ELSEVIER

Journal of Statistical Planning and Inference 83 (2000) 137–151

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Multiple comparisons between successive treatments for randomly right-censored survival data

Yuh-Ing Chen[a, *], Yunchan Chi[b]

[a] *Institute of Statistics, National Central University, Chung-Li 32054, Taiwan*
[b] *Department of Statistics, National Cheng-Kung University, Tainan 70101, Taiwan*

Received 1 April 1998; received in revised form 18 November 1998; accepted 15 February 1999

Abstract

In this paper we are concerned with the problem of comparing adjacent ordered treatments in a one-way layout where survival data are subject to random right censorship. Multiple testing procedures based on two-sample statistics, each comparing an individual treatment with the previous one, are proposed for determining the pattern of the treatment effects. The two-sample statistics under consideration are weighted logrank statistics (Fleming and Harrington, 1991, Counting Process and Survival Analysis. Wiley, New York) and weighted Kaplan–Meier statistics (Pepe and Fleming, 1989. Biometrics 45, 497–507; 1991. J. Roy. Statist. Soc. B 53, 341–352). An illustrated numerical example is reported. Finally, the comparative results of a Monte Carlo error rate and power study for small sample sizes are presented.

MSC: 62J15; 62N05

Keywords: Monte Carlo study; Multiple comparisons; One-way layout; Right-censored data

1. Introduction

The effects of a toxin or a drug are often investigated by an experiment including several increasing dose levels (treatments) of the substance. Usually, it can be reasonably assumed that the increasing dose levels produce stronger or at least equal treatment effects. However, the pattern of the monotonic dose response relationship remains unknown. To get insight into the pattern when data are normally distributed, van Eeden (1960), and Lee and Spurrier (1995a) considered multiple comparisons between neighboring dose levels to decide if a dose increase leads to an additional effect or the dose response relationship in this domain is too flat. Budde and Bauer (1989) and Lee and

* Corresponding author.

0378-3758/00/$ - see front matter
PII: S0378-3758(99)00050-6

Spurrier (1995b) suggested nonparametric procedures for comparing the adjacent treatments when data are not normally distributed. In animal carcinogenesis experiments or comparative clinical trials, however, it occurs frequently that the primary outcome of interest is time to a certain event (for example, death, tumor occurrence). Moreover, randomly right-censored data are often involved in these studies, since subjects who randomly enter the study to receive treatments may be lost to follow-up randomly, or the study may be terminated at a preassigned time owing to time limitation. Therefore, testing procedures for determining the pattern of the treatment effects with randomly right-censored survival data are needed.

For the ith sample ($i = 1,\ldots,k$), let $T_{i1},\ldots,T_{in_i}$ be independent identically distributed (i.i.d.) random variables each with a continuous distribution function F_i, and let $U_{i1},\ldots,U_{in_i}$ be i.i.d. random variables each with a continuous distribution function C_i, where U_{ij} is the censoring time associated with the survival time T_{ij}. Suppose that the k samples are independent of each other and the U_{ij} are distributed independent of T_{ij}. In such a setting, we actually only observe the bivariate vectors (X_{ij}, δ_{ij}), where $X_{ij} = \min(T_{ij}, U_{ij})$, $\delta_{ij} = 1$, if $X_{ij} = T_{ij}$, and 0, otherwise. Let $S_i = 1 - F_i$ be the survival function of the ith group, $i=1,\ldots,k$. Liu et al. (1993) based on weighted logrank statistics (Fleming and Harrington, 1991) to develop testing procedures for the null hypothesis H_0: ($S_i=S$, $i=1,2,\ldots,k$) against the ordered alternative H_1: ($S_1 \geqslant S_2 \geqslant \cdots \geqslant S_k$ with at least one strict inequality) (Barlow et al., 1972). Chi and Chen (1998) further suggested an ordered test on the basis of the weighted Kaplan–Meier statistics (Pepe and Fleming, 1989, 1991). Note that both the tests are designed for testing against the global ordered alternative H_1, but they do not provide any information about the ordered pattern of the treatment effects.

To determine the pattern of the treatment effects when survival data are subject to random right censorship, we consider multiple testing procedures between neighboring treatments on the basis of the two-sample statistics each comparing an individual treatment with the previous one. The two-sample statistics under consideration are weighted logrank statistics and weighted Kaplan–Meier statistics. The use of these testing procedures is illustrated with the numerical example assessing the effect of ovalbumin immune bone marrow cells on the transfer antitumor activity (Hornung et al., 1995). Comparative results of a Monte Carlo study investigation demonstrate the relative error rate and power performances of these testing procedures for small sample sizes. Some suggestions and conclusions are finally given.

2. Weighted logrank multiple tests

For $i = 1,\ldots,k$, let $D_i(t) = \#\{u: X_{iu} \leqslant t,\ \delta_{iu} = 1,\ u = 1,2,\ldots,n_i\}$ be the number of patients in group i who have been died by time t and let $Y_i(t) = \#\{u: X_{iu} \geqslant t,\ u = 1,2,\ldots,n_i\}$ be the number of patients in group i who are still alive and uncensored at time t. Let $N_{i+} = n_i + n_{i+1}$, $Y_{i+}(t) = Y_i(t) + Y_{i+1}(t)$ and $D_{i+}(t) = D_i(t) + D_{i+1}(t)$. Set $t_c = \text{minimum}(t_1,\ldots,t_k)$, where t_i is the last observation in group i. Using the counting

process formulation described in Gill (1980), the weighted logrank statistic comparing the $(i+1)$th treatment with the ith treatments is

$$\mathrm{WLR}_i = \int_0^{t_c} W_i(t) \frac{Y_i(t)Y_{i+1}(t)}{Y_{i+}(t)} \left\{ \frac{\mathrm{d}D_i(t)}{Y_i(t)} - \frac{\mathrm{d}D_{i+1}(t)}{Y_{i+1}(t)} \right\}. \tag{2.1}$$

Fleming and Harrington (1991) suggested to use $W_i(t) = \{\hat{S}_i(t)\}^{\rho}\{1-\hat{S}_i(t)\}^{\gamma}$ for $\rho, \gamma \geqslant 0$, where $\hat{S}_i(t)$ is the Kaplan and Meier (1958) survival estimate based on the ith and $(i+1)$th samples. Note that taking $\rho = \gamma = 0$ produces the logrank statistic (Mantel, 1966) and setting $\rho = 1$ and $\gamma = 0$ yields the Peto–Prentice statistic (Peto and Peto, 1972; Prentice, 1978). Moreover, the consistent and unbiased estimator of the variance of WLR_i is given by

$$s_{ii} = \int_0^{t_c} W_i^2(t) \frac{Y_i(t)Y_{i+1}(t)}{Y_{i+}(t)} \left\{ 1 - \frac{\Delta D_{i+}(t) - 1}{Y_{i+}(t)} \right\} \frac{\mathrm{d}D_{i+}(t)}{Y_{i+}(t)}, \tag{2.2}$$

where $\Delta D_{i+}(t) = D_{i+}(t) - D_{i+}(t-)$. Let

$$\mathrm{WLR}_i^* = \mathrm{WLR}_i / \sqrt{s_{ii}}, \quad i = 1, 2, \ldots, k-1. \tag{2.3}$$

It can be shown (see Appendix) that, under the null hypothesis H_0, the asymptotic distribution of the random vector $(\mathrm{WLR}_1^*, \mathrm{WLR}_2^*, \ldots, \mathrm{WLR}_{k-1}^*)$ is the $(k-1)$-variate normal with mean zero vector and correlation matrix $\boldsymbol{R} = \{r_{ij}\}$, where $r_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ and the σ_{ii} and σ_{ij} are stated in Eqs. (A.1) and (A.2). Note that the matrix $\boldsymbol{R}$ can be consistently estimated by $\hat{\boldsymbol{R}} = \{s_{ij}/\sqrt{s_{ii}s_{jj}}\}$, where the s_{ii} are stated in (2.2) and the s_{ij} are given by, for $i < j = 2, \ldots, k-1$,

$$s_{ij} = -\int_0^{t_c} W_i(t)W_{i+1}(t) \frac{Y_i(t)Y_{i+1}(t)Y_{i+2}(t)}{Y_{i+}(t)Y_{i+1+}(t)} \left\{ 1 - \frac{\Delta D_{i++}(t) - 1}{Y_{i++}(t) - 1} \right\} \frac{\mathrm{d}D_{i++}(t)}{Y_{i++}(t)}, \tag{2.4}$$

if $i = j-1$ and 0, otherwise, $D_{i++}(t) = D_i(t) + D_{i+1}(t) + D_{i+2}(t)$, $Y_{i++}(t) = Y_i(t) + Y_{i+1}(t) + Y_{i+2}(t)$ and $\Delta D_{i++}(t) = D_{i++}(t) - D_{i++}(t-)$. Let $(Z_1, Z_2, \ldots, Z_{k-1})$ be a $(k-1)$-variate normal vector with mean zero and the correlation matrix $\hat{\boldsymbol{R}}$, and let $\mathrm{zmax}(k-1, \alpha)$ be the upper αth percentile of the distribution of $\max(Z_1, Z_2, \ldots, Z_{k-1})$. As a generalization of the Lee and Spurrier (1995a) testing procedure, we claim

$$S_{i+1} > S_i \quad \text{if } \mathrm{WLR}_i^* \geqslant \mathrm{zmax}(k-1, \alpha) \quad \text{for } i = 1, 2, \ldots, k-1. \tag{2.5}$$

It is obvious that the experimentwise error rate, the probability of erroneously declaring at least one treatment better than its preceding one, for this procedure is approximately controlled, since

$$\begin{aligned} \alpha &\approx P\{\max(\mathrm{WLR}_1^*, \mathrm{WLR}_2^*, \ldots, \mathrm{WLR}_{k-1}^*) \geqslant \mathrm{zmax}(k-1, \alpha) \mid \mathrm{H}_0\} \\ &= P\{\mathrm{WLR}_i^* \geqslant \mathrm{zmax}(k-1, \alpha) \text{ for at least one } i \mid \mathrm{H}_0\}. \end{aligned}$$

For any z, the probability $P\{\max(Z_1, Z_2, \ldots, Z_{k-1}) \leqslant z\}$ can be computed using a program for calculating multivariate normal probabilities (Schervish, 1984). Therefore, the critical value $\mathrm{zmax}(k-1, \alpha)$ can be found such that $P\{\max(Z_1, Z_2, \ldots, Z_{k-1}) \geqslant \mathrm{zmax}(k-1, \alpha)\} = \alpha$.

Table 1
Table of summary statistics for the bone marrow transplantation-tumor data

$(i, i+1)$	WLR(Logrank)	WLR(Peto–Prentice)	WKM
(1, 2)	0.591	0.991	0.657
(2, 3)	2.189	1.961	1.879
(3, 4)	0.437	0.434	0.363
(4, 5)	−0.488	−0.553	−0.525
Correlation			
(1, 2) and (2, 3)	−0.487	−0.474	−0.627
(2, 3) and (3, 4)	−0.459	−0.481	−0.515
(3, 4) and (4, 5)	−0.502	−0.494	−0.518

Remember that weight function $\omega_i(t) = 1$ corresponds to a logrank statistic, and $\omega_i(t) = S(t)$ to a Peto–Prentice statistic. Under the assumption of equal censoring, that is, $C_i = C$, $i = 1, 2, \ldots, k$, we observe, from (A.3) and (A.4), that the correlation structure for the two weight functions is

$$r_{ii} = 1, \quad r_{ij} = \begin{cases} -\sqrt{p_i p_{i+2}/[(p_i + p_{i+1})(p_{i+1} + p_{i+2})]}, & \text{if } i = j - 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

This correlation structure is the same as that stated in Lee and Spurrier (1995a). Therefore, when sample sizes are equal and the assumption of equal censoring is tenable, we suggest to use the critical values reported in Lee and Spurrier's (1995a) Table 1 with infinite degrees of freedom.

3. Weighted Kaplan–Meier multiple tests

Note that the weighted logrank statistic in (2.1) on the basis of the difference of the estimated hazard functions is, in fact, appropriate for testing against the hypothesis of two ordered hazard functions. For constructing a procedure which is more sensitive to testing against $S_i < S_{i+1}$, Pepe and Fleming (1989) proposed a class of weighted Kaplan and Meier (1958) statistics given by

$$\mathrm{WKM}_i = \int_0^{T_c} \sqrt{n_i n_{i+1}}\, \hat{w}_i(t)\{\hat{S}_{i+1}(t) - \hat{S}_i(t)\}\, \mathrm{d}t \tag{3.1}$$

where $T_c = \sup\{t\colon \min(\hat{G}_i(t), \hat{S}_i(t),\ i = 1, \ldots, k) > 0\}$, $\hat{G}_i(t)$ be the Kaplan–Meier estimator of censoring survival distribution $G_i(t) = 1 - C_i(t)$, $\hat{S}_i(t)$ is Kaplan–Meier estimators of S_i, and $\hat{w}_i(t)$ is the random weight function which downweights the contribution of $\hat{S}_{i+1}(t) - \hat{S}_i(t)$ over later time periods if censoring is heavy so that the statistic WKM_i is stable. Moreover, if $S_i = S_{i+1} = S$, the variance of WKM_i can be estimated by

$$v_{ii} = \int_0^{T_c} \left\{ \int_t^{T_c} \hat{w}_i(u)\hat{S}(u)\, \mathrm{d}u \right\}^2 \frac{n_i \hat{G}_i(t-) + n_{i+1}\hat{G}_{i+1}(t-)}{\hat{G}_i(t-)\hat{G}_{i+1}(t-)} \frac{\mathrm{d}\hat{F}(t)}{\hat{S}(t)\hat{S}(t-)}, \tag{3.2}$$

where $\hat{S}(t)$ is the Kaplan–Meier estimator of the common survival distribution $S(t)$ based on the ith and $(i+1)$th samples and $\hat{F}(t) = 1 - \hat{S}(t)$. Let $\bar{p}_i$ be n_i/N, $i = 1, \ldots, k$. For the k-sample setting studied in this paper, we employ the following weight function

suggested in Pepe and Fleming (1989):

$$\hat{w}_i(t) = \frac{\hat{G}_i(t-)\hat{G}_{i+1}(t-)}{\bar{p}_i\hat{G}_i(t-) + \bar{p}_{i+1}\hat{G}_{i+1}(t-)}. \tag{3.3}$$

To determine the pattern of the treatment effects with randomly right-censored data, we consider the random vector $(\mathrm{WKM}_1^*, \mathrm{WKM}_2^*, \ldots, \mathrm{WKM}_{k-1}^*)$, where

$$\mathrm{WKM}_i^* = \mathrm{WKM}_i/\sqrt{v_{ii}}, \quad i = 1, 2, \ldots, k-1.$$

It can be shown in Appendix that under the null hypothesis H_0, the asymptotic distribution of the random vector $(\mathrm{WKM}_1^*, \mathrm{WKM}_2^*, \ldots, \mathrm{WKM}_{k-1}^*)$ is the $(k-1)$-variate normal with mean zero and correlation matrix $\Gamma = \{\rho_{ij}\}$, where $\rho_{ij} = \phi_{ij}/\sqrt{\phi_{ii}\phi_{jj}}$ and the ϕ_{ii} and ϕ_{ij} are stated in Eqs. (A.6) and (A.7), respectively. Note that the matrix Γ can be consistently estimated by $\hat{\Gamma} = \{v_{ij}/\sqrt{v_{ii}v_{jj}}\}$, where the v_{ii} providing consistent estimates of $N\phi_{ii}$ are stated in (3.2), and the v_{ij} as consistent estimates of $N\phi_{ij}$ are given by, for $i < j = 2, \ldots, k-1$,

$$v_{ij} = -\sqrt{n_i n_{i+2}} \int_0^{T_c} \left\{ \int_t^{T_c} \hat{w}_i(u)\hat{S}(u)\,\mathrm{d}u \right\} \left\{ \int_t^{T_c} \hat{w}_{i+1}(u)\hat{S}(u)\,\mathrm{d}u \right\} \times \frac{\mathrm{d}\hat{F}(t)}{\hat{S}(t)\hat{S}(t-)\hat{G}_{i+1}(t-)} \tag{3.4}$$

if $i = j-1$, and 0 otherwise, and $\hat{S}(t)$ is the Kaplan–Meier estimator computed from the combined samples of i, $i+1$, and $i+2$. As a resemble to the testing procedure in (2.5), we claim

$$S_{i+1} > S_i \quad \text{if } \mathrm{WKM}_i^* \geqslant \mathrm{zmax}^*(k-1, \alpha), \quad \text{for } i = 1, 2, \ldots, k-1. \tag{3.5}$$

where $\mathrm{zmax}^*(k-1, \alpha)$ is the upper αth percentile of the $(k-1)$-variate normal distribution with mean zero and correlation matrix $\hat{\Gamma}$. Note that, from (A.8) and(A.9), the correlation structure under equal censoring pattern is the same as in (2.6). Therefore, the critical values reported in Lee and Spurrier's (1995a) Table 1 with infinite degrees of freedom can, again, be used in (3.5) when the assumption of equal censoring is tenable and sample sizes are equal.

4. An example

Hornung et al. (1995) conducted a laboratory study to assess whether antigen-specific antitumor immune responses, elicited in normal donor mice by immunization with the soluble form of the surrogate tumor antigen ovalbumin (OVA), can be transferred via bone marrow transplantation into lethally irradiated, syngeneic recipient mice. In this paper, we investigate the pattern of the antitumor immune responses transferred from donors with increasing number of OVA-immune bone marrow cells.

Fifty female C57BL/6 mice bearing day-10, subcutaneous E.G7-OVA tumors were given lethal TBI, then reconstituted with various doses of pooled bone marrow cells from OVA-immune donors. The dosages of bone marrow considered in the study were:

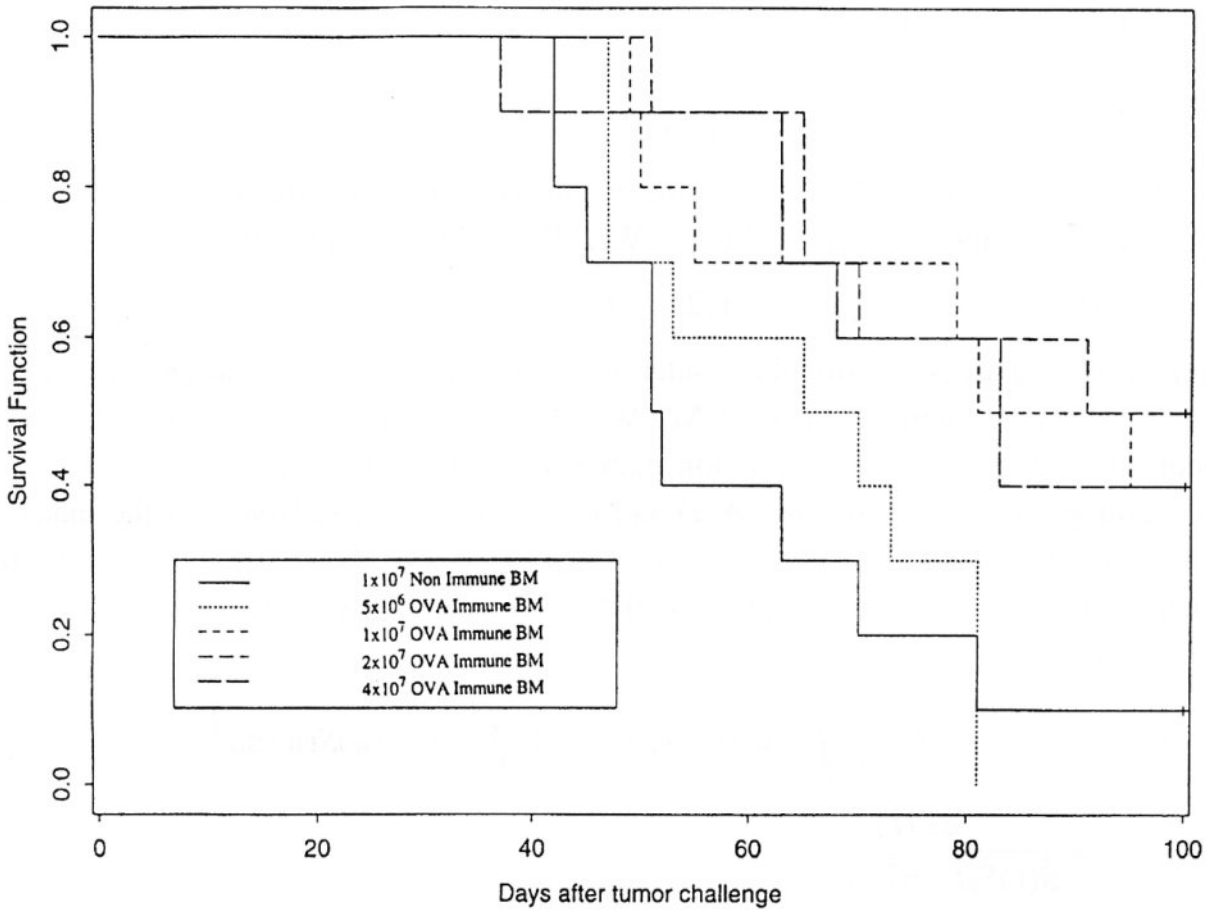


Fig. 1. The Kaplan–Meier estimates for the bone marrow transplantation-tumor data.

1×10^7 Non-immune bone marrow cells (group 1); 5×10^6 OVA-immune bone marrow cells (group 2); 1×10^7 OVA-immune bone marrow cells (group 3); 2×10^7 OVA-immune bone marrow cells (group 4); and 4×10^7 OVA-immune bone marrow cells (group 5). The measurement of record for each dosage group was the survival time after reconstituted with bone marrow cells. Mice without noticeable tumors were sacrificed and autopsied at 150 days and were considered long-term survivors, yielding censored data for their respective dosage groups. The Kaplan and Meier (1958) estimates of the survival functions for the five groups of mice are presented in Fig. 1. Since there is a monotonic relationship between the transferred OVA-immune bone marrow cells and the antitumor immune responses, we reported, in Table 1, the relevant one-sided statistics and the critical values of the proposed tests corresponding to their correlation estimates.

The approximate 5% and 10% critical values corresponding to the three sets of correlation estimates are 2.237 and 1.951, respectively. (The 5% and 10% critical values with infinite degrees of freedom reported in Lee and Spurrier (1995a) are 2.238 and 1.952, respectively.) The WLR tests based on logrank and Peto–Prentice statistics reach the same conclusion that, under the significance level $\alpha = 0.10$, there is only one significantly different pair of 5×10^6 OVA-immune bone marrow cells (group 2) and 1×10^7 OVA-immune bone marrow cells (group 3) in which group 3 produces better antitumor immune response than does group 2. The WKM test fails to detect such a difference at the same significance level.

5. Monte Carlo study

A Monte Carlo study was performed to examine the relative level and power performances of the weighted logrank (WLR) and weighted Kaplan–Meier (WKM) tests for comparing the adjacent treatments when survival data are subject to random right censorship. The WLR tests based on logrank and Peto–Prentice statistics are denoted by WLR(L) and WLR(P), respectively. Herein, we considered $k = 5$ treatment groups with sample sizes $n_1 = \cdots = n_5 = n = 10, 20, 30$ in the error rate study and $n = 20$ and 30 in the power study.

Exponential distributions and three types of piecewise exponential distributions, demonstrated in Fig. 2, were employed to be the survival distributions under the null hypothesis and a variety of alternative hypotheses corresponding to different types of

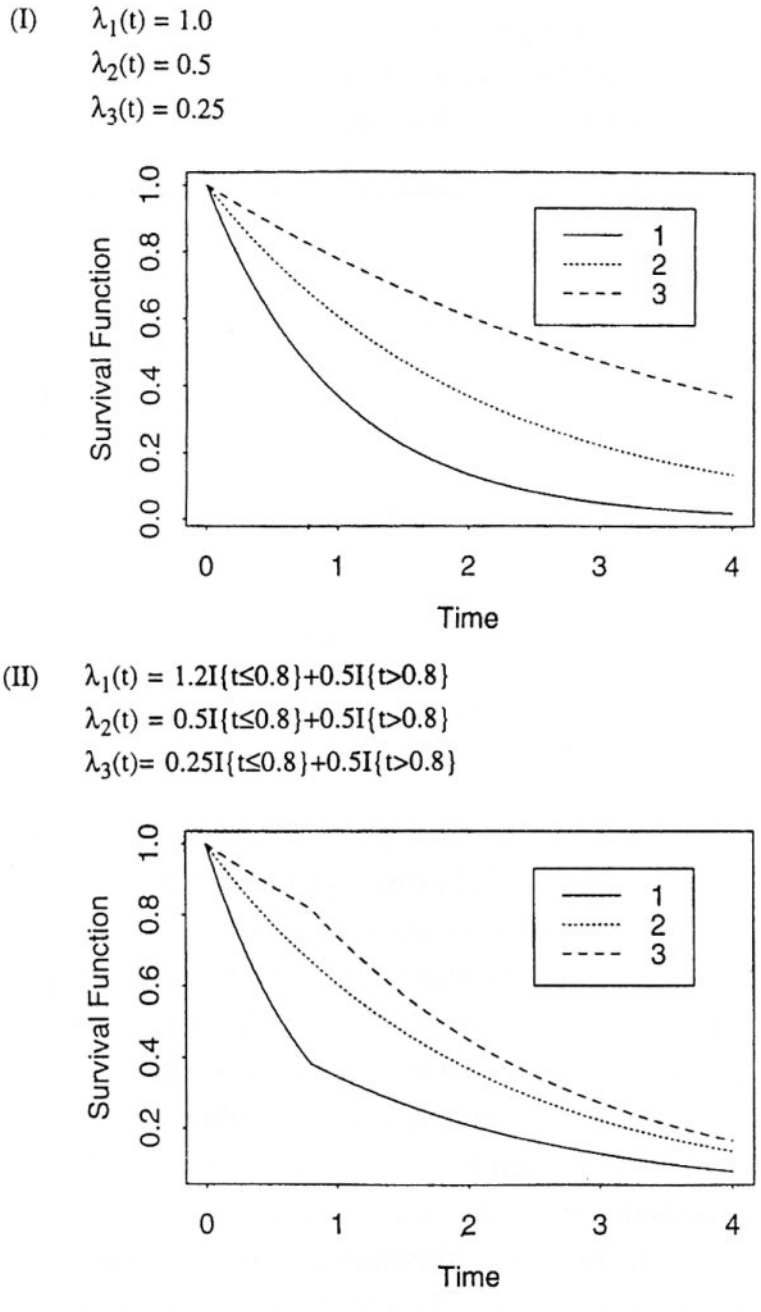
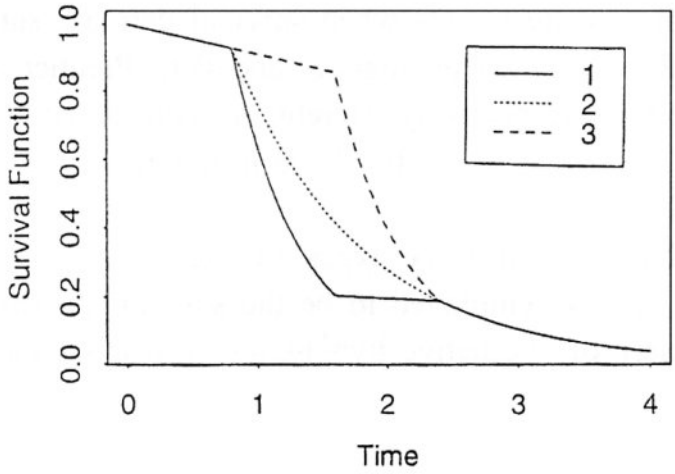


Fig. 2. Survival configurations for alternatives.

(III) $\lambda_1(t) = 0.1I\{t\le0.8\}+1.9I\{0.8<t\le1.6\}+0.1I\{1.7<t\le2.4\}+I\{t>2.4\}$
$\lambda_2(t) = 0.1I\{t\le0.8\}+1.0I\{0.8<t\le1.6\}+1.0I\{1.7<t\le2.4\}+I\{t>2.4\}$
$\lambda_3(t) = 0.1I\{t\le0.8\}+0.1I\{0.8<t\le1.6\}+1.9I\{1.7<t\le2.4\}+I\{t>2.4\}$



(IV) $\lambda_1(t) = 0.3I\{t\le0.8\}+3.5I\{t>0.8\}$
$\lambda_2(t) = 0.3I\{t\le0.8\}+2.0I\{t>0.8\}$
$\lambda_3(t) = 0.3I\{t\le0.8\}+0.5I\{t>0.8\}$

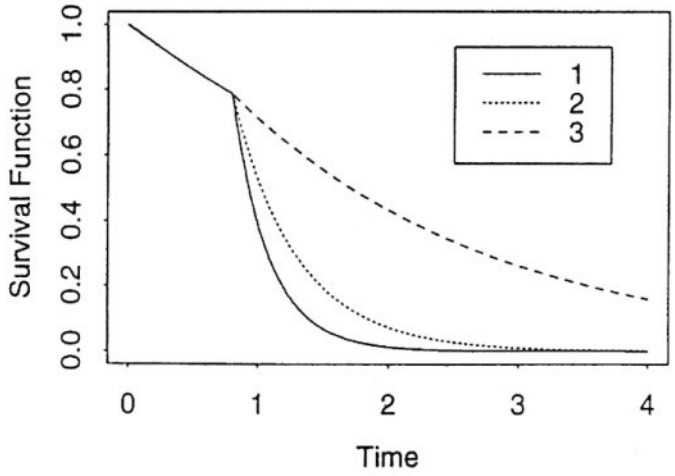


Fig. 2. (*continued*)

hazard differences. The solid line in each panel represents the common survival function under the null hypothesis. Panel (I) displays the survival distributions generated by exponential distributions with different hazard rates which correspond to proportional hazards; the survival distributions in panel (II) are generated by piecewise exponential distributions to produce an early hazard difference alternative; the survival distributions in panel (III) are generated by piecewise exponential distributions to give a middle hazard difference alternative; and the survival distributions in panel (IV) generated by piecewise exponential distributions yield a late hazard difference alternative. The ordered alternatives considered in the study are (i) $S_1 = S_2 < S_3 = S_4 = S_5$ and (ii) $S_1 = S_2 < S_3 = S_4 < S_5$. In both the alternatives, the first two groups have the same hazard rate λ_1. In alternative (i), S_3, S_4 and S_5 are survival functions corresponding to λ_3, while, in alternative (ii), λ_2 is the hazard rate of groups 3 and 4, and S_5 is

the survival function corresponding to λ_3. Uniform distribution over $(0, R)$ was used as the censoring distribution. Various values of R which correspond to the probability of censorship as 0.3 and 0.5 were considered in the error rate study, the corresponding uniform distributions were then employed as censoring distributions in the power study. Note that the censoring probabilities were fixed for each population in the error rate study, but they may be different for the populations involved in the power study due to different survival distributions.

For each of these settings, 25 000 replications were used to obtain the estimated experimentwise error rates, 10 000 replications were employed to estimate the experimentwise powers (probability of correctly detecting at least one treatment better than its preceding one) and marginal powers (probability of detecting the $(i+1)$th treatment better than the ith treatment), denoted by π_i under the nominal level $\alpha = 0.05$. Therefore, the standard error for the estimated error rate is around $0.001(\approx \sqrt{(0.05)(0.95)/25\,000})$, while the maximum standard error for the power estimate is $0.005(=\sqrt{(0.5)(0.5)/10\,000})$. The estimated error rates and powers are presented in Tables 2–4, respectively.

Table 2 clearly reveals that the WLR(P) holds its error rate well across all the simulations under consideration. In addition, the unweighted logrank test WLR(L) and the WKM test reasonably maintain their error rates when the common sample size is at least 20.

The power study in Tables 3 and 4 indicates that the WLR(L) test is more powerful than either the WLR(P) or WKM test for exponential distributions. This result is not surprising, since the WLR(L) test is the most efficient one for proportional hazards. Nevertheless, the WKM test provides with a competitor to the WLR(L) test for proportional hazards. For early hazard difference alternatives, the WLR(P) test has the best

Table 2
Estimated level for $k = 5$ and $n_1 = \cdots = n_5 = n$

n	Distribution	Censoring probability					
		0.3			0.5		
		WKM	WLR(L)	WLR(P)	WKM	WLR(L)	WLR(P)
10	(I)	0.054	0.061	0.051	0.057	0.056	0.050
	(II)	0.054	0.063	0.051	0.061	0.059	0.052
	(III)	0.040	0.066	0.053	0.048	0.067	0.053
	(IV)	0.044	0.066	0.052	0.043	0.067	0.051
20	(I)	0.051	0.056	0.050	0.054	0.054	0.049
	(II)	0.055	0.055	0.052	0.052	0.049	0.050
	(III)	0.051	0.060	0.051	0.053	0.056	0.050
	(IV)	0.050	0.059	0.050	0.051	0.057	0.050
30	(I)	0.053	0.055	0.052	0.054	0.053	0.052
	(II)	0.052	0.053	0.049	0.051	0.053	0.048
	(III)	0.050	0.055	0.050	0.055	0.056	0.051
	(IV)	0.050	0.056	0.050	0.051	0.057	0.047

Table 3
Estimated experimentwise powers for $k = 5$ and $n_1 = \cdots = n_5 = n$

n	Alternative		Censoring probability					
			0.3			0.5		
			WKM	WLR(L)	WLR(P)	WKM	WLR(L)	WLR(P)
20	(I)	(i)	0.755	0.772	0.718	0.528	0.554	0.528
		(ii)	0.410	0.420	0.395	0.265	0.279	0.260
	(II)	(i)	0.560	0.509	0.609	0.587	0.555	0.565
		(ii)	0.313	0.294	0.31	0.313	0.297	0.292
	(III)	(i)	0.427	0.399	0.623	0.483	0.498	0.542
		(ii)	0.248	0.238	0.349	0.281	0.276	0.303
	(IV)	(i)	0.496	0.632	0.353	0.170	0.316	0.178
		(ii)	0.307	0.413	0.241	0.115	0.198	0.126
30	(I)	(i)	0.921	0.931	0.904	0.734	0.764	0.740
		(ii)	0.573	0.594	0.537	0.382	0.398	0.380
	(II)	(i)	0.743	0.678	0.800	0.785	0.752	0.773
		(ii)	0.437	0.402	0.477	0.460	0.440	0.443
	(III)	(i)	0.579	0.529	0.805	0.657	0.654	0.740
		(ii)	0.332	0.299	0.488	0.373	0.363	0.430
	(IV)	(i)	0.693	0.838	0.529	0.236	0.486	0.271
		(ii)	0.437	0.600	0.355	0.153	0.289	0.179

Table 4
Estimated marginal powers for $k = 5$ and $n_1 = \cdots = n_5 = n$

n	Alternative			Censoring probability					
				0.3			0.5		
				WKM	WLR(L)	WLR(P)	WKM	WLR(L)	WLR(P)
20	(I)	(i)	π_2	0.755	0.772	0.718	0.528	0.554	0.528
		(ii)	π_2	0.280	0.285	0.239	0.177	0.188	0.172
			π_4	0.184	0.189	0.177	0.107	0.111	0.104
	(II)	(i)	π_2	0.560	0.509	0.609	0.587	0.555	0.565
		(ii)	π_2	0.255	0.243	0.272	0.249	0.242	0.234
			π_4	0.078	0.066	0.086	0.085	0.073	0.076
	(III)	(i)	π_2	0.427	0.399	0.623	0.483	0.498	0.542
		(ii)	π_2	0.077	0.097	0.108	0.091	0.104	0.092
			π_4	0.186	0.154	0.271	0.208	0.192	0.231
	(IV)	(i)	π_2	0.496	0.632	0.353	0.170	0.316	0.178
		(ii)	π_2	0.059	0.095	0.058	0.038	0.064	0.041
			π_4	0.259	0.349	0.195	0.081	0.144	0.088
30	(I)	(i)	π_4	0.921	0.931	0.904	0.734	0.764	0.740
		(ii)	π_2	0.414	0.431	0.375	0.269	0.282	0.264
			π_4	0.273	0.287	0.262	0.157	0.166	0.161
	(II)	(i)	π_2	0.743	0.678	0.800	0.785	0.752	0.773
		(ii)	π_2	0.370	0.346	0.410	0.384	0.371	0.372
			π_4	0.106	0.085	0.113	0.124	0.109	0.116
	(III)	(i)	π_2	0.579	0.529	0.805	0.657	0.654	0.740
		(ii)	π_2	0.099	0.118	0.152	0.120	0.115	0.264
			π_4	0.257	0.204	0.396	0.291	0.133	0.345
	(IV)	(i)	π_2	0.693	0.838	0.529	0.236	0.486	0.271
		(ii)	π_2	0.087	0.143	0.080	0.050	0.090	0.056
			π_4	0.383	0.535	0.299	0.109	0.219	0.131

power performance when the null censoring probability is light as 0.3, while the WKM test outperforms over the other two when the null censoring probability is about 0.5. This is because that the WKM test puts more weight on early times for heavy censored data. For middle occurring hazard difference alternatives, although the WKM test is better than the WLR(L) test as specified in Pepe and Fleming (1989), the WLR(P) test is superior to the WKM test. For late difference hazard alternatives, the WLR(L) has the highest power. The WKM test is second to the WLR(L) test when the null censoring probability is light as 0.3. However, when the null censoring probability is about 0.5, the WKM test puts less weight on late times, thereby reducing its power for detecting the late occurring hazard differences. In this case, the WKM test is even less powerful than the WLR(P) test.

6. Conclusions

To use the weighted logrank statistics (Fleming and Harrington, 1991) in constructing the class of multiple test for comparing successive treatments, the most important issue is how to choose appropriate weight functions. The logrank statistic ($\rho = \gamma = 0$) is known to be optimal under proportional hazards alternatives and the Peto–Prentice statistic ($\rho = 1$ and $\gamma = 0$) is suitable for early occurring hazard differences. Moreover, appropriate weight function would be the one corresponding to $\rho = 1$ and $\gamma = 1$ for hazard differences occurring at middle times, and $\rho = 0$ and $\gamma = 1$ for late hazard differences. Some useful plots, for example, the plot of log{-log(survival estimate)}, can be used to assess the feasibility of the proportional hazards. The Kaplan–Meier survival estimates can also be used to investigate whether the hazards differ at early, middle or late times. Furthermore, although we only consider, for simplicity, the use of the same type of weight function in this paper, we can, in fact, employ different types of weight functions for comparing different pairs of adjacent treatments in the weighted logrank multiple test.

The multipe test on the basis of weighted Kaplan–Meier statistics does not lose too much power than the logrank multiple test for proportional hazards alternatives. In addition, the power performance of the weighted Kaplan–Meier multiple test is competitive for some nonproportional hazards alternatives. However, to use the weighted Kaplan–Meier multiple test, we still need to select the weight function satisfying the constraints specified in Pepe and Fleming (1989) to ensure stability of the weighted Kaplan–Meier statistic.

According to the observations stated above, we learn that the weighted Kaplan–Meier multiple test may not be the best one, although it would not be the worst one in most cases. Moreover, the weight function previously chosen in (3.3), for example, involves the estimators for the censoring distributions, which seems to be a little bit curious. In contrast to this, the weighted logrank multiple test does not use the censoring estimator and the weight functions give the statistician the chance to make the test sensitive to the corresponding hazard differences. For these reasons, the weighted logrank multiple

test with appropriate weight functions is preferred for comparing successive treatments if the times at which adjacent hazards are different can be recognized clearly.

Acknowledgements

The work of the first author was partially supported by the National Science Council of Taiwan under Grant NSC86-2115-M-008-018. We would like to thank the editor and two referees for many helpful suggestions which improve the presentation in this paper. We would also like to thank Mr. Min-Shiao Tsai for his computing assistance.

Appendix A

A.1. Asymptotic null distribution of $N^{-1/2}(\mathrm{WLR}_1, \mathrm{WLR}_2, \ldots, \mathrm{WLR}_{k-1})$

Note that, when $S_1 = S_2 = \cdots = S_k$, using the martingale framework, the statistic WLR_i in (2.1) can be written as

$$\mathrm{WLR}_i = \int_0^{t_c} \frac{K_i(t)}{Y_i(t)}\,\mathrm{d}M_i(t) - \int_0^{t_c} \frac{K_i(t)}{Y_{i+1}(t)}\,\mathrm{d}M_{i+1}(t),$$

where $K_i(t) = W_i(t)Y_i(t)Y_{i+1}(t)/Y_{i+}(t)$, $M_i(t) = D_i(t) - \int_0^t Y_i(s)\,\mathrm{d}\Lambda(s)$ are independent zero-mean martingales and $\Lambda(\mathrm{s})$ is the common cumulative hazard function. Suppose that $N \to \infty$ in such a way that $n_i/N \to p_i$, $0 < p_i < 1$, $i=1,2,\ldots,k$. If $Y_i(t)/n_i \xrightarrow{\mathrm{p}} \pi_i(t)$, $i = 1,\ldots,k$, and $W_i(t) \xrightarrow{\mathrm{p}} \omega_i(t)$ uniformly as $N \to \infty$, then $K_i^2(t)Y_{i+}(t)]/[NY_i(t)Y_{i+1}(t)] \xrightarrow{\mathrm{p}} k_{ii}(t)$ and $K_i(t)K_{i+1}(t)/[NY_{i+1}(t)] \xrightarrow{\mathrm{p}} k_{i+}(t)$ uniformly as $N \to \infty$ for $i = 1,2,\ldots,k-1$, where

$$k_{ii}(t) = p_i p_{i+1}\omega_i^2(t)\pi_i(t)\pi_{i+1}(t)/[p_i\pi_i(t) + p_{i+1}\pi_{i+1}(t)]$$

$$k_{i+}(t) = p_i p_{i+1} p_{i+2}\omega_i(t)\omega_{i+1}(t)\pi_i(t)\pi_{i+1}(t)\pi_{i+2}(t)/\{[p_i\pi_i(t) + p_{i+1}\pi_{i+1}(t)] \times[p_{i+1}\pi_{i+1}(t) + p_{i+2}\pi_{i+2}(t)]\}.$$

Hence, the Martingale Central Limit Theorem (see, for example, Theorem 6.2.1 in Fleming and Harrington, 1991) implies that, the null (H_0) asymptotic distribution of the random vector $N^{-1/2}(\mathrm{WLR}_1, \mathrm{WLR}_2, \ldots, \mathrm{WLR}_{k-1})$ is the $(k-1)$-variate normal with mean zero and covariance matrix $\Sigma = \{\sigma_{ij}\}$, where, for $i = 1,2,\ldots,k-1$,

$$\sigma_{ii} = \int_0^\infty k_{ii}(t)\{1 - \Delta\Lambda(t)\}\,\mathrm{d}\Lambda(t), \tag{A.1}$$

and, for $i < j = 2,\ldots,k-1$,

$$\sigma_{ij} = \begin{cases} -\displaystyle\int_0^\infty k_{i+}(t)\{1 - \Delta\Lambda(t)\}\,\mathrm{d}\Lambda(t) & \text{if } i = j-1, \\ 0 & \text{otherwise,} \end{cases} \tag{A.2}$$

with $\Delta\Lambda(t)=\Lambda(t)-\Delta\Lambda(t-)$. The unbiased and consistent estimators of $N\sigma_{ii}$ and $N\sigma_{ij}$ are then given by s_{ii} and s_{ij} stated in Eqs. (2.2) and (2.3), respectively. Under the assumption of equal censoring, that is, $C_i=C$, $i=1,2,\ldots,k$, we obtain, for $i=1,2,\ldots,k-1$,

$$\sigma_{ii}=\frac{p_i p_{i+1}}{p_i+p_{i+1}}\int_0^{\infty}\omega_i^2(t)G(t)\mathrm{d}F(t) \tag{A.3}$$

and, for $i<j=2,3,\ldots,k-1$,

$$\sigma_{ij}=\frac{-p_i p_{i+1} p_{i+2}}{[p_i+p_{i+1}][p_{i+1}+p_{i+2}]}\int_0^{\infty}\omega_i(t)\omega_{i+1}(t)G(t)\,\mathrm{d}F(t), \tag{A.4}$$

if $i=j-1$, and 0, otherwise, where $F(t)=1-S(t)$ and $G(t)=1-C(t)$.

A.2. Asymptotic null distribution of $\mathbf{N^{-1/2}}$ *(*$\mathbf{WKM_1}$, $\mathbf{WKM_2,\ldots,WKM_{k-1}}$*)*

When $S_1=S_2=\cdots=S_k=S$, the weighted Kaplan–Meier statistic WKM_i can be expressed as

$$\mathrm{WKM}_i=\int_0^{T_c}\hat{W}_i(t)\,\mathrm{d}\{\hat{S}_{i+1}(t)-\hat{S}_i(t)\}$$

where $\hat{W}_i(t)=\sqrt{n_i n_{i+1}}\int_t^{T_c}\hat{w}_i(u)S(u)\,\mathrm{d}u$ is a predictable weight function. Applying Lemma 2.4.1 in Fleming and Harrington (1991), the martingale representation of $N^{-1/2}\mathrm{WKM}_i$ is given by

$$\begin{aligned}&\int_0^{T_c}H_i(t)\,\mathrm{d}M_i(t)-\int_0^{T_c}H_{i+1}(t)\,\mathrm{d}M_{i+1}(t)\\&\quad+\sqrt{\frac{n_i n_{i+1}}{N}}\int_0^{T_c}\left\{\int_t^{T_c}\hat{w}_i(u)S(u)\,\mathrm{d}u\right\}\mathrm{d}\left\{\frac{B_{i+1}(t)-B_i(t)}{S(t)}\right\}\end{aligned} \tag{A.5}$$

where

$$H_i(t)=\sqrt{\frac{n_i n_{i+1}}{N}}\left\{\int_t^{T_c}\hat{w}_i(u)S(u)\,\mathrm{d}u\right\}\frac{\hat{S}_i(t)}{S(t)}\frac{I\{Y_i(t)>0\}}{Y_i(t)}$$

and

$$B_i(t)=I\{T_c<t\}\frac{\hat{S}_i(T_c)\{S(T_c)-S(t)\}}{S(T_c)}.$$

Since the last term in (A.5) converges to zero in probability as $N\to\infty$, we observe

$$N^{-1/2}\,\mathrm{WKM}_i=\int_0^{T_c}H_i(t)\,\mathrm{d}M_i(t)-\int_0^{T_c}H_{i+1}(t)\mathrm{d}M_{i+1}(t)+\mathrm{o_p}(1).$$

Suppose that $Y_i(t)/n_i\xrightarrow{\mathrm{p}}\pi_i(t)$ and $\hat{w}_i(t)\xrightarrow{\mathrm{p}}w_i(t)$ uniformly for $i=1,2,\ldots,k-1$. Then, $H_i^2(t)Y_i(t)\xrightarrow{\mathrm{p}}h_i(t)$ uniformly, where

$$h_i(t)=p_{i+1}\left\{\int_t^{\tau}w_i(u)S(u)\,\mathrm{d}u\right\}^2\frac{S_i^2(t)}{S^2(t)}\frac{1}{\pi_i(t-)},\quad i=1,2,\ldots,k-1.$$

Hence, the Martingale Central Limit Theorem implies that, the null ($\mathrm{H_0}$) asymptotic distribution of the random vector $N^{-1/2}(\mathrm{WKM}_1,\mathrm{WKM}_2,\ldots,\mathrm{WKM}_{k-1})$ is the

$(k-1)$-variate normal with mean zero and covariance matrix $\Psi = \{\phi_{ij}\}$, where for $i = 1, 2, \ldots, k-1$,

$$\phi_{ii} = \int_0^\tau \left\{ \int_t^\tau w_i(u) S(u)\, \mathrm{d}u \right\}^2 \frac{p_i \pi_i(t-) + p_{i+1}\pi_{i+1}(t-)}{\pi_i(t-)\pi_{i+1}(t-)} \{1 - \Delta\Lambda(t)\}\, \mathrm{d}\Lambda(t), \tag{A.6}$$

and for $i < j = 2, 3, \ldots, k-1$

$$\phi_{ij} = \int_0^\tau \sqrt{p_i p_{i+2}} \left\{ \int_t^\tau w_i(u) S(u)\, \mathrm{d}u \right\} \left\{ \int_t^\tau w_{i+1}(u) S(u)\, \mathrm{d}u \right\} \times \frac{1}{\pi_{i+1}(t-)} \{1 - \Delta\Lambda(t)\}\, \mathrm{d}\Lambda(t), \tag{A.7}$$

if $i = j-1$ and 0, otherwise. Consistent estimators of $N\phi_{ii}$ and $N\phi_{ij}$ are then obtained as stated in (3.2) and (3.4), respectively. Assume that $C_i = C$, $i = 1, 2, \ldots, k$. Let $\tau = \sup\{t:\ \min(S(t), G(t)) \geqslant 0\}$, where $G(t) = 1 - C(t)$. For the weight function given in (3.3), we obtain for $i = 1, 2, \ldots, k-1$,

$$\phi_{ii} = \frac{1}{p_i + p_{i+1}} \int_0^\tau \left\{ \int_t^\tau G(u) S(u)\, \mathrm{d}u \right\}^2 \frac{\mathrm{d}F(t)}{S(t)S(t-)G(t-)} \tag{A.8}$$

and for $i < j = 2, 3, \ldots, k-1$,

$$\phi_{ij} = \frac{-\sqrt{p_i p_{i+2}}}{(p_i + p_{i+1})(p_{i+1} + p_{i+2})} \int_0^\tau \left\{ \int_0^\tau G(u) S(u)\, \mathrm{d}u \right\}^2 \frac{\mathrm{d}F(t)}{S(t)S(t-)G(t-)}, \tag{A.9}$$

if $i = j-1$, and 0, otherwise.

References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference Under Order Restrictions. Wiley, New York.

Budde, M., Bauer, P., 1989. Multiple test procedures in clinical dose finding studies. J. Amer. Statist. Assoc. 84, 792–796.

Chi, Y., Chen, Y.I., 1998. Weighted Kaplan–Meier tests for ordered alternatives with right-censored survival data. Technical no. 98-1, Department of Statistics, National Cheng-Kung University.

Fleming, T.R., Harrington, D.P., 1991. Counting Process and Survival Analysis. Wiley, New York.

Hornung, R.L., Longo, D.L., Bowersox, O.C., Kwak, L.W., 1995. Tumor antigen-specific immunization of bone marrow transplantation donors as adoptive therapy against established tumor. J. Natl. Cancer Inst. 87, 1289–1296.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimator from incomplete observations. J. Amer. Statist. Assoc. 53, 457–481.

Lee, R.E., Spurrier, J.D., 1995a. Successive comparisons between ordered treatments. J. Statist. Plann. Inference 43, 323–330.

Lee, R.E., Spurrier, J.D., 1995b. Distribution-free multiple comparisons between successive treatments. J. Nonparametric Statist. 5, 26–273.

Liu, P.Y., Green, S., Wolf, M., Crowley, J., 1993. Testing against ordered alternatives for censored survival data. J. Amer. Statist. Assoc. 188, 153–160.

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

Pepe, M.S., Fleming, T.R., 1989. Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. Biometrics 45, 497–507.

Pepe, M.S., Fleming, T.R., 1991. Weighted Kaplan–Meier statistics: large sample and optimality considerations. J. Roy. Statist. Soc. B 53, 341–352.

Peto, R., Peto, J., 1972. Asymptotically efficient rank invariant test procedures (with discussion). J. Roy. Statist. Soc. A 135, 185–206.

Prentice, R.L., 1978. Linear rank tests with right censored data. Biometrika 65, 165–179.

Schervish, M.J., 1984. Multivariate Normal Probabilities with error bound. Appl. Statist. 33, 81–94.

van Eeden, C., 1960. A class of tests for the hypothesis that k parameters $\theta_1, \ldots, \theta_k$ satisfy the inequalities $\theta_1 \leqslant \cdots \leqslant \theta_k$. Bull. Inter. Statist. Inst. 27, Book 3, 10–30.