STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:23–35
Published online 28 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1733

Multiple testing procedures based on weighted Kaplan–Meier statistics for right-censored survival data

Yunchan Chi*,†

Department of Statistics, National Cheng-Kung University, Tainan, Taiwan 701, R.O.C.

SUMMARY

In clinical trials or drug development studies, researchers are often interested in identifying which treatments or dosages are more effective than the standard one. Recently, several multiple testing procedures based on weighted logrank tests have been proposed to compare several treatments with a control in a one-way layout where survival data are subject to random right-censorship. However, weighted logrank tests are based on ranks, and these tests might not be sensitive to the magnitude of the difference in survival times against a specific alternative. Therefore, it is desirable to develop a more robust and powerful multiple testing procedure. This paper proposes multiple testing procedures based on two-sample weighted Kaplan–Meier statistics, each comparing an individual treatment with the control, to determine which treatments are more effective than the control. The comparative results from a simulation study are presented and the implementation of these methods to the prostate cancer clinical trial and the renal carcinoma tumour study are presented. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: multiple testing procedure; weighted Kaplan–Meier statistic; right-censored data

1. INTRODUCTION

When different treatments (therapies) are compared with a standard treatment or placebo in comparative clinical trials, the researchers are often interested in identifying which treatment effects exceed the placebo in prolonging the survival times of patients. However, randomly right-censored data are frequently observed in the clinical trials, since the study may be terminated at a preassigned time due to time limitation, the death may be attributed to competing risks, which is not of interest in the present study, or subjects may be randomly lost to follow-up. For example, a double blind randomized clinical trial comparing four treatments for patients with prostate cancer in stages III and IV was conducted. The data had previously been analysed by Byar and Corle [1] and Byar and Green [2]. The treatments were placebo

*Correspondence to: Yunchan Chi, Department of Statistics, National Cheng-Kung University, Tainan, Taiwan 701, R.O.C.
†E-mail: ycchi@email.stat.ncku.edu.tw

Contract/grant sponsor: National Science Council of Taiwan; contract/grant number: NSC89-2115-M-006-021

Received March 2002
Accepted August 2003

pill, 0.2 mg diethylstilbestrol (DES), 1.0 mg of DES, or 5.0 mg of DES, with all drugs administered daily by mouth. One of the goals of this study is to identify effective treatment dosage as compared with the placebo. Thus many-to-one comparisons with randomly right-censored survival data are needed.

Multiple testing procedures for the many-to-one comparisons problem with right-censored data have recently been developed. For example, Chakraborti and Desu [3] used Slepian's [4] inequality to suggest a conservative multiple testing procedure based on Gehan's [5] two-sample statistics, each comparing an individual treatment with the control. Chen [6] proposed a generalization of Steel's [7] test on the basis of the maximum of several two-sample Gehan's statistics under the special case of equal censoring distributions. Chen [8] further extended Steel's test and the closed testing procedure of Marcus *et al.* [9] for many-to-one comparisons based on two-sample weighted logrank statistics, each comparing an individual treatment with the control. Since the weighted logrank statistic is a function of the difference of Nelson [10]–Aalen [11] cumulative hazard functions, it might not be sensitive to the magnitude of the difference in survival times against a specific alternative. As a result, Pepe and Fleming [12, 13] developed a class of test statistics based on the sum of weighted differences in Kaplan–Meier [14] estimators of survival functions. Moreover, they showed that the weighted Kaplan–Meier test is competitive with the logrank test under proportional hazards alternative, and may perform better under non-proportional hazards alternatives. Therefore, in this paper we consider generalizing the Steel's test and the closed testing procedure based on two-sample Kaplan–Meier statistics, each comparing an individual treatment with the control for many-to-one comparisons.

The remainder of this paper is organized as follows. Section 2 reviews Chen's generalization of Steel's test, and Marcus *et al.* s' procedures based on two-sample weighted logrank tests. The proposed many-to-one testing procedures are developed in Section 3. Section 4 reports the comparative results of identifying treatment effects from Monte Carlo study. The implementation of the proposed method is illustrated through two examples based on real data in Section 5. Finally, the last section gives concluding remarks.

2. MULTIPLE TESTING PROCEDURES BASED ON WEIGHTED LOGRANK TESTS

Let $t_1 < t_2 < \cdots < t_{N_i}$ be the distinct ordered observed failure times based on the pooled data of the control and the ith group, $i = 1, 2, \ldots, k$. Let d_{0j} and d_{ij} be the number of failure events of the control group and the ith group occurring at time t_j, $j = 1, 2, \ldots, N_i$, respectively. Similarly, let n_{0j} and n_{ij} be the number of subjects at risk at time t_j of the control group and the ith group, respectively. Let $e_{ij} = n_{ij}(d_{0j} + d_{ij})/(n_{0j} + n_{ij})$ be the number of expected failure events occurring at time t_j of the ith group. Accordingly, the weighted logrank statistic comparing the ith treatment with the control (0th treatment) is

$$U_{0i} = \sum_{j=1}^{N_i} W_{ij}(d_{ij} - e_{ij})$$

where W_{ij} are weights chosen to let U_{0i} be more sensitive to detect certain alternatives. For example, $W_{ij} = n_{0j} + n_{ij}$ yields Gehan–Wilcoxon statistics [5] having higher power in detecting early hazard differences alternative; $W_{ij} = 1$ is logrank statistic [15] that is known to be the locally most powerful test against proportional hazards model. Furthermore, for $\rho,\gamma \geqslant 0$, $W_{ij}=\{\hat{S}_{0i}(t_j)\}^{\rho}\{1-\hat{S}_{0i}(t_j)\}^{\gamma}$ is an extension version of the weight function from Fleming and Harrington [16], where $\hat{S}_{0i}(t)$ are the Kaplan–Meier estimates of survival functions based on the pooled data of the control and the ith sample. Note that taking $\rho=\gamma=0$ produces the logrank statistic, while setting $\rho=1$ and $\gamma=0$ yields the Peto–Prentice–Wilcoxon statistic [17, 18], which is also very sensitive to early hazard differences alternative. Moreover, the consistent estimator of the variance of U_{0i} is given by

$$s_{ii}=\sum_{j=1}^{N_i} W_{ij}^2 \frac{n_{0j}n_{ij}(d_{0j}+d_{ij})(n_{0j}+n_{ij}-d_{0j}-d_{ij})}{(n_{0j}+n_{ij})^2(n_{0j}+n_{ij}-1)} \tag{1}$$

Let $U_i=U_{0i}/\sqrt{s_{ii}}$, $i=1,2,\ldots,k$. It can be shown (see Reference [8]) that, under the null hypothesis $H_0:(S_0(t)=S_i(t),\ i=1,2,\ldots,k)$, the asymptotic distribution of the random vector $U=(U_1,U_2,\ldots,U_k)$ is the k-variate normal with mean zero vector and correlation matrix R, which can be consistently estimated by $\hat{R}=\{s_{ir}/\sqrt{s_{ii}s_{rr}}\}$, where the s_{ii} are stated in (1) and s_{ir}, for $i\neq r$, are given by

$$s_{ir}=\sum_{j=1}^{N_{ir}} W_{ij}W_{rj}\frac{n_{0j}n_{ij}n_{rj}(d_{0j}+d_{ij}+d_{rj})(n_{0j}+n_{ij}+n_{rj}-d_{0j}-d_{ij}-d_{rj})}{(n_{0j}+n_{ij})(n_{0j}+n_{rj})(n_{0j}+n_{ij}+n_{rj})^2}$$

where N_{ir} denotes the largest distinct observed failure time based on the pooled data of the control, ith and rth samples.

Let $(Z_1,Z_2,\ldots,Z_k)$ be a k-variate normal vector with mean zero and correlation matrix R, and let $z\max(k,\alpha)$ be the upper αth percentile of the distribution of $\max(Z_1,Z_2,\ldots,Z_k)$ with R estimated by $\hat{R}$. Previously, Chen [8] proposed to claim $S_i(t)>S_0(t)$ for all t, if $U_i\geqslant z\max(k,\alpha)$, for $i=1,2,\ldots,k$, in order to determine which treatments are better than the control, based on the generalization of Steel's many-to-one comparison procedure (SMAX). It is clear that the experiment-wise error rate (the probability of erroneously declaring at least one treatment better than the control) for this procedure is approximately controlled, because

$$\begin{aligned}\alpha &\approx P\{\max(U_1,U_2,\ldots,U_k)\geqslant z\max(k,\alpha)|H_0\}\\ &=P\{U_i\geqslant z\max(k,\alpha)\text{ for at least one } i\,|\,H_0\}\end{aligned}$$

For any z, the probability $P\{\max(Z_1,Z_2,\ldots,Z_k)\leqslant z\}$ can be computed using a program for calculating multivariate normal probabilities [19]. Therefore, the critical value $z\max(k,\alpha)$ can be found such that $P\{\max(Z_1,Z_2,\ldots,Z_k)\geqslant z\max(k,\alpha)\}=\alpha$.

Because a stepwise comparison is generally more powerful than a single-step procedure, Chen [8] further applied the closed testing procedure proposed by Marcus *et al.* [9] to determine which treatments are more effective than the control with right-censored data. For the problem considered herein, the class of null hypotheses $H=H_0(P)$, where $H_0(P)=\{S_0(t)=S_i(t),\text{ for } i\in P\}$ and P is any subset of positive integers $\{1,2,\ldots,k\}$, is closed under intersection in the sense that both $H_0(P)$ and $H_0(Q)$ in H implies their intersection also in H. Let $U_{(1)}<U_{(2)}<\cdots<U_{(k)}$ be the ordered U_i's and let d_i be the position of $U_{(i)}$ in the vector U, that is, $U_{(i)}=U_{d_i}$ and let $z\max((i),\alpha)$ denote the upper αth percentile of the distribution of $\max(Z_{d_1},Z_{d_2},\ldots,Z_{d_i})$. A closed testing procedure (CLOSE), separating α-level tests of individual $H_0(P)$ applied in the stepdown manner, is obtained which claims $S_{d_i}(t)>S_0(t)$ for all t, if $U_{(i)} \geqslant z \max((i), \alpha)$, for $i = 1, 2, \ldots, k$, provide that $U_{(j)} \geqslant z \max((j), \alpha)$ for $j > i$. This procedure continues until $U_{(m)} < z \max((m), \alpha)$, for some $m = 1, 2, \ldots, k$. Then, Chen determined at an approximate experiment-wise error rate α, that only the treatments labelled by $d_{m+1}, \ldots, d_k$ are more effective than the control.

3. MULTIPLE TESTING PROCEDURES BASED ON WEIGHTED KAPLAN–MEIER TESTS

Considering the disadvantages of the weighted logrank statistics, Pepe and Fleming [12, 13] proposed a two-sample test based on the differences of the Kaplan–Meier survival estimators

$$\mathrm{WKM}_{0i} = \sqrt{\frac{n_0 n_i}{n_0 + n_i}} \sum_{j=1}^{N_i - 1} \hat{w}_i(t_j)(\hat{S}_i(t_j) - \hat{S}_0(t_j))(t_{j+1} - t_j)$$

where $\hat{S}_0(t)$ and $\hat{S}_i(t)$ are the Kaplan–Meier estimates of the survival functions for the control and group i, respectively, and $\hat{w}_i(t)$ is a random weight function which downweights the difference between $\hat{S}_i(t)$ and $\hat{S}_0(t)$ over later time periods for heavy censoring so that the statistic WKM_{0i} is stable. Let $\hat{C}_0(t)$ and $\hat{C}_i(t)$ be the Kaplan–Meier estimates of the survival functions based on censoring times for the control group and the ith group, respectively. Furthermore, let $A_{ij} = \sum_{m=j}^{N_i - 1} \hat{w}_i(t_m)\hat{S}_p(t_m)(t_{m+1} - t_m)$ where $\hat{S}_p(t_m)$ is the Kaplan–Meier estimate of the common survival function based on the control and ith samples. Then under the global null hypothesis $H_0 : (S_0(t) = S_i(t),\ i = 1, 2, \ldots, k)$, a consistent estimator for the variance of WKM_{0i} is given by

$$v_{ii} = \sum_{j=1}^{N_i - 1} \frac{A_{ij}^2}{\hat{S}_p(t_j)\hat{S}_p(t_{j-1})} \frac{n_0 \hat{C}_0(t_{j-1}) + n_i \hat{C}_i(t_{j-1})}{(n_0 + n_i)\hat{C}_0(t_{j-1})\hat{C}_i(t_{j-1})} (\hat{S}_p(t_{j-1}) - \hat{S}_p(t_j)) \tag{2}$$

where n_0 and n_i are sample sizes of the control and group i, respectively. Let $\hat{p}_i$ be n_i/n, $i = 0, 1, 2, \ldots, k$. The weight function

$$\hat{w}_i(t) = \frac{\hat{C}_0(t-)\hat{C}_i(t-)}{\hat{p}_0 \hat{C}_0(t-) + \hat{p}_i \hat{C}_i(t-)} \tag{3}$$

suggested by Pepe and Fleming [12] is a competitor to the logrank test for the proportional hazards alternative, and may perform better than the logrank test for crossing hazards alternative. Thus, we consider multiple test procedures based on Kaplan–Meier test with the above weight function for further comparisons in this paper.

Let $L_i = \mathrm{WKM}_{0i}/\sqrt{v_{ii}}$, $i = 1, 2, \ldots, k$. It can also be shown (see Reference [20]) that, under the global null hypothesis the asymptotic distribution of the random vector $L = (L_1, L_2, \ldots, L_k)$ is the k-variate normal with mean zero vector and correlation matrix $\Gamma = \{\rho_{ir}\}$; and it can be consistently estimated by $\hat{\Gamma} = \{v_{ir}/\sqrt{v_{ii} v_{rr}}\}$, where v_{ii} are stated in (2) and for $i \neq r$, v_{ir} are given by

$$v_{ir} = \sqrt{n_i n_r} \sum_{j=1}^{N_{ir} - 1} \frac{A_{ij} A_{rj}}{\hat{S}_p(t_j)\hat{S}_p(t_{j-1})} \frac{1}{\hat{C}_0(t_{j-1})} (\hat{S}_p(t_{j-1}) - \hat{S}_p(t_j))$$

where $\hat{S}_P(t_j)$ is the Kaplan–Meier estimator computed from the combined samples of the control, i and r.

As an extension of the SMAX testing procedure reviewed in Section 2, in order to identify which treatments are more effective than the control we claim $S_i(t) > S_0(t)$ for all t, if $L_i \geqslant z \max *(k,\alpha)$, for $i = 1,2,\ldots,k$, where $z \max *(k,\alpha)$ is the upper αth percentile of the distribution of the largest order statistic from k-variate normal vector $(Z_1, Z_2, \ldots, Z_k)$ with mean zero and the correlation matrix estimated by $\hat{\Gamma}$. Moreover, a stepwise comparison proposed by Marcus *et al.* [9] can be constructed based on weighted Kaplan–Meier statistics. Let $L_{(1)} < L_{(2)} < \cdots < L_{(k)}$ be the ordered L_i's and, again, let d_i be the position of $L_{(i)}$ in the vector L, that is, $L_{(i)} = L_{d_i}$ and let $z \max *((i),\alpha)$ denote the upper αth percentile of the distribution of $\max(Z_{d_1}, Z_{d_2}, \ldots, Z_{d_i})$. A closed stepdown testing procedure is obtained which claims $S_{d_i}(t) > S_0(t)$ for all t, if $L_{(i)} \geqslant z \max *((i),\alpha)$, for $i = 1,2,\ldots,k$, provide that $L_{(j)} \geqslant z \max *((j),\alpha)$ for $j > i$. This procedure continues until $L_{(m)} < z \max *((m),\alpha)$, for some $m = 1,2,\ldots,k$, and at an approximate experiment-wise error rate α, declares that only the treatments labelled by $d_{m+1},\ldots,d_k$ are more effective than the control.

4. SIMULATION STUDY

To investigate the performances of SMAX and CLOSE multiple testing procedures based on weighted Kaplan–Meier tests in identifying which treatments are more effective than the control, a simulation study is carried out. Table I examines experiment-wise error rates for the proposed test statistics; such as WKM with the weight function stated in (3) and WKMs with square root of (3) as the weight function, and the weighted logrank tests; including logrank test (LR), and Peto–Prentice–Wilcoxon test (PPW), under equal sample sizes with 20 observations in each group. Exponential survival distribution with scale parameter 1 and lognormal survival distribution with zero normal mean and standard deviation 0.5 are selected for examining type I error rate. Moreover, the survival functions generated by piecewise exponential distributions with different hazards at different periods are also investigated for examining type I error rate. The solid line in each panel of Figure 1 represents the common survival function under the null hypothesis. The censoring distribution is uniform $(0,b)$ with various values of b corresponding to the percentage of censoring as 0.3 and 0.5. Note that under equal sample sizes

Table I. Estimated experiment-wise error rates for $\alpha = 0.05$, uniform censoring distribution $\mathrm{U}(0,b)$ and $n_0 = n_1 = n_2 = n_3 = 20$.

Survival distribution	b	LR	PPW	WKM	WKMs
Exponential (I)	3.197	0.061	0.056	0.060	0.062
	1.593	0.055	0.053	0.056	0.058
Lognormal (II)	3.768	0.048	0.052	0.046	0.046
	2.200	0.050	0.048	0.051	0.052
Piecewise exponential (III)	3.654	0.052	0.052	0.054	0.055
	1.457	0.048	0.048	0.047	0.047
Piecewise exponential (IV)	3.119	0.048	0.048	0.047	0.046
	1.861	0.054	0.052	0.055	0.055

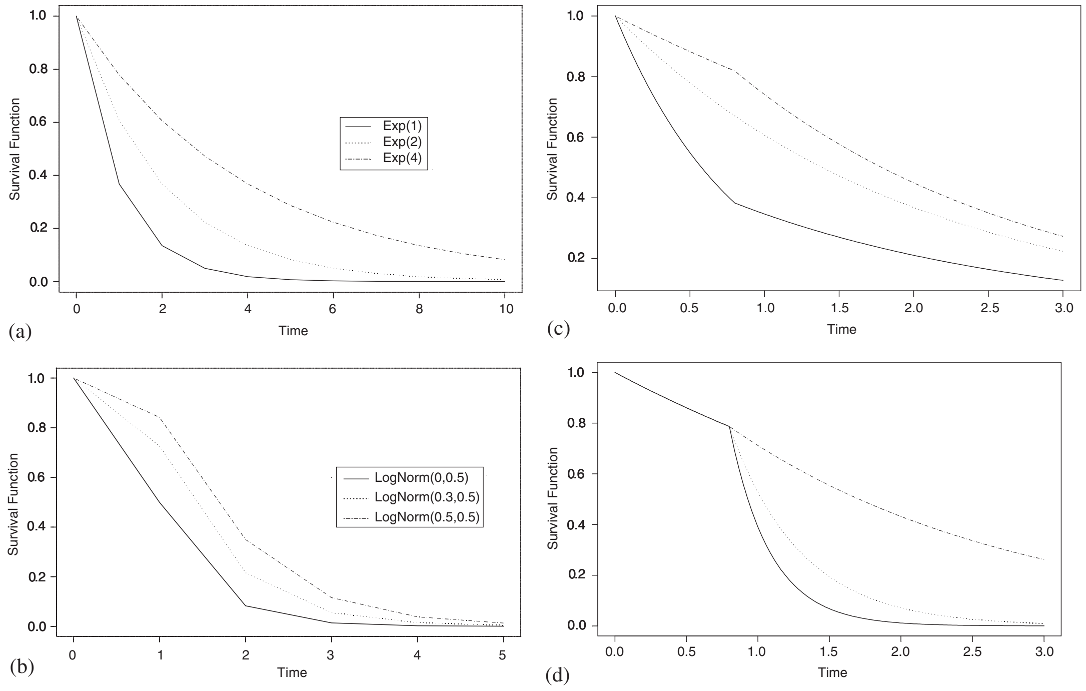

Figure 1. Survival function configurations for simulation study, (a) Exponential survival functions, (b) Lognormal survival functions, (c) $\lambda_1(t) = 1.2I\{t \leqslant 0.8\} + 0.5I\{t > 0.8\}$; $\lambda_2(t) = 0.5I\{t \leqslant 0.8\} + 0.5I\{t > 0.8\}$; $\lambda_3(t) = 0.25I\{t \leqslant 0.8\} + 0.5I\{t > 0.8\}$, (d) $\lambda_1(t) = 0.3I\{t \leqslant 0.8\} + 3.5I\{t > 0.8\}$; $\lambda_2(t) = 1.0I\{t \leqslant 0.5\} + 2.5I\{t > 0.5\}$; $\lambda_3(t) = 1.0I\{t \leqslant 0.5\} + 2.0I\{t > 0.5\}$.

and equal censoring patterns, the off-diagonal elements in correlation matrix are 0.5 for all the tests examined here. Hence, the critical values, z max$(3, 0.05) = 2.063$, z max$(2, 0.05) = 1.917$, z max$(1, 0.05) = 1.645$, found in Reference [21] are used in simulation study.

Note that the standard error based on 5000 replicates is about 0.003. Thus, from Table I, the estimated experiment-wise error rates of the LR, WKM, and WKMs tests are slightly higher than the pre-specified significance level of 0.05 under exponential distribution. But the estimated error rates of all the tests examined here are within the reasonable range under lognormal and piecewise exponential distributions.

Tables II and III present the estimated comparison-wise powers (probability of correctly detecting all treatments which are better than the control) of SMAX and CLOSE multiple testing procedures for $\alpha = 0.05$ under equal sample sizes with 20 observations in each group. Exponential distributions with different values of scale parameter θ_i's and lognormal distributions with same standard deviation 0.5 and various mean values θ_i's are selected for power comparisons, and the results are listed in Table II. Note that exponential distribution is a natural model for the proportional hazards model and lognormal distribution generates survival times from the non-proportional hazards model. In addition, piecewise exponential distributions with different hazards at different time periods corresponding to early or late hazard differences alternatives are also selected for power comparison and the results are listed in Table III. The

Table II. Estimated comparison-wise powers of SMAX or CLOSE tests for $\alpha = 0.05$, uniform censoring distribution U$(0, b)$ and $n_0 = n_1 = n_2 = n_3 = 20$.

					SMAX				CLOSE			
b	θ_0	θ_1	θ_2	θ_3	LR	PPW	WKM	WKMs	LR	PPW	WKM	WKMs
Exponential survival distribution												
3.20	1	1	1	4	0.795	0.749	0.782	0.791	0.782	0.738	0.769	0.778
	1	1	2	4	0.309	0.257	0.301	0.308	0.333	0.294	0.334	0.340
	1	1	4	4	0.692	0.625	0.677	0.686	0.713	0.653	0.699	0.711
	1	2	4	4	0.317	0.252	0.309	0.318	0.472	0.416	0.470	0.477
	1	4	4	4	0.640	0.551	0.618	0.636	0.797	0.730	0.782	0.790
1.59	1	1	1	4	0.578	0.557	0.562	0.578	0.566	0.543	0.549	0.566
	1	1	2	4	0.182	0.164	0.179	0.187	0.213	0.191	0.206	0.212
	1	1	4	4	0.451	0.418	0.424	0.445	0.488	0.451	0.453	0.476
	1	2	4	4	0.189	0.160	0.176	0.190	0.332	0.299	0.315	0.326
	1	4	4	4	0.360	0.321	0.336	0.354	0.542	0.493	0.513	0.534
Lognormal survival distribution												
3.77	0	0	0	0.5	0.634	0.667	0.678	0.662	0.621	0.653	0.663	0.645
	0	0	0.3	0.5	0.248	0.257	0.267	0.262	0.275	0.292	0.302	0.292
	0	0	0.5	0.5	0.502	0.536	0.545	0.531	0.529	0.572	0.576	0.564
	0	0.3	0.5	0.5	0.243	0.245	0.261	0.254	0.373	0.391	0.413	0.403
	0	0.5	0.5	0.5	0.450	0.472	0.488	0.475	0.613	0.652	0.670	0.659
2.20	0	0	0	0.5	0.517	0.533	0.553	0.546	0.506	0.513	0.539	0.530
	0	0	0.3	0.5	0.195	0.197	0.215	0.214	0.214	0.222	0.246	0.244
	0	0	0.5	0.5	0.377	0.394	0.414	0.408	0.408	0.421	0.449	0.443
	0	0.3	0.5	0.5	0.171	0.175	0.190	0.190	0.298	0.316	0.340	0.335
	0	0.5	0.5	0.5	0.319	0.324	0.354	0.345	0.483	0.504	0.538	0.528

survival functions of these alternatives corresponding to $S_0 = S_1 < S_2 < S_3$ are presented in Figure 1. The common censoring distribution is uniform $(0, b)$ with $b = 3.20$ (3.77) and $b = 1.59$ (2.20) corresponding to low and moderate censoring rates, respectively, for exponential (lognormal) distributions. Likewise, $b = 3.645$ (3.119) and $b = 1.457$ (1.861) correspond to low and moderate censoring rates, respectively, for early (late) hazards difference alternatives.

For all the tests investigated for the level study, the comparison-wise powers of the SMAX procedure are slightly greater than that of the CLOSE procedure for the alternative with $S_0 = S_1 = S_2 < S_3$. However, the CLOSE procedure outperforms the SMAX procedure for $S_0 = S_1 < S_2 < S_3$, $S_0 = S_1 < S_2 = S_3$, $S_0 < S_1 = S_2 = S_3$ and $S_0 < S_1 < S_2 = S_3$. This phenomenon is because the CLOSE procedure has a greater likelihood in detecting that there is more than one treatment better than the control. Under exponential distribution, the LR test is more powerful than the PPW test, while the WKMs tests are as powerful as the LR test. However, the power of the WKM test is slightly higher than that of the LR test, and it is compatible with the PPW test under lognormal distribution and low censoring rate. Moreover, the WKM and WKMs tests are superior to the PPW and LR tests for moderate censoring rate and lognormal distribution. For early hazard differences alternative, the PPW test performs better than the WKM

Table III. Estimated comparison-wise powers of SMAX or CLOSE tests for $\alpha = 0.05$, uniform censoring distribution $U(0, b)$ and $n_0 = n_1 = n_2 = n_3 = 20$.

		SMAX				CLOSE			
b	Alternatives	LR	PPW	WKM	WKMs	LR	PPW	WKM	WKMs
Early hazard difference alternative									
3.654	$S_0 = S_1 = S_2 < S_3$	0.545	0.647	0.602	0.563	0.533	0.637	0.589	0.549
	$S_0 = S_1 < S_2 < S_3$	0.223	0.279	0.251	0.230	0.250	0.310	0.280	0.252
	$S_0 = S_1 < S_2 = S_3$	0.410	0.517	0.459	0.423	0.433	0.545	0.486	0.447
	$S_0 < S_1 = S_2 = S_3$	0.221	0.254	0.237	0.220	0.347	0.411	0.381	0.359
	$S_0 < S_1 < S_2 = S_3$	0.356	0.464	0.411	0.374	0.518	0.638	0.578	0.538
1.457	$S_0 = S_1 = S_2 < S_3$	0.575	0.592	0.607	0.604	0.557	0.579	0.593	0.588
	$S_0 = S_1 < S_2 < S_3$	0.246	0.244	0.264	0.263	0.275	0.277	0.290	0.291
	$S_0 = S_1 < S_2 = S_3$	0.448	0.460	0.482	0.478	0.482	0.498	0.518	0.514
	$S_0 < S_1 = S_2 = S_3$	0.226	0.216	0.236	0.244	0.370	0.372	0.392	0.391
	$S_0 < S_1 < S_2 = S_3$	0.403	0.403	0.430	0.433	0.567	0.579	0.598	0.600
Late hazard difference alternative									
3.645	$S_0 = S_1 = S_2 < S_3$	0.660	0.406	0.531	0.589	0.647	0.311	0.520	0.576
	$S_0 = S_1 < S_2 < S_3$	0.102	0.041	0.061	0.068	0.120	0.038	0.072	0.083
	$S_0 = S_1 < S_2 = S_3$	0.499	0.190	0.356	0.421	0.540	0.145	0.396	0.467
	$S_0 < S_1 = S_2 = S_3$	0.095	0.026	0.053	0.063	0.194	0.075	0.128	0.147
	$S_0 < S_1 < S_2 = S_3$	0.387	0.109	0.267	0.329	0.599	0.226	0.445	0.523
1.457	$S_0 = S_1 = S_2 < S_3$	0.359	0.208	0.197	0.263	0.350	0.202	0.191	0252
	$S_0 = S_1 < S_2 < S_3$	0.050	0.022	0.022	0.030	0.057	0.026	0.026	0.037
	$S_0 = S_1 < S_2 = S_3$	0.180	0.066	0.075	0.112	0.212	0.085	0.089	0.136
	$S_0 < S_1 = S_2 = S_3$	0.034	0.010	0.015	0.022	0.086	0.033	0.037	0.053
	$S_0 = S_1 < S_2 = S_3$	0.111	0.027	0.035	0.062	0.226	0.074	0.085	0.144

and WKMs tests under low censoring rate, but the WKM and WKMs tests are more powerful than the PPW test under moderate censoring rate. This is because the weight functions of the WKM and WKMs tests are based on the censoring distributions. Hence the WKM and WKMs tests perform better under heavy censoring and early hazard difference alternatives. In contrast, the WKM and WKMs tests lose their power under late hazard difference alternatives. Moreover, the LR test outperforms the other tests considered here and the WKMs test performs much better than the PPW test. Nevertheless, none of the tests examined is uniformly better than the others for the alternatives considered in the simulation study.

5. EXAMPLES

In this section, the proposed method is illustrated through two examples based on real data. A double blind randomized clinical trial comparing four treatments for patients with prostate cancer was conducted. The survival data from prostate cancer patients in stage IV are used for the illustration purpose. The placebo group consisted of 53 patients, while 54 patients

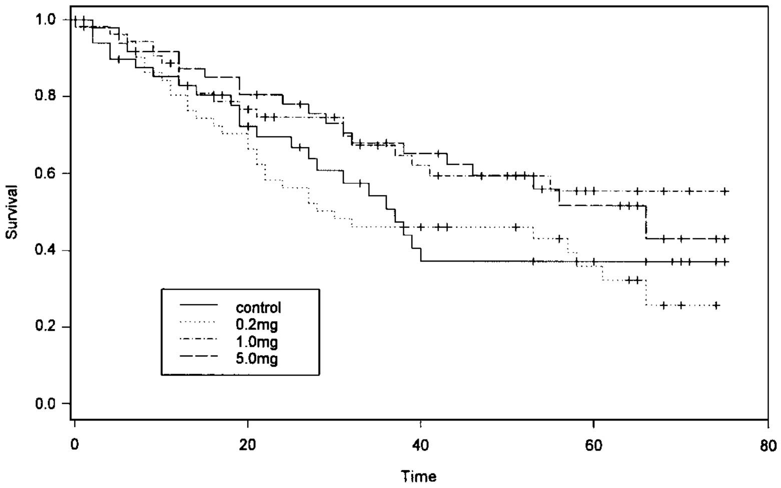

Figure 2. Estimated survival curves for prostate cancer study.

were treated with 0.2 mg diethylstilbestrol (DES), 55 patients were treated with 1.0 mg of DES, and 52 patients were treated with 5.0 mg of DES. All drugs were administered daily by mouth and the survival times until death were recorded. The patients who died from other causes and those who were still alive at the end of the trial are considered as right-censored observations. Because one of the goals of this study is to identify effective treatment dosages as compared with the placebo, the proposed multiple testing procedures are then applied. The estimated survival curves of the four groups are displayed in Figure 2. It is not easy to classify the possible alternatives for these survival curves at once. After plotting pairwise estimated survival curves (not shown), these curves for the placebo and 1.0 mg of DES groups seem to display something like late hazard differences, while the estimated survival curves of the placebo and 5.0 mg of DES groups exhibit sort of middle hazard differences.

The values of the LR, PPW, WKM, and WKMs tests for each dosage compared with placebo group are listed in Table IV. At 0.1 significance level, the critical values of the SMAX multiple testing procedure are computed by using a recently developed FORTRAN program for calculating multivariate normal probabilities [22] with estimated correlation matrix, and we obtained z max$(3, 0.10) = 1.715$, z max$(3, 0.10) = 1.720$, z max$*(3, 0.10) = 1.708$, and z max$*(3, 0.10) = 1.710$, for the LR, PPW, WKM, and WKMs tests, respectively. Thus, compared with the above critical values, none of the dosage levels show significantly better survival than the placebo. It is worth noting that the value of the WKMs test is very close to the corresponding critical value. Thus using the same FORTRAN program, the p-values of the SMAX procedure are 0.131, 0.164, 0.136, and 0.111 for the LR, PPW, WKM, and WKMs tests, respectively. Therefore, the WKMs test shows slightly stronger evidence that the survival rate of 1.0 mg of DES group is better than that of the placebo group at 0.12 significance level by the SMAX procedure as well as the CLOSE procedure. Moreover, the CLOSE procedure also concludes that the 5.0 mg of DES group has better survival rate than the control. Hence, it seems that the WKMs test has good potential for certain data to identify significant treatment effects.

Table IV. Summary statistics for Prostate cancer data set.

Test statistics	0.2 mg vs placebo	1.0 mg vs placebo	5.0 mg vs placebo
Logrank	−0.465	1.599	1.373
PPW	−0.476	1.455	1.453
WKM	−0.221	1.502	1.555
WKMs	−0.214	1.658	1.630
Logrank	$\begin{bmatrix}1 & 0.557 & 0.552\\ & 1 & 0.534\\ & & 1\end{bmatrix}$		
PPW	$\begin{bmatrix}1 & 0.550 & 0.541\\ & 1 & 0.521\\ & & 1\end{bmatrix}$		
WKM	$\begin{bmatrix}1 & 0.553 & 0.540\\ & 1 & 0.521\\ & & 1\end{bmatrix}$		
WKMs	$\begin{bmatrix}1 & 0.556 & 0.543\\ & 1 & 0.527\\ & & 1\end{bmatrix}$		

In the renal carcinoma (Renca) tumour study [23], four groups of mice injected with Renca cells were arranged to receive no treatment (control), three different anti-tumour activities (interleukin: potent immunoregulatory cytokines) IL-2 alone, IL-12 alone, and IL-2 plus IL-12, respectively. The WKM and WKMs tests are also applied to identify the treatment effect. The values of the LR, PPW, WKM, and WKMs tests for each dosage compared with the control are listed in Table V. Since there are 1 and 5 censored observations, which are the largest observed survival time, in IL-2 and IL-2 plus IL-12 groups, respectively, the WKM and WKMs procedure yield same test values. Because the off-diagonal elements of the estimated correlation matrix for the LR and PPW tests are close to 0.5 (see Table V), Chen [8] used the critical values, $z\ \max(3,0.05)=2.063$, $z\ \max(2,0.05)=1.917$, $z\ \max(1,0.05)=1.645$, found in Reference [21] to make inference. However, since the off-diagonal elements of the estimated correlation matrix for the WKM test are much higher than 0.5, we then applied the algorithm [22] to obtain the critical values, $z\ \max*(3,0.05)=2.035$, $z\ \max*(2,0.05)=1.910$, $z\ \max*(1,0.05)=1.645$, at each step for the CLOSE procedure. Note that these critical values are slightly smaller than the ones computed with the correlation matrix with 0.5 as off-diagonal elements. The WKM test and the LR test reach the same conclusion that only the combined treatment IL-2 plus IL-12 is significantly more effective than the control, whereas the PPW test shows that the IL-12 alone and IL-2 plus IL-12 treatments are more effective than the control. The simulation results in Table III and the estimated survival curves plotted in Figure 3 can be used to explain these inconsistent results. Since the vertical differences between the estimated survival curves of the IL-2 plus IL-12 treatment and the placebo groups are sufficiently clear, all the tests examined here conclude that the IL-2 plus IL-12 treatment

Table V. Summary statistics for interleukin-tumour data set.

Test statistics	IL-2 vs control	IL-12 vs control	IL-2 plus IL-12 vs control
Logrank	1.696	1.788	3.724
PPW	1.584	2.023	3.507
WKM	1.678	1.853	3.358
WKMs	1.678	1.853	3.358
Logrank	$\begin{bmatrix}1 & 0.552 & 0.527\\ & 1 & 0.524\\ & & 1\end{bmatrix}$		
PPW	$\begin{bmatrix}1 & 0.479 & 0.460\\ & 1 & 0.458\\ & & 1\end{bmatrix}$		
WKM	$\begin{bmatrix}1 & 0.547 & 0.630\\ & 1 & 0.628\\ & & 1\end{bmatrix}$		
WKMs	$\begin{bmatrix}1 & 0.547 & 0.630\\ & 1 & 0.628\\ & & 1\end{bmatrix}$		

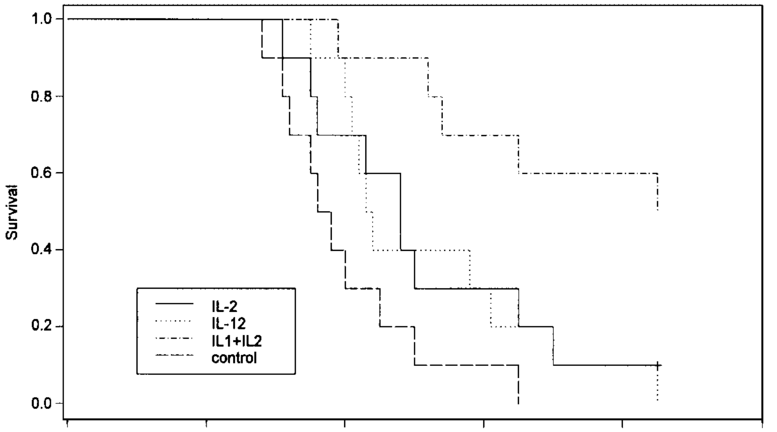

Figure 3. Estimated survival curves for Renca tumour study.

group has better survival than the control. In addition, the estimated survival curves reveal that early hazard differences may exist between the IL-12 treatment group and the control, while proportional hazards may exist between the control and IL-2 treatment groups. The simulation results in Table III show that the PPW test has highest power to detect early hazard difference alternatives with low censoring rate. Therefore, the PPW test seems to provide stronger evidence in identify treatment effect for this data set.

Nevertheless, Chen [8] suggested applying different weights in two-sample weighted logrank tests to extract more information according to the types of alternatives to expect for this data set. The LR test was applied to compare the control with IL-2 alone and IL-2 plus IL-12, respectively, and the PPW test was used to compare IL-12 with the control. Consequently, all treatments are more effective than the control at 0.05 significance level. It is worth noting that at 0.06 significance level, the critical values of the CLOSE procedure at each step for the WKM test are $z\max *(3, 0.06) = 1.961$, $z\max *(2, 0.06) = 1.83$, $z\max *(1, 0.06) = 1.565$. Hence the WKM test can also indicate that all treatments are more effective than the control. Thus the WKM test exhibits its robustness to different types of alternatives, as demonstrated in the simulation.

6. CONCLUDING REMARKS

In general, if the researchers can identify the types of alternatives to expect for a given data set, then from Tables II and III the WKM and LR tests are recommended for application to detect the treatment effects under proportional hazards model, while the WKM and WKMs tests should be applied under lognormal survival distributions. For early hazard differences alternative the PPW test should be used for low censoring rate, whereas the WKM and WKMs tests should be used for high censoring rate. Although the LR test has better power than the other tests considered here for late hazards difference alternatives, the weighted logrank test with increasing weight function, such as $W(t) = 1 - \hat{S}(t)$, is preferable to identify treatment effect for such alternatives. A multiple testing procedure may consist of several pairwise tests, each comparing an individual treatment with the control; therefore, if one can identify the types of alternatives to expect for each pair, then different weights or even different types of test statistics should be employed for each individual comparison to gain better power. However, when it is difficult to identify the possible alternatives, as illustrated in the prostate cancer study, the multiple testing procedure based on weighted Kaplan–Meier statistics may be a robust test to apply because its power seems to be more stable over various alternatives.

For practical computing, from our experience the p-value of the SMAX procedure is more easily computed than the critical value from the algorithm [22] that computes the percentile of the distribution of the largest order statistic from multivariate normal random variables with estimated correlation structure. Furthermore, since the sample sizes are unequal and the censoring patterns are unknown for most real data sets, the p-value or critical values obtained from this algorithm with estimated correlation structure tend to yield more accurate results than computations with the correlation matrix with 0.5 as off-diagonal elements. Therefore, we suggest using this algorithm to obtain p-values or critical values for making inference and drawing conclusions. This algorithm can be obtained from the author on request.

ACKNOWLEDGEMENTS

The author would like to thank Mr. Min-Shiao Tsai for his computing assistance. The author would like to thank the editor and the reviewers for many insightful comments that greatly improved this article. The work was supported by the National Science Council of Taiwan under Contract No. NSC89-2115-M-006-005.

REFERENCES

1. Byar DP, Corle DK. Selecting optimal treatment in clinical trials using covariate information. *Chronic Diseases* 1977; **30**:445–469.
2. Byar DP, Green SB. The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bulletin Du Cancer* 1980; **67**:477–488.
3. Chakraborti S, Desu MM. Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship. *Statistical Neerlandica* 1991; **45**:227–254.
4. Slepian D. The one-sided barrier problem for Gaussain noise. *Bell System Technical Journal* 1962; **41**: 463–501.
5. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; **52**:203–223.
6. Chen YI. A generalized Steel's procedure for comparing several treatments with a control under random right-censorship. *Communication in Statistics Simulation* 1994; **23**:1–16.
7. Steel RGD. A multiple comparison rank sum test: treatments versus control. *Biometrics* 1959; **15**:560–572.
8. Chen YI. Multiple comparisons in carcinogenesis study with right-censored survival data. *Statistics in Medicine* 2000; **19**:353–367.
9. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
10. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics* 1972; **14**: 945–965.
11. Aalen OO. Nonparametric inference for a family of counting processes. *Annual of Statistics* 1978; **6**:701–726.
12. Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics* 1989; **45**:497–507.
13. Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: large sample and optimality considerations. *Journal of Royal Statistical Society*, *Series B* 1991; **53**:342–352.
14. Kaplan EL, Meier P. Nonparametric estimator from incomplete observations. *Journal of American Statistical Association* 1958; **53**:457–481.
15. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; **50**:163–170.
16. Fleming TR, Harrington DP. A class of hypothesis tests for one and two-samples of censored data. *Communication in Statistics* 1981; **10**:131–139.
17. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of Royal Statistical Society*, *Series A* 1972; **135**:185–206.
18. Prentice RL. Linear rank tests with right-censored data. *Biometrika* 1978; **65**:165–179.
19. Schervish MJ. Multivariate Normal Probabilities with error bound. *Applied Statistics* 1984; **33**:81–94.
20. Chi Y. Many-to-One multiple testing procedures for right censored data. *NSC89-2118-M-006-005 Technical Report*, 2000.
21. Gupta SS. Probability integrals of multivariate normal and multivariate t. *Annals of Mathematical Statistics* 1963; **34**:792–828.
22. Drezner Z. Computation of the Multivariate normal integral. *Transactions on Mathematical Software* 1992; **19**:470–480.
23. Wigginton JM, Komschlies KL, Back TC, Franco JL, Brunda MJ, Wiltrout RH. Administration of interleukin 12 with pulse interleukin 2 and the rapid and complete eradication of murine renal carcinoma. *Journal of the National Cancer Institute* 1996; **88**:38–43.