Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 93

A New Class of Rank Tests for Interval-censored Data

Guadalupe Gomez[*] Ramon Oller Pique[†]

[*]Harvard School of Public Health, ggomez@hsph.harvard.edu
[†]Matematica i Informatica de la Universitat de Vic, ramon.oller@uvic.cat

A new class of rank tests for interval–censored data

By GUADALUPE GÓMEZ

Visiting Scientist, Department of Biostatistics, Harvard School of Public Health

Permanent affiliation: Departament d'Estadística i Investigació Operativa de la Universitat Politècnica de Catalunya

ggomez@hsph.harvard.edu

and RAMON OLLER PIQUÉ

Departament d'Economia, Matemàtica i Informàtica de la Universitat de Vic

ramon.oller@uvic.cat

Abstract

The class of weighted logrank tests proposed by Fleming & Harrington (1991) has been widely used in survival analysis and is nowadays, unquestionably, the established method to compare, nonparametrically, k different survival functions based on right–censored survival data. This paper presents a new general class of rank based tests for interval–censored data, $G^{\rho,\lambda}$, which plays an analogous role to the Fleming & Harrington class for right–censored data. We first derive the class of tests G^{ρ} which is shown to correspond to the efficient score test in a general regression model if data are discrete. We consider a Fisher information and a permutation approach to derive an asymptotic normal distribution. Then, we develop a weighted log–rank form of the test statistics, which makes explicit the analogy between our proposal and the original class for right–censored data and makes easier the understanding of the power behavior of the family. Finally, the class $G^{\rho,\lambda}$ is obtained as a natural extension of the G^{ρ} family.

Some key words: interval–censored data; treatment comparison; weighted log–rank test; permutation test.

1 Introduction

Comparison of two or more samples when data are censored is one of the topics which arises in most survival studies. The complexity of the censoring mechanism has determined the development of new survival statistical methods. While many tests have been proposed when data are right–censored, research for interval–censored data is still ongoing and lacks of an unified approach. Interval censoring often arises when individuals are inspected intermittently and the event of interest is only known to have occurred between two consecutive inspection times. Mantel (1967) and Peto and Peto (1972) were the first authors to propose testing methods for interval–censored data. These authors extend the Wilcoxon test and the log–rank test to interval–censored data and use a permutation approach to avoid the difficulty of finding the distribution of the corresponding test statistics. Finkelstein (1986) derives the log–rank test as a score statistic of a proportional hazards model. Finkelstein assumes grouped continuous data and uses the Fisher information matrix to obtain the asymptotic distribution of the test statistic instead of the permutation distribution. There is a large literature on interval–censored data associated to the extension of the Wilcoxon and log–rank tests, see for instance Fay (1996, 1999), Fay and Shih (1998), Sun (1996), Zhao and Sun (2004), Sun *et al.* (2005) and Huang *et al.* (2008). Other authors such as Petroni and Wolfe (1994), Fang *et al.* (2002) and Lim and Sun (2003) discuss generalizations of the weighted Kaplan–Meier class developed by Pepe and Fleming (1989) for right–censored data. An extensive review of k–sample methods for interval–censored data can be found in Gómez *et al.* (2004) and in Sun (2006).

A large number of k–sample methods have been proposed for right–censored data. A useful family of test statistics is the so–called class of weighted log–rank statistics and, in particular, the $G^{\rho,\lambda}$ subfamily introduced by Fleming and Harrington (1991). For this subfamily, the weight function is chosen to be the product between the Kaplan–Meier estimate of the survival function raised to power ρ and its complementary (the cumulative distribution function) raised to λ. The appropriate selection of the parameters ρ and λ gives emphasis to early, middle or late hazard differences. The $G^{\rho,\lambda}$ family contains as special cases the log-rank statistic ($\rho = 0$ and $\lambda = 0$) something as well as an statistic close to the Peto–Prentice extension of the Wilcoxon statistic ($\rho = 1$ and $\lambda = 0$). Moreover, when $\lambda = 0$, the corresponding subfamily is called the G^{ρ} family.

In this paper we propose an extension, for interval–censored data, of the $G^{\rho,\lambda}$ family. Section 2 formulates the problem and gives the basic notation. A gradual and intuitive presentation of our proposal is introduced from Section 3 to Section 6. Section 3 derives the subclass G^{ρ} as a likelihood score procedure under discrete data. Section 4 describes the permutation approach we apply for the test statistics as an alternative to the likelihood approach. Section 5 shows the equivalence between a general weighted log–rank form of the test statistics and the linear form given in the previous sections. Finally, the class $G^{\rho,\lambda}$ is obtained as a natural extension of the G^{ρ} family in Section 6. The paper continues with Section 7 where we report a simulation study which has three main goals: first, it gives guidance on the behavior of the $G^{\rho,\lambda}$ family of tests and shows their good behavior, second, by simulating 4 different configurations for the hazards, the weight function is easily interpreted, and third it compares our proposal to a different extension for the $G^{\rho,\lambda}$ family given in Sun *et al.* (2005). In Section 8 we apply the new family of tests to a real data set from an AIDS study. Section 9 concludes the paper.

2 Notation

Let T be the time to the event of interest. Assume that we have k groups of data, $G_1, \ldots, G_k$ with respective sample sizes $N_1, \ldots, N_k$ and define $S_1, \ldots, S_k$ the survival functions of T under each one of these groups. Our goal is to test the hypothesis $H_0 : S_1 = \cdots = S_k$ versus $H_a : S_j \neq S_{j'}$ for some $j \neq j'$. If data are interval–censored, the only information about the lifetime T is that it lies between two observed times, namely L and R, and we write $T \in (L, R]$. In this paper we consider that the observed intervals are half open intervals. The methods we describe below are, however, easily modifiable if we would observe closed intervals. The use of closed intervals would have the advantage that the uncensored observations would be included when $L = R$ and would accommodate grouped data. However, the use of half open intervals is more common and appears in situations where the individuals are inspected intermittently.

Under the assumption that the censoring process is noninformative, see Oller *et al.* (2004, 2007), the likelihood function for the pooled sample simplifies as follows:

$$Lik(S) = \prod_{i=1}^{n} \left\{ S(l_i) - S(r_i) \right\}, \tag{1}$$

where $n = N_1 + \cdots + N_k$ and $(l_1, r_1), \ldots, (l_n, r_n)$ are independent observations. As noted by

Peto (1973) and Turnbull (1976), the nonparametric maximum likelihood estimator (NPMLE) $\hat{S}(t)$ might not be unique. From the sets $\mathcal{L} = \{l_i, 1 \leq i \leq n\}$ and $\mathcal{R} = \{r_i, 1 \leq i \leq n\}$ we can derive all the distinct closed intervals whose left and right end-points lie in the sets $\mathcal{L}$ and $\mathcal{R}$ respectively and which contain no other members of $\mathcal{L}$ or $\mathcal{R}$ other than at their left and right endpoints respectively. Let these intervals, known as Turnbull's intervals, be written in increasing order as $(q_1, p_1], (q_2, p_2], \ldots, (q_m, p_m]$ with $q_j < p_j \leq q_{j+1}$. Then, $\hat{S}(t)$ is unspecified in each $(q_j, p_j]$ and is well defined and flat between these intervals.

Fay and Shih (1998) define an estimate of the survival function for the i^{th} individual, $\hat{S}^i(t)$, as Turnbull's overall survival $\hat{S}(t)$ truncated at the i^{th} observed interval. That is,

$$\hat{S}^i(t) = P_{\hat{S}}((t, +\infty) \mid (l_i, r_i]) = \frac{\hat{S}(l_i \vee t) - \hat{S}(r_i \vee t)}{\hat{S}(l_i) - \hat{S}(r_i)}$$

where $P_{\hat{S}}$ denotes the probability measure of T given by the survival function $\hat{S}(t)$ and $a \vee b$ stands for the minimum between a and b. These individual estimators have two relevant properties: 1) $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}^i(t)$, and 2) $\hat{S}^i(t) = \mathbb{1}_{\{t<t_i\}}$ when the i^{th} observation is not censored and $T = t_i$. $\hat{S}(t)$ and $\hat{S}^i(t)$ play an important role in the test statistics we introduce next.

3 The rank class G^ρ: Special case for discrete data

Let $\mathbf{z}_i$ correspond to a covariate k–vector of group indicators associated with the i^{th} observation, that is, $\mathbf{z}_i = \left(\alpha_i^{(1)}, \alpha_i^{(2)}, \ldots, \alpha_i^{(k)}\right)'$ and $\alpha_i^{(j)}$ is an indicator function that is equal to 1 if the individual belongs to group G_j and 0 otherwise. To test for differences among the k survival functions $S_1, \ldots, S_k$ we propose a class of test statistics of the form

$$\boldsymbol{U} = \sum_{i=1}^{n} \mathbf{z}_i c_i, \tag{2}$$

where c_i is a score value associated to each individual such that

$$c_i = \frac{(\hat{S}(l_i))^{\rho+1} - (\hat{S}(r_i))^{\rho+1}}{\rho\,(\hat{S}(l_i) - \hat{S}(r_i))} - \frac{1}{\rho}\,, \qquad \rho \geq 0 \tag{3}$$

and $\hat{S}(t)$ is Turnbull's overall survival function. This family extends, for interval–censored data, the G^ρ family given in Fleming and Harrington (1991) and it is a class of efficient score statistics in a linear transformation model when interval–censored data are discrete, as it is shown in the next Theorem. Furthermore, the statistic $\boldsymbol{U}$ reduces to the log-rank and Wilcoxon–Peto test statistics originally proposed in Peto and Peto (1972) for choices of $\rho \to 0$ and $\rho = 1$, respectively. Finkelstein (1986) and (Fay, 1996; Fay, 1999) have already shown that under discrete or grouped continuous interval–censored data, the log–rank and the Wilcoxon–Peto tests introduced by Peto and Peto (1972) could be derived as the efficient score statistics for a proportional hazards model and for a proportional odds model, respectively. Several other score statistics can be derived, in this setup, from the linear transformation model studied in Fay (1996). As a matter of fact we have used the results in Fay (1996) to extend the G^ρ family of tests.

Theorem Let $g(T_i) = -\mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i$ be a linear transformation model with g being an increasing function and ϵ_i having survival function $S_\epsilon(t) = [1 + \rho\exp(t)]^{-\frac{1}{\rho}}$. Assume that the censoring mechanism is independent of the covariates $\mathbf{z}_i$ and that the support for the observable data is finite, that is, $L, R \in \{t_0, t_1, \ldots, t_m\}$ where $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = +\infty$. Consider $S(t \,|\, \mathbf{z}_i) = P(T > t \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_j)_{j=1}^m$ is a vector of nuisance parameters such that $\theta_j = g(t_j)$. Consider the likelihood function $Lik(S)$ given in (1). Then,

(a) The maximum likelihood estimator of the nuisance parameters when $\boldsymbol{\beta} = \mathbf{0}$ gives the nonparametric maximum likelihood estimator of the survival function of the pooled sample, that is, $\hat{S}(t) = \left[P(T > t \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta})\right]_{\beta=0,\theta=\hat{\theta}}$.

Moreover, if none of the parameters is on boundary of the parameter space, that is, $1 > \hat{S}(t_1) > \cdots > \hat{S}(t_{m-1}) > 0$, then:

(b) The efficient score statistic $\boldsymbol{U} = \left[\dfrac{\partial \log(Lik(S))}{\partial \boldsymbol{\beta}}\right]_{\beta=0,\theta=\hat{\theta}}$ is given by (2) and (3).

(c) The likelihood based variance–covariance matrix V of the efficient score statistic for $\boldsymbol{\beta}$ can be explicitly given and it is postponed to the Appendix.

The proof of this result is omitted because it follows from standard statistical theory and it is analogous to Fay (1996, 1999). We remark, however, that for a general error survival function $S_\epsilon(t)$, Fay (1996) shows that the efficient score statistic for a linear transformation model is given by (2) and

$$c_i = \frac{S'_\epsilon(S_\epsilon^{-1}(\hat{S}(l_i))) - S'_\epsilon(S_\epsilon^{-1}(\hat{S}(r_i)))}{\hat{S}(l_i) - \hat{S}(r_i)} \tag{4}$$

where $S'_\epsilon(t)$ and $S_\epsilon^{-1}(t)$ are respectively the first derivative and the inverse function of $S_\epsilon(t)$. Statement (b) follows straightforwardly from (4) when we consider the survival function $S_\epsilon(t) = [1 + \rho \exp(t)]^{-\frac{1}{\rho}}$. We will use again equation (4) in the next sections of this paper.

In the formulation of the model given in the Theorem, it could seem more intuitive to use the parametrization $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$. However, we note that this model includes the well–known proportional hazards model (when $\rho \to 0$) and the proportional odds model (when $\rho = 1$), and the usual formulation of these models in terms of hazard functions takes the parametrization $\boldsymbol{\beta}$.

The asymptotic behavior of $\boldsymbol{U}$ follows standard theory for score statistics. Under the null hypothesis, the random variable $\boldsymbol{U}V^-\boldsymbol{U}'$ is asymptotically chi–squared with $k-1$ degrees of freedom, where V^- is the generalized inverse of the observed Fisher's information V. In practice, however, score tests cannot be applied with interval–censored data because the parameter estimates come near to the parameter boundary. To avoid this problem, it is common to use a permutation approach. Next section contains the main aspects of this approach.

4 Permutation distribution

In comparison with the likelihood method given above, the permutation approach is straightforward and it applies for discrete as well as for continuous data. A permutation test remains valid even if the assumed model does not hold. In this case, however, the test might not necessarily be asymptotically efficient. The main assumption to apply a permutation test is that the underlying censoring process have to be identical across groups. This restriction, though it is important, it is also necessary with other methods, for instance, in the likelihood approach given in the previous section, in the multiple imputation methods proposed in Zhao and Sun (2004) and Huang *et al.* (2008) or in the asymptotic method proposed in Sun *et al.* (2005).

The permutation approach can be applied easily to the linear form of the score statistic $\boldsymbol{U}$ given by (2). The idea behind the permutation approach is that if the null hypothesis is true, the labels on the scores c_i are exchangeable. The permutation distribution of $\boldsymbol{U}$ is then obtained by permuting the labels and recomputing the test statistic for all the possible rearranged labels. The permutation distribution can be computed exactly when the sample size is small. When n is large, a version of the Central Limit theorem for exchangeable random variables can be applied yielding a normal approximation with expectation and variance–covariance matrix given by $E(\boldsymbol{U}) = n\bar{c}\bar{\mathbf{z}}'$ and $V_0 = \frac{1}{n-1}\left(\sum_{i=1}^n c_i^2 - n\bar{c}^2\right)\left(\sum_{i=1}^n(\mathbf{z}_i\mathbf{z}_i' - \bar{\mathbf{z}}\bar{\mathbf{z}}')\right)$ respectively.

In our situation $\bar{c} = 0$ and, consequently, $E(\boldsymbol{U}) = \mathbf{0}$. Moreover, we consider $\mathbf{z}_i$ as a k–vector of group indicator functions, thus the permutation test is based on the Mahalanobis distance $\boldsymbol{U}'V_0^-\boldsymbol{U} = \frac{n-1}{\sum_{i=1}^n c_i^2}\sum_{j=1}^k n_j\bar{c}_{(j)}^2$, where V_0^- is the generalized inverse of V_0 and $\bar{c}_{(j)} = \frac{1}{n_j}\sum_{i=1}^n c_i\alpha_i^{(j)}$. We would reject the null hypothesis for extreme values of $\boldsymbol{U}'V_0^-\boldsymbol{U}$ as compared from a χ^2_{k-1} distribution. In the sequel we consider only the permutation distribution of the statistic $\boldsymbol{U}$.

5 The rank class G^ρ: The weighted log–rank form

In this section we elucidate the analogy between our proposal and the original G^ρ family proposed by Fleming and Harrington for right-censored data. We present our proposal both as a procedure to compare hazard functions between groups and as a class of statistics of the form $\sum w\,(O-E)$, that is, as a weighted summation of the observed minus the expected number of deaths. These two forms are the usual weighted log–rank formulations of the original family for right–censored data.

Throughout this section, for any function $F(t)$ and fixed value t, we denote a function increment as $dF(t) = F(t) - F(t^-)$.

Let us define the following estimator of the survival function for the j^{th} group,

$$\hat{S}^{(j)}(t) = \frac{1}{N_j} \sum_{i=1}^{n} \alpha_i^{(j)} \hat{S}^i(t). \tag{5}$$

Let us consider $d\widehat{H}(t) = \dfrac{-d\hat{S}(t)}{\hat{S}(t^-)}$ and $\widehat{H}^{(j)}(t) = \dfrac{-d\hat{S}^{(j)}(t)}{\hat{S}^{(j)}(t^-)}$ as estimators of the overall hazard function and the hazard function for the j^{th} group, respectively. Then, the G^ρ family derived in Section 3 can be written as a vector $\boldsymbol{U} = \big(U_1, \ldots, U_k\big)'$ where the j^{th} component U_j is expressed as

$$U_j = \int_0^{+\infty} w(t)\; n_{jt}\, \big[d\widehat{H}^{(j)}(t) - d\widehat{H}(t)\big], \tag{6}$$

with $w(t) = \hat{S}(t^-) \frac{(\hat{S}(t^-))^\rho - (\hat{S}(t))^\rho}{\rho(\hat{S}(t^-) - \hat{S}(t))}$ being a weight function and $n_{jt} = n\; \hat{S}^{(j)}(t^-)$ being an estimation of the number of individuals at risk at t for the group j. Alternatively, if we define $n_t = n\, \hat{S}(t^-)$ as the estimated total number of individuals at risk at t, and $d_{jt} = -n\, d\hat{S}^{(j)}(t)$ and $d_t = -n\, d\hat{S}(t)$ as the estimated number at risk, then

$$U_j = \int_0^{+\infty} w(t)\; \big[O_{jt} - E_{jt}\big] = \int_0^{+\infty} w(t)\;\; \left[d_{jt} - \frac{n_{jt}}{n_t} d_t\right].$$

The following proposition gives the equivalence between the above weighted log–rank form of the G^ρ family and the definition given in Section 3.

Proposition 5.1. *A weighted log–rank test statistic* $\boldsymbol{U} = \big(U_1, \ldots, U_k\big)'$ *with components given by (6) holds the following properties:*

(a) *For any weight function* $w(t)$*, the statistic* $\boldsymbol{U}$ *can be represented in the linear form* $\sum_{i=1}^{n} \mathbf{z}_i c_i$ *where the scores are given by*

$$c_i = \int_0^{+\infty} w(t) \left[-d\hat{S}^i(t) + \frac{\hat{S}^i(t^-)}{\hat{S}(t^-)} d\hat{S}(t) \right]. \tag{7}$$

(b) *For the weight function*

$$w(t) = \hat{S}(t^-)\; \frac{\gamma(\hat{S}(t^-)) - \gamma(\hat{S}(t))}{\hat{S}(t^-) - \hat{S}(t)} = \hat{S}(t^-)\; \frac{d\gamma(\hat{S}(t))}{d\hat{S}(t)} \tag{8}$$

where $\gamma(t)$ is a differentiable nondecreasing function ($\gamma(1) = 0$), the weighted log–rank scores (7) simplify to

$$c_i = \frac{\hat{S}(l_i)\gamma(\hat{S}(l_i)) - \hat{S}(r_i)\gamma(\hat{S}(r_i))}{\hat{S}(l_i) - \hat{S}(r_i)}.$$

(c) *The function $\gamma(t) = \frac{S'_\epsilon(S_\epsilon^{-1}(t))}{t}$ gives the general scores (4) derived by Fay (1996) under the linear transformation model.*

(d) *The function $\gamma(t) = \frac{1}{\rho}\,(t^\rho - 1)$ gives the weight function $w(t) = \hat{S}(t^-)\frac{(\hat{S}(t^-))^\rho - (\hat{S}(t))^\rho}{\rho(\hat{S}(t^-) - \hat{S}(t))}$ and the G^ρ scores (3).*

Proof:

We only proof the result given in *(b)*. The remaining results are almost immediate, indeed *(c)* and *(d)* follow straightforwardly from *(b)*.

Replacing the weight function $w(t) = \hat{S}(t^-)\,\frac{d\gamma(\hat{S}(t))}{d\hat{S}(t)}$ in equation (7), gives

$$c_i = \int_0^{+\infty} \frac{d\gamma(\hat{S}(t))}{d\hat{S}(t)} \big[\hat{S}^i(t^-)d\hat{S}(t) - \hat{S}(t^-)d\hat{S}^i(t)\big],$$

which can be equivalently written as

$$c_i = \int_0^{+\infty} \frac{d\gamma(\hat{S}(t))}{d\hat{S}(t)} \big[\hat{S}^i(t)d\hat{S}(t) - \hat{S}(t)d\hat{S}^i(t)\big].$$

Since $\hat{S}^i$ is a truncation of $\hat{S}$ at the observed interval $(l_i, r_i]$, then

$$c_i = \int_0^{l_i} d\gamma(\hat{S}(t)) + \int_{l_i}^{r_i} \frac{\hat{S}(t) - \hat{S}(r_i)}{\hat{S}(l_i) - \hat{S}(r_i)}\, d\gamma(\hat{S}(t)) - \int_{l_i}^{r_i} \frac{\hat{S}(t)}{\hat{S}(l_i) - \hat{S}(r_i)}\, d\gamma(\hat{S}(t))$$

$$= \gamma(\hat{S}(l_i)) - \gamma(1) - \frac{\hat{S}(r_i)\{\gamma(\hat{S}(r_i)) - \gamma(\hat{S}(l_i))\}}{\hat{S}(l_i) - \hat{S}(r_i)} = \frac{\hat{S}(l_i)\gamma(\hat{S}(l_i)) - \hat{S}(r_i)\gamma(\hat{S}(r_i))}{\hat{S}(l_i) - \hat{S}(r_i)}\,.$$

This completes the proof of *(b)*.

□

6 The $G^{\rho,\lambda}$ family of tests

In this section we define the $G^{\rho,\lambda}$ family of tests for interval–censored data as a procedure for testing the hypothesis $H_0 : S_1 = \cdots = S_k$ or, equivalently, $H_0 : H_1 = \cdots = H_k$, where S_j and H_j are the survival and cumulative hazard functions under each group, respectively. We propose a class of test statistics within the weighted log–rank class (6), with a weight function given by (8) and γ function which coincides with an incomplete beta function

$$\gamma(t) = -B(1-t; \lambda+1, \rho) = -\int_0^{1-t} x^{\lambda}\, (1-x)^{\rho-1}\, dx. \tag{9}$$

We define the $G^{\rho,\lambda}$ family as a class of vectors $\boldsymbol{U} = \big(U_1, \ldots, U_k\big)'$ with components $U_j = \int_0^{+\infty} w(t)\; n_{jt}\; \big[d\widehat{H}^{(j)}(t) - d\widehat{H}(t)\big]$ and

$$w(t) = \hat{S}(t^-)\; \frac{B(1-\hat{S}(t); \lambda+1, \rho) - B(1-\hat{S}(t^-); \lambda+1, \rho)}{\hat{S}(t^-) - \hat{S}(t)}\,. \tag{10}$$

Alternatively, a test statistic in the $G^{\rho,\lambda}$ family can be expressed as $\boldsymbol{U} = \sum_{i=1}^n \mathbf{z}_i c_i$ with scores c_i given by

$$c_i = \frac{\hat{S}(r_i) B(1-\hat{S}(r_i); \lambda+1, \rho) - \hat{S}(l_i) B(1-\hat{S}(l_i); \lambda+1, \rho)}{\hat{S}(l_i) - \hat{S}(r_i)}\,. \tag{11}$$

When $\lambda = 0$, this proposal reduces to the G^{ρ} family given in Section 3. Furthermore, it is a natural extension for interval–censored data of the original $G^{\rho,\lambda}$ family. To show this, we note that the weight function given by (8) has the following property

$$\lim_{\hat{S}(t^-) \longrightarrow \hat{S}(t)} w(t) = \hat{S}(t)\; \gamma'(\hat{S}(t))$$

where $\gamma'(t)$ is the first derivative function of $\gamma(t)$. This gives a characterization of the $G^{\rho,\lambda}$ weights (10) equivalent to the weights given by Fleming and Harrington (1991) for right–censored data, that is, the weights are close to $\hat{S}(t)\; \gamma'(\hat{S}(t)) = (\hat{S}(t))^{\rho}(1-\hat{S}(t))^{\lambda}$ when $\hat{S}$ is very nearly continuous at a given point t.

This extension of the $G^{\rho,\lambda}$ family reproduces the interpretation of the weight function in the original family: 1) when $\lambda = 0$, early hazard differences are stronger emphasized as ρ

increases; 2) when $\rho = 0$, late hazard differences are stronger emphasized as λ increases; and 3) when $\rho = \lambda$, hazard differences around the median are stronger emphasized as ρ and λ increase. This behavior is validated in the next simulation section.

Another interesting issue is whether or not the $G^{\rho,\lambda}$ family is a class of score statistics under the linear transformation model. From Proposition 5.1 and the γ function given by (9), it follows that the survival function of the error term S_ϵ should be a solution of the following differential equation,

$$\frac{S'_\epsilon(S_\epsilon^{-1}(t)))}{t} = -B(1-t;\lambda+1,\rho), \tag{12}$$

Except for the special case $\lambda = 0$, we note that the differential equation (12) would not generally have an analytic solution.

7 Simulation study

A large simulation study has been conducted to assess the performance of the $G^{\rho,\lambda}$ family of tests given in Section 6, to validate, in terms of power, the interpretation of the weight function given in the previous section and to compare our proposal to another generalization of the $G^{\rho,\lambda}$ class given in Sun *et al.* (2005).

The censoring mechanism for T has been simulated mimicking a longitudinal study where there is a periodical follow–up with scheduled visits following Schick and Yu's model (2000). Specifically, for an individual i, we consider a set of examination times $\{Y_{ai}, a = 1, \ldots, \tau_i\}$ which are sum of inter–follow–up times, $Y_{ai} = \sum_{b=1}^{a-1} \xi_{bi}$. The inter–follow–up times are independent and identically distributed as an exponential distribution ($E(\xi_{bi}) = \mu$). For each individual, the number of examination times satisfies that $\tau_i = \sup\{a \geq 1 : \sum_{b=1}^{a-1} \xi_{bi} \leq \tau\}$ where τ represents the length of the study. The parameters μ and τ provide a control of the length of the observed intervals and the percentage of right–censored observations, respectively. In the present simulation study, we have considered $\mu = 2$ and $\tau = 20$. The study has been based on 1000 replications and the normal approximation of the permutational distribution has been used.

We have simulated a large number of scenarios where the null hypothesis was true and in all of them the nominal significance level $\alpha = 0.05$ was roughly reached. Even for scenarios with small sample sizes, for instance a scenario of two groups and $N_1 = N_2 = 50$, the percentage of rejection was near to 0.05 . We do not present the results here but they can be provided upon request.

For the validation and the interpretation of the weight function, two groups have been considered with sample sizes $N_1 = N_2 = 50$ and 10 scenarios within an accelerated failure time (AFT) model for T and an error term distribution $S_\epsilon(t)$ holding equation (12). Each scenario has a parametrization (ρ, λ) of the function $S_\epsilon(t)$, a location parameter for each group and a common scale parameter for both groups. The location parameters have been chosen in order to have median equal to 6.5 in the first group and 7.5 in the second. The scale parameters are respectively 0.15, 0.1, 0.075, 0.06, 0.08, 0.04, 0.05, 0.035, 0.0055 and 0.0015. Table 2 gives the empirical powers of the $G^{\rho,\lambda}$ family. The value within the parenthesis gives the power ratio of our proposal to the proposal of Sun *et al.* (2005) and it is discussed at the end of this section.

The results in Table 2 show that the $G^{\rho,\lambda}$ tests have higher power when they coincide with the score statistic of the AFT model, that is, when the parameters (ρ, λ) of the test statistic and the AFT model are identical. Moreover, the power decreases as the parameters (ρ, λ) of the test statistic and the AFT model diverge. For instance, under a proportional hazards model (scenario 1), the power of the test statistics decreases as the parameters ρ or λ increase. Under a proportional odds model (scenario 2), the test statistics have higher power when $\lambda = 0$ and ρ is close to 1 and have lower power when $\rho = 0$ and λ increases. Thus, under an AFT with $S_\epsilon(t)$ holding equation (12) and a continuous censoring mechanism, the simulations show that the score test statistic is efficient, at least, within the class $G^{\rho,\lambda}$.

[TABLE 2 here]

We have simulated other scenarios for two-sample comparisons where the corresponding hazard functions differ at early times, at late times or around the median. T has been simulated as a piecewise constant hazard function, as follows: a set of points $0 = x_0 < x_1 < \cdots < x_b < x_{b+1} = +\infty$ has been fixed and for each group G_j $(j = 1, 2)$, T has been simulated with a hazard function $h_j(t) = h_{j,a}$ when $x_{a-1} \leq t < x_a$ $(a = 1, \ldots, b)$ and such that the median of the pooled sample is 5. We have consider sample sizes $N_1 = N_2 = 50$ and

$N_1 = N_2 = 200$.

For the *Early times situation* where the hazard functions differ at early times we take $b = 1$ and simulate hazard differences until the point x_1 and equal hazards ($h_{1,2} = h_{2,2} = 0.14$) thereafter. In this situation there are early hazard differences when x_1 is clearly smaller than the median, there are early and middle hazard differences when x_1 is close to the median and there are early, middle and late hazard differences when x_1 is clearly bigger than the median. We distinguish three different scenarios *High–Early* for $x_1 = 1.25$, $h_{1,1} = 0.06$ and $h_{2,1} = 0.22$, *Intermediate–Early* for $x_1 = 6.25$, $h_{1,1} = 0.12$ and $h_{2,1} = 0.16$ and *Low–Early* for $x_1 = 11.25$, $h_{1,1} = 0.12$ and $h_{2,1} = 0.16$. Analogously, for the *Late times situation* where the hazard functions do not show differences until x_1 being equal to $h_{1,2} = h_{2,2} = 0.14$ and differ thereafter, we distinguish as well three different scenarios *Low–Late* when $x_1 = 1.25$, $h_{1,2} = 0.11$ and $h_{2,2} = 0.17$, *Intermediate–Late* when $x_1 = 6.25$, $h_{1,2} = 0.10$ and $h_{2,2} = 0.18$ and *High–Late* for $x_1 = 11.25$, $h_{1,2} = 0.10$ and $h_{2,2} = 0.18$. For the *Middle times situation* where the hazard functions only show differences between x_1 and x_2, we have taken $h_{1,1} = h_{2,1} = h_{1,3} = h_{2,3} = 0.14$ and consider two scenarios *Low–Middle* for $x_1 = 1.25$, $x_2 = 8.75$, $h_{1,2} = 0.11$ and $h_{2,2} = 0.17$ and *High–Middle* when $x_1 = 3.75$, $x_2 = 6.25$, $h_{1,2} = 0.06$ and $h_{2,2} = 0.22$.

Table 3, giving the empirical powers and the power ratios of the $G^{\rho,\lambda}$ family, confirm the interpretation of the weight function given in the previous section. When we look at early, late and middle configurations, then the statistics with $\lambda = 0$, $\rho = 0$ and $\rho = \lambda \neq 0$ give the higher powers, respectively. There are two exceptions, the *Low–Early* and *Low–Late* scenarios, where the hazard pattern is less clear. Interestingly, in situations where the hazard pattern is very clear, we observe that the power increases as the suitable combination of the parameters is chosen and the parameters increase. That is the case of the *High–Early*, *High–Late* and *High–Middle* scenarios. For instance, in the scenario *High–Early*, the combination $(\rho, 0)$ increases the power of the test statistic when ρ increases.

[TABLE 3 here]

We have as well simulated observations coming from two groups where the two hazards cross each other. In this situation, and following analogous steps as before, we have considered $b = 1$, $h_{1,1} = 0.12$, $h_{2,1} = 0.16$, $h_{1,2} = 0.16$, $h_{2,2} = 0.12$. Within this we have simulated a

Crossing–Early Scenario with $x_1 = 2.5$ where the two hazards cross before the median, a *Crossing–Middle Scenario* with $x_1 = 5$ where the two hazards cross at the median and a *Crossing–Late Scenario* with $x_1 = 7.5$ and where the two hazards cross after the median.

The results in Table 4 indicate that, as expected, the $G^{\rho,\lambda}$ family behaves poorly when hazards cross. In scenario *Crossing–Early* the best powers are reached when late times are weighted, that is, when $\rho = 0$. On the other hand, in scenario *Crossing–Late* the best powers are reached when the early times are weighted, that is, when $\lambda = 0$.

[TABLE 4 here]

The last part of the simulation study attempts to compare our proposal with another extension of the $G^{\rho,\lambda}$ family given in Sun *et al.* (2005). These authors use a class of tests which has the linear form given in equation (2) and which we can write as a weighted log–rank test (6) with weight function (8) fixed by

$$\gamma(t) = \log(t)t^{\rho}(1-t)^{\lambda}.$$

When $\rho = 0$ and $\lambda = 0$ the test statistic reduces to the log–rank test proposed in Peto and Peto (1972). However, when $\rho \neq 0$ or $\lambda \neq 0$, this family differs from our extension of the $G^{\rho,\lambda}$ family and does not include the Wilcoxon–Peto test statistic.

All the scenarios discussed earlier have been used as well to calculate the power of the Sun's family of tests. Tables 2, 3 and 4 present a value in parentheses indicating the power ratio of our proposal to the proposal of Sun *et al.* (2005).

The simulation study shows a much better behavior of our class in Table 2. If we exclude the log–rank test statistics ($\lambda = 0$ and $\rho = 0$), 66% of the power ratios are greater than one. Indeed, when the parameters (ρ, λ) of the test statistic and the AFT model are identical, all the power ratios are equal or greater than one. Moreover, when the test statistics have parameter $\lambda = 0$ (this is parametrization more commonly used for right–censored data), the ratios are equal or greater than 1 in all the situations (they are often greater than 1.5). We observe a similar behavior when the test statistics have parameter $\lambda = \rho \neq 0$: the ratios are greater than 1 except for scenarios 2, 3 and 4 where the ratios lie between 0.8 and 1. Our proposal does not perform as well when $\rho = 0$, then the ratios almost always lie between 0.7 and 1. However it is worth mentioning that in the cases with ratios smaller than 1, hence, where Sun's tests are superior, the power of the test statistics are very low and neither one would be recommended.

For the situations with early, late or middle differences reported in Table 3, among the 144 evaluated cases (we exclude again the log–rank test statistics), we observe that our proposal has the higher power in 68 cases (47%) and the same power in 22 cases (15%). Our proposal is clearly better when we choose the powerful combinations of the parameters in each scenario. For the *Early times situation* and the parameters $\rho \neq 0$ and $\lambda = 0$, our proposal has the higher power in 11 cases (61%) and the same power in 2 cases (11%). For the *Late times situation* and the parameters $\rho = 0$ and $\lambda \neq 0$, our proposal has the higher power in 8 cases (44%) and the same power in 6 cases (33%). For the *Middle times situation* and the parameters $\rho = \lambda \neq 0$, our proposal has the higher power in 8 cases (67%) and the same power in 2 cases (17%).

Our family of tests, as well as Sun's family, are not suitable for crossing hazards, and in those cases, the power of the tests are quite low and no clear winner can be claimed from the comparison (see Table 4).

8 Illustration

In this section we analyse the data corresponding to a cohort of injecting drug users (IDU) attending the Germans Trias i Pujol detoxification unit (Badalona, Spain) between February 1987 and November 1997. Details from this study can be found in Gómez *et al.* (2000). We are interested in the elapsed time T, measured in months, between intravenous drug initiation and seroconversion (HIV infection). The analysis of such data distinguishes four calendar periods according to the date for starting intravenous drug use: Period 1 (P1) contains those patients who started IDU before or at 1980, Period 2 (P2) includes IDU patients who started the addiction between 1981 and 1985, the third period P3 is for patients who started IDU between 1986 and 1991 and finally P4 includes all those patients starting IDU after or at 1992. In this illustration we only analyse the data for the last three periods P2, P3 and P4, as in Gómez *et al.* (2000). In P1 most of the patients began the use of intravenous drugs earlier than 1978, when HIV infection was extremely unlikely; furthermore the elapsed time between intravenous drug initiation and HIV infection is bounded below by at least 5 years, due to the fact that HIV seropositivity could not be determined before 1985.

In our first analysis, we consider the elapsed time to seroconversion according to periods P2, P3 and P4 separately for men and women. The gender stratification is meaningful since factors such sexual risk habits and a more likely HIV transmission from men to women, indicate that is appropriate to proceed in this way.

The estimates of the survival functions given in equation (5) have been applied to these data and their plots are shown in Figure 1 and Figure 2 for men and women, respectively. In this analysis we use the statistic $\boldsymbol{U}$ with parameters $\rho = 1$ and $\lambda = 1$ which emphasize middle hazard differences, see equation (11). Concerning men, the statistic is equal to $\boldsymbol{U} = \big(4.24, -3.48, -0.76\big)$ with p-value 0.041. For women, we obtain $\boldsymbol{U} = \big(1.72, -2.04, 0.32\big)$ which has a nonsignificant p-value equal to 0.087. However, it must be noted that for period P4 there are only 14 women and their follow-up is very short. This fact affects the value of the statistic $\boldsymbol{U}$ and henceforth its significance.

The second analysis takes into account the age in which patients have started to use drugs, since it is very likely that this could be a risk factor for HIV infection. We center this analysis in period P3 taking into account that the median age for starting IDU in this period is 20 years. We split the 240 patients in P3 into two groups: individuals younger than or exactly 21 years old and individuals older than 21 years. Figure 3 shows the estimated survival functions considering the two age groups in period P3. The statistic $\boldsymbol{U}$ with parameters $\rho = 1$ and $\lambda = 0$ emphasize early hazard differences and it is equal to $\boldsymbol{U} = \big(1.72, -1.72\big)$ with p–value equal to 0.043.

In Table 1 we provide the p-values of $\boldsymbol{U}$ for other choices of the parameters of the $G^{\rho,\lambda}$ family. We note that the log–rank statistic ($\rho = 0$ and $\lambda = 0$), the most popular one, fails to reject the null hypothesis in all scenarios. In this illustration we have done a post hoc choice of the parameters. However, it is obvious that there are many situations where data do not accommodate proportional hazard differences, for instance treatment effects studies. In these cases, the goal is to infer whether the treatment effect appears in the early, late or middle phase of the study.

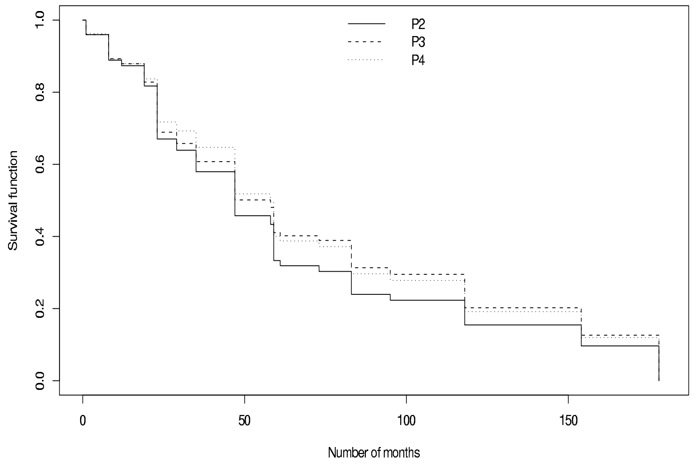

Figure 1: Elapsed time to seroconversion since starting intravenous drug use for men entering at risk either in calendar period P2 ($N_1 = 300$), or in period P3 ($N_2 = 240$) or in period P4 ($N_3 = 73$).

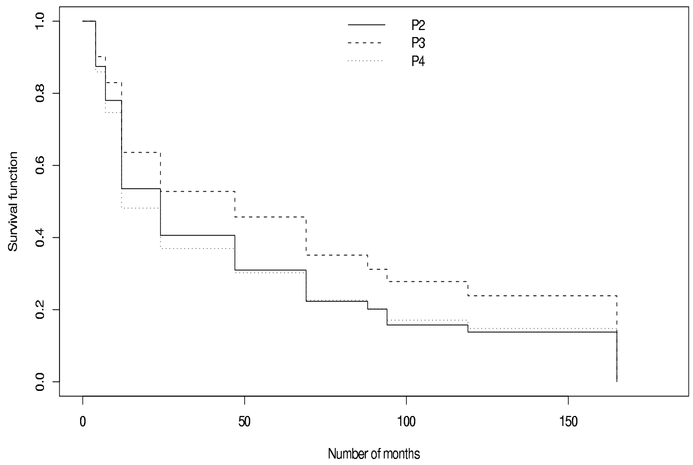

Figure 2: Elapsed time to seroconversion since starting intravenous drug use for women entering at risk either in calendar period P2 ($N_1 = 74$), or in period P3 ($N_2 = 66$) or in period P4 ($N_3 = 14$).

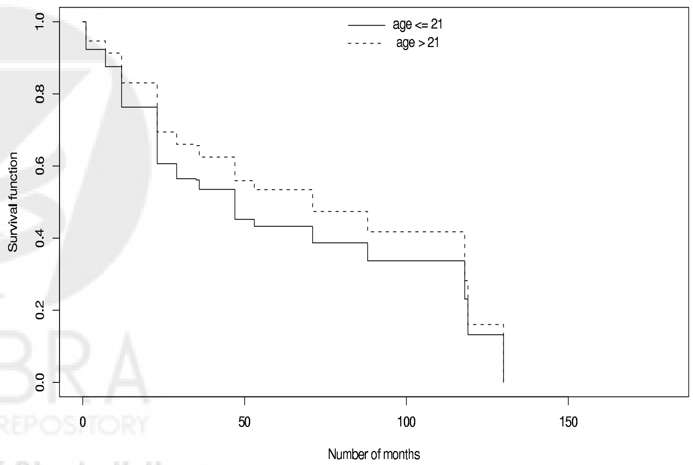

Figure 3: Elapsed time to seroconversion since starting intravenous drug use for individuals younger than 21 ($N_1 = 192$) and individuals older than 21 ($N_2 = 114$) entering at risk in calendar period P3.

Table 1: *Illustration: summary of p–values of the $G^{\rho,\lambda}$ family.*

	Parameters (ρ, λ) of the test statistic									
Scenarios	$(0,0)$	$(1,0)$	$(2,0)$	$(3,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(1,1)$	$(2,2)$	$(3,3)$
P2, P3 & P4 (for men)	.089	.166	.284	.402	.042	.057	.105	.041	.028	.022
P2, P3 & P4 (for women)	.086	.093	.111	.132	.111	.159	.215	.087	.089	.091
Age ≤ 21 & > 21 (in P3)	.061	.043	.038	.036	.177	.372	.615	.100	.128	.148

9 Conclusions

This paper proposes a new class of test statistics for interval–censored data. We have shown that this class extends the $G^{\rho,\lambda}$ family given in Fleming and Harrington (1991) and that presents a better behavior than the extension considered in Sun *et al.* (2005). However, several issues deserve further comments.

First, it is interesting to consider the application of our proposal to right–censored data. In this situation, our family does not coincide with the original family. The score values c_i for exact observations would be equivalent in both families, however they will differ for right–censored observations. This difference was also noted by Peto and Peto (1972) for the log–rank test and is due to the fact that our estimation of the hazard functions does not coincide with the usual estimation for right–censored data. Zhao and Sun (2004) propose a multiple imputation approach and another generalization of the log–rank test which seems to be useful when data are interval–censored data and there is a high percentage of right–censored observations. Sun (2006) gives an sketch of how to generalize this method to a class of weighted log–rank tests. The comparison of our proposal and Sun's method (2006), including the new multiple imputation method proposed by Huang *et al.* (2008), remains as a future research question.

Another issue is that the permutation approach we have applied is in fact a conditional approach since the distribution of the test statistic is computed conditional on the observed intervals. It is not obvious whether the permutation approach gives power properties similar to an unconditional approach. With right–censored data, Heimann and Neuhaus (1998) show that the permutation version of the log–rank test and the unconditional version are asymptotically equivalent even under unequal censoring. With interval–censored data, the comparison of the asymptotic behavior of the permutation distribution of $\boldsymbol{U}$ with an unconditional distribution, for instance the likelihood distribution, deserves further attention. In the particular situation of case II interval–censored data, Sun *et al.* (2005) give the asymptotic unconditional distribution of the statistic $\boldsymbol{U}$. A careful look at the estimation of the asymptotic variance given by these authors shows that it coincides with the permutation variance given above except for the use of the fraction $\frac{1}{n}$ instead of $\frac{1}{n-1}$. Thus, for case II interval–censored data, the conditional distribution of the test statistic given by the permutation approach is asymptotically equivalent to the unconditional distribution given by Sun *et al.* (2005).

Acknowledgements

The authors are grateful to the GRASS group for their fruitful discussions. This research was partially supported by Grants MTM2005-08886 and MTM2008-06747-C02-00 from the Ministerio de Educación y Ciencia. Part of this paper has been written under the support of Grant AI24643 from the National Institute of Allergy and Infectious Diseases.

References

Fang, H–B., Sun, J. and Lee, M–L T. (2002). Nonparametric survival comparisons for interval–censored continuous data. *Statistica Sinica*, 12, 1073–1083.

Fay, M. P. (1996). Rank invariant tests for interval-censored data under the grouped continuous model. *Biometrics*, 52, 811–822.

Fay, M. P. (1999). Comparing several score tests for interval-censored data. *Statistics in Medicine*, 18, 273–285.

Fay, M. P. and Shih, J. H. (1998). Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, 93, 387–396.

Finkelstein, D. M. (1986). A proportional hazards models for interval-censored failure time data. *Biometrics*, 42, 845–854.

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

Gómez, G., Calle, M. L. and Oller, R. (2004). Frequentist and bayesian approaches for interval–censored data. *Statistical Papers*, 45, 139–173.

Gómez, G., Calle, M. L., Egea, J. M. and Muga, R. (2000). Risk of HIV infection as a function of the duration of intravenous drug use: a non–parametric Bayesian approach. *Statistics in Medicine*, 19, 2641–2656.

Heimann, G. and Neuhaus, G. (1998). Permutational distribution of the log–rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics*, 54, 168–184.

Huang, J., Chinsan, L. and Yu, Q. (2008). A generalized log–rank test for interval–censored failure time data via multiple imputation. *Statistics in Medicine*, 27, 3217–3226.

Lim, H. J. and Sun, J. (2003). Nonparametric tests for interval–censored failure time data. *Biometrical Journal*, 45, 263–276.

Mantel, N. (1967). Ranking procedures for arbitrarily restricted observation. *Biometrics*, 23, 65–78.

Oller, R., Gómez, G. and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, 32, 315–326.

Oller, R., Gómez, G. and Calle, M. L. (2007). Interval censoring: identifiability and the constant–sum property. *Biometrika*, 94, 61–70.

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 45, 497–507.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, 135, 185–207.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, 22, 86–91.

Petroni, G. R. and Wolfe, A. (1994). A two sample test for stochastic ordering with interval-censored data. *Biometrics* 50, 77–87.

Schick, A. and Yu, Q. (2000). Consistency of the GMLE with mixed case interval–censored data. *Scandinavian Journal of Statistics*, 27, 45–55.

Sun. J. (1996). A non–parametric test for interval–censored failure time data with application to AIDS studies. *Statistics in Medicine*, 15, 1387–1395.

Sun, J., Zhao, Q. and Zhao, X. (2005). Generalized log–rank tests for interval–censored failure time data. *Scandinavian Journal of Statistics*, 32, 49–57.

Sun, J. (2006). *The statistical analysis of interval–censored failure time data.* Springer, New York.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38, 290–295.

Zhao, Q. and Sun. J. (2004). Generalized log–rank test for mixed interval-censored failure time data. *Statistics in Medicine*, 23, 1621–1629.

Appendix: Likelihood based variance–covariance matrix

The likelihood based variance–covariance matrix V of the efficient score statistic for $\boldsymbol{\beta}$ is given by

$$V = -\left[\frac{\partial^2 \log(Lik(S))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} - \left(\frac{\partial^2 \log(Lik(S))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}'}\right)\left(\frac{\partial^2 \log(Lik(S))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)^{-1}\left(\frac{\partial^2 \log(Lik(S))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}'}\right)\right]_{\beta=0,\theta=\hat{\theta}}$$

where for arbitrary parameters ψ_u and ψ_v

$$\frac{\partial^2 \log(Lik(S))}{\partial\psi_u\partial\psi_v} = \sum_{i=1}^{n} \frac{1}{\triangle_i S}\left[\left(\frac{\partial^2 \triangle_i S}{\partial\psi_u\partial\psi_v}\right) - \frac{1}{\triangle_i S}\left(\frac{\partial \triangle_i S}{\partial\psi_u}\right)\left(\frac{\partial \triangle_i S}{\partial\psi_v}\right)\right],$$

with

$$\triangle_i S = S(r_i \mid \mathbf{z}_i) - S(l_i \mid \mathbf{z}_i)$$

and

$$\left[\frac{\partial S(t_j \mid \mathbf{z}_i)}{\partial\beta_u}\right]_{\beta=0,\theta=\hat{\theta}} = -\frac{1}{\rho}\{\hat{S}(t_j)\}\{1-(\hat{S}(t_j))^{\rho}\}\,\alpha_i^{(u)}$$

$$\left[\frac{\partial S(t_j \mid \mathbf{z}_i)}{\partial\theta_u}\right]_{\beta=0,\theta=\hat{\theta}} = -\frac{1}{\rho}\{\hat{S}(t_j)\}\{1-(\hat{S}(t_j))^{\rho}\}\,\mathbf{1}_{\{j=u\}}$$

$$\left[\frac{\partial^2 S(t_j \mid \mathbf{z}_i)}{\partial\beta_u\,\partial\beta_v}\right]_{\beta=0,\theta=\hat{\theta}} = \frac{1}{\rho^2}\{\hat{S}(t_j)\}\{1-(\hat{S}(t_j))^{\rho}\}\{1-(\rho+1)(\hat{S}(t_j))^{\rho}\}\,\alpha_i^{(u)}\alpha_i^{(v)}$$

$$\left[\frac{\partial^2 S(t_j \mid \mathbf{z}_i)}{\partial\beta_u\,\partial\theta_v}\right]_{\beta=0,\theta=\hat{\theta}} = \frac{1}{\rho^2}\{\hat{S}(t_j)\}\{1-(\hat{S}(t_j))^{\rho}\}\{1-(\rho+1)(\hat{S}(t_j))^{\rho}\}\,\alpha_i^{(u)}\mathbf{1}_{\{j=v\}}$$

$$\left[\frac{\partial^2 S(t_j \mid \mathbf{z}_i)}{\partial\theta_u\,\partial\theta_v}\right]_{\beta=0,\theta=\hat{\theta}} = \frac{1}{\rho^2}\{\hat{S}(t_j)\}\{1-(\hat{S}(t_j))^{\rho}\}\{1-(\rho+1)(\hat{S}(t_j))^{\rho}\}\mathbf{1}_{\{j=u=v\}}$$

The proof of this result is omitted because it follows from standard statistical theory and it is analogous to Fay (1999).

Table 2: *Simulation study: estimated powers under several accelerated failure failure time models.*

	AFT model										
Scenario	Parameters	Parameters (ρ, λ) of the test statistic									
number	(ρ, λ)	$(0, 0)$	$(1, 0)$	$(2, 0)$	$(3, 0)$	$(0, 1)$	$(0, 2)$	$(0, 3)$	$(1, 1)$	$(2, 2)$	$(3, 3)$
1	$(0, 0)$	.906 (1.0)	.865 (1.2)	.812 (1.8)	.751 (3.0)	.816 (1.0)	.648 (0.9)	.460 (0.7)	.890 (1.0)	.865 (1.2)	.845 (1.3)
2	$(1, 0)$	.818 (1.0)	.875 (1.1)	.860 (1.4)	.837 (1.8)	.493 (0.8)	.272 (0.7)	.128 (0.5)	.750 (0.9)	.717 (1.0)	.684 (1.0)
3	$(2, 0)$	.697 (1.0)	.833 (1.0)	.848 (1.2)	.841 (1.5)	.336 (0.8)	.154 (0.6)	.059 (0.4)	.634 (0.8)	.588 (0.9)	.558 (0.9)
4	$(3, 0)$	.630 (1.0)	.824 (1.0)	.854 (1.1)	.860 (1.3)	.287 (0.8)	.124 (0.6)	.045 (0.4)	.553 (0.8)	.499 (0.8)	.480 (0.8)
5	$(0, 1)$	.671 (1.0)	.426 (2.5)	.272 (4.1)	.189 (3.9)	.800 (1.0)	.738 (0.9)	.634 (0.9)	.724 (1.4)	.718 (1.9)	.711 (2.4)
6	$(0, 2)$	.576 (1.0)	.300 (2.9)	.171 (3.2)	.124 (1.8)	.853 (1.0)	.863 (1.0)	.803 (0.9)	.700 (1.9)	.712 (2.9)	.710 (4.0)
7	$(0, 3)$	.401 (1.0)	.165 (2.7)	.086 (1.3)	.068 (1.0)	.656 (1.1)	.727 (1.0)	.720 (1.0)	.410 (3.1)	.388 (6.6)	.362 (7.0)
8	$(1, 1)$	.728 (1.0)	.584 (1.8)	.439 (4.0)	.337 (5.5)	.659 (0.9)	.472 (0.8)	.301 (0.7)	.769 (1.1)	.747 (1.3)	.729 (1.4)
9	$(2, 2)$	.769 (1.0)	.605 (2.0)	.432 (6.5)	.300 (8.1)	.790 (1.0)	.608 (0.8)	.392 (0.7)	.866 (1.1)	.877 (1.2)	.877 (1.3)
10	$(3, 3)$	.632 (1.0)	.441 (2.4)	.301 (6.7)	.203 (4.2)	.656 (0.9)	.483 (0.8)	.310 (0.7)	.768 (1.2)	.790 (1.4)	.789 (1.6)

Remark: Scenarios 1 and 2 are a proportional hazards model and a proportional odds model, respectively.

Table 3: *Simulation study: estimated powers under scenarios with early, late and middle hazard differences.*

Piecewise exponential scenarios		Parameters (ρ, λ) of the test statistic									
		$(0,0)$	$(1,0)$	$(2,0)$	$(3,0)$	$(0,1)$	$(0,2)$	$(0,3)$	$(1,1)$	$(2,2)$	$(3,3)$
High–Early	50	.146 (1.0)	.241 (0.7)	.318 (0.8)	.357 (0.9)	.060 (1.0)	.049 (1.0)	.050 (1.0)	.099 (0.8)	.073 (0.7)	.065 (0.8)
$x_1 = 1.25$	200	.414 (1.0)	.723 (0.8)	.840 (0.9)	.896 (1.0)	.072 (0.8)	.051 (0.8)	.048 (1.0)	.185 (0.5)	.125 (0.5)	.088 (0.4)
Intermediate–Early	50	.142 (1.0)	.177 (1.0)	.179 (1.6)	.171 (2.2)	.086 (0.9)	.066 (0.9)	.049 (0.8)	.133 (0.8)	.130 (0.8)	.126 (0.8)
$x_1 = 6.25$	200	.408 (1.0)	.521 (1.1)	.514 (1.7)	.481 (2.9)	.167 (0.8)	.082 (0.8)	.060 (0.8)	.385 (0.8)	.358 (0.7)	.349 (0.8)
Low–Early	50	.199 (1.0)	.211 (1.4)	.177 (2.5)	.152 (2.5)	.132 (0.9)	.092 (0.9)	.074 (0.8)	.195 (1.0)	.201 (1.3)	.196 (1.6)
$x_1 = 11.25$	200	.633 (1.0)	.647 (1.4)	.576 (2.7)	.503 (4.0)	.447 (0.9)	.258 (0.7)	.182 (0.8)	.620 (1.0)	.600 (1.4)	.560 (1.8)
Low–Late	50	.324 (1.0)	.222 (2.2)	.159 (3.3)	.125 (2.4)	.322 (1.0)	.263 (0.9)	.222 (0.9)	.338 (1.6)	.315 (2.2)	.301 (2.5)
$x_1 = 1.25$	200	.865 (1.0)	.702 (2.5)	.509 (8.3)	.354 (7.5)	.860 (1.0)	.785 (0.9)	.676 (0.9)	.879 (1.3)	.868 (1.9)	.842 (2.6)
Intermediate–Late	50	.163 (1.0)	.077 (1.4)	.056 (1.0)	.052 (1.0)	.278 (1.1)	.302 (1.0)	.304 (1.0)	.154 (2.4)	.132 (2.2)	.116 (1.6)
$x_1 = 6.25$	200	.514 (1.0)	.165 (3.4)	.071 (0.8)	.055 (0.8)	.783 (1.1)	.848 (1.0)	.844 (1.0)	.495 (5.7)	.428 (5.2)	.375 (2.7)
High–Late	50	.070 (1.0)	.049 (0.8)	.058 (1.1)	.059 (1.2)	.095 (1.1)	.116 (1.2)	.123 (1.1)	.065 (1.3)	.059 (1.0)	.056 (0.9)
$x_1 = 11.25$	200	.130 (1.0)	.063 (1.1)	.057 (1.0)	.054 (1.0)	.243 (1.2)	.316 (1.2)	.365 (1.2)	.088 (1.6)	.065 (1.0)	.057 (0.9)
Low–Middle	50	.154 (1.0)	.163 (1.5)	.141 (2.4)	.111 (2.3)	.123 (0.9)	.085 (0.8)	.064 (0.8)	.196 (0.9)	.205 (1.2)	.206 (1.3)
$x_1 = 1.25$	200	.590 (1.0)	.597 (1.7)	.460 (5.1)	.350 (7.1)	.386 (0.8)	.221 (0.8)	.136 (0.7)	.667 (1.0)	.683 (1.2)	.680 (1.3)
$x_2 = 8.75$											
High–Middle	50	.172 (1.0)	.161 (1.8)	.123 (2.2)	.092 (1.5)	.130 (0.8)	.078 (0.8)	.054 (0.7)	.232 (0.9)	.268 (1.1)	.289 (1.2)
$x_1 = 3.75$	200	.556 (1.0)	.459 (2.5)	.297 (5.2)	.185 (1.7)	.458 (0.8)	.261 (0.7)	.154 (0.7)	.740 (1.0)	.806 (1.3)	.837 (1.4)
$x_2 = 6.25$											

Table 4: *Simulation study: estimated powers under crossing hazards scenarios.*

Piecewise		Parameters (ρ, λ) of the test statistic									
exponential scenarios		$(0, 0)$	$(1, 0)$	$(2, 0)$	$(3, 0)$	$(0, 1)$	$(0, 2)$	$(0, 3)$	$(1, 1)$	$(2, 2)$	$(3, 3)$
Crossing–Early	50	.089 (1.0)	.061 (0.8)	.064 (0.5)	.072 (0.6)	.131 (1.0)	.144 (1.0)	.136 (0.9)	.105 (1.7)	.113 (2.1)	.116 (2.2)
$x_1 = 2.5$	200	.183 (1.0)	.052 (0.4)	.074 (0.2)	.101 (0.3)	.457 (1.0)	.469 (1.0)	.422 (0.9)	.285 (3.8)	.324 (5.9)	.334 (5.5)
Crossing–Middle	50	.058 (1.0)	.084 (0.6)	.107 (0.7)	.117 (1.1)	.076 (1.2)	.100 (1.1)	.105 (1.0)	.059 (0.6)	.056 (0.4)	.053 (0.4)
$x_1 = 5$	200	.057 (1.0)	.205 (0.5)	.311 (0.7)	.345 (1.2)	.145 (1.3)	.241 (1.2)	.282 (1.0)	.055 (0.2)	.057 (0.1)	.062 (0.1)
Crossing–Late	50	.079 (1.0)	.138 (0.8)	.169 (1.3)	.159 (1.8)	.050 (1.0)	.068 (1.1)	.077 (1.1)	.083 (0.5)	.090 (0.5)	.093 (0.5)
$x_1 = 7.5$	200	.197 (1.0)	.475 (0.9)	.513 (1.5)	.482 (2.6)	.054 (0.9)	.073 (1.3)	.129 (1.3)	.219 (0.4)	.233 (0.4)	.259 (0.4)