\big) - \big(\Gamma_2(X_{k,i},X_{k,j}) - \gamma_2(X_{k,i},X_{k,j})\big)\\ &+ \big(\gamma_1(X_{k,i},X_{k,j}) - \gamma_2(X_{k,i},X_{k,j})\big)\Big]^2\end{aligned}$$

From Lemma 1, as $n \longrightarrow \infty$,

$$\begin{aligned}\mathcal{D}_k &\xrightarrow{a.s.} E\Big[\frac{\delta_{k,1}\delta_{k,2}}{\bar{G}_k(X_{k,1})\bar{G}_k(X_{k,2})}\big(\gamma_1(X_{k,1},X_{k,2}) - \gamma_2(X_{k,1},X_{k,2})\big)^2\Big]\\ &= E_{T_{k,1},T_{k,2}}\left\{E\Big[\frac{\delta_{k,1}\delta_{k,2}}{\bar{G}_k(X_{k,1})\bar{G}_k(X_{k,2})}\big(\gamma_1(X_{k,1},X_{k,2}) - \gamma_2(X_{k,1},X_{k,2})\big)^2 | T_{k,1},T_{k,2}\Big]\right\}\\ &= E_{T_{k,1},T_{k,2}}\Big[\gamma_1(T_{k,1},T_{k,2}) - \gamma_2(T_{k,1},T_{k,2})\Big]^2\\ &= D_k.\end{aligned}$$

Therefore, $\mathcal{D}_1 + \mathcal{D}_2 \xrightarrow{a.s.} D(F_1,F_2)$ is proved. □

Proof of Theorem 2. Denoted by $1 - \widehat{F}_{k,n_k}$ the Kaplan-Meier estimator for $1 - F_k$ and $\varphi(t)$ a Borel-measurable function on the real line such that $\int |\varphi| dF_k < \infty$, it is shown that $\int \varphi(t)\widehat{F}_{k,n_k}(dt)$ converges almost surely to $\int \varphi(t)F_k(dt)$ for the continuous distribution $F_k(t)$[36]. Furthermore, the strong consistency property still holds for two-sample U-statistic with kernel function $h(x,y)$ such that $\int h(x,y)F_1(dx)F_2(dy) < \infty$[37], i.e.,

$$\int h(x,y)\widehat{F}_{1,n_1}(dx)\widehat{F}_{2,n_2}(dy) \xrightarrow{a.s.} \int h(x,y)F_1(dx)F_2(dy).$$

Let $\varphi(t) = \eta(X_{k(i)}, X_{k(j)}, t)$, then we have $\int \eta(X_{k(i)}, X_{k(j)}, t)F_k(dt) \xrightarrow{a.s.} \int \eta(X_{k(i)}, X_{k(j)}, t)F_k(t) = \gamma_k(X_{k(i)}, X_{k(j)})$.

$$
\begin{aligned}
\mathcal{D}_k^A &= \sum_{i=1}^{n_k}\sum_{j=1}^{n_k} W_{k(i)}W_{k(j)} \left[\sum_{u=1}^{n_1} W_{1(u)}\eta(X_{k(i)}, X_{k(j)}, X_{1(u)}) - \sum_{v=1}^{n_2} W_{2(v)}\eta(X_{k(i)}, X_{k(j)}, X_{2(v)})\right]^2 \\
&= \sum_{i=1}^{n_k}\sum_{j=1}^{n_k} W_{k(i)}W_{k(j)} \left[\Big(\sum_{u=1}^{n_1} W_{1(u)}\eta(X_{k(i)}, X_{k(j)}, X_{1(u)}) - \int \eta(X_{k(i)}, X_{k(j)}, t)F_1(dt)\Big)\right. \\
&\quad \left. - \Big(\sum_{v=1}^{n_2} W_{2(v)}\eta(X_{k(i)}, X_{k(j)}, X_{2(v)}) - \int \eta(X_{k(i)}, X_{k(j)}, t)F_2(dt)\Big) + \Big(\gamma_1(X_{k(i)}, X_{k(j)}) - \gamma_2(X_{k(i)}, X_{k(j)})\Big)\right]^2 \\
&\xrightarrow{a.s.} \int \Big(\gamma_1(x, y) - \gamma_2(x, y)\Big)^2 F_k(dx)F_k(dy) = D_k,
\end{aligned}
$$

Therefore, $\mathcal{D}_1^A + \mathcal{D}_2^A \xrightarrow{a.s.} D(F_1, F_2)$ is proved. □

How to cite this article: You N., He X., Dai H. and Wang X. (2022), Ball divergence for the equality test of crossing survival curves, *Statistics in Medicine*, *2017;00:1–6*.