行政院國家科學委員會專題研究計畫　成果報告

異質性族群單調時變療效之逐次分析

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-039-001-
執行期間：91年08月01日至92年07月31日
執行單位：中國醫藥大學公共衛生學系

計畫主持人：吳宏達

計畫參與人員：吳宏達

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92年10月28日

A Partial Score Test for Difference among Heterogeneous Populations

Hong-Dar Isaac Wu,
School of Public Health, China Medical University,
91 Hsueh-Shih Rd., Taichung 40443, Taiwan.
E-mail: honda@mail.cmu.edu.tw

2003 August

A Partial Score Test for Differences among Heterogeneous Populations

Summary

In event time data analysis, comparisons between distributions are made by the logrank test. When the data appear to contain crossing hazards phenomena, nonparametric weighted logrank statistics are usually suggested to accommodate different-weighted functions to increase the power. However, the gain in power by imposing different weights has its limits since differences before and after the crossing point may balance each other out. In contrast to the weighted logrank tests, we propose a score-type statistic based on the semiparametric-, heteroscedastic-hazards regression model of Hsieh (2001, Journal of the Royal Statistical Society, Series B **63**,63-79), by which the nonproportionality is explicitly modeled. Our score test is based on estimating functions derived from partial likelihood under the heteroscedastic model considered herein. Simulation results show the benefit of modeling the heteroscedasticity and power of the proposed test to two classes of weighted logrank tests (including Fleming-Harrington's test and Moreau's locally most powerful test), a Renyi-type test, and the Breslow's test for acceleration. We also demonstrate the application of this test by analyzing actual data in clinical trials.

Key Words: heteroscedasticity; crossing hazards; proportional hazards; weighted logrank test.

1 Introduction

When one is dealing with event time data in comparative trials, the proportional hazards (PH) model (Cox, 1972) is usually used to estimate the relative effect of a treatment adjusted for prognostic factors. However, when 'nonproportionality' (or more specifically, 'crossing hazards') is present, the PH model and the companying estimation based on the partial likelihood leads to a biased estimate of the 'effect-measure', e.g., hazards ratio. In this situation, the logrank test for equality between distributions may have poor power against the crossing hazards alternatives (Andersen et al., 1993, page 390). To detect a substantial difference, weighted logrank tests are often used to improve the power (Gill, 1980; Harrington and Fleming, 1982; Moreau et al., 1992; Kosorok and Lin, 1999). In particular, Fleming and Harrington (1991, Chapter 7) presented a class of weighted logrank statistics, hereafter referred to as the $G^{\rho,\gamma}$-statistic, by which different weights can be used to stress early, middle, or late differences among groups by imposing different ρ-γ configurations. Nevertheless, the weighted logrank test has a drawback in that if a specific weight is chosen, it is not globally valid for other cases or datasets. If there is a possible hazards crossing and it is a result of a general 'acceleration' (not necessarily a multiplicative effect on 'time'), a complementary statistic to the ordinary logrank statistic can be considered to accommodate for the cross-effect (Breslow, Edler, and Berger, 1984). In this article, however, we study a quite-general model as an alternative which explicitly accounts for the crossing hazards as a result of *heterogeneity*. We assume the null hypthesis ($\mathcal{H}_0$) that the hazards (or distributions) of different groups are equal; the alternative hypothesis ($\mathcal{H}_a$) is the class of heteroscedastic hazards regression (HHR) model (Hsieh, 2001). In terms of cumulative hazards, the HHR model is expressed as

$$\Lambda(t; Z, X) = \{\Lambda_0(t)\}^{\exp(\phi' X)} \exp(\beta' Z), \tag{1}$$

where X and Z are two vectors of predictable time-dependent covariates, β and ϕ are the associated parameters of interests, and $\Lambda_0(t)$ is an unknown baseline cumulative hazard function with $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. If X and Z are not time dependent, model (1) can also be expressed in terms of the hazard function:

$$\lambda(t; Z, X) = \lambda_0(t)\text{exp}(\phi' X)\{\Lambda_0(t)\}^{\text{exp}(\phi' X)-1}\text{exp}(\beta' Z). \tag{2}$$

In view of model (2), it can also accommodate time dependent X and Z. To avoid complexity of exposition, we assume that $X = Z$ throughout the following context. It is the same model used as an 'alternative' by Quantin et al. (1996) and Devarajan and Ibrahimi (2002) to test the validity of the proportional hazards model. In this paper, however, we investigate the performance of a score-type test for differences among heterogeneous populations, based on a set of estimating functions. The estimating functions are derived from the 'partial likelihood' appearing in the first factor of a decomposition of the full likelihood (Johansen, 1983) with either model (1) or (2).

In contrast to the HHR model, a simpler *semiparametric* alternative was studied by Moreau et al. (1992) to produce a locally most powerful (LMP) test for $\mathcal{H}_0$:

$$\Lambda(t; X) = \{\Lambda_0(t)\}^{\text{exp}(\phi' X)}, \tag{3}$$

which includes the Weibull class, $\Lambda(t; \eta, \phi' X) = (\eta t)^{\text{exp}(\phi' X)}$, as a special case. The difference between (1) and (3) is that from the viewpoint of the linear transformation model (see §2), model (3) considers only the change in 'scale', whereas model (1) accommodates both of 'location-shift' and 'scale-change'. It is thus appealing to compare the performance of tests for equal distributions constructed on models with this *nested* structure: $\mathcal{H}_0$ $\subset$model (3)$\subset$ model (1) (or $\mathcal{H}_a$). Heuristically, if a more-extended model, parametric or semiparametric, fits the data well, the statistical test constructed from a likelihood based on these models will generally outperform the nonparametric tests as well as the tests constructed on the basis of a narrower class.

In §2, Hsieh's estimating functions together with the Breslow-type estimate of the baseline cumulative hazard are presented. With the regression setting of (1), a score-type test is proposed in §3 based on the estimating functions which are treated as real scores. As one may be concerned with the performance of the proposed test, the corresponding two-sample expression is also derived. For the two-sample study, simulations and data analyses are reported in §4 to compare the proposed test with two classes of weighted logrank tests (including Fleming-Harrington's $G^{\rho,\gamma}$-statistics and Moreau's LMP test), a Renyi-type test which only captures the supremum discrepancy between observed and expected realizations of a process, and Breslow's acceleration test designed for a possible crossing-hazards phenomenon between groups. Implementation of the HHR model and the proposed test are illustrtated through analyses of actual data published in the literature. Finally, we provide some discussion on model formulation, the concern of bias when heterogeneity is neglected, and the goodness-of-fit problem.

2 The Heteroscedastic Model and Estimating Functions

2.1 Model genesis

Recently, the linear transformation model has attracted much attention, because it attempts to provide a very general framework for survival data analysis (Dabrowska and Doksum, 1988; Cheng, Wei, and Ying, 1995, 1997). Consider the following model:

$$h(T) = -\beta' z + \epsilon, \tag{4}$$

where $h(\cdot)$ is an unknown function of the random variable T, and the distribution of ϵ, $F(t)$, is specified. When $F(t) = 1 - \exp\{-\exp(t)\}$, which is an extreme value distribution, (4) corresponds to the proportional hazards model. In contrast, Hsieh (2001) considers the error terms subject to heteroscedasticity:

$$h(T) = -\beta' z + \sigma\epsilon, \tag{5}$$

where $\sigma = \exp(\phi' X)$. If $Z = X$, it means that the heteroscedasticity is also a result of the covariate Z itself. In terms of cumulative hazards, the corresponding model of (5) is the HHR model (1). As a parametric example, Hsieh (1996) demonstrated that two-parameter Weibull distributions with different scales and different shapes satisfy the formulation of model (1).

2.2 Partial score equations and the computational algorithm

Assume that there are n observed failures or right-censored times $T_1, T_2, \ldots, T_n$. Without loss of generality, let $T_1 < T_2 < \ldots < T_n$. In this subsection, notations and estimating functions are introduced. Let $N_i(t)$ be the counting process of individual i associated with the intensity

$$h_i(t) = Y_i(t)\lambda_0(t)\exp\{(\beta+\phi)' Z_i\}\{\Lambda_0(t)\}^{\exp(\phi' Z_i)-1},$$

where $Y_i(t)$ is the at-risk indicator at time t. Further, denote

$$S_J(t) = (1/n)\sum Y_i(t)J_i(t)\exp\{(\beta+\phi)' Z_i\}\{\Lambda_0(t)\}^{\exp(\phi' Z_i)-1},$$

for a predictable process $J(t)$. According to either (1) or (2), the full likelihood, L_F, with Johansen's decomposition (Johansen, 1983) is

$$L_F(\theta, \Lambda_0) = \Pi \int_0^\tau \frac{h_i(u)dN_i(u)}{S_1(u)} \cdot \Pi \int_0^\tau S_1(u)\lambda_0(u)dN_i(u) \cdot \exp\{-\int_0^\tau nS_1(u)du\}$$

for a maximal observation time τ, $\theta = (\beta', \phi')'$. We have the $\sqrt{n}$-scaled partial log-likelihood

$$l_p = (1/\sqrt{n})\sum \log\{\int_0^\tau \frac{h_i(u)dN_i(u)}{S_1(u)}\}.$$

The estimating functions can be derived from taking partial derivatives of l_p with respect to β and ϕ:

$$\mathrm{E}_\beta = (1/\sqrt{n})\sum \int_0^t \{Z_i - \frac{S_Z(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)}\}dN_i(u) \text{ and} \tag{6}$$

$$\mathrm{E}_{\phi} = (1/\sqrt{n}) \sum \int_0^t \{V_i - \frac{S_V(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)}\} dN_i(u), \tag{7}$$

where $V_i(t) = Z_i(t)\exp(\phi' Z_i)\log\{\Lambda_0(t)\}$. The baseline cumulative hazard $\Lambda_0(t)$ in (7) makes it identifiable from (6). To estimate $\Lambda_0(t)$, the Breslow-type estimator can be considered:

$$\Lambda_0(t) = \sum \int_0^t [\sum Y_i(u)\exp\{(\beta+\phi)' Z_i\}\{\Lambda_0(u)\}^{\exp(\phi' Z_i)-1}]^{-1} dN_i(u). \tag{8}$$

Instead of solving (8) directly, however, a finite-dimensional sieve approximation of $\Lambda_0(t)$ is used:

$$\Lambda_{0m}(t) = \int_0^t \sum_1^m \alpha_i 1\{\tau_{i-1} < u \leq \tau_i\} du, \tag{9}$$

where τ_j's are appropriate cutoff points, and m, a 'smoothing' factor, is the dimension of Λ_{0m} chosen to approximate the function Λ_0. Note that the estimating functions (6) and (7) can also be obtained from some algebraic transformation of the primary estmating functions, derived by Hsieh (2001), based on the nonparametric maximum likelihood estimation introduced by Gill (1989) and Gill and van der Vaart (1993), along with a martingale structure defined.

An algorithm, other than that suggested by Hsieh (2001, §5.1, steps 1∼3), can be used to compute the 'first-step' estimates of parameters β, ϕ, and $\{\alpha_i\}_1^m$ for the iteration procedure in the *over-identified estimating equation* (OEE) approach. In the j-th step iteration,

$$\Lambda_{0m}^{(j)}(t) = \sum \int_0^t [\sum Y_i(u)\exp[\{\beta^{(j-1)} + \phi^{(j-1)}\} Z_i]\{\Lambda_{0m}^{(j-1)}(u)\}^{\sigma_i^{(j-1)}-1}]^{-1} dN_i(u), \tag{10}$$

where $\Lambda_{0m}^{(j)}(t)$, $\beta^{(j)}$, $\phi^{(j)}$, and $\sigma_i^{(j)} = \exp\{\phi^{(j)} Z\}$ denote the j-th step iterated values, $j = 0, 1, 2, 3, \ldots$. We can choose the initial guess of β and Λ_0 (i.e., $\beta^{(0)}$ and $\Lambda_{0m}^{(0)}$) as the estimates of the conventional Cox's model where $\gamma^{(0)} = 0$. The first-step estimates thus obtained are consistent and asymptotically normal if some regularity conditions are assumed. However, they are less efficient than the estimates obtained from the OEE method. In addition, computational problems also will be encountered since the 'surface' of the likelihood function with a sieve approximation is complex.

3 Proposed Test

3.1 Regression

The HHR model is a nonproportional hazards model under which the hazards according to heterogeneous populations can cross at some point(s). By assuming the HHR model in our problem, the null hypothesis is $\mathcal{H}_0 : \beta = \phi = 0$; and the alternative is $\mathcal{H}_a : \beta$ and ϕ are both arbitrary and finite. Although model adequacy is crucial for making a score-type test plausible, we can only suggest in this paper that a good fit of the HHR model is judged by *visual diagnostics*. A formal goodness-of-fit test for model validity based on a large sample consideration has been studied by Wu, Hsieh, and Chen (2002). When visual discrepancy between estimated survivals based on the Kaplan-Meier method and the HHR model is 'small', we expect that a **local test** constructed on the basis of the HHR model can be a more powerful one for detecting the difference. We now treat E_β and E_ϕ as if they were the true score functions. Under model (1) or (2), a score-type test statistic can be written as

$$\mathcal{W} = \{E_\beta, E_\phi\}\mathcal{I}^{-1}\{E_\beta, E_\phi\}', \tag{11}$$

which is evaluated at $(\beta, \phi, \Lambda_0(\tau)) = (0, 0, \hat{\Lambda}_0(\tau))$. It is important to note that $\hat{\Lambda}_0(\cdot)$ is obtained as an estimate in the parameter space of $\mathcal{H}_a$ (not under the null hypothesis $\mathcal{H}_0$) using the method of §2.2 with sieve approximation, and thus is consistent in the entire parameter space of $\mathcal{H}_0 \bigcup \mathcal{H}_a$. The information matrix, $\mathcal{I}$, defined as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\phi} \\ \mathcal{I}_{\phi\beta} & \mathcal{I}_{\phi\phi} \end{pmatrix},$$

has the components

$$\mathcal{I}_{\beta\beta} = (1/n) \sum \int_0^\tau \{Z_i - \frac{S_Z(u; \Lambda_0, \theta)}{S_1(u; \Lambda_0, \theta)}\}^{\otimes 2} dN_i(u),$$

$$\mathcal{I}_{\phi\phi} = (1/n) \sum \int_0^\tau \{V_i - \frac{S_V(u; \Lambda_0, \theta)}{S_1(u; \Lambda_0, \theta)}\}^{\otimes 2} dN_i(u),$$

and

$$\mathcal{I}_{\beta\phi} = \mathcal{I}_{\phi\beta} = (1/n)\sum \int_0^\tau \{Z_i - \frac{S_Z(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)}\}\{V_i - \frac{S_V(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)}\}dN_i(u).$$

The symbol $A^{\otimes 2}$ denotes the product of column vector A and its transpose A'. Asymptotically, $\mathcal{W}$ is distributed as χ^2_{2k}, where k is the dimension of Z.

3.2 Two-sample study

We assume that, in the two-sample case, an individual does not switch between the groups over time. Let D_{ji} be the number of failures in group j $(j = 0, 1)$ at time t_i. Assuming no ties, $D_{1i} = 1$ if the individual who failed is a member of group 1 and $D_{1i} = 0$ otherwise. Further, $\overline{Y}_{ji}$ is the risk set size of group j at time t_i. The estimating functions corresponding to (6) and (7) are reduced to

$$E_\beta = (1/\sqrt{n})\sum(\mathcal{D}_{1i} - \mathcal{E}_{1i}) \text{ and}$$
$$E_\phi = (1/\sqrt{n})\sum \Delta_i(\mathcal{D}_{1i} - \mathcal{E}_{1i}), \tag{12}$$

where Δ_i =log$\Lambda_0(t_i)$; and $\mathcal{E}_{1i} = \overline{Y}_{1i}/\overline{Y}_i$, $\overline{Y}_i = \overline{Y}_{1i} + \overline{Y}_{0i}$, is the empirical probability calculated at t_i when the person who failed is a member of group 1 under the 'equility assumption' $\lambda_1 = \lambda_0$. Entries of the corresponding information matrix are

$$\mathcal{I}_{\beta\beta} = (1/n)\sum(\mathcal{D}_{1i} - \mathcal{E}_{1i})^2,$$
$$\mathcal{I}_{\beta\phi} = (1/n)\sum \Delta_i(\mathcal{D}_{1i} - \mathcal{E}_{1i})^2, \text{ and}$$
$$\mathcal{I}_{\phi\phi} = (1/n)\sum \Delta_i^2(\mathcal{D}_{1i} - \mathcal{E}_{1i})^2. \tag{13}$$

In practice, Δ_i is substituted by $\hat{\Delta}_i = \log\hat{\Lambda}_0(t_i)$. However, the estimate $\hat{\Lambda}_0(t_i)$ **cannot** be derived from (8) under $\mathcal{H}_0$ as $\hat{\Lambda}_0(t_i) = \sum_{k\le i}\{\overline{Y}_k\}^{-1}$. Instead, bacause the HHR-class is the alternative $(\mathcal{H}_a)$, $\hat{\Lambda}_0(t_i)$ must be solved from (8), with β and γ being solved simultaneously from (6) and (7) (in their two-sample setting of (12)); it is dependent on the values of $\hat{\beta}$ and $\hat{\gamma}$, and thus can be expressed as $\hat{\Lambda}_0(t_i) = \hat{\Lambda}_0(t_i;\hat{\beta},\hat{\gamma})$.

The two-sample counterparts of the $\mathcal{W}$-statistic in (11) is distributed as χ^2_2 under $\mathcal{H}_0$ because $\dim(Z) = 1$. A similar two-sample result from the empirical process approach with *strong approximations* to the receiver operating characteristic (ROC) curves, followed by a least square method, can be found in Hsieh (1996). We are interested in comparing the performance of $\mathcal{W}$ to the class of weighted logrank tests (Klein and Moeschberger, 1997, pages 191-198; Fleming and Harrington, 1991, Chapter 7) with a predictable weight process $\mathcal{K}(t)$:

$$T_{\mathcal{K}} = \frac{\{\sum_1^n \mathcal{K}(t_i)(\mathcal{D}_{1i} - \mathcal{E}_{1i})\}^2}{\sum_1^n \mathcal{K}^2(t_i)\mathcal{E}_i(1-\mathcal{E}_i)}. \tag{14}$$

In particular, we consider two choices of $\mathcal{K}(t)$. (I) The function $\mathcal{K}(t_i) = \{\hat{S}(t_i-)\}^\rho\{1-\hat{S}(t_i-)\}^\gamma$, weighs differences in early, middle, and late stages between the two groups according to $(\rho,\gamma) = (1,0), (1,1)$, and $(0,1)$, respectively. The quantity $\hat{S}(t-)$ is the Kaplan-Meier (K-M) estimate of survivor function of the *pooled sample* just before time t, under $\mathcal{H}_0$. When $(\rho,\gamma) = (0,0)$, the $G^{0,0}$-statistic corresponds to the ordinary logrank statistic. (II) If the weight function $\mathcal{K}(t_i) = 1 + \log[-\log\{\Pi_{k=1}^i \overline{Y}_k/(1+\overline{Y}_k)\}]$, this results in Moreau's LMP statistic, $\mathcal{M}$, when the parameter space of $\mathcal{H}_a$ is further restricted to model (3) by ignoring the location parameter β (Moreau et al., 1992). In addition, we also compare the performance of $\mathcal{W}$-, $\mathcal{M}$-, and $G^{\rho,\gamma}$-statistics with a Renyi-type test $\mathcal{R}$ (Klein and Moeschberger, 1997, Chapter 7) and a test ($\mathcal{B}$) proposed in Breslow et al. (1984) designed for general 'acceleration'. The Renyi-type test captures the supremum of the process

$$(1/\sqrt{n})\sum \mathcal{K}(t_i)(\mathcal{D}_{1i} - \mathcal{E}_{1i})$$

and is robust to the crossing-hazards alternative. The Breslow-type test employed a complementary statistic for detecting crossing-hazards. In this paper, we adopt $\mathcal{B} = T_1 + T_Q$, where $Q(t_i) = \hat{\Lambda}_0(t_i)$, for comparison. Under $\mathcal{H}_0$, the $G^{\rho,\gamma}$- and $\mathcal{M}$-statistics are both distributed as χ^2_1; the $\mathcal{B}$-statistic is distributed as χ^2_2; and the distribution of $\mathcal{R}$ can be approximated by the distribution of the supremum of standard Brownian motion. In the next section, power comparisons between $G^{\rho,\gamma}$-, $\mathcal{M}$-, $\mathcal{R}$-, $\mathcal{B}$-, and $\mathcal{W}$-statistics are made through simulation studies. Finally, it is worth noting again that these comparisons were made between various statistics constructed on *nested* spaces of models (also refers to §1).

4 Numerical Study

4.1 Simulations

Let T^*_z be the failure time random variable distributed as F_z=Weibull(a, b) with $a = \exp(\beta' z)$ and $b = \exp(\phi' z)$; the cumulative hazard function is $\Lambda(t) = at^b$. To make comparisons, we choose half of the sample according to $z = 0$ and half to $z = 1$. In any condition, the baseline group ($z = 0$) is chosen to be distributed as Weibull(1,1) (or exponential(1)), and the other group to be distributed as Weibull$(\exp(\beta), \exp(\phi))$. Consider the situations when the cumulative hazards, and thus the survival functions, of these two groups cross at some point t_c such that Prob$(T^*_0 \geq t_c)$=Prob$(T^*_1 \geq t_c) = 1 - \overline{s}, 0 < \overline{s} < 1$. When $\overline{s}$ takes a value close to 0 or 1, this means that the cumulative hazards cross at an early or late stage of observations, respectively. In our simulations, the parameter configurations of β and ϕ are: $\beta = 0, \log 2$, and $-\log 2$, $\phi = 0, \log 2$, and $-\log 2$. Because the HHR model only permits *monotonic* hazards ratios in time between two groups (this can be verified from (2) by taking $X = Z$ and $Z = z_1$ versus $Z = z_0$ as stated in Wu et al. (2002)), the case of hazards-crossing associated with hump-shaped or bathtub-shaped hazards ratios is not considered in this paper. The sample size is 100 for each study. The censoring mechanism, C_g, is chosen to produce 25% failures censored in the following way. Let C_g be distributed as G=Weibull(a^*, b); that is, the shape parameter b is chosen to be the same for both T_z and C_g to simplify the situation. It is then easy to compute from $\int G(u)dF(u) = 0.25$ that $a^* = a/3$. Nonetheless, the results may depend on the distribution of C_g.

To check the behavior of $\mathcal{W}$, compared with other tests under $\mathcal{H}_0$, Table 1 gives empirical upper-tailed probabilities ($\hat{p}$) according to 1000 simulations with the true $p = 0.01$, 0.05, and 0.1. The upper-tailed probability is defined as follows. If a statistic (say, $\mathcal{T}$) has the sampling distribution χ^2_ν, when Prob($\chi^2_\nu > q$)=p, the tailed probability of $\mathcal{T}$ in these 1000 simulations is

$$\hat{p} = (1/1000) \sum_{i=1}^{1000} 1(\mathcal{T}_i > q),$$

where $\mathcal{T}_i$ is the ith realization of $\mathcal{T}$. In our study, the Renyi-type statistic also has a weight process $\mathcal{K}(\cdot)$ which needed to be chosen. We use the same weight as $G^{1,1}$ because it has the best performance in almost all situations. Table 1 shows that, under $\mathcal{H}_0$, the empirical distributions of the proposed statistic $\mathcal{W}$ and the weighted logrank statistics $G^{\rho,\gamma}$ and $\mathcal{M}$ all have satisfactory tail behavior in both of the 0%- and 25%-censored cases. However, the $\mathcal{R}$- and $\mathcal{B}$-tests have larger type I errors than the nominal level. Under $\mathcal{H}_a$, rejection proportions are reported in Table 2.

[Tables 1 and 2 about here.]

In Table 2, the ordinary logrank test ($G^{0,0}$) and the Breslow's test ($\mathcal{B}$) have the best performance in power when $\phi = 0$, which corresponds to the proportional hazards model; Moreau's LMP test $\mathcal{M}$ has the best performance when $\beta = 0$. However, even in these two cases, the statistical power of the proposed test $\mathcal{W}$ still is comparable with those of the former two. In other cases, $\mathcal{W}$ has greater power for both cases of 0%- and 25%-censoring. If there is early crossing, e.g. when $\overline{s} = 0.221$ or $(\beta, \phi) = (-\log 2, -\log 2)$, the $G^{1,0}$-test has the lowest power since it puts greater weight on early-stage observations. For late crossings when $\overline{s} = 0.982$ and 0.865, the powers of $G^{1,1}$- and $G^{0,1}$-tests are both poor because weights are leaned on in the middle and late stages. However, the results of $\overline{s} = 0.632$ and 0.393 have to be interpreted with more care. At first glance, they may be considered as crossing near the middle stage, if 'middle' is recognized as representing the common survival of 0.5. The former case, $\overline{s} = 0.632$, includes two combinations: $(\beta, \phi) = (0, \log 2)$ and $(0, -\log 2)$. In this case, $G^{1,0}$ has the lowest power for the '0%-censoring' (rejection propabilities of 0.124 and 0.135) and $G^{0,0}$ has the lowest power for the '25%-censoring' (rejection propabilities of 0.047 and 0.065). For the latter case, $\overline{s} = 0.393$, which corresponds to $(\beta, \phi) = (\log 2, \log 2)$, $G^{1,0}$ again has the lowest power. We learn from these results that *the interpretation of 'early', 'middle', or 'late' crossings are not dogmatic.* Rather, it depends on the interrelation between the groups, including the value of $\overline{s}$ (1 minus the common survival), the curvature of (estimated) survivor functions, and even the censoring mechanism. In some situations, for example, censoring can delay the point of intersection of hazards. So, a test which puts weight on a late-stage may have good or moderate power for a middle-stage crossing in the case of 0%-censoring but a much smaller power in the case of 25%-censoring. The two cases of $\overline{s} = 0.632$ illustrate this situation. Finally, Moreau's $\mathcal{M}$ is not sensitive enough to reject the null hypothesis when there is only 'location-shift' but no 'scale-change', i.e., when $\beta \neq 0$ and $\phi = 0$. In general, the power of $\mathcal{W}$ is superior to the weighted logrank tests, $G^{\rho,\gamma}$ and $\mathcal{M}$, as well as to the $\mathcal{R}$- and $\mathcal{B}$-tests, revealing the benefits gained by considering the current semiparametric model. If the nonproportionality cannot be modelled by the HHR model, on the other hand, nonparametric tests are still recommended.

4.2 Actual data analysis

In this section, the data listed in Piantadosi (1997, Chapter 19, pages 483-488) concerning the survival times of lung cancer patients and those analyzed in Stablein and Koutrouvelis (1985) and Hsieh (2001) concerning a set of gastric carcinoma patients are used to illustrate the implementation of model (1) and the $\mathcal{W}$-statistic in the case of a two-sample problem, compared with other statistics. For both datasets, *event time* is defined as the 'survival' time of cancer patients. It is crucial that model (1) is appropriate in order that the $\mathcal{W}$-statistic can be applied to test for differences. In the literature concerning the proportional hazards model, checking model adequacy can be accomplished by 'omnibus' tests, including those of Schoenfeld

(1980), Wei (1984), Gill and Schumacher (1987), and Lin (1991). Regarding the HHR model (1), a chi-square goodness-of-fit test is suggested in Hsieh (2001) and Wu et al. (2002) for a similar purpose, based on the OEE approach. As mentioned in a previous context, however, we suggest that data practitioners diagnose the fits of model (1) by simply plotting the estimated survivor curves of different groups ($\hat{S}_{HH}$) according to distinct covariate values and comparing them with the Kaplan-Meier (K-M) estimates ($\hat{S}_{KM}$). (See Fig. 1 for lung cancer data and refer to Hsieh (2001) for gastric cancer data.) The rationale is that when the model is adequate, and consistent estimators $\hat{\beta}$, $\hat{\phi}$, and $\hat{\Lambda}_0$ have been solved from §2, the estimated survivals based on the *semiparametric* HHR model,

$$\hat{S}_{HH}(t;z) = \exp[-\{\hat{\Lambda}_0(t)\}^{\exp(\hat{\phi}'z)} \exp(\hat{\beta}'z)], \tag{15}$$

must be close to the *nonparametric* K-M estimates. Table 3 gives the results of the tests after estimation for the datasets mentioned above.

For gastric cancer data, 90 patients were randomized into two groups, each containing 45 individuals receiving chemotherapy and chemo- plus radiotherapy, respectively (Stablein and Koutrouvelis, 1985). There is a cross at around 2.7 years between the two groups. At that place, the survivor estimate for both groups is around 0.2 ($\overline{s} = 0.8$), which can be interpreted as a late or middle stage. The p-values of the $G^{1,1}$- and $G^{0,1}$-tests are 0.902 and 0.150, respectively, for they put weights on middle and late stages, respectively. The $G^{1,0}$-test, as well as Moreau's test $\mathcal{M}$ and the proposed test $\mathcal{W}$, give significant results of testing for the group difference at the 0.05 α-level. Neither results of $\mathcal{R}$- or $\mathcal{B}$-tests are significant. However, Moreau's test is the most significant due to the fact that if we apply the HHR model, the parameter estimates of (β, ϕ) are $(\hat{\beta}, \hat{\phi}) = (0.3251, -0.7933)$, with corresponding p-values of 0.3267 and 0.0559. The estimated heteroscedasticity parameter $\hat{\phi}$ is nearly significant, but the estimate $\hat{\beta}$ is not. This is a situation when Moreau's LMP test, $\mathcal{M}$, performs the best. However, the proposed test $\mathcal{W}$ also performs well in this case. Furthermore, if the estimated $\hat{\beta}$ becomes more significant, not necessarily at the customary 0.05 level, the $\mathcal{W}$-test will be more powerful, as the following dataset shows.

[Table 3 about here.]

The situation is more complicated for lung cancer data. There were 164 patients divided in two groups who received radiotherapy (sample size of 86) or radiotherapy plus 'CAP' (sample size of 78). From Fig. 1, there are two places at which nonproportionality is apparent. First, the Kaplan-Meier estimates *tend to cross* at a time of around 7 months. Second, they *cross* at around 33 months, and at the intersection, the common survival of these two groups is 0.26 ($\overline{s} = 0.74$), which also can be recognized as a late- or middle-stage crossing. For this sort of data, seeking a model which accounts for multiple crossings is a forthcoming effort. In the current analysis, however, applying the HHR model can *only* model the dataset as if it has a *single crossing*, which appears near the second place. We observe that, for most places, the fit of $\hat{S}_{HH}$ is fine, except for those near the early stage in the radiotherapy-plus-CAP group. In our analysis, it is expected that $G^{0,1}$, which places weight on the late stage, is poor at detecting the difference (p-value of 0.994) and $G^{1,1}$, which puts weight on middle stage, is also not capable of reaching this goal (p-value of 0.555). Contrastingly, the $G^{1,0}$-test has a p-value of 0.075; Moreau's $\mathcal{M}$ and the proposed $\mathcal{W}$ state a more-significant difference between the two groups than the $G^{\rho,\gamma}$-statistic. Likewise, neither the $\mathcal{R}$- nor the $\mathcal{B}$-tests give significant results. The estimated parameters $(\hat{\beta}, \hat{\phi}) = (0.2781, -0.4914)$ have p-values of 0.181 and 0.085, which can be used to compare $\mathcal{W}$ and $\mathcal{M}$. We can see from this example that the estimate $\hat{\beta}$ becomes more significant than the former dataset. The result can be interpreted to mean that the $\mathcal{W}$-test has a smaller p-value (0.041) than Moreau's $\mathcal{M}$-test (0.046). In conclusion, analyses of these two datasets, along with the simulations, reveal benefits of modeling nonproportionality, **if possible**, rather than choosing different weights for a nonparametric class.

[Fig. 1 about here.]

5 Discussion

In model (1), there are two components which are expressed as linear combinations of two sets of covariates: the component $\exp(\beta' Z)$ appears as the risk function as in Cox's model and $\exp(\phi' X)$, referred to as the heteroscedasticity component, appears as the power of the baseline cumulative hazard. Generally, $\exp(\phi' X)$ determines the 'shape' of cumulative hazards. If the heterogeneity (modelled by the heteroscedasticity component) is not accounted for, the estimate of effect-measure associated with Z is biased. This is similar to the arguments given by Gail, Wieand, and Piantadosi (1984) which dealt with the case of omitting an important explanatory variable in a *nonlinear* regression. If the true model is the current HHR model, analytical computation of bias when it is misspecified as the PH model can be obtained from mimicking the calculations in the paper by Gail et al. (1984) along with those in Tsiatis's (1981) work. Here, relaxing the simple disposition that $Z = X$, model (1) allows the vector X to share common components with Z, and the common part is the 'mechanism' which results in *heterogeneity*. As an illustration, let $X = (X_1', Z_1')'$ and $Z = (Z_1', Z_2')'$. Omitting X_1 and/or Z_1 in the heteroscedasticity component can induce bias, even if X_1 is independent of Z_1 and Z_2. If there is hazards crossing when analyzing actual life-time data, positive and negative differences may balance each other out *before* and *after* the place where the cross-effect occurs (Marubini and Valsecchi, 1995; Chapter 4). Given this concern, logrank or weighted logrank statistics are not globally capable of capturing the genuine difference. Moreover, the locally most powerful test constructed in Moreau et al. (1992) is only sensitive for detecting differences in 'shape'. With the HHR model, however, differences between groups appearing as a location shift or scale change can be captured by the system of score-type estimating functions (6) and (7). An important step is to provide a formal test other than visual diagnostics. In this regard, Hsieh's (2001) test is an *omnibus* one and can be viewed as a by-product of his OEE approach. In contrast, it is also worth noting that the tests of Quantin et al. (1996), Bagdonavičius, Hafdi, and Nikulin (2002), and Devarajan and Ibrahimi (2002) are only designed to test for validity of the PH model ($\mathcal{H}_0$) which is nested in the class of HHR models ($\mathcal{H}_a$).

With Hsieh's test of model validity, a smoothing factor m (the number of cutoff segments) needs to be decided on according to the order $m = O(n^{1/3})$, where n is the total sample size. In the procedure of data analysis, the selected cutoff points should be **adapted** in some cases to the 'pattern' of estimated survivor functions. For example, the cutoff points for lung cancer data can either be chosen as the $\{20, 40, 60, 80\}$-th percentiles or the $\{4, 8, 18, 65\}$-th percentiles. The former gives an equal-number partition which results in a p-value of 0.052 for the $\mathcal{W}$-test; the latter was reported in the last section. As a result, the second set of cutoff points gives a better fit for both groups using the HHR model. Thus far, a unified rule is still lacking for the choice of cutoff points and/or the value of m in order for an *optimal* result to be obtained when the total sample size n is moderate or less.

ACKNOWLEDGEMENT

This research was partly supported by grant NSC 91-2118-M-039-001 from the National Science Council of Taiwan.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes.* New York: Springer-Verlag.

Bagdonavičius, V., Hafdi, M. A., and Nikulin, M. (2002). The generalized proportional hazards model and its application for statistical analysis of the Hsieh model. *The Second Euro-Japanese Workshop on Stochastic Modelling for Finance, Insurance, Production and Reliability*, Chamonix, France.

Breslow, N.E., Edler, L., and Berger, J. (1984). A two-sample censored-data rank test for acceleration. *Biometrics* **40**, 1049-1062.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835-845.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association* **92**, 227-235.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1-23.

Devarajan, K. and Ebrahimi, N. (2002). Goodness-of-fit testing for the Cox proportional hazards model. *Goodness-of-Fit Tests and Model Validity: Chapter 18* (Edited by Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., and Mesbah, M.) Boston: Birkhäuser.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis.* New York: Wiley.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. *Biometrika* **71**, 431-444.

Gill, R. D. (1980). *Censoring and stochastic integrals.* Amsterdam: Mathematisch Centrum.

Gill, R. D. and Schumacher, M. (1987) A simple test of the proportional hazards assumption. *Biometrika* **74**, 289-300.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimation and the von Mises method–I. *Scandinavian Journal of Statistics* **16**, 97-128.

Gill, R. and van der Vaart, A. W. (1993). Non- and semi-parametric maximum likelihood estimation and the von Mises method–II. *Scandinavian Journal of Statistics* **20**, 271-288.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 133-143.

Hsieh, F. (1996). A transformation model for two survival curves: an empirical process approach. *Biometrika* **83**, 519-528.

Hsieh, F. (2001). On heteroscedastic Cox's regression models and its applications. *Journal of the Royal Statistical Society, Series B* **63**, 63-79.

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 165-174.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated Data.* New York: Springer-Verlag.

Kosorok, M. R. and Lin, C.-Y. (1999). The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association* **94**, 320-332.

Lin, D. Y. (1991). Goodness of fit for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association* **86**, 725-728.

Marubini, E. and Valsecchi, M. G. (1995). *Analysing survival data from clinical trials and observational studies.* Chichester: Wiley.

Moreau, T., Maccario, J., Lellouch, J., and Huber, C. (1992). Weighted log rank statistics for comparing two distributions. *Biometrika* **79**, 195-198.

Piantadosi, S. (1997). *Clinical trials: a methodologic perspective.* New York: Wiley.

Quantin, C., Moreau, T., Asselain, B., Maccario, T., and Lellouch, J. (1996). A regression survival model for testing the proportional hazards hypothesis. *Biometrics* **52**, 874-885.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145-153.

Stablein, D. M. and Koutrouvelis, I. A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics* **41**, 643-652.

Tsiatis, A. A. (1981) A large sample study of Cox's regression model. *The Annals of Statistics* **9**, 93-108.

Wei, L. J. (1984). Testing goodness-of-fit for the proportional hazards model with censored observations. *Journal of the American Statistical Association* **79**, 649-652.

Wu, H.-D. I, Hsieh, F., and Chen, C.-H. (2002). Validation of a heteroscedastic hazards regression model. *Lifetime Data Analysis* **8**, 21-34.

Table 1: Upper-tailed probabilities in 1000 replicates of Fleming-Harrington's $G^{\rho,\gamma}$-statistics, a Renyi-type test $\mathcal{R}$, Breslow's acceleration test $\mathcal{B}$, Moreau's $\mathcal{M}$-statistic, and the proposed partial-score test $\mathcal{W}$. The value p is the specified probability of the theoretical sampling distribution at which the tailed behavior of each test's statistics is to be investigated. The $G^{\rho,\gamma}$- and $\mathcal{M}$-statistics both have a χ_1^2 sampling distribution; Breslow's $\mathcal{B}$ and the proposed partial-score statistic $\mathcal{W}$ are distributed as χ_2^2; whereas the distribution of $\mathcal{R}$-statistic is approximated by the supremum of standard Brownian motion. In each simulation, we chose the censoring distribution so that 0% and 25% of the sample was censored.

0% censored								
p	$G^{0,0}$	$G^{1,0}$	$G^{1,1}$	$G^{0,1}$	Renyi	Breslow	Moreau	Proposed
					$\mathcal{R}$	$\mathcal{B}$	$\mathcal{M}$	$\mathcal{W}$
0.100	0.106	0.104	0.094	0.100	0.128	0.130	0.098	0.106
0.050	0.047	0.049	0.050	0.057	0.071	0.076	0.059	0.052
0.010	0.005	0.012	0.006	0.013	0.017	0.022	0.011	0.008
25% censored								
0.100	0.095	0.083	0.095	0.119	0.168	0.131	0.114	0.105
0.050	0.050	0.042	0.045	0.057	0.105	0.081	0.066	0.049
0.010	0.012	0.008	0.011	0.012	0.041	0.029	0.011	0.008

Table 2: Empirical rejection probability of Fleming-Harrington's $G^{\rho,\gamma}$-statistics, a Renyi-type test $\mathcal{R}$, Breslow's acceleration test $\mathcal{B}$, Moreau's $\mathcal{M}$-statistic, and the proposed partial-score statistic ($\mathcal{W}$) under various β-ϕ configurations with the HHR model; $\overline{s}$ is the 'probability' at which the survivor functions of the two groups cross ($\overline{s}$ means '1-survival'). When $\phi = 0$, there are no crossings; we denote this case by the symbol '$-$'. There are 1000 replications in cases of 0% and 25% censored samples.

β	0			log2			$-\log 2$		
ϕ	0	log2	$-\log 2$	0	log2	$-\log 2$	0	log2	$-\log 2$
$\overline{s}$	$-$	0.632	0.632	$-$	0.393	0.982	$-$	0.865	0.221
0% censored									
$G^{0,0}$	0.047	0.139	0.138	0.922	0.824	0.893	0.934	0.228	0.990
$G^{1,0}$	0.049	0.124	0.135	0.821	0.201	0.996	0.860	0.791	0.802
$G^{1,1}$	0.050	0.354	0.347	0.884	0.927	0.763	0.878	0.108	0.995
$G^{0,1}$	0.057	0.811	0.798	0.844	0.997	0.256	0.846	0.162	1.000
Renyi-type ($\mathcal{R}$)	0.071	0.352	0.364	0.894	0.920	0.930	0.889	0.410	0.995
Breslow ($\mathcal{B}$)	0.076	0.768	0.761	0.938	0.992	0.842	0.935	0.344	1.000
Moreau ($\mathcal{M}$)	0.059	0.988	0.986	0.101	0.994	0.686	0.082	0.915	0.969
Proposed test ($\mathcal{W}$)	0.052	0.972	0.978	0.858	0.999	0.998	0.876	0.976	1.000
25% censored									
$G^{0,0}$	0.050	0.047	0.065	0.829	0.413	0.965	0.831	0.421	0.904
$G^{1,0}$	0.042	0.189	0.203	0.769	0.086	0.997	0.746	0.824	0.622
$G^{1,1}$	0.045	0.163	0.189	0.751	0.746	0.785	0.771	0.162	0.979
$G^{0,1}$	0.057	0.391	0.440	0.691	0.913	0.396	0.720	0.087	0.991
Renyi-type ($\mathcal{R}$)	0.105	0.283	0.332	0.842	0.837	0.961	0.881	0.547	0.983
Breslow ($\mathcal{B}$)	0.081	0.426	0.480	0.839	0.883	0.910	0.834	0.387	0.992
Moreau's $\mathcal{M}$	0.066	0.900	0.891	0.094	0.944	0.585	0.102	0.810	0.937
Proposed $\mathcal{W}$	0.049	0.883	0.878	0.754	0.946	0.992	0.738	0.953	0.993

Table 3: Analyses of gastric cancer data and lung cancer data with the weighted logrank tests of Fleming-Harrington's $G^{\rho,\gamma}$- and Moreau's $\mathcal{M}$-statistics, a Renyi-type test $\mathcal{R}$, Breslow's acceleration test $\mathcal{B}$, and the proposed test $\mathcal{W}$.

Test statistic	$G^{0,0}$	$G^{1,0}$	$G^{1,1}$	$G^{0,1}$	Renyi $\mathcal{R}$	Breslow $\mathcal{B}$	Moreau $\mathcal{M}$	Proposed $\mathcal{W}$
Gastric cancer								
Realization	0.222	3.963	0.015	2.071	1.857	3.435	9.054	7.941
(p-value)	(0.637)	(0.046)	(0.902)	(0.150)	(0.127)	(0.179)	(0.003)	(0.019)
Lung cancer								
Realization	1.275	3.177	0.349	0.001	1.255	1.458	3.976	6.394
(p-value)	(0.259)	(0.075)	(0.555)	(0.994)	(0.419)	(0.482)	(0.046)	(0.041)

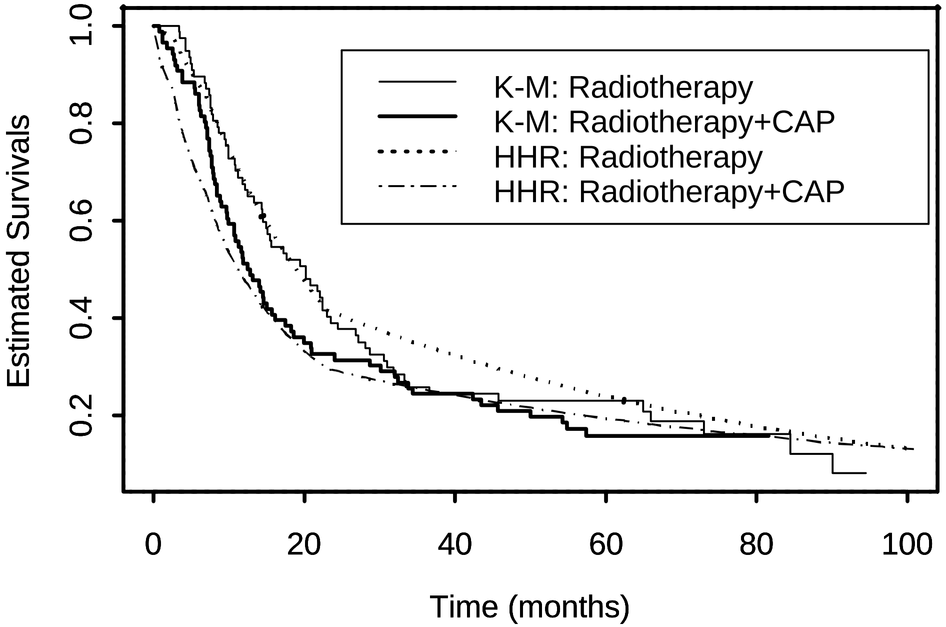

Figure 1: Comparison between Kaplan-Meier estimates and those of survivor functions obtained from the HHR model for lung cancer data.