STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2509–2523
Published online 25 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2125

Pairwise multiple comparison adjustment in survival analysis

Brent R. Logan[1,*,†], Hong Wang[2] and Mei-Jie Zhang[1]

[1]*Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509, U.S.A.*
[2]*Department of Biostatistics, University of Pittsburgh, 201 N. Craig Street, Suite 325, Pittsburgh, PA 15213, U.S.A.*

SUMMARY

Many clinical studies have as their endpoint the time until some event (such as death) occurs. Often in such studies researchers are interested in comparing several treatment or prognostic groups with one another in terms of their survival curves. When many such pairwise group comparisons are done, the chance of finding a false significance among all of the comparisons is inflated above the usual desired significance level. This paper investigates methods of adjusting the survival analysis for the number of comparisons being made. These methods are applied to a retrospective study conducted by the International Bone Marrow Transplant Registry and compared in a simulation study in terms of the power to detect actual differences in the survival curves between the groups. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Bonferroni procedure; step-down procedure; closed test procedure; log-rank test

1. INTRODUCTION

Many biological, epidemiological, and cancer clinical studies have as their endpoint the time until some event occurs. Often in such studies researchers are interested in comparing several treatment or prognostic groups with one another. When this is done, the chance of making at least one type I error, or finding a falsely significant difference between any two groups, is inflated above the desired α level. Often such analyses are done without any adjustment for multiple comparisons, resulting in an excess of type I errors. A more appropriate criterion to control when making several comparisons is the familywise error (FWE) rate, which is the chance of making at least one type I error among all treatment comparisons being made.

*Correspondence to: Brent R. Logan, Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509, U.S.A.
†E-mail: blogan@mcw.edu

Contract/grant sponsor: American Cancer Society and Medical College of Wisconsin Cancer Center
Contract/grant sponsor: National Cancer Institute; contract/grant number: 2 R01 CA54706-10

Received March 2004
Accepted October 2004

The simplest way to adjust for multiple pairwise comparisons is to apply the Bonferroni method, where instead of performing each pairwise log-rank test at the α-level, one uses an α/C level instead, where C is the number of pairwise comparisons being made. While this method is simple, it has been found to be conservative when applied to other endpoints besides survival. This means that it has low power to detect actual survival differences between treatment groups. There are two reasons for this conservatism. First, the Bonferroni method ignores the correlations between the tests which is present because each treatment group is being compared to each other treatment. Second, the Bonferroni method makes no allowance for situations where some treatments have clearly different survival curves, in which case less multiplicity adjustment is needed.

Multiple comparison procedures which control the FWE have been proposed for several types of endpoints, including normally distributed response variables and binary response variables [1, 2]. However, comparing several groups in terms of their survival curves has received much less attention in the literature. Koziol and Reid [3] consider methods for adjusting pairwise log-rank tests which are sharper than the Bonferroni adjustment, by using the Sidak inequality. This accounts for some of the correlation among the tests; however, their method is still conservative. Chen [4] considers methods which account for correlation for the simpler multiple comparisons with a control setting. We focus on appropriate adjustments for correlation in the general setting of pairwise comparisons of survival curves using the weighted log-rank tests. In Section 2, we lay out the notation and background. Section 3 reviews simple Bonferroni-type procedures. In Section 4 we discuss procedures which directly account for the correlation among the test statistics. Section 5 discusses an application of the closed test procedure to survival endpoints. Simulations are conducted to compare the various procedures in Section 6. Finally, these methods are applied to a retrospective research study in allogeneic bone marrow transplantation in Section 7.

2. NOTATION AND BACKGROUND

Let the observations on subject ℓ of treatment group i be $\{T_{i\ell}, D_{i\ell}\}$ where $D_{i\ell}=0$ if subject (i,ℓ) is censored, $D_{i\ell}=1$ otherwise, and $T_{i\ell}$ is the observation time of subject (i,ℓ), for $i=1,\ldots,K$ and $\ell=1,\ldots,n_i$.

Let $N_{i\ell}(u)=I\{T_{i\ell}\leqslant u, D_{i\ell}=1\}$ be the counting process and $Y_{i\ell}(u)=\mathscr{I}\{T_{i\ell}\geqslant u\}$ be the indicator of whether the ℓth individual is at risk at time u and is in the ith treatment group. Assume that the counting process $N_{i\ell}$ has intensity $\lambda_{i\ell}(t)=Y_{i\ell}(t)\,\lambda_i(t)$. The processes $M_{i\ell}(t)=N_{i\ell}(t)-\int_0^t \lambda_{i\ell}(s)\,\mathrm{d}s$ are local martingales. Denote $N_{i+}=\sum_\ell N_{i\ell}$, $M_{i+}=\sum_\ell M_{i\ell}$, $Y_{i+}=\sum_\ell Y_{i\ell}$, $N_{ij+}=N_{i+}+N_{j+}$ and $Y_{ij+}=Y_{i+}+Y_{j+}$ for treatment groups i and j.

We are interested in all pairwise hypotheses, $H_{ij}:\lambda_i(t)=\lambda_j(t)$, $\forall t$, where $1\leqslant i<j\leqslant K$, and K is the number of treatment groups. To test these hypotheses, we consider pairwise weighted log-rank tests

$$X_{ij}=\int_0^\tau K_{ij}(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\left\{\frac{\mathrm{d}N_{i+}(u)}{Y_{i+}(u)}-\frac{\mathrm{d}N_{j+}(u)}{Y_{j+}(u)}\right\}$$

where $K_{ij}(t)$ is the weight function. Note that this leads to the standard log-rank test when $K_{ij}(t)=1$. Alternatively, Fleming and Harrington [5] suggested a class of weights

$$K_{ij}(t)=\{\widehat{S}_{ij}(t)\}^{\rho}\{1-\widehat{S}_{ij}(t)\}^{\gamma} \tag{1}$$

where $\widehat{S}_{ij}$ is the Kaplan–Meier estimated survival function from the pooled ith and jth samples, and $0\leqslant\rho,\gamma\leqslant 1$. When more than two groups are being compared, one may wish to define a consistent weight function across all pairwise comparisons of the form $K_{ij}(t)=\{\widehat{S}_{p}(t)\}^{\rho}\{1-\widehat{S}_{p}(t)\}^{\gamma}$, where $\widehat{S}_{p}$ is the Kaplan–Meier estimated survival function pooled across all samples.

Note that

$$X_{ij}=\int_0^{\tau}K_{ij}(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\left\{\frac{\mathrm{d}M_{i+}(u)}{Y_{i+}(u)}-\frac{\mathrm{d}M_{j+}(u)}{Y_{j+}(u)}\right\}$$
$$+\int_0^{\tau}K_{ij}(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\{J_i(u)\lambda_i(u)-J_j(u)\lambda_j(u)\}\,\mathrm{d}u$$

where $J_i(u)=I\{Y_{i+}(u)>0\}$. Therefore, under H_{ij} and some regularity conditions [6], the test statistic X_{ij} is asymptotically equivalent to the random process

$$W_{ij}=\int_0^{\tau}K_{ij}(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\left\{\frac{\mathrm{d}M_{i+}(u)}{Y_{i+}(u)}-\frac{\mathrm{d}M_{j+}(u)}{Y_{j+}(u)}\right\} \tag{2}$$

which asymptotically converges to a normal distribution with mean zero and variance of

$$\sigma_{ij}^2=\int_0^{\tau}K_{ij}^2(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\lambda_i(u)\,\mathrm{d}u$$

The variance can be consistently estimated by

$$\widehat{\sigma}_{ij}^2=\int_0^{\tau}K_{ij}^2(u)\frac{Y_{i+}(u)Y_{j+}(u)}{(Y_{ij+}(u))^2}\,\mathrm{d}N_{ij+}(u)$$

Then the usual weighted log-rank test rejects H_{ij} if

$$Z_{ij}=\frac{|X_{ij}|}{\sqrt{\widehat{\sigma}_{ij}^2}}>Z_{\alpha/2}$$

or equivalently if

$$\chi_{ij}^2=Z_{ij}^2>\chi_{1,\alpha}^2. \tag{3}$$

3. BONFERRONI-TYPE PROCEDURES

When there are K groups, applying the usual weighted log-rank test at level α to each H_{ij} results in an inflation of the FWE above α. Next, we consider procedures which adjust each individual test to maintain a FWE of α or less.

The Bonferroni procedure tests each one by reducing the individual error rate applied to each test by a factor of C, so that we reject H_{ij} if

$$Z_{ij} > Z_{\alpha/2/C}$$

The Bonferroni procedure can be conservative because it ignores information in the other pairwise test statistics when setting the critical value. For example, when some of the hypotheses are clearly false, the multiplicity problem can be reduced. This is because the FWE is the probability of at least one incorrect rejection of a null hypothesis. If some hypotheses are clearly false (i.e. highly significant p-values), then the incorrect rejection must occur from one of the remaining hypotheses. Since there are fewer candidates for incorrectly rejected null hypotheses, then the multiplicity adjustment can be smaller. This indicates that stepwise procedures can be more efficient, so that as hypotheses are rejected, the multiplicity adjustment for subsequent hypotheses is reduced. The Bonferroni method is considered a single-step (SS) procedure, because it uses the same multiplicity adjustment (α/C) for all pairwise comparisons. It is well known that stepwise procedures or other procedures which account for the significance of the other tests can be substantially more powerful to reject subsequent hypotheses than single-step procedures. Therefore, we also consider such stepwise procedures which are appropriate for comparing several survival curves.

The simplest stepwise (specifically step-down (SD)) method is Holm's [7] sequentially rejective procedure. In this procedure, one orders the Z-statistics from the log-rank tests so that $|Z|_{(1)} \leqslant \cdots \leqslant |Z|_{(C)}$, and the corresponding hypotheses are $H_{(1)}, \ldots, H_{(C)}$. First, check if $|Z_{(C)}| \leqslant Z_{\alpha/2/C}$. If so, then all hypotheses are accepted. If not, then reject $H_{(C)}$ and proceed to test $H_{(C-1)}$. If $|Z_{(C-1)}| \leqslant Z_{\alpha/2/(C-1)}$ then accept $H_{(C-1)}, \ldots, H_{(1)}$ and stop. Otherwise, reject $H_{(C-1)}$ and proceed to test $H_{(C-2)}$ and so on. Suppose that hypothesis $H_{(k)}$ represents the first hypothesis where $|Z_{(k)}| < Z_{\alpha/2/k}$. Then using Holm's procedure, one accepts all hypotheses $H_{(k)}, \ldots, H_{(1)}$ and rejects all hypotheses $H_{(C)}, \ldots, H_{(k+1)}$. Holm's procedure is more powerful than the Bonferroni procedure, because each test is only adjusted for the number of tests which have not already been rejected (α/k), rather than the entire set of tests (α/C). However, Holm's procedure does not take into account the correlation among the tests, and so there is room for further improvement.

4. PROCEDURES WHICH ACCOUNT FOR CORRELATION

To account for correlation while controlling the FWE, one can determine the critical value based on the maximum of the standardized weighted log-rank statistics. Therefore, we need the $(1 - \alpha)$ percentile of the distribution of $\max |Z_{ij}|$. This section proposes two ways of determining this critical value. First, we derive the joint asymptotic distribution of the pairwise log-rank statistics, and simulate from this multivariate normal (MVN) distribution to determine the appropriate percentile. An alternative method discussed subsequently is to simulate the martingales directly under the null hypothesis.

4.1. Joint asymptotic distribution of $\boldsymbol{X}$

To account for correlation among the test statistics, we can use the joint asymptotic distribution of the weighted log-rank statistics under the overall null hypothesis $H_0 : \cap_{i,j} H_{ij}$, where all

groups have the same hazard rates. The test statistics form a vector $\boldsymbol{X}=(X_{12},\ldots,X_{K-1,K})'$. Under the overall null hypothesis, we have shown that any two comparisons X_{ij} and $X_{i'j'}$ are asymptotically equivalent to

$$W_{ij}=\int_0^\tau K_{ij}(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\left\{\frac{\mathrm{d}M_{i+}(u)}{Y_{i+}(u)}-\frac{\mathrm{d}M_{j+}(u)}{Y_{j+}(u)}\right\}$$

$$W_{i'j'}=\int_0^\tau K_{i'j'}(u)\frac{Y_{i'+}(u)Y_{j'+}(u)}{Y_{i'j'+}(u)}\left\{\frac{\mathrm{d}M_{i'+}(u)}{Y_{i'+}(u)}-\frac{\mathrm{d}M_{j'+}(u)}{Y_{j'+}(u)}\right\}$$

Then the predictable variation process of these two statistics is

$$\langle X_{ij},X_{ij}\rangle_t=\int_0^t K_{ij}^2(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)}\lambda_i(u)\,\mathrm{d}u$$

$$\langle X_{ij},X_{ij'}\rangle_t=\int_0^t K_{ij}(u)K_{ij'}(u)\frac{Y_{j+}(u)Y_{j'+}(u)}{Y_{ij+}(u)Y_{ij'+}(u)}Y_{i+}(u)\lambda_i(u)\,\mathrm{d}u\quad\text{for } j\neq j'$$

$$\langle X_{ij},X_{i'j}\rangle_t=\int_0^t K_{ij}(u)K_{i'j}(u)\frac{Y_{i+}(u)Y_{i'+}(u)}{Y_{ij+}(u)Y_{i'j+}(u)}Y_{j+}(u)\lambda_j(u)\,\mathrm{d}u\quad\text{for } i\neq i'$$

$$\langle X_{ij},X_{jj'}\rangle_t=-\int_0^t K_{ij}(u)K_{jj'}(u)\frac{Y_{i+}(u)Y_{j'+}(u)}{Y_{ij+}(u)Y_{jj'+}(u)}Y_{j+}(u)\lambda_j(u)\,\mathrm{d}u$$

$$\langle X_{ij},X_{i'j'}\rangle_t=0\quad\text{for } i\neq i',j\neq j',j\neq i',i\neq j'$$

Thus we can show that under some regularity conditions and the overall null hypothesis, the vector of test statistics $\boldsymbol{X}$ converges to a MVN process with mean zero and the variance–covariance matrix $\boldsymbol{\Sigma}$, which can be estimated consistently by $\widehat{\boldsymbol{\Sigma}}$ where

$$\widehat{\boldsymbol{\Sigma}}_{ij,ij}=\int_0^\tau K_{ij}^2(u)\frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}^2(u)}\,\mathrm{d}N_{ij+}(u)$$

$$\widehat{\boldsymbol{\Sigma}}_{ij,ij'}=\int_0^\tau K_{ij}(u)K_{ij'}(u)\frac{Y_{i+}(u)Y_{j+}(u)Y_{j'+}(u)}{Y_{ij+}(u)Y_{ij'+}(u)Y_{ijj'+}(u)}\,\mathrm{d}N_{ijj'+}(u)\quad\text{for } j\neq j'$$

$$\widehat{\boldsymbol{\Sigma}}_{ij,i'j}=\int_0^\tau K_{ij}(u)K_{i'j}(u)\frac{Y_{i+}(u)Y_{i'+}(u)Y_{j+}(u)}{Y_{ij+}(u)Y_{i'j+}(u)Y_{iji'+}(u)}\,\mathrm{d}N_{iji'+}(u)\quad\text{for } i\neq i'$$

$$\widehat{\boldsymbol{\Sigma}}_{ij,jj'}=-\int_0^\tau K_{ij}(u)K_{jj'}(u)\frac{Y_{i+}(u)Y_{j+}(u)Y_{j'+}(u)}{Y_{ij+}(u)Y_{jj'+}(u)Y_{ijj'+}(u)}\,\mathrm{d}N_{ijj'+}(u)$$

$$\widehat{\boldsymbol{\Sigma}}_{ij,i'j'}=0\quad\text{for } i\neq i',j\neq j',j\neq i',i\neq j'\tag{4}$$

Equivalently, the standardized weighted log-rank statistics Z_{ij} converge to a multivariate normal process with mean zero and covariance matrix $\boldsymbol{R}$, where $R_{ij,ij} = 1$ and

$$R_{ij,i'j'} = \frac{\widehat{\boldsymbol{\Sigma}}_{ij,i'j'}}{\hat{\sigma}_{ij}\hat{\sigma}_{i'j'}}$$

To control the FWE, we set the critical value at the $(1-\alpha)$ percentile of the distribution of $\max|Z_{ij}|$. This can be obtained by parametric resampling in the following way: Generate B random samples from this multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{R}$, $\boldsymbol{Z}^b = (Z^b_{12}, \ldots, Z^b_{K-1,K})$, for $b = 1, \ldots, B$. Compute $Z^b_{\max} = \max_{1 \leqslant i < j \leqslant K} |Z^b_{ij}|$. Then the appropriate critical value D_α is the $(1-\alpha)$100th percentile of $Z^b_{\max}$.

Furthermore, we can use this method of generating multiplicity adjusted critical values within a more powerful step-down procedure. Suppose that the log-rank statistics are ordered such that $|Z_{(1)}| \leqslant \cdots \leqslant |Z_{(C)}|$. Then a step-down procedure tests these statistics in sequence, starting with $Z_{(C)}$, using a series of critical values $D_{(1),\alpha}, \ldots, D_{(C),\alpha}$ corresponding to each ordered test statistic $|Z_{(1)}|, \ldots, |Z_{(C)}|$, rather than the same critical value D_α for each one. First check if $|Z_{(C)}| \leqslant D_{(C),\alpha}$. If so, then all hypotheses are accepted. If not, then reject $H_{(C)}$ and proceed to test $H_{(C-1)}$. If $|Z_{(C-1)}| \leqslant D_{(C-1),\alpha}$ then accept $H_{(C-1)}, \ldots, H_{(1)}$ and stop. Otherwise, reject $H_{(C-1)}$ and proceed to test $H_{(C-2)}$ and so on. Suppose that hypothesis $H_{(k)}$ represents the first hypothesis where $|Z_{(k)}| < D_{(k),\alpha}$. Then using this step-down procedure, one accepts all hypotheses $H_{(k)}, \ldots, H_{(1)}$ and rejects all hypotheses $H_{(C)}, \ldots, H_{(k+1)}$. At each stage, the critical value $D_{(k),\alpha}$ can be obtained in the following manner. First re-order the covariance matrix $\boldsymbol{R}$ to match the ordering $Z_{(1)}, \ldots, Z_{(C)}$. Generate B random samples from this multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{R}$ as above, $\boldsymbol{Z}^b = (Z^b_{(1)}, \ldots, Z^b_{(C)})$, for $b = 1, \ldots, B$. Compute $Z^b_{\max,(k)} = \max_{1 \leqslant i \leqslant k} |Z_{(i)}|^b$. Then the appropriate critical value $D_{(k),\alpha}$ is the $(1-\alpha)$100th percentile of $Z^b_{\max,(k)}$. Note that D_α from the single step procedure is identical to $D_{(C),\alpha}$ in the step-down procedure. Also note that computationally, the random samples only need to be generated once, and can be reused in determination of the critical value at each step by simply computing $Z^b_{\max,(k)}$ over a successively smaller index k. This step-down procedure is expected to be more powerful than the Holm procedure because it utilizes the correlation among the test statistics to generate the appropriate sequence of critical values, rather than relying upon the conservative Bonferroni critical value at each step.

4.2. Accounting for correlation: simulated martingales

An alternative way of accounting for correlation is to directly simulate the testing processes under the null hypotheses in order to generate the multiplicity adjusted critical values. This strategy may be useful for example in more complicated settings such as adjusting for covariates, dealing with non-proportional hazards, or using a test other than a weighted log-rank test, where the direct asymptotic results are more difficult to obtain.

As mentioned previously, under the overall null hypothesis, $\boldsymbol{X}$ is asymptotically equivalent to $\boldsymbol{W} = (W_{12}, \ldots, W_{K-1,K})'$, where W_{ij} is given in (2). By the results in Lin *et al.* [8], if one replaces $\{M_{i\ell}(u)\}$ with $\{G_{i\ell}N_{i\ell}(u)\}$ in $\boldsymbol{W}$ to yield $\widetilde{\boldsymbol{W}}$ with component

$$\widetilde{W}_{ij} = \int_0^\tau K_{ij}(u) \frac{Y_{i+}(u)Y_{j+}(u)}{Y_{ij+}(u)} \left\{ \frac{\sum_\ell G_{i\ell}\, \mathrm{d}N_{i\ell}(u)}{Y_{i+}(u)} - \frac{\sum_\ell G_{j\ell}\, \mathrm{d}N_{j\ell}(u)}{Y_{j+}(u)} \right\}$$

where $G_{i\ell}$ are independent standard normal random variables, then $\boldsymbol{W}$ and $\widetilde{\boldsymbol{W}}$ will have the same asymptotic distribution. This can be done as follows: Generate B sets of i.i.d. N(0, 1) random variables $G^b_{i\ell}$ for $b = 1, \ldots, B$, $i = 1, \ldots, K$, and $\ell = 1, \ldots, n_i$. Compute $\widetilde{W}^b_{ij}$ for each b, i, j using the replacement discussed above. Note that these statistics $\widetilde{W}^b_{ij}$ must be standardized before using them to generate the critical value. While asymptotically, one could standardize each by $\hat{\sigma}_{ij}$, we have found in simulations that this doesn't work as well for moderate sample sizes. Alternatively, one can estimate the variability using the sample variance of the simulated samples,

$$\widetilde{\sigma}^2_{ij} = \frac{\sum_{b=1}^{B} (\widetilde{W}^b_{ij} - \overline{\widetilde{W}}_{ij})^2}{B - 1}$$

where $\overline{\widetilde{W}}_{ij}$ is the sample mean of $\widetilde{W}^b_{ij}$ over b. In simulations this has been shown to work much better in standardizing the simulated statistics as

$$\widetilde{Z}^b_{ij} = \frac{\widetilde{W}^b_{ij}}{\widetilde{\sigma}_{ij}}$$

Compute $\widetilde{Z}^b_{\max} = \max_{1 \leqslant i < j \leqslant K} |\widetilde{Z}^b_{ij}|$. Then the appropriate critical value $\widetilde{D}_\alpha$ is the $(1 - \alpha)100$th percentile of $\widetilde{Z}^b_{\max}$.

This method can also be extended to a step-down procedure by computing the sequence of critical values $\widetilde{D}_{(1),\alpha}, \ldots, \widetilde{D}_{(C),\alpha}$ corresponding to the ordered log-rank statistics $|Z|_{(1)} \leqslant \cdots \leqslant |Z|_{(C)}$ as follows. First, re-order the statistics based on the simulated martingales so that $\widetilde{Z}^b_{(k)}$ use the same ordering as the observed log-rank statistics for each b. Compute $\widetilde{Z}^b_{\max,(k)} = \max_{1 \leqslant i \leqslant k} |\widetilde{Z}_{(i)}|^b$. Then the appropriate critical value $\widetilde{D}_{(k),\alpha}$ is the $(1 - \alpha)100$th percentile of $\widetilde{Z}^b_{\max,(k)}$.

5. GENERAL CLOSED TEST PROCEDURE

To take the correlation into account in addition to the significances from the other pairwise tests, one can implement a closed test procedure (CTP) [9]. First, form a closed family of hypotheses by including in the family all intersections of hypotheses in the family. These intersections will take one of two forms. Some will be joint hypotheses such as

$$H_P : \lambda_i = \lambda \quad \text{for } i \in P$$

where $P \subseteq \{1, \ldots, K\}$. For example $H_{12} \cap H_{13} = H_{123} : \lambda_1 = \lambda_2 = \lambda_3$. Other intersection hypotheses will be disjoint hypotheses such as $H_{12} \cap H_{34}$. Then the closure procedure rejects any hypothesis H_P iff every set Q such that $P \subseteq Q \subseteq \{1, \ldots, K\}$ is also rejected by an α-level test. For example, if we have four groups, then we reject H_{12} if H_{1234}, H_{123}, H_{124}, $H_{12} \cap H_{34}$, and H_{12} are all rejected with their corresponding α level (unadjusted) tests. This controls the FWE at level α, accounts for correlation among the log-rank tests, and uses the results from other log-rank tests in adjusting for multiplicity. In fact, Holm's procedure can be shown to be a closed test procedure with each intersection hypothesis tested using a Bonferroni adjustment.

The issue with applying a closed test procedure in survival analysis is in determining what test to apply to each intersection hypothesis. For the joint intersection hypotheses H_P, the simplest test would be the multiple group weighted log-rank test. We obtain a subvector of weighted log-rank tests $\boldsymbol{X}$ for comparisons in the set P given by $\boldsymbol{X}_P = (X_{ij} \mid i, j \in P, i < j)$. The dimension of $\boldsymbol{X}_P$ is equal to $\binom{|P|}{2}$, where $|P|$ is the number of groups in the set P. For example, to test H_{123}, the subvector $\boldsymbol{X}_{123} = (X_{12}, X_{13}, X_{23})$ with dimension 3. The estimated covariance matrix of $\boldsymbol{X}_P$, $\hat{\boldsymbol{\Sigma}}_P$, is estimated by the submatrix of $\hat{\boldsymbol{\Sigma}}$ (given in (4)) corresponding to $X_{ij} \in \boldsymbol{X}_P$. Then we can construct a χ^2 test of H_P using

$$\chi^2_P = \boldsymbol{X}'_P \hat{\boldsymbol{\Sigma}}^-_P \boldsymbol{X}_P \tag{5}$$

where $\hat{\boldsymbol{\Sigma}}^-_P$ is the generalized inverse of the estimated covariance matrix of $\boldsymbol{X}_P$. This χ^2 statistic has degrees of freedom equal to $|P| - 1$. Note the generalized inverse is used because $\hat{\boldsymbol{\Sigma}}_P$ is not of full rank. Another way to compute the test statistic is to use a subvector of $\boldsymbol{X}_P$ so that the corresponding covariance matrix is of full rank. For example, to test H_{123}, rather than using the vector $\boldsymbol{X}_{123}$ instead use the first two components given by $\boldsymbol{X}^*_{123} = (X_{12}, X_{13})$ with dimension 2. The corresponding 2×2 covariance matrix of $\boldsymbol{X}^*_{123}$, denoted by $\hat{\boldsymbol{\Sigma}}^*_{123}$, is invertible and is obtained by deleting the rows and columns corresponding to the deleted component of $\boldsymbol{X}_{123}$. Then the test statistic of H_{123} can be computed as

$$\chi^2_P = (\boldsymbol{X}^*_{123})' (\hat{\boldsymbol{\Sigma}}^*_{123})^{-1} \boldsymbol{X}^*_{123}$$

which follows a χ^2_2 distribution.

Note that a slightly different alternative test statistic can be used to test a joint hypothesis H_P, by comparing the hazard functions in each group to the hazard function obtained by pooling all of the groups in the set P, as described in Andersen *et al.* [6]. However, these are expected to be very similar, and in subsequent simulations we use (5) because it requires minimal additional calculations beyond what is already done for the other procedures under investigation.

To test a disjoint intersection hypothesis such as $H_{12} \cap H_{34}$, we can sum the corresponding χ^2 tests for each disjoint component because of independence, resulting in another χ^2 test with degrees of freedom equal to the sum of the component test degrees of freedom. For example, both H_{12} and H_{34} can be tested with 1 d.f. weighted log-rank χ^2 tests, χ^2_{12} and χ^2_{34}, as in (3), so the disjoint intersection hypothesis can be tested with a 2 d.f. χ^2 test equal to the sum of the component χ^2 tests,

$$\chi^2 = \chi^2_{12} + \chi^2_{34}$$

To apply the closed test procedure, one needs to be able to efficiently enumerate all the possible joint and disjoint hypotheses. The following algorithm provides such an enumeration:

1. Set all pairwise comparison flags to 1 (reject pairwise hypothesis).
2. Define a vector of length K with elements $(i_1, \ldots, i_K)$, and set $i_1 = 1$.
3. Set up a series of successive do loops for $i_2, \ldots, i_K$ as follows:

$$\text{Do } i_2 = 1 \text{ to } \left(\max_{j<2} i_j + 1\right) \quad (=2);$$

$$\vdots$$

$$\text{Do } i_k = 1 \text{ to } \left(\max_{j<k} i_j + 1\right)$$

$$\vdots$$

Each vector $(i_1, \ldots, i_K)$ describes a joint or disjoint hypothesis as follows: All elements i_k with common values are being tested for equality. If there are multiple values which match more than one i_k, these represent disjoint groups which are being tested for equality. Values which only occur once are ignored. For example, with $K = 4$ the representation $(1,2,2,2)$ indicates null hypothesis $H_{234} : \lambda_2 = \lambda_3 = \lambda_4$, while the vector $(1,2,1,2)$ indicates the disjoint hypothesis $H_{13} \cap H_{24} : \lambda_1 = \lambda_3$ and $\lambda_2 = \lambda_4$.

4. Given a vector $(i_1, \ldots, i_K)$, one can set up the test statistic as follows:
   (a) Initialize $\chi^2_{\text{Tot}} = 0$ and $\text{df}_{\text{Tot}} = 0$.
   (b) For each possible value of i_k, denoted INDX, count the number of elements of the vector that match that value, denoted CNT[INDX].
   (c) If CNT[INDX] > 1 then compute the χ^2 statistic for all groups k such that $i_k = \text{INDX}$, denoted χ^2_{INDX}. Compute the degrees of freedom (df) for this test as $\text{df}_{\text{INDX}} = \text{CNT[INDX]} - 1$.
   (d) Update $\chi^2_{\text{Tot}} = \chi^2_{\text{Tot}} + \chi^2_{\text{INDX}}$ and $\text{df}_{\text{Tot}} = \text{df}_{\text{Tot}} + \text{df}_{\text{INDX}}$.
5. If $\chi^2_{\text{Tot}} < \chi^2_{\text{df Tot}, 1-\alpha}$, then accept the hypothesis corresponding to $(i_1, \ldots, i_K)$ and set the pairwise comparison flags of all corresponding subsets to 0 (accept pairwise hypothesis).

Note that it is possible to omit some of the tests because of the hierarchical ordering of the closed test procedure. A hypothesis test need not be performed if the pairwise comparison flags for all subset hypotheses are already 0 (i.e. all subset hypotheses have already been accepted by implication).

6. SIMULATION STUDY

A simulation study was performed to examine the operating characteristics of each of these methods. A significance level or FWE of 5 per cent was used throughout.

To study the FWE under the global null hypothesis (also called the EWE or Experimentwise Error rate), survival times were generated assuming an exponential survival distribution with parameter $\lambda = 1.5$ for each of $K = 4$ or $K = 6$ groups. An independent exponential distribution with parameter $\lambda_C = 0.7$ was used as a censoring distribution, resulting in approximately 30 per cent censoring. Additional scenarios were run with unequal censoring distributions, but these had similar results and are omitted for brevity. In each case, 10 000 data sets were simulated. The estimated FWE's are given in Table I for each method and for sample sizes of 50,

Table I. FWE rates of procedures based on unweighted log-rank tests: $K=4$ and $K=6$, for $\alpha=0.05$ and exponential survival distribution.

	$K=4$			$K=6$		
Method	50	150	250	50	150	250
Unadjusted	0.215	0.211	0.198	0.388	0.374	0.373
Bonferroni	0.049	0.042	0.039	0.048	0.040	0.041
Mart. sim.	0.058	0.052	0.047	0.061	0.052	0.053
MVN sim.	0.056	0.052	0.047	0.060	0.050	0.052
CTP	0.050	0.046	0.045	0.025	0.027	0.028

150, and 250. For the simulated martingale and the simulated MVN approaches, $B=3000$ simulated samples for each data set were used. For brevity, the SD procedures are omitted because their FWE's will be the same as that of the corresponding SS procedure.

The martingale simulation and MVN simulation methods appear to be anticonservative for small sample sizes of 50 per group. This is perhaps because the multiplicity adjustment is based on the distribution of the maximum log-rank test, and the rate of convergence may be slower for the maximum. However, the simulation based methods work well for moderate to large sample sizes of 150–250 per group. We have found also that the results are sensitive to the number of simulated samples for each data set B, and values less than 3000 did not perform well. However, this will not be a concern in practice, because one can use much larger values of B than is available within a simulation study. Finally, note that the CTP is conservative in estimating the FWE under the global null hypothesis for $K=6$. This is because the FWE in Table I is the probability of falsely rejecting a pairwise null hypothesis. In order for the closed test procedure to reject a pairwise null hypothesis, it must also reject all hypotheses implying that pairwise hypothesis including the global null hypothesis. Since each of these is tested at (unadjusted) level α, this results in a FWE for the pairwise hypotheses less than α.

We further examined the control of the FWE under partial null hypothesis configurations where some of the hypotheses are true and some are false. A sample size of 250 per group was used, under a variety of hypothesis configurations denoted by the vector λ of exponential hazards for each group. The parameters for the exponential censoring distributions in each group were chosen so that each group had approximately 30 per cent censoring. Each scenario used $K=4$ or $K=6$ groups, 5000 simulated data sets, and $B=3000$ simulated samples for each data set. The martingale simulation and the MVN simulation methods perform similarly, so only the MVN simulation method is reported. These are given in Table II. Note that as expected, the single-step procedures are overly conservative while the step-down and CTP procedures are closer to the targeted level of 0.05. Also the CTP is generally less conservative for the partial null configurations than for the overall null hypothesis, because of the hierarchical ordering of the hypotheses being tested.

Next, the power of each of these methods to detect false hypotheses was examined in terms of three criteria: the any-pairs power (ANYPOW), which is the probability that any false hypothesis is rejected, the all-pairs power (ALLPOW), which is the probability that all false hypotheses are rejected, and the average per-pairs power (AVGPOW), which is the average of the individual powers for each false hypothesis. The power was investigated

Table II. FWE rates under various partial null configurations for procedures based on unweighted log-rank tests: $K = 4$ and $K = 6$, for $\alpha = 0.05$ and exponential survival distribution.

			Single-step		Step-down		
K	$(\lambda_1, \ldots, \lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	(2.25, 1.50, 1.50, 1.50)	0.120	0.023	0.029	0.040	0.047	0.050
	(2.25, 2.25, 1.50, 1.50)	0.097	0.018	0.023	0.039	0.041	0.046
	(2.25, 1.75, 1.75, 1.25)	0.052	0.012	0.015	0.025	0.027	0.049
6	(2.25, 1.50, 1.50, 1.50, 1.50, 1.50)	0.290	0.027	0.034	0.035	0.045	0.041
	(2.25, 2.25, 1.50, 1.50, 1.50, 1.50)	0.250	0.022	0.028	0.036	0.041	0.037
	(2.25, 2.25, 2.25, 1.50, 1.50, 1.50)	0.227	0.018	0.024	0.032	0.040	0.030
	(2.50, 2.50, 2.00, 2.00, 1.50, 1.50)	0.150	0.010	0.013	0.020	0.025	0.035

Table III. Average power for procedures based on unweighted log-rank tests: $K = 4$ and $K = 6$, for $\alpha = 0.05$ and exponential survival distribution.

			Single-step		Step-down		
K	$(\lambda_1, \ldots, \lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	(2.25, 1.50, 1.50, 1.50)	0.967	0.874	0.889	0.896	0.906	0.924
	(2.25, 2.25, 1.50, 1.50)	0.969	0.876	0.889	0.911	0.919	0.957
	(2.25, 1.75, 1.75, 1.25)	0.807	0.629	0.648	0.696	0.710	0.762
	(2.50, 2.00, 1.50, 1.00)	0.881	0.778	0.788	0.863	0.865	0.878
6	(2.25, 1.50, 1.50, 1.50, 1.50, 1.50)	0.966	0.802	0.823	0.820	0.838	0.819
	(2.25, 2.25, 1.50, 1.50, 1.50, 1.50)	0.966	0.803	0.826	0.841	0.857	0.872
	(2.25, 2.25, 2.25, 1.50, 1.50, 1.50)	0.965	0.798	0.822	0.845	0.859	0.873
	(2.50, 2.50, 2.00, 2.00, 1.50, 1.50)	0.771	0.522	0.542	0.566	0.584	0.616
	(3.50, 3.00, 2.50, 2.00, 1.50, 1.00)	0.857	0.729	0.739	0.794	0.799	0.832

under the same scenarios as were used to study the FWE under the partial null configurations, plus some additional ones. The results for AVGPOW and ALLPOW are given in Tables III and IV. The results for ANYPOW are close to 1 for all procedures for the configurations studied and so are omitted.

In general, the simulation based SS and SD methods offer only marginal improvements over the corresponding Bonferroni or Holm procedures, on the order of 1–3 per cent improvement in average per-pairs power or all-pairs power. The step-down procedures such as the Holm procedure can improve average power by 2–8 per cent over the single-step counterparts, and result in even greater improvement (4–28 per cent) in all-pairs power. The CTP is consistently the best performing procedure for $K = 4$; it also performs best for most configurations when $K = 6$. However, it gets prohibitive to implement as K gets larger and may become less powerful relative to the other SD procedures. Finally, we note that these results are similar when there are unequal censoring rates across group.

We also conducted additional simulations where the survival times came from a log-normal distribution with mean parameter μ_k for group k and common standard deviation parameter $\sigma = 1$, which results in a hazard function which hits a peak early and then tapers at later times. In these simulations, a weight function as in (1) with $(\rho = 1, \gamma = 0)$ was used

Table IV. All-pairs power for procedures based on unweighted log-rank tests: $K=4$ and $K=6$, for $\alpha=0.05$ and exponential survival distribution.

			Single-step		Step-down		
K	$(\lambda_1,\ldots,\lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	(2.25, 1.50, 1.50, 1.50)	0.918	0.736	0.765	0.788	0.803	0.835
	(2.25, 2.25, 1.50, 1.50)	0.897	0.672	0.699	0.771	0.784	0.892
	(2.25, 1.75, 1.75, 1.25)	0.367	0.102	0.118	0.225	0.240	0.355
	(2.50, 2.00, 1.50, 1.00)	0.351	0.059	0.075	0.337	0.340	0.351
6	(2.25, 1.50, 1.50, 1.50, 1.50, 1.50)	0.882	0.515	0.557	0.565	0.596	0.563
	(2.25, 2.25, 1.50, 1.50, 1.50, 1.50)	0.819	0.367	0.406	0.470	0.500	0.538
	(2.25, 2.25, 2.25, 1.50, 1.50, 1.50)	0.800	0.323	0.365	0.444	0.473	0.492
	(2.50, 2.50, 2.00, 2.00, 1.50, 1.50)	0.066	0.000	0.005	0.005	0.005	0.012
	(3.50, 3.00, 2.50, 2.00, 1.50, 1.00)	0.005	0.000	0.000	0.000	0.001	0.005

Table V. FWE rates of procedures based on weighted log-rank tests: $K=4$ and $K=6$, for $\alpha=0.05$ and log-normal survival distribution.

	$K=4$			$K=6$		
Method	50	150	250	50	150	250
Unadjusted	0.210	0.206	0.205	0.357	0.363	0.365
Bonferroni	0.043	0.045	0.043	0.042	0.038	0.039
Mart. sim.	0.051	0.053	0.052	0.053	0.049	0.052
MVN sim.	0.051	0.052	0.051	0.051	0.048	0.052
CTP	0.049	0.052	0.047	0.030	0.030	0.032

Table VI. FWE rates under various partial null configurations for procedures based on weighted log-rank tests: $K=4$ and $K=6$, for $\alpha=0.05$ and log-normal survival distribution.

			Single-step		Step-down		
K	$(\lambda_1,\ldots,\lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	$(0.35,0,0,0)$	0.115	0.022	0.026	0.036	0.044	0.048
	$(0.35,0.35,0,0)$	0.095	0.015	0.017	0.032	0.036	0.043
	$(0.3,0,0,-0.3)$	0.047	0.009	0.011	0.024	0.027	0.046
6	$(0.35,0,0,0,0,0)$	0.288	0.028	0.036	0.036	0.047	0.040
	$(0.35,0.35,0,0,0,0)$	0.234	0.019	0.026	0.031	0.038	0.033
	$(0.35,0.35,0.35,0,0,0)$	0.238	0.020	0.027	0.036	0.041	0.034
	$(0.3,0.3,0,0,-0.3,-0.3)$	0.144	0.010	0.014	0.025	0.029	0.039

to reflect interest in early events. This weight is similar to the Peto and Peto test [10]. The censoring distribution was also selected from the log-normal $(\mu_k^C,1)$ distribution, with μ_k^C selected for each group to result in approximately 30 per cent censoring. Table V gives the estimates of the FWE under the global null configuration, where $\mu_k=0$ for all k, while Tables VI–VIII give estimates of the FWE, average power, and any-pairs power under

Table VII. Average power for procedures based on weighted log-rank tests: $K = 4$ and $K = 6$, for $\alpha = 0.05$ and log-normal survival distribution.

			Single-step		Step-down		
K	$(\lambda_1, \ldots, \lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	$(0.35, 0, 0, 0)$	0.952	0.835	0.850	0.857	0.869	0.894
	$(0.35, 0.35, 0, 0)$	0.950	0.831	0.848	0.873	0.882	0.929
	$(0.3, 0, 0, -0.3)$	0.896	0.738	0.758	0.815	0.825	0.872
	$(0.45, 0.15, -0.15, -0.45)$	0.934	0.834	0.845	0.926	0.926	0.933
6	$(0.35, 0, 0, 0, 0, 0)$	0.950	0.745	0.773	0.766	0.790	0.774
	$(0.35, 0.35, 0, 0, 0, 0)$	0.953	0.744	0.771	0.786	0.807	0.826
	$(0.35, 0.35, 0.35, 0, 0, 0)$	0.948	0.741	0.769	0.791	0.810	0.829
	$(0.3, 0.3, 0, 0, -0.3, -0.3)$	0.912	0.706	0.730	0.780	0.792	0.823
	$(0.75, 0.45, 0.15, -0.15, -0.45, -0.75)$	0.956	0.852	0.862	0.941	0.942	0.954

Table VIII. All-pairs power for procedures based on weighted log-rank tests: $K = 4$ and $K = 6$, for $\alpha = 0.05$ and log-normal survival distribution.

			Single-step		Step-down		
K	$(\lambda_1, \ldots, \lambda_K)$	No. adj.	Bonf	MVN	Holm	MVN	CTP
4	$(0.35, 0, 0, 0)$	0.882	0.664	0.692	0.719	0.738	0.774
	$(0.35, 0.35, 0, 0)$	0.851	0.586	0.619	0.696	0.713	0.843
	$(0.3, 0, 0, -0.3)$	0.606	0.249	0.279	0.445	0.460	0.597
	$(0.45, 0.15, -0.15, -0.45)$	0.624	0.208	0.242	0.612	0.613	0.624
6	$(0.35, 0, 0, 0, 0, 0)$	0.828	0.413	0.459	0.464	0.501	0.465
	$(0.35, 0.35, 0, 0, 0, 0)$	0.756	0.272	0.314	0.370	0.402	0.430
	$(0.35, 0.35, 0.35, 0, 0, 0)$	0.725	0.229	0.277	0.355	0.379	0.397
	$(0.3, 0.3, 0, 0, -0.3, -0.3)$	0.412	0.021	0.035	0.121	0.131	0.196
	$(0.75, 0.45, 0.15, -0.15, -0.45, -0.75)$	0.455	0.008	0.014	0.422	0.425	0.455

alternative configurations of μ_k. In general, the comparative conclusions about the various procedures are consistent with what was found for the exponential distribution and unweighted log-rank test.

7. EXAMPLE

Oh *et al.* [11] present the findings of a research study on the effect of ethnicity on outcomes after allogeneic bone marrow transplantation, in which one objective is to compare survival among patients from five racial groups (North American Caucasian, African-American, Irish, Scandinavian, and Japanese) who received matched, related donor allotransplants. The rationale for this comparison is that island racial populations will tend to have less heterogeneity in their genotypes and will tend to match the recipients better. In turn this will lead to less acute graft versus host disease and a better survival prognosis. African-Americans are considered to have greater genotypic heterogeneity than North American Caucasians and are therefore

Table IX. Descriptive statistics.

Race	Alive	Dead	Total	Median follow-up (Survivors) (months)
North American Caucasians	530	299	829	48
North American African-Americans	42	29	71	33
Scandinavian Caucasians	135	57	192	60
Irish Caucasians	66	29	95	42
Japanese	372	190	562	65

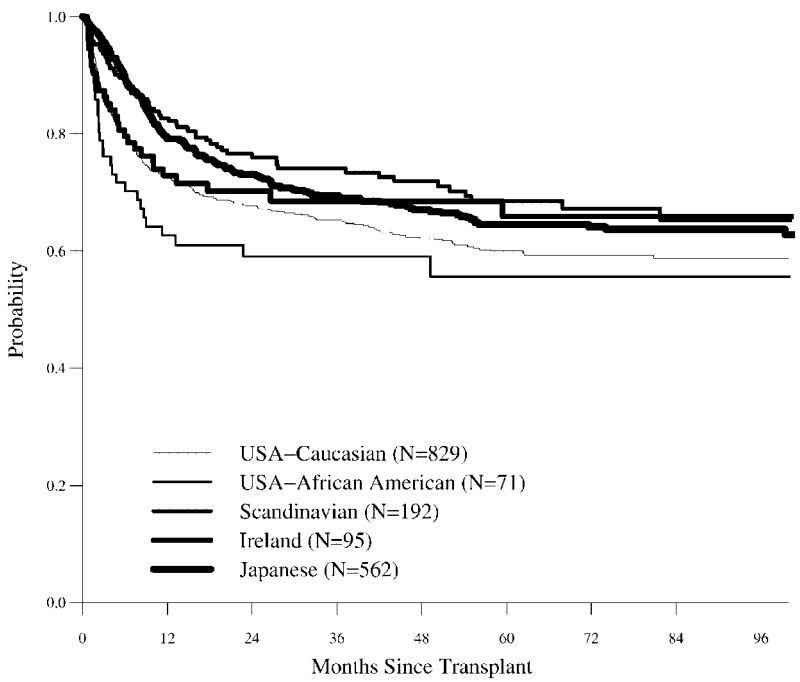

Figure 1. Kaplan–Meier estimates of overall survival for each racial group.

expected to have a poorer prognosis. Descriptive statistics are given in Table IX, and survival curves are presented in Figure 1.

The overall 4 d.f. log-rank test is significant ($\chi^2 = 12.807$, $p = 0.012$). Each of the pairwise comparisons, their log-rank tests, and the critical values used in each multiple comparisons procedure, are given in Table X. These statistics are ordered, and so critical values are only given until each multiple comparisons procedure stops. Note that the CTP does not result in critical values for each comparison, but only an accept/reject decision.

The Bonferroni and Holm procedures do not find any significant differences among the groups, the simulation based SS and SD approaches (both MVN and martingale) find two of the comparisons significant, while the CTP finds four comparisons significant. An analysis

Table X. Ordered log-rank statistics and critical values generated using proposed methods.

			Single-step			Step-down			
Grp 1	Grp 2	$\|Z\|$	Bonf.	MVN	Mart	Holm	MVN	Mart	CTP
AFA	Scan.	2.752	2.807	**2.699**	**2.673**	2.807	**2.699**	**2.673**	**Sig.**
AFA	Jap.	2.712		**2.699**	**2.673**		**2.674**	**2.652**	**Sig.**
US-Wh	Jap.	2.472		2.699	2.673		2.635	2.639	**Sig.**
US-Wh	Scan.	2.360							**Sig.**
AFA	Irish	1.472							
US-Wh	AFA	1.374							
Scan.	Irish	0.853							
Scan.	Jap.	0.803							
US-Wh	Irish	0.774							
Irish	Jap.	0.464							

with no adjustment for multiple comparisons would also find four comparisons significant. The relative performance of the procedures on this real data set is similar to what was seen in the simulation study in Section 7.

ACKNOWLEDGEMENTS

Dr Logan's and Dr Wang's research was supported by a grant from the American Cancer Society and the Medical College of Wisconsin Cancer Center. Dr Zhang's research was supported by National Cancer Institute Grant 2 R01 CA54706-10. We thank the International Bone Marrow Transplant Registry for providing us the data for the example.

REFERENCES

1. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
2. Tamhane AC. Multiple Comparisons. In *Handbook of Statistics*, vol. 13. Ghosh S, Rao CR (eds). Elsevier Science: New York, 1996; 587–630.
3. Koziol JA, Reid N. On multiple comparisons among k samples subject to unequal patterns of censorship. *Communications in Statistics*: *Theory and Methodology* 1977; **12**:1149–1164.
4. Chen Y-I. Multiple comparisons in carcinogenesis study with right-censored survival data. *Statistics in Medicine* 2000; **19**:353–367.
5. Fleming TR, Harrington DP. A class of hypothesis tests for one and two samples of censored survival data. *Communications in Statistics* 1981; **10**:763–794.
6. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer: New York, 1993.
7. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**: 65–70.
8. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**:73–81.
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. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society* 1972; **135**:185–206.
11. Oh H, Zhang M-J, Akiyama H, Asai T, Barrett AJ, Loberiza FR, Miyawaki S, Okamoto S, Ringden O, Horowitz MM. Comparison of graft-versus-host disease (GVHD) and survival in different ethnic populations: collaborative study by the Japan Adult Leukemia Study Group (JALSG) and the International Bone Marrow Transplant Registry (IBMTR). Abstract 1621 presented at the *American Society of Hematologists Meeting*, San Diego, CA, 2003.

