STATISTICS IN MEDICINE
Statist. Med. **19**, 353–367 (2000)

MULTIPLE COMPARISONS IN CARCINOGENESIS STUDY WITH RIGHT-CENSORED SURVIVAL DATA

YUH-ING CHEN*

Institute of Statistics, National Central University, Chung-Li, Taiwan 320, R.O.C.

SUMMARY

This paper considers the practical problem in animal carcinogenesis experiments where several treatment groups are compared with a control group in a one-way layout and the observed survival data are subject to random right-censorship. Proposed herein are multiple testing procedures based on two-sample weighted logrank statistics, each comparing an individual treatment with the control, for determining which treatments are more effective than the control. The associated p-value of claiming a certain treatment is more effective than the control is also discussed. A test-based confidence set for the scale changes between each treatment and the control is then obtained. The comparative results of a Monte Carlo error rate and power study for small sample sizes are presented. Finally, a numerical example involving renal carcinoma in mice demonstrates the feasibility of the proposed multiple testing procedures and test-based confidence set.

1. INTRODUCTION

Statistical procedures for comparing several treatments with a control have been extensively discussed when data are completely observed in a one-way layout (see, for example, Hochberg and Tamhane[1]). In animal carcinogenesis experiments, different therapies (treatments) are usually compared with a standard therapy or placebo (control) to evaluate the effect of the treatments on the prolongation of the survival time of animals with a certain established carcinoma. Of particular interest are those treatments in which effects exceed that of the control. However, in these studies, survival data are frequently subject to random right-censorship, since the study may be terminated at a pre-assigned time owing to time limitation, or the death may be attributed to a competing risk which is not of interest. Therefore, statistical procedures are needed for the many-to-one comparisons with randomly right-censored survival data.

Multiple testing procedures for the many-to-one comparisons problem with right-censored data have been recently developed. For instance, Chakraborti and Desu[2] used Slepian's[3] inequality to suggest a conservative multiple testing procedure based on Gehan's[4] two-sample statistics each comparing an individual treatment with the control. Chen[5] proposed a

* Correspondence to: Yuh-Ing Chen, Institute of Statistics, National Central University, Chung-Li, Taiwan 320, R.O.C.
E-mail: ychen@stat.ncu.edu.tw

Contract/grant sponsor: National Science Council of Taiwan
Contract/grant number: NSC 86-2115-M-008-018

CCC 0277–6715/2000/030353–15$17.50

Received April 1998
Accepted March 1999

generalization of Steel's[6] test on the basis of the maximum of Gehan's[4] statistics for the special case of equal censoring. However, the logrank statistic (Mantel[7]) is probably the most commonly used two-sample statistic and Gehan's statistic is a member of the general class of weighted logrank statistics (see, for example, Fleming and Harrington[8]). Chen[9] further extended the Chakraborti–Desu testing procedure on the basis of the two-sample weighted logrank statistics.

In this paper we consider an alternative generalization of Steel's test for the many-to-one comparisons based on the two-sample weighted logrank statistics each comparing an individual treatment with the control. Based on the generalized Steel's test, a confidence set for the ratios of scale parameters of each treatment and the control in a scale-change model is obtained (Wei and Gail[10]). In addition, a closed testing procedure (Marcus *et al.*[11]) is suggested in which the generalized Steel's test is modified. Three special weighted logrank statistics under consideration are Gehan's statistic, the logrank statistic and the Peto–Prentice statistic (Peto and Peto[12] and Prentice[13]). Moreover, for the conclusion that a certain treatment is more effective than the control reached by the proposed testing procedure, the associated p-value defined to be the smallest overall significance level for obtaining the conclusion is discussed. Comparative results of a Monte Carlo study investigation demonstrate the relative error rate and power performances of the proposed testing procedures for small sample sizes. Finally, the use of these testing procedures is illustrated with the numerical example involving renal carcinoma in mice described in Section 2.

2. A RENAL CARCINOMA EXAMPLE

Interleukin 2 (IL-2) and interleukin 12 (IL-12) are potent immunoregulatory cytokines that exhibit anti-tumour activity (see, for example, Gately[14] and Maas *et al.*[15]). Preliminary evidence further suggests that combined administration of IL-12 and IL-2 may yield greater anti-tumour activity than that observed with either agent alone (see, for instance, Rossi *et al.*[16]). To evaluate the ability of IL-2, IL-12 and combination regimens of IL-2 and IL-12 to induce regression of established primary and metastatic murine renal carcinoma (Renca) tumours, Wigginton *et al.*[17] conducted an animal experiment. Forty BALB/c mice administered an internal injection of 1×10^5 Renca cells and developing subcutaneous primary Renca tumours were divided into four groups to receive no treatment (control), IL-2 (300 000 IU given twice daily one day per week) alone, IL-12 (0·5 μg given on a daily basis) alone, or IL-12 in combination with IL-2. The measurement of record for each group was the survival time after tumour injection. However, the mice survived and experienced tumour regression at the end of the study, yielding censored data.

In this study, the anti-tumour activity of IL-12 in combination with IL-2, IL-2 or IL-12 alone on regression of the established Renca tumours needs to be evaluated. Whether treatment with IL-12 plus IL-2 displays greater anti-tumour activity against the established Renca tumours than either IL-12 or IL-2 alone also needs to be investigated. To this end, testing procedures preserving a correct overall significance level are needed for determining simultaneously which of IL-2 alone, IL-12 alone, and the combined administration of IL-12 and IL-2, comparing to the control, can yield anti-tumour activity against the Renca tumours. Testing procedures holding a correct overall significance level are also required for simultaneously comparing the combined IL-12 and IL-2 with IL-12 alone and IL-2 alone. Note that the involving multiple comparisons are, in fact, corresponding to the many-to-one comparisons.

3. PROPOSED TESTING PROCEDURES

Let S_i be the survival function of the ith group, $i = 0, 1, \ldots, m$. Suppose that the zero population ($i = 0$) is the control and the other m populations are treatments. In such a setting with right-censored data, we only observe the bivariate vectors (X_{iu}, δ_{iu}), $u = 1, \ldots, n_i$, $i = 0, 1, \ldots, m$, where X_{iu} is the minimum of the survival time and the associated censoring time, and δ_{iu} is the indicator of censorship, which is one if the observation is not censored, and zero otherwise. In this paper, specifically, we are concerned with the many-to-one multiple comparisons problem of determining the treatments which are more effective than the control; that is, $S_i > S_0$, $i = 1, \ldots, m$, when survival data are subject to random right-censorship.

Let $T_1 < \cdots < T_L$ denote the ordered observed distinct death times in the sample formed by combining the ith and control groups. Let d_{uk} and Y_{uk}, $k = 1, \ldots, L$, $u = 0, i$, denote the number of observed deaths and number at risk, respectively, in sample u at time T_k. Set $d_{+k} = d_{0k} + d_{ik}$ and $Y_{+k} = Y_{0k} + Y_{ik}$. Note that, given d_{+k}, Y_{0k} and Y_{ik}, the conditional distribution of d_{0k} is the hypergeometric distribution with mean $e_{0k} = d_{+k}Y_{0k}/Y_{+k}$ and variance $v_k = d_{+k}Y_{0k}Y_{ik}(Y_{+k} - d_{+k})/\{(Y_{+k})^2(Y_{+k} - 1)\}$. For some appropriate weight functions $W_i(T_k)$, let

$$U_{0i} = \sum_{k=1}^{L} W_i(T_k)(d_{0k} - e_{0k}) \tag{1}$$

and

$$s_{ii} = \sum_{k=1}^{L} \{W_i(T_k)\}^2 v_k. \tag{2}$$

Then, under significance level α, the two-sample weighted logrank tests conclude $S_i > S_0$, if

$$U_i = U_{0i}/\sqrt{s_{ii}} \geqslant z(\alpha)$$

where $z(\alpha)$ is the upper αth percentile of the standard normal distribution. Note that taking $W_i(T_k) = Y_{+k}/(n_0 + n_i)$ to be the proportion at risk at time T_k produces Gehan's statistic, taking $W_i(T_k) = 1$ yields the logrank statistic, and setting $W_i(T_k)$ to be the Kaplan–Meier[18] survival estimate based on the pooled samples from the ith and the control groups gives the Peto–Prentice statistic. These three special cases of the weighted logrank statistics are of general interest, since they approximate many practical applications and are available in most statistical packages.

To obtain a generalized Steel's many-to-one comparisons procedure which controls the overall significance level or experimentwise error rate (probability of erroneously declaring at least one treatment more effective than the control), we need to find the percentiles of the distribution of $\max(U_1, U_2, \ldots, U_m)$ under the null hypothesis H_0:$(S_i = S_0,, i = 1, 2, \ldots, m)$. Note that, as specified in the Appendix, the asymptotic null distribution of $(U_1, U_2, \ldots, U_m)$ is an m-variate normal with mean zero and covariance matrix $\Sigma = (\rho_{ij})$, which can be consistently estimated by $\mathbf{S}$. Let $(Z_1, Z_2, \ldots, Z_m)$ be an m-variate normal vector with mean zero and covariance matrix $\mathbf{S}$. As a generalization of Steel's testing procedure, we suggest

$$\text{SMAX: claim } S_i > S_0, \text{ if } U_i \geqslant z\max(m, \alpha), \quad \text{for } i = 1, 2, \ldots, m \tag{3}$$

where $z\max(m, \alpha)$ is the upper αth percentile of the distribution of $\max(Z_1, Z_2, \ldots, Z_m)$. Obviously, the experimentwise error rate for the procedure (3) is approximately controlled, since

$$\alpha \approx P\{\max(U_1, U_2, \ldots, U_m) \geqslant z\max(m, \alpha) | H_0\}$$

$$= P\{U_i \geqslant z\max(m, \alpha) \text{ for at least one } i = 1, 2, \ldots, m | H_0\}.$$

Note that, for any z, the probability $P\{\max(Z_1, Z_2, \ldots, Z_m) \leqslant z\}$ can be computed using a program for calculating multivariate normal probabilities (Schervish[19]). Therefore, the critical value $z\max(m, \alpha)$ can be found such that $P\{\max(Z_1, Z_2, \ldots, Z_m) \geqslant z\max(m, \alpha)\} = \alpha$. However, when the $m + 1$ groups under consideration have the same censoring distribution, for the three weight functions considered herein, the result in Chen[9] implies that $\rho_{ij} = 1$, if $i = j$, and $\sqrt{\{n_i n_j/(n_0 + n_i)(n_0 + n_j)\}}$ otherwise. Moreover, for the particular case of treatment-balanced design with $n_0 = c$ and $n_1 = \cdots = n_k = n$, we observe $\rho_{ij} = n/(n + c)$, if $i \neq j$. Therefore, when the assumption of equal censoring is tenable, we recommend using the critical value corresponding to the known covariance matrix $\Sigma = (\rho_{ij})$ which can be computed by using the program in Schervish[19] or found in Gupta[20] for the treatment-balanced design. Also note the conservative procedure based on Slepian's inequality proposed by Chen:[9]

$$\text{CHEN: claims } S_i > S_0, \text{ if } U_i \geqslant z(b),\ i = 1, 2, \ldots, m \tag{4}$$

where $b = 1 - (1 - \alpha)^{1/m}$.

Suppose that the survival times in the $m + 1$ groups satisfy the scale-change model; that is, $F_i(x) = F(x/\theta_i)$, with $F_i = 1 - S_i$, $i = 0, 1, \ldots, m$. Set $\gamma_i = \theta_i/\theta_0$ to be the ratios of the scale parameters of the ith treatment group and the control. Let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})$ and $\Delta_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{in_i})$, $i = 0, 1, \ldots, m$. Denote the U_{0i} in (1) and s_{ii} in (2) as $h(\mathbf{X}_0, \mathbf{X}_i, \Delta_0, \Delta_i)$ and $V(\mathbf{X}_0, \mathbf{X}_i, \Delta_0, \Delta_i)$, respectively. Therefore, by applying the results in Wei and Gail,[10] an approximate $(1 - \alpha) \times 100$ per cent confidence set for $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)$ based on the test in (3) is given by

$$C(\gamma) = \{(\gamma_1, \gamma_2, \ldots, \gamma_m): \gamma_i \geqslant \gamma_i, \qquad i = 1, 2, \ldots, m\} \tag{5}$$

where γ_i is the smallest value of a such that

$$h(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i) \leqslant z\max(m, \alpha)\sqrt{\{V(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i)\}}.$$

If only a significance testing procedure is of interest, a stepwise comparison is then required which is generally more powerful than a single step procedure. For the problem considered herein, the class of null hypotheses $H = \{H_0(P)\}$, where $H_0(P)$: ($S_i = S_0$ for i in P) and P is any subset of $\{1, 2, \ldots, k\}$, is closed under intersection in the sense that both $H_0(P)$ and $H_0(Q)$ in H implies the intersection of $H_0(P)$ and $H_0(Q)$ also in H. Therefore, to determine which treatments are more effective than the control with right-censored survival data, a closed testing procedure (Marcus *et al.*[11]) can be constructed which includes separate α-level tests of individual $H_0(P)$ applied in a step-down manner. Let $U_{(1)} < U_{(2)} < \cdots < U_{(k)}$ be the ordered U_i's and let d_i be the antirank of i, that is, $U_{(i)} = U_{d_i}$. Denote by $z\max((i), \alpha)$ the upper αth percentile of the distribution of $\max(Z_{d_1}, Z_{d_2}, \ldots, Z_{d_i})$. Consequently, a closed testing procedure is obtained:

$$\text{CLOSE:} \quad \text{claims } S_{d_i} > S_0 \text{ if } U_{(i)} \geqslant z\max(i), \alpha) \tag{6}$$

provided that $U_{(j)} \geqslant z\max(j)), \alpha)$ for $j > i$. This procedure continues until $U_{(k)} < z\max((k), \alpha)$ for some $k = 1, 2, \ldots, m$. Finally, we declare at an approximate experimentwise error rate α, that only the treatment labelled by $d_{k+1}, \ldots, d_m$ are more effective than the control.

Note that the p-value of a single test is the smallest significance level which leads to a rejection of the null hypothesis. For the many-to-one comparisons procedure with the conclusion that a certain treatments are better than the control, the associated p-value is then defined to be the smallest experimentwise error rate at which the testing procedure reaches the conclusion. Let $u_{(1)} < u_{(2)} < \cdots < u_{(m)}$ be the observed values of $U_{(1)} < U_{(2)} < \cdots < U_{(m)}$. Recall that $u(i) = u_{d_i}$, $i = 1, \ldots, m$. The p-value of the generalized Steel's procedure in (3) with the conclusion that the treatments labelled by $d_{k+1}, \ldots, d_m$ $(k < m)$ are more effective than the control is then given by $P\{\max(U_1, U_2, \ldots, U_m) \geqslant u_{d_{k+1}} | H_0\}$ which can be approximated by

$$p(\text{SMAX}; d_{k+1}, \ldots, d_m) = 1 - P\{Z_1 \leqslant u_{d_{k+1}}, \ldots, Z_m \leqslant u_{d_{k+1}}\} \tag{7}$$

where the vector $(Z_1, Z_2, \ldots, Z_m)$ is, again an m-variate normal vector with mean zero and covariance matrix $\mathbf{S}$. Following the results in Dunnett and Tamhane,[21] let

$$p'(d_i) = 1 - P\{Z_{d_1} \leqslant u_{d_i}, \ldots, Z_{d_i} \leqslant u_{d_i}\}, \qquad i = 1, 2, \ldots, m.$$

Then, the approximate p-value of the closed test in (6) with the conclusion that the treatments labelled by $d_{k+1}, \ldots, d_m$ are more effective than the control is obtained as

$$p(\text{CLOSE}; d_{k+1}, \ldots, d_m) = \max\{p'(d_{k+1}), p'(d_{k+2}), \ldots, p'(d_m)\}. \tag{8}$$

Also note that procedures in (3) and (6) have the same experimentwise error rate and experimentwise power (probability of correctly detecting at least one treatment which is better than the control). Meanwhile, since $z\max((i), \alpha)$ is increasing in i, the closed testing procedure in (6) would have a greater comparisonwise power (probability of correctly declaring all treatments which are better than the control) than does the generalized Steel's procedure in (3). However, the closed testing procedure, as a stepwise procedure, cannot be used for constructing a confidence set for the ratios of scale parameters in the scale-change model. Furthermore, procedures (3)–(6) are allowed to employ two-sample weighted logrank statistics with different weight functions.

All the procedures mentioned above are appropriate for the situation where the treatments are expected to be at least as good as the control (right-sided comparisons). However, when it is expected that the control is at least as good as the treatments (left-sided comparisons) or there is at least one treatment which is different from the control (two-sided comparisons), the procedures in (3)–(6) should be modified as stated in the Appendix.

4. MONTE CARLO STUDY

4.1. Discussion of study

A Monte Carlo study was performed to examine the relative level and power performances of the generalized Steel's testing procedure SMAX in (3), the conservative test CHEN in (4), and the closed testing procedure CLOSE in (6) for comparing several treatments with a control when observations are subject to random right-censorship. For simplicity, we only considered procedures based on two-sample weighted logrank statistics with the same type of weight functions. The level performances of these tests were evaluated by the experimentwise error rate and their power performances were assessed by both the experimentwise and comparisonwise powers.

In this study, we considered $k = 3$ treatments with sample sizes $n_0 = c$, $n_1 = \cdots = n_3 = n$. We employed $(c, n) = (10, 10)$, $(20, 20)$, $(30, 30)$ and $(30, 20)$ in the level study, and $(c, n) = (20, 20)$ and $(30, 20)$ in the power study. Exponential and log-normal distributions were considered as survival time distributions and the uniform distribution over $(0, R)$ was used as the censoring distribution. Appropriate uniform, normal and exponential variates were generated by using the IMSL routines RNUN, RNNOR and RNEXP, respectively. The necessary log-normal variates were then given by the exponential of the normal variates. In the level study, the standard exponential distribution and the log-normal distribution with zero normal mean and normal standard deviation $\sigma = 1/2$ were considered. In the power study, we used exponential distributions with various values of scale parameters θ_i's and log-normal distributions with normal standard deviation $\sigma = 1/2$ but different values of normal means θ_i's. Note that the exponential distribution represents the proportional hazards model and the log-normal distribution corresponds to location shifts in log survival times. In fact, the log-normal distributions considered in the power study have different hazards at early times. Various values of R which correspond to the probability of censorship (the probability that survival time is greater than the censoring time), p, as 0·10 and 0·30 were considered in the level study, the corresponding uniform distributions were then employed as censoring distributions in the power study. For example, when survival time distribution is the standard exponential and $p = 0{\cdot}1$, $R = 9{\cdot}901$. For the log-normal distribution with zero normal mean and normal standard deviation $\sigma = 1/2$, $R = 3{\cdot}756$ corresponds to $p = 0{\cdot}3$. Note that the censoring probabilities were fixed for each population in the level study, while they may be different for the populations involved in the power study due to different life time distributions.

For each of these settings, 5000 replications were used to obtain the estimated experimentwise error rates or both the experimentwise and comparisonwise powers under the nominal level $\alpha = 0{\cdot}05$. Therefore, the maximum standard error for the estimate is around 0·007 ($\approx \sqrt{\{(0{\cdot}5\,(0{\cdot}5)/5000\}}$). In fact, under the nominal level $\alpha = 0{\cdot}05$, the standard error for the estimated error rate is about 0·003 ($\approx \sqrt{(0{\cdot}05)\,(0{\cdot}95)/5000\}}$). We then indicate, by $+\,(-)$ signs, whenever the estimated error rate is two or more standard deviations above (below) 0·05. Note that both the SMAX and CLOSE tests have the same experimentwise error rate and power. Therefore, Table I reports the simulated error rates for the SMAX (or CLOSE) and CHEN tests only. Since the CHEN test was found to have a relatively conservative level performance, the simulated experimentwise powers for SMAX (or CLOSE) are presented in Tables II and the simulated comparisonwise powers for SMAX and CLOSE are reported in Tables III and IV.

4.2. Discussion of results

It can be seen from Table I that the logrank (LR) version of the SMAX or CLOSE test reasonably maintains its level only when the sample size in each group is at least 20. In addition, the SMAX or CLOSE test based on Gehan's (GH) and the Peto–Prentice (PP) statistics hold their levels well across all the situations under consideration. Obviously, the CHEN test tends to be conservative, especially, for the cases with 20 or more observations in each sample.

The power study in Tables II, III and IV indicates that the logrank test is more powerful than either Gehan's or the Peto–Prentice test for exponential distributions. Meanwhile, both the Gehan and Peto–Prentice tests are superior to the logrank test for log-normal distributions. Moreover, although the Peto–Prentice test is better than Gehan's test for exponential distributions, the two tests appear to perform rather similarly for log-normal distributions.

Table I. Estimated experimentwise error rates for $\alpha = 0{\cdot}05$, uniform censoring distribution $U(0, R)$ and $n_0 = c$, $n_1 = n_2 = n_3 = n$

c, n	R	SMAX (or CLOSE)			CHEN		
		LR	GH	PP	LR	GH	PP
Exponential survival distribution							
10, 10	9·901	0·065 +	0·051	0·051	0·059 +	0·045	0·047
	3·185	0·060 +	0·049	0·050	0·053	0·044	0·044
20, 20	9·901	0·056	0·053	0·054	0·052	0·045	0·046
	3·185	0·052	0·046	0·046	0·046	0·039 −	0·040 −
30, 30	9·901	0·055	0·053	0·054	0·050	0·050	0·049
	3·185	0·054	0·049	0·050	0·047	0·042 −	0·043 −
30, 20	9·901	0·051	0·045	0·046	0·045	0·035 −	0·035 −
	3·185	0·048	0·044	0·045	0·040 −	0·038 −	0·038 −
Log-normal survival distribution							
10, 10	11·219	0·060 +	0·050	0·051	0·052	0·044	0·042 −
	3·756	0·062 +	0·056	0·056	0·056	0·049	0·049
20, 20	11·219	0·053	0·047	0·048	0·045	0·041 −	0·042 −
	3·756	0·056	0·051	0·052	0·055	0·044	0·044
30, 30	11·219	0·049	0·047	0·047	0·042 −	0·042 −	0·042 −
	3·756	0·055	0·051	0·053	0·050	0·043 −	0·045
30, 20	11·219	0·056	0·050	0·049	0·053	0·045	0·045
	3·756	0·051	0·048	0·048	0·046	0·043	0·044

+ (−): at least two standard deviations above (below) $\alpha = 0{\cdot}05$

The comparisonwise power study further demonstrates that the CLOSE test is more powerful than the SMAX test when there are two or more treatments better than the control. The apparent superiority of the CLOSE test over the SMAX test occurs when all the treatments are more effective than the control. The SMAX test is slightly better than the CLOSE test when there is only one treatment better than the control and the rest are equally effective as the control. This is because that, in this case, the CLOSE test has a greater likelihood of erroneously claiming that there is more than one treatment better than the control.

5. DATA ANALYSIS

In the renal carcinoma (Renca) tumours study mentioned in Section 2, four groups of mice injected with Renca cells were arranged to receive no treatment (control), IL-2 alone (treatment 1), IL-12 alone (treatment 2) and IL-2 plus IL-12 (treatment 3), respectively. Figure 1 presents the Kaplan–Meier[18] estimates of the survival functions for the four groups of mice.

To assess whether IL-2, IL-12 or IL-2 plus IL-12 displays a greater anti-tumour activity than that exhibited without any agent, the two-sample weighted logrank statistics with the same type of weight function for comparing IL-2, IL-12 or IL-2 plus IL-12 with the control are computed and listed in Table V. However, the plot of the $\log\{-\log(\text{Kaplan–Meier survival estimate})\}$ in Figure 2 suggests that the proportional hazards model may not be appropriate for the pair of

Table II. Estimated experimentwise powers of SMAX or CLOSE tests for $\alpha = 0{\cdot}05$, uniform censoring distribution $U(0, R)$ and $n_0 = c$, $n_1 = n_2 = n_3 = 20$

R	θ_0	θ_1	θ_2	θ_3	$c = 20$			$c = 30$		
					LR	GH	PP	LR	GH	PP
Exponential survival distribution										
9·901	1	1	1	2·5	0·676	0·570	0·585	0·749	0·626	0·635
	1	1	1·8	2·5	0·736	0·637	0·651	0·806	0·696	0·711
	1	1	2·5	2·5	0·841	0·761	0·774	0·897	0·818	0·828
	1	1·8	2·5	2·5	0·849	0·775	0·787	0·915	0·845	0·857
	1	2·5	2·5	2·5	0·909	0·844	0·852	0·949	0·893	0·901
3·185	1	1	1	2·5	0·536	0·448	0·482	0·589	0·481	0·519
	1	1	1·8	2·5	0·593	0·508	0·542	0·655	0·568	0·607
	1	1	2·5	2·5	0·691	0·620	0·654	0·772	0·668	0·727
	1	1·8	2·5	2·5	0·715	0·640	0·671	0·790	0·715	0·752
	1	2·5	2·5	2·5	0·773	0·704	0·735	0·847	0·782	0·815
Log-normal survival distribution										
11.219	0	0	0	0·5	0·748	0·787	0·787	0·822	0·859	0·861
	0	0	0·3	0·5	0·771	0·815	0·818	0·841	0·885	0·888
	0	0	0·5	0·5	0·873	0·910	0·910	0·918	0·949	0·949
	0	0·3	0·5	0·5	0·877	0·921	0·921	0·931	0·954	0·953
	0	0·5	0·5	0·5	0·911	0·947	0·947	0·958	0·974	0·973
3·756	0	0	0	0·5	0·662	0·688	0·688	0·724	0·758	0·763
	0	0	0·3	0·5	0·711	0·741	0·741	0·765	0·801	0·802
	0	0	0·5	0·5	0·806	0·842	0·844	0·867	0·900	0·900
	0	0·3	0·5	0·5	0·814	0·844	0·845	0·879	0·913	0·915
	0	0·5	0·5	0·5	0·865	0·900	0·904	0·924	0·942	0·944

IL-12 and the control, since the vertical distance between the two plots is not a constant. The survival plot in Figure 1 further suggests that the hazards of IL-12 and the control differ at early times. Therefore, we employ the two-sample Peto–Prentice statistic for comparing IL-12 and the control, but utilize the unweighted logrank statistics for comparing IL-12 and the control as well as IL-2 plus IL-12 and the control. The relevant statistics are also reported in Table V. In addition, to evaluate whether the combined administration of IL-2 and IL-12 yields greater anti-tumour activity against the established Renca tumours than either IL-12 or IL-2 alone, two-sample weighted logrank statistics for comparing IL-2 plus IL-12 (control) with IL-2 along (treatment 1) and IL-12 alone (treatment 2), respectively, are calculated and reported in Table V.

An empirical observation from Table V is that the estimated correlations are all relatively close to 0.5, which accounts for why we compare the U_i's with the appropriate level $\alpha = 0{\cdot}05$ critical values ($z\max(3, 0{\cdot}05) \approx 2{\cdot}06$, $z\max(2, 0{\cdot}05) \approx 1{\cdot}92$ and $z\max(1, 0{\cdot}05) = 1{\cdot}65$) found in Gupta.[17] The critical value used in the CHEN test is $z(0{\cdot}017) = 2{\cdot}12$ for three treatments versus a control, while it is $z(0{\cdot}025) = 1{\cdot}96$ for comparing two treatments with a control. Both the SMAX and CHEN tests claim that, under $\alpha = 0{\cdot}05$, when compared with the no treatment control, only IL-2 plus IL-12 has an effect in regression of Renca tumour, since 3·724, 3·552 and 3·508 are greater than 2·12, but 1·788, 2·004 and 2·024 are all less than 2·06. The CLOSE test based on logrank statistics ($1{\cdot}788 < 1{\cdot}92$) reaches the same conclusion as that of the SMAX and CHEN tests, while the one based on the Gehan or Peto–Prentice statistics ($2{\cdot}004 > 1{\cdot}92$ and $2{\cdot}024 > 1{\cdot}92$, but

Table III. Estimated comparisonwise powers for $\alpha = 0{\cdot}05$, exponential survival distribution, uniform censoring distribution $U(0, R)$ and $n_0 = c$, $n_1 = n_2 = n_3 = 20$

R	θ_0	θ_1	θ_2	θ_3	SMAX			CLOSE		
					LR	GH	PP	LR	GH	PP
c = 20										
9·901	1	1	1	2·5	0·641	0·534	0·547	0·630	0·523	0·535
	1	1	1·8	2·5	0·273	0·198	0·202	0·307	0·226	0·231
	1	1	2·5	2·5	0·485	0·364	0·375	0·521	0·399	0·412
	1	1·8	2·5	2·5	0·269	0·171	0·177	0·425	0·310	0·316
	1	2·5	2·5	2·5	0·412	0·268	0·280	0·599	0·447	0·460
3·185	1	1	1	2·5	0·497	0·408	0·443	0·486	0·396	0·431
	1	1	1·8	2·5	0·189	0·135	0·152	0·212	0·151	0·170
	1	1	2·5	2·5	0·335	0·240	0·270	0·363	0·266	0·298
	1	1·8	2·5	2·5	0·162	0·099	0·114	0·282	0·200	0·221
	1	2·5	2·5	2·5	0·251	0·166	0·195	0·418	0·314	0·343
c = 30										
9·901	1	1	1	2·5	0·716	0·609	0·620	0·702	0·596	0·606
	1	1	1·8	2·5	0·327	0·227	0·234	0·364	0·263	0·271
	1	1	2·5	2·5	0·594	0·433	0·448	0·621	0·474	0·488
	1	1·8	2·5	2·5	0·305	0·189	0·198	0·470	0·339	0·350
	1	2·5	2·5	2·5	0·510	0·329	0·341	0·689	0·519	0·533
3·185	1	1	1	2·5	0·545	0·452	0·485	0·533	0·441	0·473
	1	1	1·8	2·5	0·216	0·152	0·174	0·250	0·184	0·204
	1	1	2·5	2·5	0·389	0·271	0·310	0·424	0·309	0·348
	1	1·8	2·5	2·5	0·180	0·108	0·128	0·327	0·225	0·255
	1	2·5	2·5	2·5	0·294	0·181	0·217	0·479	0·346	0·392

$1{\cdot}564 < 1{\cdot}65$ ad $1{\cdot}585 < 1{\cdot}65$) claims that both IL-12 alone and IL-2 in combination with IL-12 exhibit anti-tumour activity against the Renca tumours. However, it is not the case that all the hazard differences occur at early times. Therefore, according to the procedure based on different types of two-sample weighted logrank statistics ($3{\cdot}724 > 2{\cdot}06$, $2{\cdot}024 > 1{\cdot}92$ and $1{\cdot}696 > 1{\cdot}65$), we conclude that IL-2, IL-12 and IL-2 plus IL-12 all display a greater anti-tumour activity than that exhibited without any agent. Note that, for such a conclusion reached by the SMAX test based on different types of weighted logrank statistics ($d_1 = 1$, $d_2 = 2$, $d_3 = 3$), the approximate p-value is p (SMAX; 1, 2, 3) $= 1 - P\{Z_1 \leqslant 1{\cdot}696, Z_2 \leqslant 1{\cdot}696, Z_3 \leqslant 1{\cdot}696\} = 0{\cdot}108$. However, we observe $p'(1) = 1 - P\{Z_1 \leqslant 1{\cdot}696\} = 0{\cdot}045$, $p'(2) = 1 - P\{Z_1 \leqslant 2{\cdot}024, Z_2 \leqslant 2{\cdot}024\} = 0{\cdot}039$ and $p'(3) = 1 - P\{Z_1 \leqslant 3{\cdot}724, Z_2 \leqslant 3{\cdot}724, Z_3 \leqslant 3{\cdot}724\} = 0{\cdot}0002$. The approximate p-value of the closed test in (6) with the conclusion that all the three treatments are more effective than the control is then obtained as p(CLOSE; 1, 2, 3) $= \max\{p'(d_1), p'(d_2), p'(d_3)\} = 0{\cdot}045$.

To assess the appropriateness of the scale-change models for the three treatments versus a control setting, Figure 3 displays the quantile-quantile plots (Wei and Gail[10]) for IL-2 versus the control, IL-12 versus the control, and IL-2 plus IL-12 versus the control. Note that quantile-quantile pairs do not deviate from linearity. Let T_0, T_1, T_3 represent the survival times for the control, IL-2, IL-12 and IL-2 plus IL-12 groups, respectively. The least squared fits

Table IV. Estimated comparisonwise powers for $\alpha = 0{\cdot}05$, log-normal survival distribution, uniform censoring distribution $U(0, R)$ and $n_0 = c$, $n_1 = n_2 = n_3 = 20$

R	θ_0	θ_1	θ_2	θ_3	SMAX			CLOSE		
					LR	GH	PP	LR	GH	PP
c = 20										
11·219	0	0	0	0·5	0·702	0·751	0·754	0·698	0·739	0·742
	0	0	0·3	0·5	0·306	0·323	0·324	0·325	0·360	0·361
	0	0	0·5	0·5	0·614	0·653	0·655	0·640	0·683	0·684
	0	0·3	0·5	0·5	0·305	0·313	0·315	0·457	0·483	0·485
	0	0·5	0·5	0·5	0·563	0·598	0·600	0·718	0·763	0·763
3·576	0	0	0	0·5	0·639	0·670	0·671	0·626	0·655	0·657
	0	0	0·3	0·5	0·258	0·263	0·267	0·281	0·298	0·300
	0	0	0·5	0·5	0·515	0·533	0·537	0·538	0·566	0·568
	0	0·3	0·5	0·5	0·257	0·256	0·257	0·393	0·422	0·427
	0	0·5	0·5	0·5	0·444	0·454	0·459	0·605	0·635	0·635
c = 30										
11·219	0	0	0	0·5	0·778	0·817	0·817	0·767	0·801	0·801
	0	0	0·3	0·5	0·343	0·370	0·371	0·375	0·413	0·413
	0	0	0·5	0·5	0·677	0·730	0·731	0·700	0·761	0·761
	0	0·3	0·5	0·5	0·332	0·358	0·359	0·494	0·545	0·546
	0	0·5	0·5	0·5	0·624	0·676	0·676	0·778	0·825	0·825
3·576	0	0	0	0·5	0·700	0·736	0·737	0·686	0·723	0·726
	0	0	0·3	0·5	0·280	0·291	0·295	0·311	0·332	0·335
	0	0	0·5	0·5	0·560	0·595	0·602	0·595	0·640	0·641
	0	0·3	0·5	0·5	0·261	0·276	0·279	0·425	0·450	0·452
	0	0·5	0·5	0·5	0·493	0·520	0·530	0·677	0·716	0·718

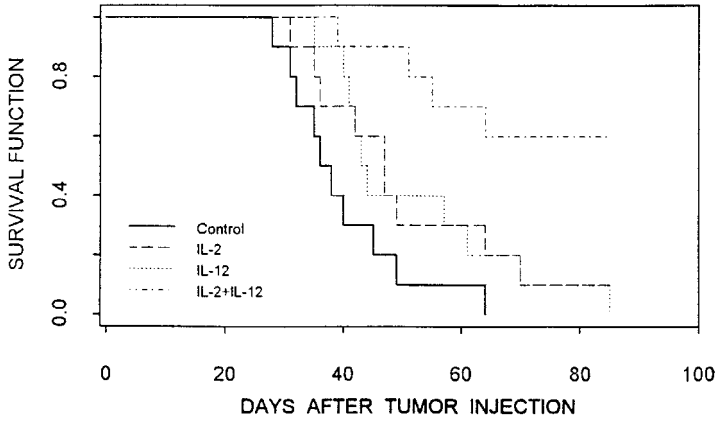

Figure 1. Kaplan–Meier estimates for the interleukin–tumour example

Table V. Summary statistics for the interleukin–tumour example

Statistics	Logrank		Gehan		Peto–Prentice	
(a) 0 for no treatment, 1 for IL-2, 2 for IL-12 and 3 for IL-2 + IL-12						
U_1 (U_{01}, s_{11})	1·696	(3·270, 3·715)	1·564	(2·050, 1·719)	1·585	(1·854, 1·368)
U_2 (U_{02}, s_{22})	1·788	(3·560, 3·962)	2·004	(2·600, 1·683)	2·024	(2·394, 1·399)
U_3 (U_{03}, s_{33})	3·724	(5·975, 2·575)	3·552	(4·350, 1·500)	3·508	(4·014, 1·309)
	1·000	0·553 0·527	1·000	0·476 0·459	1·000	0·479 0·460
S		1·000 0·524		1·000 0·453		1·000 0·458
		1·000		1·000		1·000
U_1: Logrank	1·696			1·000 0·478	0·527	
U_2: Peto–Prentice	2·024	**S** =		1·000	0·456	
U_3: Logrank	3.724				1·000	
(b) 0 for IL-2 + IL-12, 1 for IL-2 and 2 for IL-12						
U_1 (U_{10}, s_{11})	2·523	(4·252, 2·840)	2·491	(3·150, 1·599)	2·456	(2·867, 1·353)
U_2 (U_{20}, s_{22})	2·819	(4·941, 3·074)	2·424	(3·100, 1·635)	2·389	(2·853, 1·427)
	1·000	0·451	1·000	0·493	1·000	0·519
S	1·000	1·000		1·000		

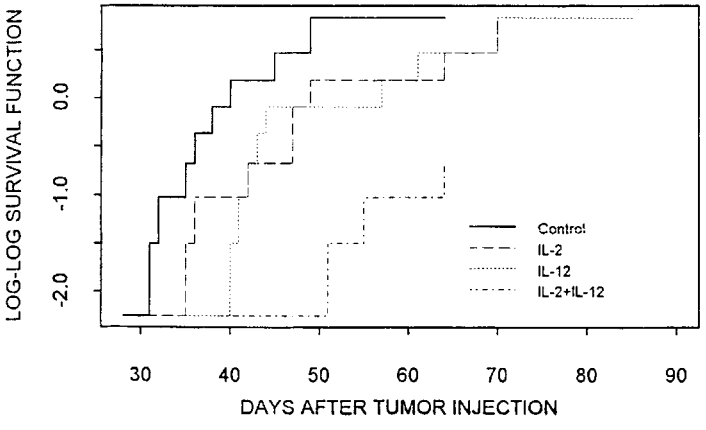

Figure 2. The log{ − log(Kaplan–Meier estimate)} for the interleukin–tumour example

suggested scale-change models for (T_0-27, T_1-27), (T_0, T_2) and T_0-24, T_3-24). Let γ_1, γ_2 and γ_3 be the ratios of the scale parameters for T_1-27 and T_0-27, T_2 and T_0, and T_3-25 and T_0-25, respectively. A 95 per cent confidence set for (γ_1, γ_2, γ_3), $\{(\gamma_1, \gamma_2, \gamma_3)$: $\gamma_1 \geqslant \boldsymbol{\gamma}_1, \gamma_2 \geqslant \boldsymbol{\gamma}_2, \gamma_3 \geqslant \boldsymbol{\gamma}_3\}$ based on the procedure with difference types of two-sample weighted logrank statistics then yields (γ_1, γ_2, γ_3) = (0·91, 0·99, 2·51). Note that these findings are in a good agreement with the results of the SMAX test (3·724 > 2·06, but 2·024 < 2·06).

Note that the log-log survival plot in Figure 2 reveals that the proportional hazards model is reasonable for the pair of IL-2 alone and IL-2 plus IL-12 or IL-12 alone and IL-2 plus IL-12. Hence, based on two-sample unweighted logrank statistics, the SMAX and CLOSE tests claim, at

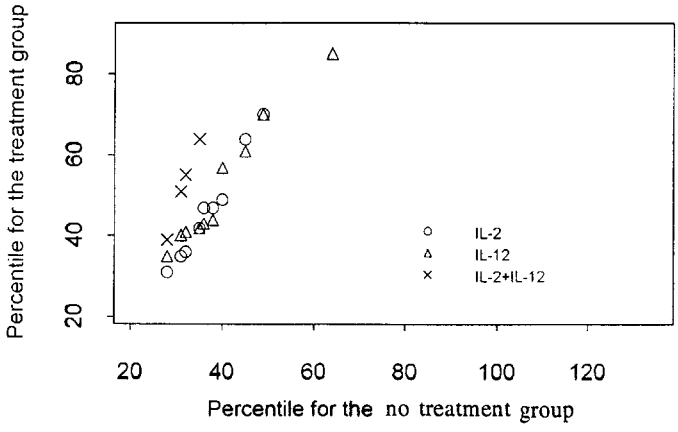

Figure 3. Quantile-quantile plot for the interleukin–tumour example

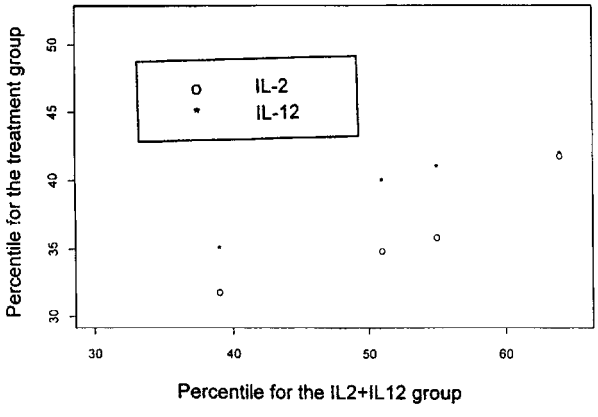

Figure 4. Quantile-quantile plot for the interleukin–tumour example

$\alpha = 0{\cdot}05$, that IL-2 plus IL-12 yields better anti-tumour activity against the Renca tumours than IL-2 or IL-12 alone ($2{\cdot}819 > 1{\cdot}92$ and $2{\cdot}525 > 1{\cdot}92$). Moreover, for such a conclusion, the approximate p-value of the SMAX test is $p(\text{SMAX};\ 1,\ 2) = 1 - P\{Z_1 \leqslant 2{\cdot}525, Z_2 \leqslant 2{\cdot}525\} = 0{\cdot}011$, while the values of $p'(1) = 1 - P\{Z_1 \leqslant 2{\cdot}525\} = 0{\cdot}006$ and $p'(2) = 1 - P\{Z_1 \leqslant 2{\cdot}819, Z_2 \leqslant 2{\cdot}819\} = 0{\cdot}005$ produce an approximate p-value of the closed test in (6) as $p(\text{CLOSE};\ 1,\ 2) = \max\{p'(d_1), p'(d_2)\} = 0{\cdot}006$.

The quantile-quantile plots for IL-2 versus IL-2 plus IL-12 versus IL-2 plus IL-12 given in Figure 4 also support the scale-change model. Let T_0, T_1 and T_2 represent the survival times for the IL-2 plus IL-12, IL-2 and IL-12 groups, respectively. The least squared fits suggested scale-change models for (T_0-26, T_1-26), (T_0-34, T_2-34). Let γ_1 and γ_2 be the ratios of the scale parameters for T_0-26 and T_1-26, and T_0-34 and T_2-34, respectively. A95 per cent confidence set for (γ_1, γ_2), $\{(\gamma_1, \gamma_2 : \gamma_1 \geqslant \gamma_1, \gamma_2 \geqslant \gamma_2\}$ based on the SMAX test with unweighted logrank statistics then

yields $(\gamma_1, \gamma_2) = (1{\cdot}34, 1{\cdot}88)$. These findings confirm, again, the results based on the unweighted logrank SMAX test for comparing IL-2 and IL-12 with IL-2 plus IL-12.

6. CONCLUSION

This paper presents a generalized Steel's test (SMAX) and a closed multiple testing procedure (CLOSE), which is actually a modification of the generalized Steel's test, based on two-sample weighted logrank statistics to determine which treatments are more effective than (or different from) the control when survival data are subject to random right-censorship. When the scale-change model is appropriate for such comparisons, a confidence set on the basis of the SMAX test for the scale-changes of each treatment group and the control is suggested. However, if only a significance testing procedure is required, the CLOSE test is recommended, since it has a better comparisonwise power performance than the SMAX test in most practical situations. The *p*-values of the testing procedures are also provided which measure the strength of the evidence for the conclusion that a certain treatment is more effective that (or different from) the control.

The appropriate choice of the two-sample weighted logrank statistics in the treatments versus a control setting is the same as that in the two-sample problem which has been extensively discussed (see, for example, Fleming and Harrington[9]). The unweighted logrank statistic should be used when the assumption of proportional hazards is tenable. For testing against early hazard differences, both the Gehan and Peto–Prentice statistics are better than the logrank statistic. However, according to Anderson *et al.*,[22] the weight function used in the Peto–Prentice test depends only on the survival experience, meanwhile the Gehan test uses a weight function that depends on survivals as well as censorings. Therefore, we suggest employing the two-sample Peto–Prentice statistic in comparing the corresponding treatment with the control when their hazards are different at early times. For the situation where hazard difference occurs at late times, according to the suggestions made by Fleming and Harrington,[9] the appropriate weight function $W_i(T_k)$ used in (1) and (2) should be $\{1 -$ the Kaplan–Meier[18] survival estimate$\}$. Finally, the log–log survival plot shown in Figure 2 is usually employed for checking the proportional hazards model. The survival plots presented in Figure 1 should be helpful for assessing if the hazard difference occurs at early or late times.

APPENDIX

Asymptotic null distribution of $(U_1, U_2, \ldots, U_m)$

Let $T_1 < \cdots < T_L$ denote the ordered observed distinct death times in the sample formed by combining the ith, jth and control groups, Let d_{uk} and Y_{uk}, $k = 1, \ldots, L$, $u = 0, i, j$, denote the number of observed deaths and number at risk, respectively, in sample u at time T_k. Set $Y_{+1k} = Y_{0k} + Y_{1k}$, $1 = i, j$, $Y_{++k} = Y_{0k} + Y_{ik} + Y_{jk}$, and $d_{++k} = d_{0k} + d_{ik} + d_{jk}$. For the three weight functions considered herein and under some regular conditions. Chen[9] showed that the null (H_0) asymptotic distribution of $(U_1, U_2, \ldots, U_m)$ is an m-variate normal with mean zero and covariance matrix $\Sigma = (\rho_{ij})$, which can be consistently estimated by $\mathbf{S} = (s_{ij}/\sqrt{s_{ii}\, s_{ij}})$, where s_{ii} are stated in (2) and

$$s_{ij} = \sum_{k=1}^{L} W_i(T_k)\, W_j(T_k)\, v_k, \qquad i \neq j = 1, 2, \ldots, m$$

with $v_k = d_{++k} Y_{0k} Y_{ik} Y_{jk} (Y_{++k} - d_{++k}) / \{ Y_{+ik}\, Y_{+jk}\, Y_{++k} (Y_{++k} - 1) \}$, $k = 1, 2, \ldots, L$.

Description of left-sided and two-sided comparisons

When the control is expected to be at least as good as the control, the two-sample weighted logrank statistics employed in the multiple testing procedures in (3), (4) and (6) would be the ones which compare the control with the ith treatment, namely

$$U_{i0} = \sum_{k=1}^{L} W_i(T_k)\,(d_{ik} - e_{ik}), \qquad i = 1, 2, \ldots, m$$

where $e_{ik} = d_{+k}Y_{ik}/Y_{+k}$. Let $\gamma_i = \theta_0/\theta_i$, $i = 1, 2, \ldots, m$. Then, an approximate $(1-\alpha)\times 100$ per cent confidence set for the $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)$ can be obtained as specified in (5), but replacing $h(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i)$ and $V(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i)$ with $h(\mathbf{X}_i, \mathbf{X}_0/a, \Delta_i, \Delta_0)$ and $V(\mathbf{X}_i, \mathbf{X}_0/a, \Delta_i, \Delta_0)$, respectively. Under the two-sided alternative that there is at least one treatment which is different from the control, the appropriate statistics in the generalized Steel's test are $|U_i|$, $i = 1, 2, \ldots, m$, where $|x|$ is the absolute value of x. The corresponding critical value, denoted by $|z|\max(m, \alpha)$, is the upper αth percentile of the distribution of $\max(|Z_1|, |Z_2|, \ldots, |Z_m|)$, where $(Z_1, Z_2, \ldots, Z_m)$ is, again, an m-variate normal vector with mean zero and covariance matrix $\mathbf{S}$. For a common censoring distribution and equal sample sizes, some values of the $|z|\max(m, \alpha)$ can be found in Dunnett.[23] Let $|U|_{(1)} < |U|_{(2)} < \cdots < \cdots |U|_{(m)}$ be the order statistics of the $|U_i|$'s. The closed testing procedure for the two-sided alternative is then the same as in (6) except that the $U_{(i)}$ and $z\max((i), \alpha)$ are replaced by $|U|_{(i)}$ and $|z|\max((i)\,\alpha)$, respectively. Moreover, an approximate $(1-\alpha)\times 100$ per cent test-based confidence set for the ratios of the scale parameters $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)$, where $\gamma_i = \theta_i/\theta_0$, $i = 1, 2, \ldots, m$, in a scale-change model is given by

$$C(\gamma) = \{(\gamma_1, \gamma_2, \ldots, \gamma_m)\colon \underline{\gamma}_i \leqslant \gamma_i \leqslant \bar{\gamma}_i, \qquad i = 1, 2, \ldots, m\}$$

where $\underline{\gamma}_i$ is the smallest value of a such that

$$h(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i) \leqslant |z|\max(m, \alpha)\sqrt{\{V(\mathbf{X}_0, \mathbf{X}_i/a, \Delta_0, \Delta_i)\}}$$

and $\bar{\gamma}_i$ is the largest value of b satisfying

$$h(\mathbf{X}_0, \mathbf{X}_i/b, \Delta_0, \Delta_i) \geqslant -|z|\max(m, \alpha)\sqrt{\{V(\mathbf{X}_0, \mathbf{X}_i/b, \Delta_0, \Delta_i)\}}.$$

ACKNOWLEDGEMENTS

The author would like to thank the editor and referees for many suggestions which improved the presentation of this paper. This work was supported by the National Science Council of Taiwan under contract number NSC86-2115-M-008-018.

REFERENCES

1. Hochberg, Y. and Tamhane, A. C. *Multiple Comparison Procedures*, Wiley, New York, 1987.
2. Chakraborti, S. and Desu, M. M. 'Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship', *Statistica Neerlandica*, **45**, 227–254 (1991).
3. Slepian, D. The one-sided barrier problem for Gaussan noise', *Bell System Technical Journal*, **41**, 463–501 (1962).
4. Gehan, E. A. 'A generalized Wilcoxon test for comparing arbitrarily singly-censored samples', *Biometrika*, **52**, 203–223 (1965).
5. Chen, Y. I. 'A generalized Steel's procedure for comparing several treatments with a control under random right-censorship', *Communications in Statistics – Simulation and Computation*, **23**, 1–16 (1994).
6. Steel, R. G. D. 'A multiple comparison rank sum test: treatments versus control', *Biometrics*, **15**, 560–572 (1959).

7. Mantel, N. 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports*, **50**, 163–170 (1966).
8. Fleming, T. R. and Harrington, D. P. *Counting Process nad Survival Analysis*, Wiley, New York, 1991.
9. Chen, Y. I. 'Simple-tree weighted logrank tests for right-censored data', *Annals of the Institute of Statistical Mathematics*, **50**, 311–324 (1998).
10. Wei, L. J. and Gail, M. H. 'Nonparametric estimation for a scale-change with censored observations', *Journal of the American Statistical Association*, **78**, 382–388 (1983).
11. Marcus, R., Peritz, E. and Gabriel, K. R. 'One closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **63**, 655–660 (1976).
12. Peto, R. and Peto, J. 'Asymptotically efficient rank invariant test procedures (with discussion)', *Journal of the Royal Statistical Society*, *Series A*, **135**, 185–206 (1972).
13. Prentice, R. L. 'Linear rank tests with right censored data', *Biometrika*, **65**, 165–179 (1978).
14. Gately, M. K. 'Interleukin-12: a recently discovered cytokine with potential for enhancing cell-mediated immune responses to tumours', *Cancer Investigations*, **11**, 500–506 (1993).
15. Maas, R. A., Dullens, H. F. and Den Otter, W. 'Interleukin-2 in cancer treatment: disappointing or (still) promising? A review', *Cancer Immunology Immunotherapy*, **36**, 141–148 (1993).
16. Rossi, A. R., Pericle, F., Rashleigh, S., Janier, J. and Djeu, J. Y. 'Lysis of neuroblastoma cell lines by human natural killer cells activated by interleukin-2 and interleukin-12, *Blood*, **83**, 323–328 (1994).
17. Wigginton, J. M., Komschlies, K. L., Back, T. C., Franco, J. L., Brunda, M. J. and Wiltrout, R. H. 'Administration of interleukin 12 with pulse interleukin 2 and the rapid and complete eradication of murine renal carcinoma', *Journal of the National Cancer Institute*, **88**, 38–43 (1996).
18. Kaplan, E. L. and Meier, P. 'Nonparametric estimator from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
19. Schervish, M. J. 'Multivariate Normal Probabilities with error bound', *Applied Statistics*, **33**, 81–94 (1984).
20. Gupta, S. S. 'Probability integrals of multivariate normal and multivariate *t*', *Annals of Mathematical Statistics*, **34**, 792–828 (1963).
21. Dunnett C. W. and Tamhane, A. C. 'Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts', *Statistics in Medicine*, **10**, 939–947 (1991).
22. Anderson, P. K., Borgan, O., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993.
23. Dunnett, C. W. 'New tables for multiple comparisons with a control', *Biometrics*, **20**, 482–491 (1964).

