A Nonparametric Test to Compare Survival Distributions with Covariate Adjustment

Glenn Heller and E.S.Venkatraman

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

1275 York Avenue, New York, New York 10021, U.S.A.

heller@biosta.mskcc.org venkat@biosta.mskcc.org

Summary. The analysis of covariance is a technique used to improve the power of a k-sample test by adjusting for concomitant variables. If the endpoint is survival time, and some observations are right censored, the score statistic from the Cox proportional hazards model is the most commonly used method to test the equality of conditional hazard functions. In many situations, however, the proportional hazards model assumptions are not satisfied. Specifically, the relative risk function is not time invariant or represented as a loglinear function of the covariates. We propose an asymptotically valid k-sample test statistic to compare conditional hazard functions, which does not require the assumption of proportional hazards, a parametric specification of the relative risk function, or randomization of group assignment. Simulation results indicate that the performance of this statistic is satisfactory. The methodology is demonstrated on a data set in prostate cancer.

KEYWORDS: Covariate adjusted k-sample test; Kernel smoothing; Nonparametric test statistic; Survival data; U-statistic

1. Introduction

The analysis of covariance is a technique used to improve the power of a k-sample test by adjusting for concomitant variables. In the classic normal linear model, this technique is used for increasing the precision of the k-sample test statistic derived from a randomized study or to adjust for sources of bias in observational studies. The power of the adjusted test statistic is a function of the strength of the association between the covariates and the response. However, the increase in power comes at the cost of model specification. The assumption of a specific linear regression function and a normal error distribution are imposed on the model structure. If either of these assumptions are incorrect, the resulting inference drawn from the adjusted test is questionable. As with any model based methodology, its appropriate use requires that the assumptions used to develop the model are satisfied.

If the endpoint is survival time, and some observations are right censored, the Cox proportional hazards model (Cox, 1972) is the most commonly used method to adjust for covariates. The conventional proportional hazards specification of the conditional hazard function is

$$\lambda(t|z, \boldsymbol{x}) = \lambda_0(t)\exp\{\beta z + \boldsymbol{\gamma}'\boldsymbol{x}\}$$

where Z represents the group classification variable of interest, $\boldsymbol{X}$ is the vector of concomitant variables, $h_0(t)$ is the baseline hazard function, and $(\beta, \boldsymbol{\gamma})$ are the regression parameters associated with $(Z, \boldsymbol{X})$. The parameter of interest β, represents the group effect on survival time.

For the proportional hazards model, the primary assumptions are that the relative risk function $\lambda(t|z, \boldsymbol{x})/\lambda_0(t)$ is 1) time invariant and 2) represented by a loglinear

function of the covariates. Since its introduction, there has been considerable interest in the development of diagnostic techniques, including test statistics and graphical summaries, to determine the adequacy of these assumptions to the data at hand. For the analysis of covariance survival problem, incorrect model specification can lead to erroneous inference on the group effect. A good overview of diagnostic techniques for the proportional hazards model is found in Therneau and Grambsch (2000).

In order to accommodate data that does not satisfy the proportional hazards model assumptions, models with increasing levels of generality have been developed. The earliest extension is the stratified proportional hazards model, which allows for non-proportional hazards between a finite set of strata (Kalbfleisch and Prentice, 1980). Sun and Yang (2000) extend this model, enabling the relative risk function to vary between strata. Application of these approaches may require ad-hoc partitioning of a continuous covariate and may become infeasible with more than one continuous covariate.

Sasieni (1992) proposed a continuously stratified Cox model

$$\lambda(t|z, \boldsymbol{x}) = \lambda^{(1)}(t|\boldsymbol{x})\exp(z\beta).$$

Here the baseline hazard changes for each value of the vector $\boldsymbol{x}$, while maintaining the same proportional hazards relationship for each point stratum. Although no inferential procedure has been developed for the adjusted group effect parameter β in this model, an adjusted test has been developed in the special case of the partly linear Cox model

$$\lambda(t|z, \boldsymbol{x}) = \lambda_0(t)\exp\{z\beta + g(\boldsymbol{x})\}$$

where g is an unknown smooth function. Inference on β in the presence of an infinite

dimensional nuisance parameter g was developed through a reparameterization of g orthogonal to β (Heller, 2001). Although the partly linear model relaxes the loglinear specification of the relative risk function, it still requires the relative risk function to remain time invariant.

Lin and Wei (1989), Kong and Slud (1997), and DiRienzo and Lagakos (2001a,b) have proposed covariate adjusted k-sample tests, using robust variance estimates for the proportional hazards score statistic to account for possible misspecification of the conditional hazard function. The formulation of their tests is based on a working proportional hazards model

$$\lambda(t|z, \boldsymbol{x}) = \lambda_0(t)\exp\{z\beta + h(\boldsymbol{\gamma}'\boldsymbol{x})\}$$

where h is specified. Lin and Wei (1989) choose h as the identity function, whereas Kong and Slud (1997) and DiRienzo and Lagakos (2001a,b) allow a wider range of working model specifications. If Z is independent of $\boldsymbol{X}$, as is the case in a randomized clinical trial design, Lin and Wei (1989) demonstrate that the covariate adjusted score statistic, using a robust variance estimate, is asymptotically valid. Kong and Slud (1997) propose an alternative score test that is asymptotically valid under the more general condition that at any point in the study, the distribution of $\boldsymbol{X}$ among subjects still at risk is independent of Z. DiRienzo and Lagakos (2001b) introduce a bias corrected score test that may be appropriate when $\boldsymbol{X}$ and Z are dependent, however they have yet to determine the conditions under which this statistic is asymptotically valid. Du et al. (2003) developed a totally nonparametric approach for an analysis of covariance using a single continuous covariate. The approach is an extension of a method for testing main effects and interaction effects within a factorial design

(Akritas and Brunner, 1997).

In this paper, we propose a test statistic to compare k groups, which adjusts for concomitant covariates, does not require the assumption of proportional hazards or a parametric specification of the relative risk function, and is asymptotically valid under conditions found in observational studies. The test statistic is developed for survival time data where right censoring is present. In Section 2, the test statistic and its asymptotic distribution is developed. Section 3 presents simulation results demonstrating the small sample operating characteristics of the statistic and Section 4 provides an application of the methodology to a prostate cancer dataset. The paper concludes with some remarks in Section 5.

2. Test Statistic

We develop a nonparametric test for the equality of the conditional hazard functions between groups

$$\lambda_1(t|\boldsymbol{x}) = \lambda_2(t|\boldsymbol{x}) = \ldots = \lambda_k(t|\boldsymbol{x}) \qquad \text{for all } (t, \boldsymbol{x}).$$

At the outset, it is assumed that the group variable of interest z is binary, taking on the values 0 and 1. The generalization to the adjusted k-sample test will be presented after the development of the two-sample test. It is assumed that conditional on Z and the covariate vector $\boldsymbol{X}$, the latent failure and censoring variables are independent. Although the most prominent application of the proposed methodology is an adjusted randomized treatment comparison, the methodology is applicable for nonrandomized comparisons as well. Examples of adjusted tests in survival analyses include: the uncontrolled comparison of an experimental treatment to a prior

conventional treatment; accounting for concomitant information in an epidemiologic case-control survival study; and examination of the importance of newly determined factors on survival time.

For each subject, define $N_i(t) = I(T_i \leq t, \delta_i = 1)$ as the counting process and $Y_i(t) = I(T_i \geq t)$ as the at risk process, where the observed survival time T_i is the minimum of the latent failure time (T_i^0) and censoring time (C_i), and δ_i is the censoring indicator ($\delta_i = 1$ signifying the failure time is smaller). It is assumed that the individual copies of the random vector $(T^0, C, Z, \boldsymbol{X})$ are independent and identically distributed.

Under the assumption that the failure time and censoring time are independent conditional on the covariates $(\boldsymbol{X}, Z)$, we can without loss of generality define the relationship between the counting process, the at risk process, and the covariates by

$$\mathrm{E}\{dN(t)|Y(t), \boldsymbol{X}, Z\} = Y(t)Z\lambda_1(t|\boldsymbol{X})dt + Y(t)(1-Z)\lambda_0(t|\boldsymbol{X})dt \tag{1}$$

where $\lambda_j(t|\boldsymbol{x})$ represents the conditional hazard for a subject with group variable $Z = j$. Building upon equation (1), it follows that

$$\mathrm{E}\{ZdN(t)|Y(t), \boldsymbol{X}, Z\} = Y(t)Z\lambda_1(t|\boldsymbol{X})dt. \tag{2}$$

To develop the nonparametric test statistic for testing $\lambda_1(t|\boldsymbol{x}) = \lambda_0(t|\boldsymbol{x})$ for all $(t, \boldsymbol{x})$, the expectation with respect to Z is computed on both sides of (2), and this relationship is considered under the hypothesis of equal conditional hazards. Performing these operations yields

$$\mathrm{E}\{ZdN(t)|Y(t), \boldsymbol{X}\} = Y(t)\lambda(t|\boldsymbol{X})\mathrm{E}\{Z|Y(t), \boldsymbol{X}\}dt \tag{3}$$

where the subscript from the conditional hazard has been omitted to denote a single population under the null hypothesis. Equation (3) is a restatement of the null hypothesis that conditional on the subject remaining at risk and the covariate value, the failure process is independent of group assignment. Thus a test statistic, which has asymptotic mean zero under the null hypothesis $\lambda_1(t|\boldsymbol{x}) = \lambda_0(t|\boldsymbol{x})$, can be created using empirical estimates of the left and right hand sides of equation (3).

Estimation of the conditional expectation $\mathrm{E}\{Z|Y(t) = 1, \boldsymbol{X} = \boldsymbol{x}_i\}$ is obtained through kernel smoothing.

$$\hat{E}\{Z|Y(t) = 1, \boldsymbol{X} = \boldsymbol{x}_i\} = \frac{\sum_j Y_j(t) z_j K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\sum_j Y_j(t) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}$$

If the dimension of the covariate $\boldsymbol{X}$ is greater than one, the multivariate kernel used for this estimate is the product kernel function, composed of p one dimensional symmetric kernel functions, $K(\boldsymbol{u}) = \prod_{l=1}^{p} k(u_l)$. The rescaled version of K is denoted by $K_{\boldsymbol{b}}(u) = \prod_{l=1}^{p} b_l^{-1} k(b_l^{-1} u_l)$, where $\boldsymbol{b} = (b_1, b_2, \ldots, b_p)^T$ and the bandwidth b_l controls the degree of smoothing over the l^{th} covariate. It now follows from equation (3), that a nonparametric test statistic that incorporates right censored survival times and adjusts for the concomitant variates $\boldsymbol{X}$ is

$$S_n = \sum_i \int z_i dN_i(t) - \sum_i \int \frac{\sum_j Y_j(t) z_j K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\sum_j Y_j(t) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)} dN_i(t) \tag{4}$$

Note that when $K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i) = 1$ for all i, j, i.e. the covariate is ignored, the test statistic reduces to the logrank statistic. As a result, we call the statistic in (4) the conditional logrank statistic. An algebraically equivalent representation of this statistic is

$$\sum_i \int w(\boldsymbol{x}_i, t) \left\{ \frac{dN_{1i}(t)}{a_{11}(\boldsymbol{x}_i, t)} - \frac{dN_{0i}(t)}{a_{10}(\boldsymbol{x}_i, t)} \right\}, \tag{5}$$

where $dN_{zi} = I_{[Z_i=z]}dN_i$ represents the group specific $(z = 0, 1)$ counting process,

$$a_{11}(\boldsymbol{x}, t) = \sum_j Y_j(t) z_j K_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{x}_j); \qquad a_{10}(\boldsymbol{x}, t) = \sum_j Y_j(t)(1 - z_j) K_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{x}_j) \tag{6}$$

and

$$w(\boldsymbol{x}, t) = \{a_{11}(\boldsymbol{x}, t) + a_{10}(\boldsymbol{x}, t)\}^{-1} a_{11}(\boldsymbol{x}, t) a_{10}(\boldsymbol{x}, t)$$

is a predictable weight function. Thus, the statistic can be interpreted as a function of the difference in the estimated conditional hazards between groups. The following theorem provides the asymptotic mean of $n^{-1}S_n$ and the limiting distribution of $n^{-1/2}S_n$ under the null hypothesis $\lambda_1(t|\boldsymbol{x}) = \lambda_0(t|\boldsymbol{x})$ for all $(t, \boldsymbol{x})$. Proof of this theorem is provided in the appendix.

Theorem 1: Suppose k is a bounded kernel function symmetric about zero, and the vector bandwidth $\boldsymbol{b}$ is chosen such that $\lim_{n\to\infty} \max_{1\le l\le p} n b_l^4 = 0$. In addition, the dimension of the covariate vector $p \le 3$. Then using the conditions stated in the appendix,

(i) $n^{-1}S_n$ converges in probability to

$$n^{-1} \sum_i \int_{s<\tau} E \left[\begin{array}{r} Y_i(s) E(Z_i|Y_i(s) = 1, \boldsymbol{X}_i)\{1 - E(Z_i|Y_i(s) = 1, \boldsymbol{X}_i)\} \\ \times\{d\Lambda_1(s|\boldsymbol{X}_i) - d\Lambda_0(s|\boldsymbol{X}_i)\} \end{array} \right]$$

(ii) $n^{-1/2}S_n$ converges in distribution under the null hypothesis to a normal random variable with mean zero and asymptotic variance V. The asymptotic variance is consistently estimated by

$$\sum_i \hat{v}_{ii}^2 + \sum_{i\ne j} \hat{v}_{ij} \{2\hat{v}_{ii} + 2\hat{v}_{jj} + \hat{v}_{ij} + \hat{v}_{ji}\} + \sum_{i\ne j\ne l} (\hat{v}_{ij}\hat{v}_{il} + \hat{v}_{ij}\hat{v}_{li} + \hat{v}_{ij}\hat{v}_{jl} + \hat{v}_{ij}\hat{v}_{lj}) \tag{7}$$

where

$$\hat{v}_{ij} = \delta_i \left[z_i - \frac{a_1(\boldsymbol{x}_i, t_i)}{a_0(\boldsymbol{x}_i, t_i)} - \frac{I(t_j \ge t_i) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{a_0(\boldsymbol{x}_i, t_i)} \left\{ z_j - \frac{a_1(\boldsymbol{x}_i, t_i)}{a_0(\boldsymbol{x}_i, t_i)} \right\} \right] \tag{8}$$

and $a_j(\boldsymbol{x}, t) = a_{j1}(\boldsymbol{x}, t) + a_{j0}(\boldsymbol{x}, t) \quad j = 0, 1.$

It follows from part (i) of the theorem that the equality of the conditional hazard functions, is a sufficient but not a necessary condition for the test statistic S_n to have zero mean. When the conditional hazard functions cross, the asymptotic mean may be zero depending on the conditional survival and censoring distributions in the two groups. The test is consistent when there is a uniform non-negative difference in the conditional hazard functions, with strict positivity for a subset of the time axis that is observed with non-zero probability. Similar to the unconditional case, the ordered survival alternative $S_1(t|\boldsymbol{x}) > S_0(t|\boldsymbol{x})$ is not sufficient to produce a consistent test.

The proposed conditional logrank test statistic requires the choice of a kernel and bandwidth. Any value of the bandwidth chosen such that $\max_{1 \le l \le p} nb_l^4 \to 0$ will provide an asymptotically valid test for the null hypothesis of no difference between the groups. However for sample sizes that one encounters in practice there are two opposing forces. If the bandwidth chosen is too small then the asymptotic variance formula may provide an estimate that is biased. In addition the small effective sample size will result in an approximation to the reference distribution that is not accurate. A large bandwidth on the other hand will dampen the ability of the test to identify the effect of the concomitant variable, resulting in a less powerful test especially when the covariate is attributable to a large part of the variation in the failure time. In simulations presented in the next section, we found the choice of a truncated Gaussian kernel with bandwidth $b = n^{-0.26}$, with the covariates standardized to have unit variance, provided good size and power properties for our test statistic. An alternative approach, proposed in the uncensored data case by Akritas et al. (2000),

is to permute the group indicator a large number of times, and select the constant in the bandwidth $b = (\text{constant})n^{-0.26}$ that on average provides the most accurate test size under the null hypothesis.

To extend the test statistic to the k-sample case, we define the random vector $\boldsymbol{Z}$ of length $k-1$, where the l^{th} component is the indicator function $Z_l = I_{[Z=l]}$ $(l = 1, \ldots, k-1)$. The vector valued test statistic $\boldsymbol{S_n}$ is derived directly from equation (4), replacing the scalar Z with the vector $\boldsymbol{Z}$. The statistic $\boldsymbol{S_n}$ is a multivariate U-statistic and has an asymptotic multivariate normal distribution. Its variance-covariance matrix is consistently estimated using a generalization of equation (7),

$$\begin{aligned}\boldsymbol{V_n} &= \sum_i \hat{\boldsymbol{v}}_{ii}\hat{\boldsymbol{v}}'_{ii} + \sum_{i\neq j}\left(\hat{\boldsymbol{v}}_{ii}\hat{\boldsymbol{v}}'_{ij} + \hat{\boldsymbol{v}}_{jj}\hat{\boldsymbol{v}}'_{ij} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{ii} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{jj} + \hat{\boldsymbol{v}}_{ij}\hat{v}'_{ij} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{ji}\right) \\ &\quad + \sum_{i\neq j\neq l}\left(\hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{il} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{li} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{jl} + \hat{\boldsymbol{v}}_{ij}\hat{\boldsymbol{v}}'_{lj}\right),\end{aligned}$$

where $\boldsymbol{v}$ is obtained by substituting the vector $\boldsymbol{z}$ in equations (8) and (6). The resulting quadratic form $\boldsymbol{S'_n V_n^{-1} S_n}$ has an asymptotic chi-square distribution with $k-1$ degrees of freedom and can be used to test the equality of the k groups.

3. Simulation Study

3.1 Simulation Structure

A series of monte carlo simulations were conducted to compare the operating characteristics of the conditional logrank statistic against those of the univariate (unconditional) logrank statistic, the score test from the Cox model, and the robust score tests proposed by Lin and Wei (1989) and Kong and Slud (1997). In contrast to the Cox model, the robust score tests do not require correct covariate specification, but

do postulate a form of independence between the group indicator Z and the covariates $\boldsymbol{X}$. Under independence, the DiRienzo and Lagakos (2001a) robust score test is the same as the Kong and Slud (1997) robust score test.

The relationship between the log failure time T^0, the concomitant variables $\boldsymbol{X}$ and the group indicator Z is given by

$$\log(T^0) \quad = \quad \beta Z + g(\boldsymbol{X}) + \epsilon$$

The working model used for the Cox score test and the robust score tests of the hypothesis $\beta = 0$ was $\lambda(t|z, \boldsymbol{x}) = \lambda_0(t) \exp(\beta z + \boldsymbol{\gamma}'\boldsymbol{x})$. The stochastic error ϵ, was generated from either a standard normal or a standard extreme value distribution. The extreme value error distribution provides a proportional hazards model and the normal error distribution results in a non-proportional hazards model. The parameter β determines the difference in the survival distributions for the two groups (both conditionally and unconditionally). The level of association between the covariate $\boldsymbol{X}$ and the failure time is given by R^2, the proportion of variance of $\log(T^0)$ explained by $\boldsymbol{X}$, *i.e.* $R^2 = \text{var}\{g(\boldsymbol{X})\}/[\text{var}\{g(\boldsymbol{X})\} + \text{var}\{\epsilon\}]$. The function g was chosen to give R^2 values of 0, 0.25, 0.50, and 0.75 in the simulations. A uniform $(0, c)$ censoring random variable was used to generate censoring times. The upper limit c of the censoring distribution was chosen to produce censoring proportions 0, 0.25, 0.50, or 0.75 when $\beta = 0$. The sample size used for all simulations was 50 per group. The size and power estimates were based on 10,000 replications. In the following we report the results of nominal 5% tests for a two sample comparison with a single covariate in Sections 3.2 and 3.3, the two covariate case in Section 3.4 and the three sample case in Section 3.5. A tabulation of all the simulation results may be found on the

Royal Statistical Society website *http://www.rss.org.uk.*

3.2 Group Variable Independent Of Covariate

In this section a single variable X was generated from a uniform distribution on the interval (0,1) and the group indicator Z was generated independently of X. Three specifications for the concomitant covariate function g were used: $g_1(x) = \psi x$, $g_2(x) = \psi \log(70x + 1)$, and $g_3(x) = \psi(x - .5)^2$. The constant ψ determined R^2. Note that g_1 and g_2 are monotonic in x whereas g_3 is non-monotonic. The choice of the extreme value error distribution and g_1, results in a correctly specified working proportional hazards model; all other error distribution/concomitant covariate specifications result in a misspecified Cox model when R^2 is non-zero. Five pairs of censoring proportions were used: (0,0), (0.25,0.25), (0.50,0.50), (0.75,0.75) and (0.25,0.75). These combinations produce 120 sets of results each for the size and the power of the tests being compared.

We first consider the size of the tests. The results are displayed as boxplots in Figure 1. The boxplots are divided into 4 sets, corresponding to four levels of association. The unconditional logrank test, which ignores the association between the concomitant variable X and failure time, attains a nominal size across all levels of association. The score test based on the Cox model performs well when there is no association between the covariate and failure time, but becomes progressively anti-conservative as the level of association increases. Although the results are not separated out, the bias is more pronounced when the proportional hazards model is incorrect. Both the Lin and Wei (1989) and Kong and Slud (1997) tests were anti-conservative, with the bias in the Lin and Wei (1989) test increasing as the

level of association increases. The results from Kong and Slud (1997) are based on a single working model, as opposed to their recommendation of choosing from a family of models, the choice based on maximizing the efficiency of the test. It is possible that implementation of their selection criterion may attenuate the bias. The conditional logrank test is unbiased across all levels of association, but is in general more variable than the other tests. The additional variability stems from the kernel smoothing procedure, where the effective sample size due to smoothing is smaller than the unconditional test statistics.

Figure 2 presents the results of the power simulations as scatterplots. In Figure 2a, the power of the unconditional logrank test is plotted against the power of the conditional logrank test. A line that corresponds to equality of the powers is also drawn in each of the plot. The scatterplot shows that there can be a substantial gain in the power (points above the line of equality) of the conditional test over the unconditional test and that the loss (points below the line) in power due to adjusting for an unassociated variable is minimal. In Figures 2b-2d, the power of the Cox model score test, the Lin and Wei robust score test (1989), and the Kong and Slud robust score test (1997) are plotted against the power of the conditional logrank test. In general, there is a power gain using the conditional logrank test, though the advantage is not uniform over all simulations. When the proportional hazards model was satisfied, and therefore the working model in the robust score tests was properly specified, the score tests were more powerful than the conditional logrank test. When the working proportional hazards model was incorrect, due to either non-proportionality or a covariate misspecification, the power of the conditional logrank test was superior to that of the score tests, especially in view of the earlier observation

that the score tests were anti-conservative. The advantage of the conditional logrank test was more pronounced for high R^2 entries. Also, as the censoring proportion was increased, the relative power advantage for the conditional logrank test was reduced. We conjecture that the flexibility gained using local neighborhoods of x to accurately estimate g, was offset by the smaller effective sample size.

3.3 Dependence Case

In this section simulations are conducted where the covariate distribution can depend on the group assignment and the censoring distribution can depend on both the group assignment and the covariate. The simulation is representative of an uncontrolled treatment comparison, where patient risk is a function of treatment assignment and the poor risk patients are offered early access to experimental treatments. Provided the censoring time does not depend on the failure time, the conditional logrank test is appropriate for this case, whereas the Lin and Wei (1989) and Kong and Slud (1997) robust versions of the Cox model are not valid in this observational study setting. DiRienzo and Lagakos (2001b), using the working Cox model formulation, have proposed a bias corrected score test to compare groups in observational studies, but have yet to provide asymptotic justification for this test.

The covariate X was generated using a conditional density function $f(x|z) = 1 + 2\theta(z - 0.5)\sin(2\pi x),\ 0 \leq x \leq 1$, where the constant θ is chosen between -1 and 1. The censoring distribution is uniform on the interval $(0, h(z, x))$, where $h(z, x) = c_{z0} + c_{z1}x + c_{z2}x^2$. The survival times were generated as $\log(T^0) = \gamma X + \epsilon$ where γ is chosen to corresponding to R^2 of 0, 0.25, 0.5 and 0.75 and ϵ is either a standard normal or an extreme value random variable. We set $\theta = 0.75$ and chose 5 sets of

values for (c_{z0}, c_{z1}, c_{z2}) corresponding to different levels of censoring. The achieved significance levels for these 40 simulations provide empirical evidence that the Cox score test (median: 0.058, range: 0.050 - 0.101) and the robust score tests of Lin and Wei (median: 0.023, range: 0.004 - 0.068) and Kong and Slud (median: 0.024, range: 0.004 - 0.076) are not valid in this dependence scenario. Although the results are not separated out, the level of degradation increases with R^2. The conditional logrank test (median: 0.051, range: 0.034 - 0.067) retains its nominal level over all values of R^2.

3.4 Multiple Covariates - Group Variable Independent Of Multiple Covariates

In order to demonstrate that the proposed method possesses desirable properties for multiple covariates, a set of monte carlo simulations for the conditional and unconditional logrank tests were conducted. As before, T^0 and Z denote the failure time and group indicator. There are now two concomitant covariates X_1 and X_2, which were generated from uniform distributions independent of Z. The failure time data were generated by

$$\log(T^0) \quad = \quad \beta Z + \phi(X_1 + X_2 - 2\eta X_1 X_2) + \epsilon$$

where ϵ is the error random variable. The constant ϕ was chosen to achieve different levels of association. The parameter values $\eta = 0$ or 1, control the interaction between X_1 and X_2. The error distributions, censoring proportions, sample size, and number of replications were the same as the single covariate simulation.

Similar to the single covariate case, the unconditional logrank test attained nominal size (median: 0.052, range: 0.046 - 0.058), as did the conditional logrank test (median: 0.049, range: 0.037 - 0.062) but was more variable. In addition, the condi-

tional logrank test showed increasing power gain relative to the unconditional logrank test as R^2 increased. In summary, when group assignment is independent of the covariates, the simulation results demonstrate that the conditional logrank test is a robust method to account for concomitant covariates when comparing survival distributions across groups.

3.5 Three sample comparison

A summary of results for the three-sample test, with a sample size of 50 in each group, is presented. The failure and censoring times are generated as described in section 3.2, and the censoring proportions used for the three groups are (0,0,0), (0.25,0.25,0.25), (0.50,0.50,0.50), (0.75,0.75,0.75) and (0.25,0.50,0.75). As previously stated, these combinations produce 120 sets of achieved significance levels for a nominal 5% test. The simulations show that the size of the three-sample conditional logrank test is nominal (median: 0.047; range: 0.032 - 0.074) though more variable than the unconditional three-sample logrank test (median: 0.054; range: 0.046 - 0.060), due to the smaller effective sample size. The conditional test, however, showed substantial gain in power over the unconditional test as the level of association between covariate and failure time increased.

4. Analysis of Prostate Cancer Data

A database of 363 metastatic prostate cancer patients treated at Memorial Sloan-Kettering Cancer Center from 1989 through 2000 was created with the purpose of determining a set of factors that influenced survival duration in this patient population. The analysis was exploratory and intended to generate hypotheses for future

clinical trial research. One subgoal of the analysis was to explore whether patients who presented with soft-tissue and bone metastasis were at greater risk of death relative to patients with bone metastasis alone. An initial analysis of the data was suggestive that soft-tissue and bone metastasis patients were at greater risk, with the logrank statistic generating a p-value of 0.138. The Kaplan-Meier estimates of the probability of survival for the two groups are presented in Figure 3a.

A further exploration of this data indicated that patients with soft-tissue and bone metastasis had lower levels of alkaline phosphatase. Alkaline phosphatase is an enzyme found in both the bone and the liver and it is used as a marker in metastatic prostate cancer; high levels indicate an increased tumor burden located in the bone. The imbalance in alkaline phosphatase levels between bone metastasis groups, required an adjustment for alkaline phosphatase levels in the comparative survival analysis. A Cox proportional hazards model containing bone metastasis and alkaline phosphatase is typically employed to adjust for the imbalance. However, the adequacy of the Cox model was questioned when we examined a diagnostic statistic based on a test of association between the scaled Schoenfeld residuals of log alkaline phosphatase and survival time (Grambsch and Therneau, 1994). The statistic, when applied to the data with both variables in the model, indicated that the proportional hazards assumption was violated (p=0.046). To account for the possibility that high alkaline phosphatase values, occuring primarily in the bone and soft tissue group, were responsible for the lack of fit, the Cox model was refit with the high alkaline phosphatase observations removed. The diagnostic test remained indicative for a model violation (p = 0.020).

As a result, we performed the adjusted bone metastasis group comparison using

the nonparametric analysis of covariance methodology developed in this manuscript. We used a normal kernel with bandwidth equal to $\hat{\sigma}_x n^{-0.26}$, where $\hat{\sigma}_x$ is the sample standard deviation of the log alkaline phosphatase values. A truncation of the highest and lowest alkaline phosphatase values avoided potential bias in the smoothing computations at the boundary, although it is noted that the truncation had little effect on the conclusion of the adjusted analysis.

The adjusted analysis produced a p-value equal to 0.005, indicating that subjects with soft-tissue and bone metastasis had an increased risk of death over subjects with bone metastasis alone. Graphical evidence of this increased risk is depicted in Figure 3b, which presents as solid lines, smoothed Kaplan-Meier estimates of the median survival time, conditional on log alkaline phosphatase value and bone metastasis group. For low values of alkaline phosphatase, the median survival times of the bone metastasis groups are comparable. But for higher alkaline phosphatase values, patients with soft-tissue bone metatasis are predicted to have a reduced median survival time relative to patients presenting with bone metastasis alone. Also in Figure 3b, we plotted as dotted lines the predicted median survival times based on the Cox proportional hazards model. The disparity between the non-model based Kaplan-Meier median survival estimates and the Cox model estimates, particularly for low alkaline phosphatase values, confirms the nonproportional relationship when both covariates are included in the analysis.

5. Discussion

Typically in survival analysis, a model based score test is applied to test a group

(perhaps treatment) effect, adjusting for other covariates. The conditional logrank test proposed in this manuscript is not model based, and makes minimal assumptions on the survival data beyond random censoring, i.e. the survival time and censoring time are independent conditional on the group variable and concomitant variables. The test statistic features kernel smoothing, in order to incorporate continuous covariates into the adjustment without reliance upon a model.

Lin and Wei (1989) and Kong and Slud (1997), and DiRienzo and Lagakos (2001a) have all proposed versions of adjusted logrank (score) tests, which also do not rely on a model to attain their validity. Importantly, each of these proposals require a form of independence between the group variable of interest and all other covariates under consideration. Independence occurs, for example, in a randomized clinical trial design, where the group variable is treatment assignment. However, the need for an adjusted treatment effect analysis is less pressing in this setting, since the randomization mechanism enables an approximate balance between treatments for each confounding factor. As a result, an adjusted analysis is often considered supplemental to the primary randomized comparison.

Historically, adjusted group comparisons have played a critical role in the analysis of observational studies. For observational data, the likelihood that all confounding factors are balanced between the levels of the group variable is small, and hence there is a need to adjust for these confounders in order to assess a group effect. The proposed conditional log rank test provides a non-model based application of analysis of covariance in the survival setting, in this general non-independence situation.

The asymptotic normality of the conditional logrank statistic was demonstrated by establishing its asymptotic equivalence to a U-statistic of degree 2. Interestingly,

the conditional Gehan statistic

$$\sum_i \sum_j \int_{s<\tau} Y_j(s)(z_i - z_j) K_{\boldsymbol{b}}(\boldsymbol{x}_i, \boldsymbol{x}_j) dN_i(s)$$

is directly a U-statistic of degree 2 and thus there is no need to formalize its asymptotic equivalence. We are currently examining whether the G^ρ family of test statistics are asymptotic U-statistics.

The conditional logrank test statistic, as it is currently proposed, restricts the number of covariates to a maximum of three. We are currently exploring the use of a single index model approach (Carroll et al., 1997), which will allow a greater number of covariates into the adjusted test by incorporating a projection of the p covariates onto a one-dimensional space, within the adjusted test statistic calculation.

Acknowledgements

The authors would like to thank the editor and referees for suggestions which led to improvements in this paper. This research is supported by the National Cancer Institute, award CA 73848.

Appendix

Proof of the asymptotic normality of the two-sample test statistic and calculation of its asymptotic variance under the null hypothesis.

We require the following conditions.

(i) The covariate vector $\boldsymbol{X}$ lies in a p-dimensional bounded rectangle $\mathcal{X}$, with dimension p less than or equal to 3. It is assumed without loss of generality that the volume of $\mathcal{X}$ is equal to 1.

(ii) The survival time is truncated at time τ, such that

$$\int_{s<\tau} \lambda(s|\boldsymbol{x})ds < \infty$$

for all $\boldsymbol{x}$. The conditional hazard $\lambda(s|\boldsymbol{x})$ and its first and second derivative with respect to $\boldsymbol{x}$ are bounded and continuous in $(s, \boldsymbol{x})$.

(iii) The product kernel function K is defined as

$$K(\boldsymbol{u}) = \prod_{l=1}^{p} k(u_l)$$

where the univariate kernel k is a density function with finite support, symmetric about zero, and has a Lipschitz continuous second derivative on its support.

(iv) Let $a_r(\boldsymbol{x}, s) = n^{-1}\sum_j Y_j(s)z_j^r K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x})$, then $a_r(\boldsymbol{x}, s) \to \alpha_r(\boldsymbol{x}, s)$ uniformly in $(\boldsymbol{x}, s)$, for $r = 0, 1$. The functions $\alpha_r(\boldsymbol{x}, s)$ are bounded in $(\boldsymbol{x}, s)$ with bounded and continuous first and second derivatives with respect to $\boldsymbol{x}$, and $\alpha_0(\boldsymbol{x}, s)$ is bounded away from zero.

(v) The bandwidth vector $\boldsymbol{b}$ is chosen such that as $n \to \infty$, $n\tilde{b}^p \to \infty$ and $n\tilde{b}^4 \to 0$, where $\tilde{b} = \max_{1\le l\le p} b_l$ and $p \le 3$.

The asymptotic normality of the test statistic

$$S_n(\tau) = \sum_i \int_{s<\tau} z_i dN_i(s) - \sum_i \int_{s<\tau} \frac{\sum_j Y_j(s)z_j K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\sum_j Y_j(s)K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)} dN_i(s)$$

is demonstrated by showing that it is asymptotically equivalent to a U-statistic of degree 2.

We can express

$$\begin{aligned} n^{-1/2}S_n(\tau) &= n^{-1/2}\sum_i \int_{s<\tau} \{z_i - \bar{z}(\boldsymbol{x}_i, s)\}\, dM_i(s) \\ &\quad + n^{-1/2} \sum_i \int_{s<\tau} \{z_i - \bar{z}(\boldsymbol{x}_i, s)\}\, Y_i(s) d\Lambda(s|\boldsymbol{x}_i) \end{aligned}$$

where $\bar{z}(\boldsymbol{x}, s) = a_1(\boldsymbol{x}, s)/a_0(\boldsymbol{x}, s)$ and $M_i(s) = N_i(s) - \int Y_i(s) d\Lambda(s|\boldsymbol{x}_i)$ is a subject specific martingale. This Doob-Meyer type decomposition of $n^{-1/2}S_n(\tau)$ is denoted as $A_n(\tau) + B_n(\tau)$.

We first show that $B_n(\tau)$ is asymptotically negligible. Note that

$$\begin{aligned} B_n(\tau) &= n^{-1/2}\sum_i \int_{\boldsymbol{x}^*} \int_{s<\tau} \{z_i - \bar{z}(\boldsymbol{x}^*, s)\}\, Y_i(s) K_{\boldsymbol{b}}(\boldsymbol{x}^*, \boldsymbol{x}_i) d\Lambda(s|\boldsymbol{x}^*) d\boldsymbol{x}^* \\ &+ n^{-1/2}\sum_i \int_{\boldsymbol{x}^*} \int_{s<\tau} \{z_i - \bar{z}(\boldsymbol{x}_i, s)\}\, \{Y_i(s) d\Lambda(s|\boldsymbol{x}_i) - Y_i(s) K_{\boldsymbol{b}}(\boldsymbol{x}^*, \boldsymbol{x}_i) d\Lambda(s|\boldsymbol{x}^*)\}\, d\boldsymbol{x}^* \\ &+ n^{-1/2}\sum_i \int_{\boldsymbol{x}^*} \int_{s<\tau} \{\bar{z}(\boldsymbol{x}^*, s) - \bar{z}(\boldsymbol{x}_i, s)\}\, Y_i(s) K_{\boldsymbol{b}}(\boldsymbol{x}^*, \boldsymbol{x}_i) d\Lambda(s|\boldsymbol{x}^*) d\boldsymbol{x}^* \end{aligned}$$

The first term of $B_n(\tau)$ is zero. Letting

$$g_i(\boldsymbol{x}^*) = \int_{s<\tau} \{\bar{z}(\boldsymbol{x}^*, s) - \bar{z}(\boldsymbol{x}_i, s)\}\, Y_i(s) d\Lambda(s|\boldsymbol{x}^*),$$

the third term may be written as $n^{-1/2}\sum_i \int_{\boldsymbol{x}^*} K_{\boldsymbol{b}}(\boldsymbol{x}^*, \boldsymbol{x}_i) g_i(\boldsymbol{x}^*) d\boldsymbol{x}^*$. Using conditions (i) through (iv), a two-term Taylor expansion produces

$$\int_{\boldsymbol{x}^*} K_{\boldsymbol{b}}(\boldsymbol{x}^*, \boldsymbol{x}_i) g_i(\boldsymbol{x}^*) d\boldsymbol{x}^* = g_i(\boldsymbol{x}_i) + O(\tilde{b}^2)$$

uniformly in $\boldsymbol{x}$. Since $g_i(\boldsymbol{x}_i) = 0$ the third term is $O(n^{1/2}\tilde{b}^2)$. The uniformity follows from the bounded and Lipschitz conditions (i)-(iii). Using a similar argument, the second term is also $O(n^{1/2}\tilde{b}^2)$. It follows from condition (v) that $B_n(\tau) \to 0$.

We now examine the process $A_n(\tau)$. A three-term Taylor expansion produces

$$
\begin{aligned}
& n^{-1/2} \sum_i \int_{s<\tau} \left\{ z_i - \frac{a_1(\boldsymbol{x}_i, s)}{a_o(\boldsymbol{x}_i, s)} \right\} dM_i(s) \\
&= n^{-1/2} \sum_i \int_{s<\tau} \left[z_i - \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_o(\boldsymbol{x}_i, s)} - \frac{1}{\alpha_0(\boldsymbol{x}_i, s)} \left\{ a_1(\boldsymbol{x}_i, s) - \alpha_1(\boldsymbol{x}_i, s) \right\} \right] dM_i(s) \\
&\quad + n^{-1/2} \sum_i \int_{s<\tau} \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_0^2(\boldsymbol{x}_i, s)} \left\{ a_0(\boldsymbol{x}_i, s) - \alpha_0(\boldsymbol{x}_i, s) \right\} dM_i(s) \; - \; r_n
\end{aligned}
$$

where $r_n = r_{n1} - r_{n2}$ is

$$
\begin{aligned}
& n^{-1/2} \sum_i \int_{s<\tau} \frac{a_1(\boldsymbol{x}_i, s)}{\alpha_0^2(\boldsymbol{x}_i, s) a_0(\boldsymbol{x}_i, s)} \left\{ a_0(\boldsymbol{x}_i, s) - \alpha_0(\boldsymbol{x}_i, s) \right\}^2 dM_i(s) \\
& -n^{-1/2} \sum_i \int_{s<\tau} \frac{1}{\alpha_0(\boldsymbol{x}_i, s) a_0(\boldsymbol{x}_i, s)} \left\{ a_1(\boldsymbol{x}_i, s) - \alpha_1(\boldsymbol{x}_i, s) \right\} \left\{ a_0(\boldsymbol{x}_i, s) - \alpha_0(\boldsymbol{x}_i, s) \right\} dM_i(s).
\end{aligned}
$$

It is demonstrated below that $r_n = o_p(1)$. First, $a_r(\boldsymbol{x}, s) = \alpha_r(\boldsymbol{x}, s) + O_p(n\tilde{b}^p)^{-1/2}$ uniformly in $(\boldsymbol{x}, s) \in \mathcal{X} \times [0, \tau]$. The uniform convergence follows from the compactness of the time and covariate space, the Lipschitz continuity condition, and the bandwidth conditions. Using these conditions and the fact that (M_i, M_j) are orthogonal martingales for $i \neq j$,

$$
E\left(r_{n2}^2\right) = E\left[n^{-1} \sum_i \int_{s<\tau} \frac{\left\{ a_1(\boldsymbol{x}_i, s) - \alpha_1(\boldsymbol{x}_i, s) \right\}^2 \left\{ a_0(\boldsymbol{x}_i, s) - \alpha_0(\boldsymbol{x}_i, s) \right\}^2}{\alpha_0^2(\boldsymbol{x}_i, s) a_0^2(\boldsymbol{x}_i, s)} dM_i^2(s) \right]
$$

is $O\left\{ (n\tilde{b}^p)^{-2} \right\}$. Therefore by Markov's inequality, for $p \leq 3$, $0 < \epsilon < 4 - p$, and condition (v), it follows that for $\tilde{b} = O\left\{ n^{-1/(4-\epsilon)} \right\}$, $r_{n2} = O_p\left\{ n^{(p-4+\epsilon)/(4-\epsilon)} \right\} = o_p(1)$. A similar argument produces $r_{n1} = o_p(1)$.

Now $A_n(\tau)$ can be rewritten as

$$
n^{-3/2} \sum_i \sum_j \int_{s<\tau} \left[z_i - \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_o(\boldsymbol{x}_i, s)} - \frac{Y_j(s) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\alpha_0(\boldsymbol{x}_i, s)} \left\{ z_j - \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_o(\boldsymbol{x}_i, s)} \right\} \right] dM_i(s) + o_p(1)
$$

and since each term has mean zero (M_i are independent martingales), $A_n(\tau)$ is an asymptotic U-statistic of degree 2. Therefore, from the asymptotic distribution theory of U-statistics and Slutsky's theorem, $n^{-1/2}S_n(\tau)$ is asymptotically normal with mean zero.

We now compute the asymptotic variance of $S_n(\tau)$. Denote the kernel of the U-statistic by

$$v_{ij} = \int_{s<\tau} \left[z_i - \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_o(\boldsymbol{x}_i, s)} - \frac{Y_j(s) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\alpha_0(\boldsymbol{x}_i, s)} \left\{ z_j - \frac{\alpha_1(\boldsymbol{x}_i, s)}{\alpha_o(\boldsymbol{x}_i, s)} \right\} \right] dM_i(s)$$

and its symmetrized version by $v_{ij}^S = v_{ij} + v_{ji}$. The test statistic $S_n(\tau)$ can be written as $\sum_i v_{ii} + \sum_{i<j} v_{ij}^S$ and its variance is

$$\text{var}\{S_n(\tau)\} = \text{var}\left(\sum_i v_{ii} \right) + 2 \times \text{cov}\left(\sum_i v_{ii}, \sum_{i<j} v_{ij}^S \right) + \text{var}\left(\sum_{i<j} v_{ij}^S \right).$$

Note that the variances and covariances can be estimated consistently by the corresponding sample sums of squares and products. Thus the first variance term is

$$\sum_i v_{ii}^2 \tag{C1}$$

and the covariance term is twice

$$\sum_{i \neq j} (v_{ii} + v_{jj}) v_{ij} \tag{C2}$$

The second variance term is split into variance and covariance components (Lehmann 1975, pp. 336-7). The variance part is given by

$$\sum_{i<j} \left(v_{ij}^S \right)^2 = \sum_{i \neq j} \left\{ (v_{ij})^2 + v_{ij} v_{ji} \right\} \tag{C3}$$

and the covariance is

$$\sum_{i \neq j \neq l} (v_{ij}v_{il} + v_{ij}v_{li} + v_{ij}v_{jl} + v_{ij}v_{lj}) \qquad (C4)$$

Putting these terms together, a consistent estimate of the variance of the test statistic $S_n(\tau)$ is obtained by replacing v_{ij} with

$$\hat{v}_{ij} = \delta_i \left[z_i - \frac{a_1(\boldsymbol{x}_i, y_i)}{a_o(\boldsymbol{x}_i, y_i)} - \frac{I(y_j \geq y_i) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{a_o(\boldsymbol{x}_i, y_i)} \left\{ z_j - \frac{a_1(\boldsymbol{x}_i, y_i)}{a_o(\boldsymbol{x}_i, y_i)} \right\} \right]$$

for v_{ij} into $(C1) + 2(C2) + (C3) + (C4)$.

For computational purposes, we can reduce the n^3 operations required to compute the C4 term down to n^2 operations. Let $v_{\cdot j}$ and $v_{i\cdot}$ denote the row and column sums of the v matrix. Then

$$\sum_i (v_{\cdot i}^2 + 2v_{\cdot i}v_{i\cdot} + v_{i\cdot}^2) = 4(C1) + 4(C2) + 2(C3) + (C4)$$

and therefore only the row sums, column sums, their inner products, and the terms C1, C2 and C3, are needed to compute the asymptotic variance.

Calculation of the asymptotic mean of the test statistic $n^{-1}S_n$

From equation (5) in the text,

$$n^{-1}S_n(\tau) = n^{-1} \sum_i \int_{s<\tau} w(\boldsymbol{x}_i, s) \left\{ \frac{dN_{1i}(s)}{a_{11}(\boldsymbol{x}_i, s)} - \frac{dN_{0i}(s)}{a_{10}(\boldsymbol{x}_i, s)} \right\}.$$

Substitution of the smoothing result

$$\frac{\sum_j Y_j(s) z_j K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\sum_j Y_j(s) K_{\boldsymbol{b}}(\boldsymbol{x}_j, \boldsymbol{x}_i)} = E(Z_i | Y_i(s) = 1, \boldsymbol{X} = \boldsymbol{x}_i) + o_p(1)$$

and the martingale decomposition $dN(s) = Y(s)d\Lambda(s|\boldsymbol{x}) + dM(s)$ into this equation, along with an application of the law of large numbers, produces the result that $n^{-1}S_n(\tau)$ converges in probability to

$$n^{-1}\sum_i \int_{s<\tau} E\left[\begin{array}{r} Y_i(s)E(Z_i|Y_i(s)=1,\boldsymbol{X}_i)\{1-E(Z_i|Y_i(s)=1,\boldsymbol{X}_i)\} \\ \times\{d\Lambda_1(s|\boldsymbol{X}_i)-d\Lambda_0(s|\boldsymbol{X}_i)\} \end{array}\right]$$

References

Akritas, M. G., Arnold, S. F. and Du, Y. (2000) Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika*, **87**, 507-526.

Akritas, M. G. and Brunner, E. (1997) Nonparametric methods for factorial designs with censored data. *J. Am. Statist. Ass.*, **92**, 568-576.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M.P. (1997) Generalized partially linear single-index models. *J. Am. Statist. Ass.*, **92**, 477-489.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187-202.

DiRienzo, A. G. and Lagakos, S. W. (2001a) Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *J. R. Statist. Soc.* B, **63**, 745-757.

DiRienzo, A. G. and Lagakos, S. W. (2001b) Bias correction for score tests under misspecified regression models. *Biometrika*, **88**, 421-434.

Du, Y., Akritas, M. G. and Keilegom, I. V. (2003) Nonparametric analysis of covariance for censored data. *Biometrika*, **90**, 269-287.

Grambsch, P. M. and Therneau, T. M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

Heller, G. (2001) The Cox proportional hazards model with a partly linear relative risk function. *Lifetime Data Analysis*, **7**, 255-277.

Kong, F. H. and Slud, E. (1997) Robust covariate-adjusted logrank tests. *Biometrika*, **84**, 847-862.

Lehmann, E. L. (1975) *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Lin, D. Y. and Wei, L. J. (1989) The robust inference for the Cox proportional hazards model. *J. Am. Statist. Ass.*, **84**, 1074-1078.

Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Sasieni, P. (1992) Information bounds for the conditional hazard ratio in a nested family of regression models. *J. R. Statist. Soc.* B, **54**, 627-635.

Sun, J. and Yang, I. (2000) Nonparametric tests for stratum effects in the Cox model. *Lifetime Data Analysis* **6**, 321-330.

Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data: Extending the Cox Model.* New York: Springer-Verlag.

Fig. 1. Boxplots of the sizes for the unconditional logrank test (LR), score test based on proportional hazards model (Cox), the robust score tests (LW,KS), and the conditional logrank test (CLR). The simulations are based on a single covariate. The level of association between covariate and time are given above each group.

Fig. 2. Scatterplots of the power of the conditional logrank test against (a) the unconditional logrank test and (b-d) the score tests along with the line of equal power. The points above and below the line correspond to situations where the conditional test is more and less powerful respectively. The vertical distance to the line is the magnitude of the difference.

Fig. 3. (a) Kaplan-Meier curves for overall survival of patients with bone metastasis only (solid line) and bone and soft-tissue metastases (dashed line) (b) Predicted median survival times as a function of log alkaline phosphatase. The solid lines are obtained by non-parametric smoothing of Kaplan-Meier curves and the dashed lines are from Cox model fit.