Modifying Weighted Kaplan-Meier Test for Two-Sample Survival Comparison

S. Lee and E. Lee

Abstract

This paper presents an approach to improving the weighted Kaplan-Meier test statistics in order to make it a more useful tool for a long-term comparison of two underlying survival distributions in the presence of right-censored data. The procedures are based on the use of some weight function that involves the percentage of censored data as a component. Some versatile procedures for the alternative, not pre-specified, are also discussed. Numerical simulations are conducted to investigate the performance of the proposed procedures. For illustration, the procedures are applied to real-world data in clinical trials, where patients with tongue cancer are divided into two groups according to tumor DNA.

Index Terms

Censored data, Kaplan-Meier estimator, Log-rank test, Survival distribution, Weighted Kaplan-Meier statistics.

I. Introduction

In analyzing survival data, especially in clinical trials with two different treatment groups, comparing survival differences between the two groups is of interest. Comparing the distributions will allow initially indistinguishable difference to emerge, and give insight into an accurate description of the relationship between them. Rank-based tests are often utilized for checking the equality of two underlying survival distributions with arbitrarily right-censored data. Examples include the log-rank (Mantel [13], Cox [4]) and the generalized Wilcoxon (Gehan [6], Peto and Peto [16]) statistics. As noted in Fleming and Harrington [5, p.266], these statistics use integrated weighted differences in hazard functions, thus they are sensitive to the ordered hazards alternative. In particular, the log-rank test is sensitive to detect alternatives and has optimum power when the assumption of proportional hazard rates is held. However, the sensitivity is not necessarily achieved for the direct comparison of two survival curves. So the rank-based tests, such as the log-rank test, may not be sensitive to the magnitude of survival differences, leading to erroneous conclusions. To address this problem, Pepe and Fleming [14] proposed the weighted Kaplan-Meier (WKM) statistic which is based directly on estimated survival functions (Kaplan and Meier [8]), rather than based on ranks, in the two-sample problem. By directly comparing the survival functions, the WKM statistic has more prognostic ability in detecting survival differences between two groups than the rank-based statistics. The literature on the WKM statistic includes Shen and Cai [18], Logan et al. [12] and Lee et al. [11], among others. Shen and Cai [18] proposed a class of maximum WKM test statistics to construct versatile test procedures. Logan et al. [12] considered the WKM statistic in a hypothesis formation to compare late differences of survival curves, and Lee et al. [11] proposed the use of a process to improve the power property of the WKM statistics.

This paper proposes a modification of the WKM statistic that leads to more powerful tests in the long-term survival comparison task. Motivated by Buyske et al. [3] that develops G^ρ statistics for low event rate survival data, the new test uses some weight function adjusted by a censoring level, facilitating a range of alternatives of survival difference with right-censored data. For example, the WKM test to assess the equality of two survival curves would lose power when the survival curves cross. The new weight function, on the other hand, is shown to be a useful tool to deal with such a non-proportional hazards circumstance, modifying the effect of the right-censoring that may cause the statistical power to decrease. Some versatile tests with the proposed statistic are also developed, which will be effective when there is no prior knowledge of the potential behavior of survival differences over a period of time. Numerical simulations demonstrate enhanced properties of the new test compared to the WKM test, showing that the test detects a broad range of alternatives of survival differences.

This paper is organized as follows: In Section 2, the WKM statistic is briefly reviewed, the new method is described, and some versatile test procedures are presented. Section 3 discusses the finite-sample properties of the new test in terms of numerical simulations, and gives an application to real data for illustration. Section 4 concludes this paper.

ISSN: 2736-5484
DOI: http://dx.doi.org/10.24018/ejmath.2022.3.1.93
Published on February 20, 2022.
S. Lee, Department of Mathematics, Illinois Wesleyan University, Bloomington, IL 61701 USA (corresponding e-mail: slee2@iwu.edu).
E. Lee, Department of Mathematics and Computational Sciences, Millikin University, Decatur, IL 62522 USA (e-mail: elee@millikin.edu).

II. METHODS

A. Notation

For two-sample censored data of sizes n_1 and n_2, suppose that the survival time, also referred to as the failure/death time, is continuous and independent of the censoring time in each group. Let T_{ij}, $i = 1, 2,\ j = 1, \ldots, n_i$ be independent, positive random variables denoting the survival time, with distribution functions $F_i(t) = P(T_{ij} \leq t)$. Let C_{ij}, $i = 1, 2$, $j = 1, \ldots, n_i$, be independent censoring variables, with distribution functions $L_i(t) = P(C_{ij} \leq t)$. The observed data are the pairs (X_{ij}, δ_{ij}), where $X_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$ in which I is the censoring indicator that is 1 if uncensored, i.e., $T_{ij} \leq C_{ij}$, and 0 otherwise. Let $S_i = 1 - F_i = P(T_{ij} > t)$ and $G_i = 1 - L_i = P(C_{ij} > t)$ be survival functions of T_{ij} and C_{ij}, respectively. The counting process approach has been developed into a general framework for analyzing censored survival time data (X_{ij}, δ_{ij}) since it was first introduced by Aalen [1], [2]. Define the counting process by $N_{ij}(t) = I(X_{ij} \leq t, \delta_{ij} = 1)$ and the at-risk process by $Y_{ij}(t) = I(X_{ij} \geq t)$. For $i = 1, 2$, let

$$N_i(t) = \sum_{j=1}^{n_i} N_{ij}(t),\ Y_i(t) = \sum_{j=1}^{n_i} Y_{ij}(t).$$

Note that $N_i(t)$ denotes the number of observed deaths in group i occurred by time t, and $Y_i(t)$ represents the number of individuals at risk (i.e., not dead and not censored) in group i at time $t-$ (i.e., just prior to time t). The product-limit estimator (Kaplan and Meier [8]), also referred to as Kaplan-Meier estimator, of survival in group i is defined as

$$\hat{S}_i(t) = \prod_{s \leq t} \left(1 - \frac{\Delta N_i(s)}{Y_i(s)}\right),$$

where $\Delta N_i(s) = N_i(s) - N_i(s-)$ which denotes the number of failures at time s. The survival distribution in the pooled sample is defined as

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right),$$

where $\Delta N(s) = N(s) - N(s-)$ with $N(s) = \sum_{i=1}^{2} N_i(s)$ and $Y(s) = \sum_{i=1}^{2} Y_i(s)$. The common distribution $F(t)$ can then be estimated by $1 - \hat{S}(t)$.

B. Weighted Kaplan-Meier Test

In studies with right-censored data, the weighted Kaplan-Meier test statistic (Pepe and Fleming [14]) is a valid test of the null hypothesis of the equality of two underlying survival distributions. Let $H_0 : S_1(t) = S_2(t)$ for all t versus $H_1 : S_1(t) \neq S_2(t)$ for some t. Based on the integrated weighted differences of the Kaplan-Meier estimators, the WKM statistic is defined as

$$W_{K_1} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{\tau} K_1(t)\{\hat{S}_1(t) - \hat{S}_2(t)\}dt, \tag{1}$$

where $\tau = \sup\{t : \min(G_1(t), G_2(t)) > 0\}$ the minimum of two groups' largest survival time, $n = n_1 + n_2$, and $K_1(t)$ is a data-dependent weight function to be chosen such that the difference $\hat{S}_1(t) - \hat{S}_2(t)$ is downweighted over later time periods. Pepe and Fleming [14], [15] proposed the following weight fucntion

$$K_1(t) = \frac{\hat{G}_1(t-)\hat{G}_2(t-)}{\hat{p}_1\hat{G}_1(t-) + \hat{p}_2\hat{G}_2(t-)},$$

where $\hat{p}_i = n_i/n$, and $\hat{G}_i$ is the Kaplan-Meier estimator (Kaplan and Meier [8]) of the censoring survival function G_i. Note that the weight function depends only on the censoring distributions, $\hat{G}_1(t-)$ and $\hat{G}_2(t-)$, not the survival distributions. The weight $K_1(t)$ stabilizes the behavior of the statistic (1) in late observation periods by downweighting its variance there when censoring is heavy (Pepe and Fleming [14]). The estimated pooled variance of W_{K_1} is

$$\hat{\sigma}^2_{W_{K_1}} = -\int_0^{\tau} \frac{\{\int_t^{\tau} K_1(t)\hat{S}(u)du\}^2}{\hat{S}(t)\hat{S}(t-)} \frac{\hat{p}_1\hat{G}_1(t-) + \hat{p}_2\hat{G}_2(t-)}{\hat{G}_1(t-)\hat{G}_2(t-)} d\hat{S}(t).$$

Then, under H_0,

$$W_{K_1}/\hat{\sigma}_{W_{K_1}} \xrightarrow{d} N(0, 1),$$

where "$\xrightarrow{d} N(0, 1)$" denotes convergence in distribution to the standard normal distribution with mean 0 and variance 1. An asymptotic level α test is to reject H_0 if

$$|W_{K_1}/\hat{\sigma}_{W_{K_1}}| > z_{\alpha/2}.$$

C. Proposed Test

The WKM statistic directly assesses the difference of two survival functions. Therefore, in testing the equality of the survival distributions, this non-rank-based statistic would be more sensitive to the magnitude of the difference and so it would be more effective for comparing the survival distributions of two groups, when compared to the rank-based statistics, such as the log-rank statistic (Pepe and Fleming [14]). Nevertheless, some problematic issues arise when the WKM statistic is utilized. In order for the heavy censoring to less affect variance, the weight function, $K_1(t)$, in (1) downweights the difference of $\hat{S}_1(t)$ and $\hat{S}_2(t)$ at a later time period during which the unstable behavior of $\hat{S}_1(t) - \hat{S}_2(t)$ could occur without it. However, the weight function only stabilizes the variance of the statistic toward the end of the period (Pepe and Fleming [14], [15], Shen and Cai [18]). Since the WKM statistic is a method to assess the equality of the entire survival curves, it may lead to misleading results if the difference is not always either positive or negative. Several approaches to tackling the problem have been proposed in the literature. Examples include Yang and Fleming [20], Shen and Cai [18], and Lee et al. [11], among others. A simple and reliable way to improve the performance of the WKM statistic to handle the situation would be to make use of a weight function consisting of the Kaplan-Meier estimator $\hat{S}(t)$. We propose the following weighted WKM statistic in which the weight function accounts for the censoring level of data as a component,

$$W_{K_2} = \sqrt{\frac{n_1 n_2}{n}} \int_0^\tau K_2(t)\{\hat{S}_1(t) - \hat{S}_2(t)\}dt, \tag{2}$$

where

$$K_2(t) = \psi(t,a)\frac{\hat{G}_1(t-)\hat{G}_2(t-)}{\hat{p}_1\hat{G}_1(t-) + \hat{p}_2\hat{G}_2(t-)}$$

with

$$\psi(t,a) = \{\hat{S}(t-) - (1-a)\hat{S}(\tau-)\}^\rho$$

in which $(1-a)$ indicates the censoring proportion, i.e., $100 \times a\%$ shows the percentage of the observed survival time. Note that in order to force the weight function, $\hat{S}^\rho(t)$, of the G^ρ statistic to decrease, Buyske et al. [3] proposed the weight function, $\{\hat{S}(t-) - \hat{S}(\tau-)\}^\rho$. By using this weight function which subtracts its value at the end τ from the Kaplan-Meier estimate, the test will better compare the survival curves, especially when the failure rate is rare. However, it might yield some interference that results from the presence of the non-proportional hazards, and thus inaccurate comparisons of survival curves. Heavy censoring would also affect the contrast between the survival curves. Since the heavy censoring makes the survival curves not plateau at a late stage in time, the comparison result may deviate from the actual difference. The weight function, $\{\hat{S}(t-) - \hat{S}(\tau-)\}^\rho$, subtracts the survival at the end point from the Kaplan-Meier estimate at any point t. For heavily censored data, the Kaplan-Meier survival function decreases slowly, not dropping to near-zero toward the end of the observation period. So, the weight function proposed by Buyske et al. [3] may inflate the difference when one of the curves stays close to 1 or the curves cross. The component, $\psi(t,a)$, in (2) is slightly different from this. This component weighs in on the weight function as the censoring proportion increases, adjusting the censoring effect on the weight. This way, the $\psi(t,a)$ better deals with the non-proportional situation when it becomes more difficult to handle by heavy censoring. So, the weight function $K_2(t)$ has a broader range of flexibility when compared to the weight function without that component.

Note that the $\psi(t,a)$ decreases from 1 to

$$\left(\frac{a\hat{S}(\tau-)}{1-(1-a)\hat{S}(\tau-)}\right)^\rho$$

when scaled by $\psi(0,a)$ for $a > 0$, $\rho > 0$ and $\hat{S}(\tau-) < 1$. The ratio of $\psi(t,a)$ and $\psi(t,0)$, i.e., $\psi(t,a)/\psi(t,0)$,

$$\left(\frac{\hat{S}(t-) - (1-a)\hat{S}(\tau-)}{1-(1-a)\hat{S}(\tau-)}\right)^\rho$$

indicates that $\psi(t,a)$ is a decreasing function for $t > 0$ and $\rho = 1$.

It is wise to obtain a simpler integral to deal with than the statistic W_{K_2} in (2) we started with. Using integration by parts, the statistic W_{K_2} can be written

$$\begin{aligned} W_{K_2} &= \sqrt{\frac{n_1 n_2}{n}} \int_0^\tau \{\int_t^\tau K_2(u)du\} d(\hat{S}_1(t) - \hat{S}_2(t)) \\ &= -\sqrt{\frac{n_1 n_2}{n}} \int_0^\tau \{\hat{S}_1(t) - \hat{S}_2(t)\} d(\int_t^\tau K_2(u)du). \end{aligned}$$

Then, by Pepe and Fleming [14] and Shen and Cai [18], we get

$$W_{K_2} \xrightarrow{d} N(0, \sigma^2_{W_{k_2}})$$

where

$$\sigma^2_{W_{k_2}} = -\int_0^\tau \frac{\{\int_t^\tau k_2(u)S(u)du\}^2}{S(t)S(t-)} \frac{p_1G_1(t-)+p_2G_2(t-)}{G_1(t-)G_2(t-)}dS(t),$$

with $k_2(t)$ the deterministic version of $K_2(t)$ such that $\sup_{t\in[0,\tau]}|K_2(t)-k_2(t)|\to 0$ in probability, and $p_i = \lim_{n\to\infty} n_i/n$, $i=1,2$. A consistent estimator of $\sigma^2_{W_{k_2}}$ is then

$$\hat{\sigma}^2_{W_{K_2}} = -\int_0^\tau \frac{\{\int_t^\tau K_2(u)\hat{S}(u)du\}^2}{\hat{S}(t)\hat{S}(t-)} \frac{\hat{p}_1\hat{G}_1(t-)+\hat{p}_2\hat{G}_2(t-)}{\hat{G}_1(t-)\hat{G}_2(t-)}d\hat{S}(t),$$

which is a pooled version of the estimator. In fact, the weight function $K_2(t)$ is not predictable with respect to the σ-filtration generated by $(N_{ij}(t), Y_{ij}(t))$. So the standard martingale central limit theorem (Rebolledo [17], Gill [7]) cannot be applied to show the asymptotic normality of the statistic W_{K_2}. Nevertheless, martingale results of W_{K_2} are still legitimate to show the asymptotic normality, since the limit of the distribution of W_{K_2} associated with appropriate estimators remains valid by Buyske et al. [3] and Wu and Gillbert [19]. Thus, the weight function K_2 can be replaced by its deterministic limit $k_2(t)$, with no change in the limiting distribution of W_{K_2} under H_0. So, $W_{K_2}/\hat{\sigma}_{W_{K_2}}$ converges to the standard normal distribution. An asymptotic level α test is then to reject H_0 when

$$|W_{K_2}/\hat{\sigma}_{W_{K_2}}| > z_{\alpha/2}.$$

D. Versatile Tests

The WKM statistic directly uses estimates of the survival functions, and is sensitive to the magnitude of the survival difference. However, the weight function $K_1(t)$ in the WKM statistic is only for stabilizing the statistic in late observation periods when heavy censoring. It may underperform under the hazard difference alternatives there due to less weight, especially when the difference is non-negligible. The proposed weight function, $K_2(t)$, on the other hand, keeps the statistic to inflate the survival difference at a later period in time by putting more weight, adapting to the censoring level. In order to bring advantage of each weight function, versatile procedures that combine the statistics are considered. In this work, we utilize a linear combination and the maximum of the WKM and proposed statistics. They are two commonly used procedures in practice for the purpose of obtaining a robust statistic that shows good performance in a broad range of alternatives.

Let (W_{K_1}, W_{K_2}) be a bivariate test statistic. Then, by the argument of Shen and Cai [18],

$$(W_{K_1}, W_{K_2}) \xrightarrow{d} (Z_1^*, Z_2^*)$$

under H_0, where (Z_1^*, Z_2^*) have jointly asymptotic normal distribution with mean zero and covariance between them is given by

$$\sigma_{l,m} = -\int_0^\tau \frac{\{\int_t^\tau k_l(u)S(u)du\}\{\int_t^\tau k_m(u)S(u)du\}}{S(t)S(t-)} \frac{p_1G_1(t-)+p_2G_2(t-)}{G_1(t-)G_2(t-)}dS(t),\ l,m=1,2$$

that can be consistently estimated by

$$\hat{\sigma}_{l,m} = -\int_0^\tau \frac{\{\int_t^\tau K_l(u)\hat{S}(u)du\}\{\int_t^\tau K_m(u)\hat{S}(u)du\}}{\hat{S}(t)\hat{S}(t-)} \frac{\hat{p}_1\hat{G}_1(t-)+\hat{p}_2\hat{G}_2(t-)}{\hat{G}_1(t-)\hat{G}_2(t-)}d\hat{S}(t),\ l,m=1,2.$$

The standardized statistics are

$$\tilde{Z}_l = \frac{W_{K_l}}{\sqrt{\hat{\sigma}_{l,l}}},\ l=1,2.$$

Based on the above, we combine the tests to develop the procedures that have a broad range of sensitivity properties in power. Under H_0, the standardized statistics $(\tilde{Z}_1, \tilde{Z}_2)$ have an asymptotically bivariate normal distribution with the standard normal as marginals and correlation coefficient matrix $(\rho_{l,m})_{2\times 2}$ where $\rho_{l,m}$ is the correlation coefficient between W_{K_1} and W_{K_2} that can be consistently estimated by

$$\hat{\rho}_{l,m} = \frac{\hat{\sigma}_{l,m}}{\sqrt{\hat{\sigma}_{l,l}\hat{\sigma}_{m,m}}}.$$

Note that, as mentioned in Yang et al. [21] and Lee [10], when combining tests using W_{K_1} and W_{K_2}, one statistic with larger variance dominates the other, and this results in poor performance of the combined tests. We first consider a linear combination of $\tilde{Z}_1$ and $\tilde{Z}_2$ with equal weights, i.e., $(\tilde{Z}_1+\tilde{Z}_2)/2$. Under H_0, the linear combination of $\tilde{Z}_1$ and $\tilde{Z}_2$ follows a univariate normal distribution with mean zero and variance that can be consistently estimated by $\sum_{l,j=1}^2 \hat{\rho}_{l,m}/4$. Then, an asymptotic level α test rejects the null hypothesis $H_0: S_1 = S_2$ if $|\tilde{Z}_1+\tilde{Z}_2|/2 \geq \tilde{Z}_{\alpha/2}\sqrt{\sum_{l,m=1}^2 \hat{\rho}_{l,m}/4}$. We now look at the maximum of the WKM and proposed statistics, which provides a powerful test if one of the two statistics does. The statistic, $\max(\tilde{Z}_1, \tilde{Z}_2)$, is asymptotically distributed as $\max(Z_1, Z_2)$, where (Z_1, Z_2) is a bivariate normal vector with mean zero and the covariance estimator $\hat{\rho}_{1,2}$. Under H_0, the asymptotic distribution of $\max(Z_1, Z_2)$ is equivalent to that of the maximum order statistic

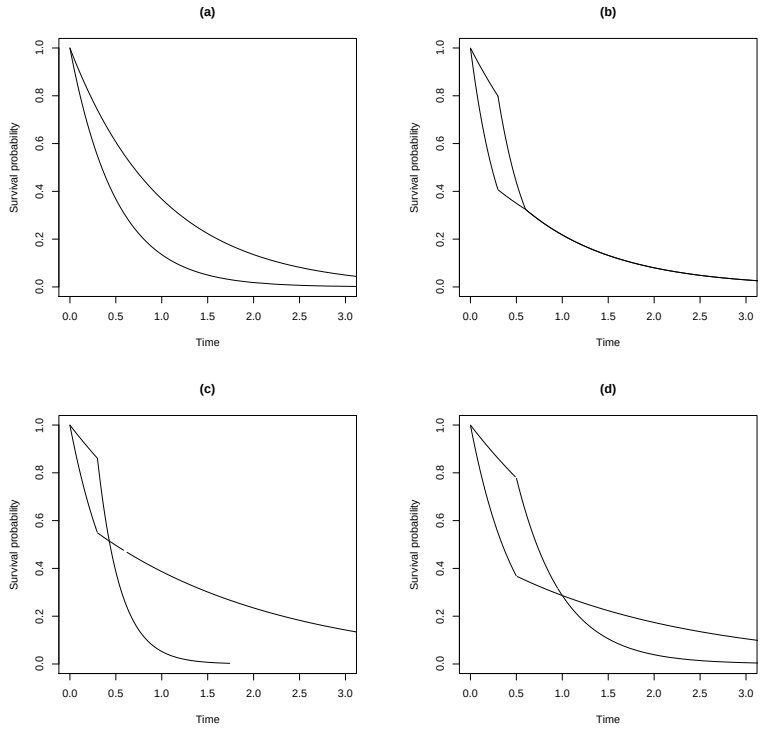

Fig. 1. Survival configurations for simulations.

of Z_1 and Z_2. Thus, for any constant c, $P(\max(Z_1, Z_2) \geq c) = 1 - P(Z_1 < c, Z_2 < c)$, which can be calculated using numerical integration of the bivariate Gaussian distribution. So, the critical value c_α, such that $P(\max(\tilde{Z}_1, \tilde{Z}_2) \geq c_\alpha) = \alpha$, can be obtained based on the asymptotic normality of $(\tilde{Z}_1, \tilde{Z}_2)$ and the estimated asymptotic correlation coefficient $\hat{\rho}_{1,2}$. A $100(1-\alpha)\%$ test rejects H_0 when $\max(\tilde{Z}_1, \tilde{Z}_2) \geq c_\alpha$. Under $H_1 : S_1 \neq S_2$, $\int_0^\tau k_2(t)\{S_1(t) - S_2(t)\}dt$ is non-zero since the variance $\hat{\sigma}_{l,l}$ is bounded (Pepe and Fleming [15]). It follows that the maximum of the tests is consistent against any alternative with the weight $k_2(t)$.

III. Numerical Studies

A. Simulation

We conducted Monte Carlo simulations to assess the performance of the tests. With data artificially censored at various levels, the WKM, log-rank, proposed and combined tests were compared in terms of the type I error rates (size) and power. For the power simulation, two-tailed tests were performed and its results were obtained at the $\alpha = 0.05$ significance level. Different types of alternatives could be seen in a study of establishing a difference that may exist between two treatment effects to find a better way to treat cancer. Our simulation study was designed to explore/compare the behaviors of the test statistics under (a) the proportional hazards case and the cases not stochastically ordered in which the log rank test could lose power to detect survival differences. They are as follows. Higher survival rates of one group than the other at early stage are possible, i.e., (b) early survival differences. Two survival curves could cross at some time point, i.e., higher (lower) survival rates of a group are observed than the other at early (late) time periods. To observe such early and late occurring survival differences with different types, (c) early crossing survival curves and (d) mid crossing of survival curves were considered. Similar alternatives for testing equality of survival distributions are often used in the literature, including Pepe and Fleming [14] and Lee [10]. In each alternative, survival times were generated from the piecewise exponential distributions with hazards rates λ_i, $i = 1, 2$, for group i. With different values of λ_i, the four types of survival scenarios for the alternative were obtained. Specifically, the power simulations were conducted for the following values of λ_i: $\lambda_1 = 1$ and $\lambda_2 = 2$ are set for (a), $\lambda_1 = 3, 0.75, 1$ and $\lambda_2 = 0.75, 3, 1$ for $t < 0.3$, $0.3 \leq t < 0.6$, $t \geq 0.6$, respectively, for (b), $\lambda_1 = 2, 0.5, 0.5$ and $\lambda_2 = 0.5, 4, 4$ for $t < 0.3$, $0.3 \leq t < 0.6$, $t \geq 0.6$, respectively, for (c), and $\lambda_1 = 2, 0.5, 0.5$ and $\lambda_2 = 0.5, 2, 2$ for $t < 0.5$, $0.5 \leq t < 0.8$, $t \geq 0.8$, respectively, for (d). For the size simulation, the survival times were generated from the unit exponential with $\lambda_1 = \lambda_2 = 2$.

Censoring distributions were taken from $c \times \text{Uniform}(0, 1)$, where the constant c is chosen to obtain the desired censoring proportions. Figure 1 depicts the survival configurations (a)-(d).

	[a]30				50				70			
	[b]30%	40%	50%	60%	30%	40%	50%	60%	30%	40%	50%	60%
						Null						
WKM	4	4.4	4.2	4.4	4.2	4.6	4.4	4.4	3.8	3.6	4.4	5.8
LR	4.8	3.6	4.2	4.8	4.4	5	4.8	4.4	3.8	4.2	4.2	5.6
NEW	4.4	3.4	3.6	4.4	4.6	4.2	5.2	4.6	4.4	4.2	5	5.6
LIN	3.6	3.8	3.6	4.6	4.6	4.2	4.6	4.8	3.2	4	4.6	5.6
MAX	6	5.8	4.6	5.6	5.4	6.2	6	5.2	6	5.6	5.2	6.2
						Configuration (a)						
WKM	53	47.6	38.6	34	76.2	72.8	65.4	49.8	89.6	88.2	80.2	69.8
LR	60.4	53	43.4	36.6	80.4	77.6	69.4	57.2	91	89.8	82.6	73.2
NEW	48.2	43.6	35	30.6	71.2	64.6	60.6	46.2	83.6	83.2	74.6	66.8
LIN	51	46	36.8	32.6	74.8	68.6	63.2	47.8	87.2	86.4	78	68.2
MAX	59.4	52.2	42.8	36.2	80.2	76.6	67.4	53.2	90.8	89.4	82.2	72.2
						Configuration (b)						
WKM	26.6	35.8	52.8	70.8	36.6	53.2	75.8	88.2	43	65.2	86.6	97
LR	26	27.8	36.6	53	33.4	41.6	50	77.6	41.2	53	68.8	86
NEW	53.6	59.8	69.2	71.8	72.8	80	87.2	90.4	83	90	96	98.2
LIN	38.8	47.2	59.8	71	56	68.2	83.2	88.8	65.4	81	92.4	97.8
MAX	58.8	63.4	71.2	74.6	75.4	82.4	88.6	91.8	85.4	90.8	96.6	98.4
						Configuration (c)						
WKM	9.2	6.2	12.6	29.6	15.6	6.6	14.8	38.2	25.6	7.6	14.8	44.8
LR	27.8	15	6.8	12.6	91.6	2.3	6.2	71	57.8	26.8	5.4	10.6
NEW	12.8	19.8	30.8	47.4	17.6	25.6	44	60.6	17.4	29.4	61.8	73.2
LIN	9.8	10.4	20.8	37.8	7.4	9.8	26.6	49.4	8	8.8	35.8	60.2
MAX	23.8	24.2	34.6	50.2	37.4	32.6	46	62.4	46.6	40	64.6	75.4
						Configuration (d)						
WKM	36.6	51	68.6	73	45	66.4	85.2	93.6	51.6	78.6	94	98
LR	22.2	27.6	46.4	58.8	21.6	42.4	62.4	81.6	30.6	48.6	75.8	92.2
NEW	65	71.2	78.6	76	82	89	93.2	93.6	92	96.4	97.8	98.8
LIN	49.6	61.6	74.6	74.4	66.8	79	89.6	93.8	76.4	90	96.8	98.6
MAX	67.6	73.2	80.4	77.2	84.6	90.6	93.6	95.2	93.6	96.6	97.8	98.8

[a] $n_1 = n_2$.
[b] % censored.

TABLE I
SIZE AND POWER OF TEST FOR SURVIVAL CONFIGURATION, $\alpha = 0.05$.

Simulations were performed at censoring proportions of 30%, 40%, 50%, and 60%, where two groups were compared with sample sizes of $n_1 = n_2 =$30, 50, and 70. For each sample size drawn from the unit exponential distribution with λ_i, $i = 1, 2$, the number of iterations in a simulation was 500. Based on the 500 iterations, Table 1 presents results of empirical size and power of the tests under various survival configurations depicted in Figure 1. In the table, WKM, LR, NEW, LIN and MAX denote the weighted Kaplan-Meier test, the log-rank test, the proposed test, the linear combination of the proposed and WKM tests, and the maximum of the proposed and WKM tests.

The random experiments can be modeled as a sequence of Bernoulli trials, based on a number of simulations. In our case, since the number of simulations is 500, the estimated error rate is about $\sqrt{(1 - 0.95) \times 0.95/500} \approx 0.01$ at the $\alpha = 0.05$ significance level. Thus, the overall results in Table 1 indicate that under the null hypothesis, all the tests appear to provide acceptable empirical significance levels that are close to the nominal level $\alpha = 0.05$ in various situations.

It is well-known that the log-rank test has optimum power when the hazard rates of two groups are proportional to each other, so it should be quite sensitive to the proportional hazards alternative with the weight distributed evenly. As presented from the results of the proportional hazards case, Table 1(a) suggests that the log-rank test (LR) is the most powerful test in detecting the proportional hazards alternatives, having the highest power among all the tests. It seems that MAX is comparable to LR for all samples of any size at every censoring level, but MAX has slightly lower power consistently than LR, although some are slightly off from the nominal level. From the results in Table 1(b) with the early survival difference alternative, the performance of MAX is seen to be superior to all the comparator tests, though not substantially better than NEW. It is plausible to say that the proposed test seems as powerful as MAX, outperforming WKM, LR and LIN. Survival curves cross at an early time in the configuration (c) of Figure 1, so, late survival differences are relatively more pronounced than early survival differences. In line with this, Table 1(c) presents that for the WKM and log-rank tests, power decreases as the censoring percentages varies from 30% to 40%, and then increases. It is not surprising that for right-censored data, which is our case, late survival differences are becoming less apparent as the censoring percentage increases. These phenomena could occur more notedly when data are heavily censored, since the later time period contributes almost nothing to the power by right-censoring. On the other hand, it seems that NEW, LIN, and MAX are unaffected significantly by the crossing of the survival curves. When data are lightly

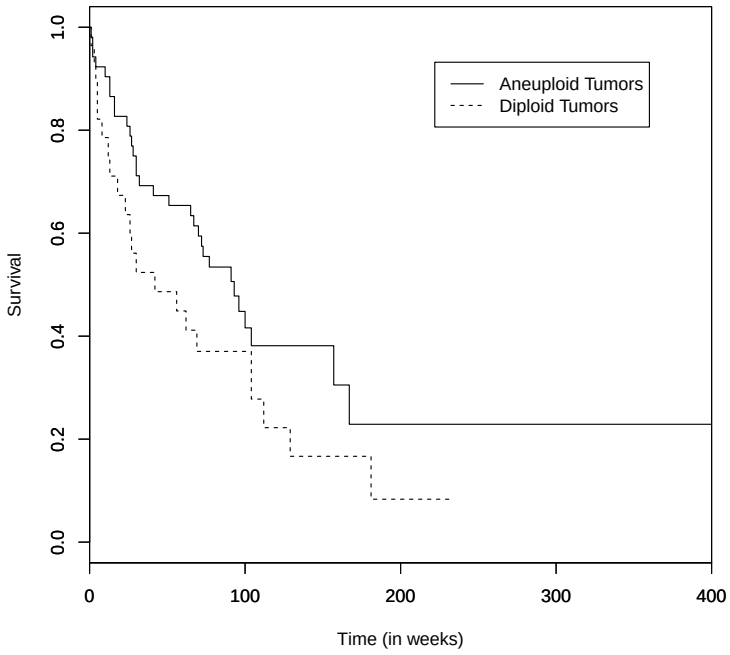

Fig. 2. Estimated survival functions for the tongue cancer study.

censored (30%), the log-rank test performs well. However, NEW and MAX tend to have higher power than other tests as the censoring proportion increases. Overall, MAX appears preferable to NEW, but their differences in power seem to be not substantial. In Table 1(d), crossing survival curves is present at a quite middle time point, which implies prolonged survival in one group relative to the other. The results suggest that LR is inferior to all other tests. This is clearly related to the violation of the proportional hazards assumption. However, when censoring is heavy in the presence of crossing survival curves, the assumption would be somewhat met. Our results show that these phenomena tend to be conspicuous when the sample size is large. For example, with the sample size of $n = 70$, LR has power of 92.2% at the 60% censoring level, which suggests that it is comparable to other tests. The tests, NEW and MAX, appear to achieve higher power than other tests across a broad range of survival differences. Though the results show the superior performance of MAX to NEW, their difference in power is not substantial. In summary, the results in Table 1 indicate that overall, NEW and MAX perform better than WKM, LR, and LIN for various situations considered, and can be used as a useful tool, especially when survival differences of two groups are not able to be specified in advance.

B. Application

The procedures based on the proposed weight function were applied to a real-world clinical data set from the literature. The data set concerns death times (in weeks) of patients with tongue cancer. Tongue cancer is a type of mouth cancer that can develop in any part of the mouth, forming a tumor by its accumulation. A study of the prognostic effect of ploidy on the survival of patients diagnosed with tongue cancer was conducted. In this study, 80 patients are divided into two groups ($n_1 = 52$, $n_2 = 28$) according to tumor DNA index, one having an aneuploid (abnormal) DNA profile and the other with the diagnosis of a diploid (normal) DNA. The censoring percentage is 27/80 $\approx$ 34%. Details about the study, including data, can be found in Klein and Moeschberger [9]. Figure 2 depicts Kaplan-Meier survival functions of the two groups, $\hat{S}_1(t)$ and $\hat{S}_2(t)$, with aneuploid and diploid, respectively. The estimated survival curves in Figure 2 indicate that the group with the aneuploid DNA may have a higher survival rate than the group with the diploid DNA, suggesting that the hazard ratio may change over time. Our interest in this study is to test the hypothesis that the survival rates of the two groups with aneuploid and diploid tumors are the same against the alternative telling the difference. The LR, WKM, NEW tests yield $p-$values of 0.098, 0.076 and 0.067, respectively. With the estimated correlation being equal to 97%, the LN and MAX tests give the $p-$values of 0.071 and 0.063, respectively. The $p-$value, 0.098, of the LR test tells us that it is likely to fail to detect the difference of two survival curves, leading to the conclusion that the two groups have the same survival rate. On the other hand, the NEW, LN

and MAX tests substantiate the conjecture that somewhat differences in the survival rates of the groups may exist. The result from the WKM test is relatively less affirmative to validate the difference, when compared to the proposed methods.

IV. Remarks

There may exist several different types of survival difference, such as crossing survival functions. Of interest in medical research is a long-term comparison of two underlying survival distributions with right-censored data. This paper improves the WKM statistic in order for it to be an effective tool to use in such case. The procedures are based on some weight function incorporated with the censoring level of the sample data as a component. Some versatile test procedures that are robust and sensitive to various survival differences are also considered. Simulation results demonstrate the superior behaviors of the new method to the WKM statistics, yielding higher power than the WKM test across a broad range of survival differences. The procedures with the proposed weight function for comparing the equality of two survival curves are illustrated in a tongue cancer data set.

References

[1] Aalen, O. (1975). *Statistical Inference for a Family of Counting Process*, Ph.D. diss., Department of Statistics, University of California, John Wiley & Sons, Berkeley.

[2] Aalen, O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics*, **6**, 701–726.

[3] Buyske, S.; Fagerstrom, R. and Z. Ying, Z. (2000). A Class of weighted log-rank tests for survival data when the event is rare, *Journal of the American Statistical Association*, **95**, 249–258.

[4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society*, **34**, 187–220.

[5] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

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

[7] Gill, R.D. (1980). *Censoring and Stochastic Integrals: MC Tract 124*, Amsterdam Mathematical Centre.

[8] Kaplan, E. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, **53**, 457–481.

[9] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis*, Springer-Verlag, New York.

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

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

[12] Logan, B.R.; Klein, J.P. and Zhang, M.J. (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation, *Biometrics*, **64**, 733–740.

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

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

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

[16] Peto, R. and Peto, R. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society, Series A*, **135**, 185–207.

[17] Rebolledo, R. (1980). Central limit theorems for local martingales, *Zeitschrift fur Wahrscheinlichkeitstheorie verw Gebiete*, **51**, 269-286.

[18] Shen, Y. and Cai, J. (2001). Maximum of the weighted Kaplan-Meier tests with applications to cancer prevention and screening trials, *Biometrics*, **57**, 837–843.

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

[20] Yang, P. and Fleming, T.R. (2006). Simultaneous use of weighted logrank and standardized Kaplan-Meier statistics, *Journal of Biopharmaceutical Statistics*, **16**, 241–252.

[21] Yang, S.; Hsu, L. and Zhao, L. (2005). Combining asymptotically normal tests: case studies in comparison of two groups, *Journal of Statistical Planning and Inference*, **133**, (2005), 139–158.