The Canadian Journal of Statistics
Vol. 40, No. 3, 2012, Pages 501–516
La revue canadienne de statistique

A generalized Fleming and Harrington's class of tests for interval-censored data

Ramon OLLER[1]* and Guadalupe GÓMEZ[2]

[1]*Departament d'Economia i Empresa, Universitat de Vic, Sagrada Família 7, 08500 Vic, Spain*
[2]*Departament d'Estadística i Investigació Operativa de la Universitat Politècnica de Catalunya, Jordi Girona 1-3, Edifici C5, Campus Nord, 08034 Barcelona, Spain*

Key words and phrases: Interval-censored data; permutation test; treatment comparison; weighted log-rank test.

MSC 2010: Primary 62N01; secondary 62G10.

Abstract: The class $G^{\rho,\lambda}$ of weighted log-rank tests proposed by Fleming & Harrington [Fleming & Harrington (1991) *Counting Processes and Survival Analysis*, Wiley, New York] 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 extends the $G^{\rho,\lambda}$ class to interval-censored data. First we introduce a new general class of rank based tests, then we show the analogy to the above proposal of Fleming & Harrington. The asymptotic behaviour of the proposed tests is derived using an observed Fisher information approach and a permutation approach. Aiming to make this family of tests interpretable and useful for practitioners, we explain how to interpret different choices of weights and we apply it to data from a cohort of intravenous drug users at risk for HIV infection. *The Canadian Journal of Statistics* 40: 501–516; 2012 © 2012 Statistical Society of Canada

Résumé: La classe $G^{\rho,\lambda}$ des tests log-rang pondérés proposée par Fleming et Harrington (1991) sont très largement utilisés en analyse de survie, et de nos jours, elle est une méthode non paramétrique qui a fait ses preuves pour comparer k fonctions de survie différentes pour les données de survie censurées à droite. Cet article généralise la classe $G^{\rho,\lambda}$ pour les données censurées par intervalles. Dans un premier temps, nous proposons une nouvelle classe de tests basés sur les rangs, et par la suite, nous faisons une analogie avec les tests de Fleming et Harrington (1991). Le comportement asymptotique des tests proposés est obtenu en utilisant l'approche de la quantité d'information de Fisher observé et une approche par permutation. Afin que les utilisateurs puissent interpréter cette famille de tests et qu'elles leur soient utiles, nous expliquons comment interpréter différents choix pour les poids et nous l'appliquons à un jeu de données sur le risque d'infection au VIH pour une cohorte d'utilisateurs de drogues intraveineuses. *La revue canadienne de statistique* 40: 501–516; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION

Comparison of two or more distributions based on censored data is a topic which arises in most survival studies. While many tests have been proposed when the data are right-censored, research for interval-censored data is still ongoing and lacks a 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. Peto & Peto (1972) were among the first

* *Author to whom correspondence may be addressed.*
E-mail: ramon.oller@uvic.cat

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 with the extension of the Wilcoxon and log-rank tests, see for instance Fay (1996, 1999), Fay & Shih (1998), Zhao & Sun (2004), Sun, Zhao, & Zhao (2005), and Huang, Lee, & Yu (2008). Other authors such as Lim & Sun (2003) discuss generalizations of the weighted Kaplan–Meier class developed by Pepe & Fleming (1989) for right-censored data. An extensive review of k-sample methods for interval-censored data can be found in Gómez, Calle, & Oller (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 class of weighted log-rank statistics and, in particular, the $G^{\rho,\lambda}$ subfamily introduced by Fleming & Harrington (1991). For this subfamily, the weight function is chosen to be $[\hat{S}(t-)]^{\rho}[1-\hat{S}(t-)]^{\lambda}$ where $\hat{S}(t)$ is the Kaplan–Meier estimate of the survival function. 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 $\lambda = 0$) as well as a 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 (for further details, see Lawless (2003)).

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. Section 3 introduces our proposal and shows that it is a natural extension of the $G^{\rho,\lambda}$ family for right-censored data. Section 4 derives the subclass G^{ρ} as a likelihood score procedure for discrete data. Section 5 describes the permutation approach to be used for inferential purposes. The paper continues with Section 6 where we report a simulation study which gives guidance on the behaviour of the $G^{\rho,\lambda}$ family of tests. In Section 7 we apply the new family of tests to a real data set from an AIDS study. Section 8 provides a summary of the results presented in this 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 $S^{(1)}, \ldots, S^{(k)}$ to be the survival functions of T for each of these groups. Our goal is to test the hypothesis $H_0 : S^{(1)} = \cdots = S^{(k)} = S$ versus $H_a : S^{(j)} \neq S^{(j')}$ for some $j \neq j'$. If the 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 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 null hypothesis H_0 and the assumption that the censoring process is noninformative (Oller, Gómez, & Calle, 2004, 2007; Lawless, 2004), the likelihood function for the pooled sample simplifies as follows:

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

where $n = n^{(1)} + \cdots + n^{(k)}$ and $(l_1, r_1), \ldots, (l_n, r_n)$ are independent observations. Denote by $\mathcal{L}$ and $\mathcal{R}$ the sets $\mathcal{L} = \{l_i, 1 \leq i \leq n\}$ and $\mathcal{R} = \{r_i, 1 \leq i \leq n\}$. Following Peto (1973) and Turnbull (1976), we can derive all the distinct intervals such that their left and right end-points lie in $\mathcal{L}$ and $\mathcal{R}$, respectively, and they do not contain other members of $\mathcal{L}$ or $\mathcal{R}$. 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, the nonparametric maximum likelihood estimator $\hat{S}(t)$ of $S(t)$ is unspecified in each $(q_j, p_j]$ and is well defined and constant between these intervals.

There are several algorithms to determine $\hat{S}(t)$, see Gómez et al. (2009) for an exhaustive description of them. If the EM algorithm is used, an estimate of the survival function for the ith individual, $\hat{S}_i(t)$, is obtained from Turnbull's overall survival $\hat{S}(t)$ truncated at the ith observed interval (Fay & Shih, 1998). That is,

$$\hat{S}_i(t) = P_{\hat{S}}((t, +\infty) \mid (l_i, r_i]) = \frac{\hat{S}(l_i \vee t) - \hat{S}(r_i \vee t)}{\hat{S}(l_i) - \hat{S}(r_i)} \tag{2}$$

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 maximum between a and b. The maximization step of the EM algorithm yields $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_i(t)$. The survival function $\hat{S}_i(t)$ is unspecified inside Turnbull's intervals, as a consequence of the unspecification of $\hat{S}(t)$. The survival functions $\hat{S}(t)$ and $\hat{S}_i(t)$ play an important role in the test statistics we introduce next. In what follows, we consider $\hat{S}(t)$ to be one of the survival functions in the equivalence class of survival functions which assigns the same probability mass to Turnbull's intervals. This convention has no effect on the final value of the test statistics.

3. THE WEIGHTED LOG-RANK CLASS $G^{\rho,\lambda}$

In this section we present our proposal for a class of weighted log-rank statistics to test $H_0 : S^{(1)} = \cdots = S^{(k)}$ versus $H_a : S^{(j)} \neq S^{(j')}$ for some $j \neq j'$. The terms in the $G^{\rho,\lambda}$ class can be interpreted as a weighted sum of observed minus expected number of events under the null hypothesis of identical survival curves. This class is a natural extension to interval-censored data of the original $G^{\rho,\lambda}$ family proposed by Fleming and Harrington for right-censored data.

Throughout this section, for any step function $F(t)$ and fixed value t, we denote a function increment as $\mathrm{d}F(t) = F(t) - F(t-)$ where as usual $F(t-) = \lim_{x\uparrow t} \mathrm{F(x)}$. Note that $\mathrm{d}F(t) = 0$ except at points of discontinuity of F.

Our proposed test statistic is a vector $\boldsymbol{U} = \left(U^{(1)}, \ldots, U^{(k)}\right)'$ with components

$$U^{(j)} = \int_0^{+\infty} w(t; \lambda, \rho) \left[d_t^{(j)} - \frac{n_t^{(j)}}{n_t} d_t \right], \quad j = 1, \ldots, k \tag{3}$$

where $n_t^{(j)} = n^{(j)} \hat{S}^{(j)}(t-)$ is the (expected) number of individuals at risk at time t from group j; $n_t = n\,\hat{S}(t-)$ is the (expected) number for all groups together; and $d_t^{(j)} = -n^{(j)} \mathrm{d}\hat{S}^{(j)}(t)$ and $d_t = -n \mathrm{d}\hat{S}(t)$ stand for the estimated number of failures in the jth group and for all groups together, respectively. In the above expressions the survival function for the jth group is estimated as

$$\hat{S}^{(j)}(t) = \frac{1}{n^{(j)}} \sum_{i=1}^{n} \alpha_i^{(j)} \hat{S}_i(t), \tag{4}$$

where $\alpha_i^{(j)}$ denotes the indicator function of the group $G^{(j)}$, and $\hat{S}_i(t)$ is the estimate of the survival function for the ith individual defined in Equation (2). The weights $w(t; \lambda, \rho)$, denoted from now on as $w(t)$ for ease of notation, are defined as

$$w(t) = \hat{S}(t-)\frac{B\big(1-\hat{S}(t);\lambda+1,\rho\big) - B\big(1-\hat{S}(t-);\lambda+1,\rho\big)}{\hat{S}(t-)-\hat{S}(t)} \tag{5}$$

where $\rho \geq 0, \lambda \geq 0$ and $B(t;a,b) = \int_0^t x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$ is an incomplete beta function. Asymptotically, as $\mathrm{d}\hat{S}(t) \to 0$ and $\hat{S}$ is nearly continuous at the point t, the weight function $w(t)$ resembles the weights of the Fleming-Harrington family, that is, $[\hat{S}(t-)]^{\rho}[1-\hat{S}(t-)]^{\lambda}$. This result will be proved in Proposition 1.

The statistic $\boldsymbol{U}$ reduces to the log-rank and Wilcoxon–Peto test statistics originally proposed in Peto & Peto (1972) for choices of $(\rho,\lambda) = (0,0)$ and $(\rho,\lambda) = (1,0)$, respectively. As for many of the usual nonparametric tests for the comparison of survival functions, we can study our proposed class of statistics from varying perspectives. We present below the $\boldsymbol{U}$ statistic as a sum, over all times, of weighted differences in the estimated hazards. In Section 3.2 we regard it as a linear scores form.

3.1. Integrated Difference of Hazards Form

Consider $\mathrm{d}\widehat{H}(t) = -\mathrm{d}\hat{S}(t)/\hat{S}(t-)$ and $\mathrm{d}\widehat{H}^{(j)}(t) = -\mathrm{d}\hat{S}^{(j)}(t)/\hat{S}^{(j)}(t-)$ $(j = 1,\ldots,k)$ as estimators of the overall hazard function and the hazard function for the jth group, respectively. Then, the jth component $U^{(j)}$ of $\boldsymbol{U}$ given in (3) is expressed equivalently as

$$U^{(j)} = \int_0^{+\infty} w(t)\, n_t^{(j)}\big[\mathrm{d}\widehat{H}^{(j)}(t) - \mathrm{d}\widehat{H}(t)\big]. \tag{6}$$

This formulation shows that this family is geared to detect an alternative hypothesis where the hazards between groups differ but do not cross.

The interpretation of the weights given by (5) reproduces the interpretation of the weights $[\hat{S}(t-)]^{\rho}[1-\hat{S}(t-)]^{\lambda}$ in the original right-censored family. The choice of the weights $w(t)$, that is, of the parameters λ and ρ, is a relevant part of the Statistical Analysis Plan since different choices would provide answers to different departures from the null hypothesis. For instance, in a given clinical trial, if one would like to assess whether the effect of a treatment or therapy on the survival is stronger at the earlier phases of the therapy, we should choose $\lambda = 0$, with increasing values of ρ emphasizing stronger early differences. If there were a clinical reason to believe that the effect of the therapy would be more pronounced towards the middle or the end of the follow-up period, it would make sense to choose $\rho = \lambda > 0$ or $\rho = 0$ respectively, with increasing values of λ emphasizing stronger middle or late differences. The choice of the weights has to be made prior to the examination of the data and taking into account that they should provide the greatest statistical power, which in turns depends on how it is believed the null is violated.

3.2. Linear Scores Form

As an alternative, the statistic $\boldsymbol{U}$ given in (3) can also take the following linear form:

$$\boldsymbol{U} = \sum_{i=1}^{n} \mathbf{z}_i c_i, \tag{7}$$

where $\mathbf{z}_i = \left(\alpha_i^{(1)}, \alpha_i^{(2)}, \ldots, \alpha_i^{(k)}\right)'$ is a covariate vector of group indicators and c_i is a score value associated with each individual defined for $\rho, \lambda \geq 0$ as

$$c_i = \begin{cases} \frac{\hat{S}(r_i)B(1-\hat{S}(r_i);\lambda+1,\rho)-\hat{S}(l_i)B(1-\hat{S}(l_i);\lambda+1,\rho)}{\hat{S}(l_i)-\hat{S}(r_i)} & \text{if } \hat{S}(r_i) \neq 0 \\ -B\left(1-\hat{S}(l_i);\lambda+1,\rho\right) & \text{if } \hat{S}(r_i) = 0 \end{cases} \tag{8}$$

This linear formulation allows us to introduce the permutation distribution of the statistic $\boldsymbol{U}$ in Section 5. The following proposition gives the equivalence between the weighted log-rank form given by (5) and (6) and the above linear form given by Equations (7) and (8). The proof of this proposition is given in the Appendix.

Proposition 1. *A weighted log-rank test statistic* $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(k)})'$ *with components given by (6) has the following properties:*

(a) 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*

$$c_i = \int_0^{+\infty} w(t)\left[-\mathrm{d}\hat{S}_i(t) + \frac{\hat{S}_i(t-)}{\hat{S}(t-)}\mathrm{d}\hat{S}(t)\right]. \tag{9}$$

(b) For a weight function

$$w(t) = \hat{S}(t-)\frac{\gamma(\hat{S}(t-)) - \gamma(\hat{S}(t))}{\hat{S}(t-) - \hat{S}(t)} = \hat{S}(t-)\frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)} \tag{10}$$

for $\gamma(t)$ *a differentiable nondecreasing function* ($\gamma(1) = 0$)*, the* c_i *scores given in (9) 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) For a given point t, the weight function (10) satisfies $\lim_{\mathrm{d}\hat{S}(t)\to 0}\{w(t) - \hat{S}(t)\gamma'(\hat{S}(t))\} = 0$*, where* $\gamma'(t)$ *is the first derivative of* $\gamma(t)$.

(d) The function $\gamma(t) = -B(1-t;\lambda+1,\rho)$ *yields the weights (5) and the scores (8). Moreover, for a given point t, the weights (5) satisfy* $\lim_{\mathrm{d}\hat{S}(t)\to 0}\{w(t) - [\hat{S}(t)]^{\rho}[1-\hat{S}(t)]^{\lambda}\} = 0$.

4. SCORE VECTORS UNDER DISCRETE OR GROUPED DATA

For discrete or grouped continuous interval-censored data, Finkelstein (1986) and Fay (1996, 1999) show that the log-rank and the Wilcoxon–Peto test statistics can be derived as the efficient score statistics for a proportional hazards model and for a proportional odds model, respectively. In this setup, several other score statistics can be derived from the linear transformation model studied in Fay (1996).

When $\lambda = 0$, the rank class $G^{\rho,\lambda}$ for interval-censored data is an extension of the so-called G^{ρ} family. The next theorem shows that the rank class subfamily $G^{\rho,0}$ is a class of efficient score statistics in a linear transformation model.

Proposition 2. *Let* $g(T_i) = -\mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i$ *be a linear transformation model with g being an unknown increasing function,* $\mathbf{z}_i = \left(\alpha_i^{(1)}, \alpha_i^{(2)}, \ldots, \alpha_i^{(k)}\right)'$ *and* ϵ_i *having survival function* $S_{\epsilon}(t) = [1 + \rho\exp(t)]^{-1/\rho}$*. Assume that the parameter* ρ *is known and the support for the observable data is finite, that is,* $L, R \in \{t_0, t_1, \ldots, t_m\}$*, where* $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = +\infty$*. Let the survival function for* T_i *be* $S(t \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta})$*, 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* $Lik(\boldsymbol{\beta}, \boldsymbol{\theta}) =$

$\prod_{i=1}^{n} \left\{ S(l_i \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta}) - S(r_i \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta}) \right\}$ *and let* $\hat{\boldsymbol{\theta}}_0$ *be the maximum likelihood estimator of the nuisance parameters when* $\boldsymbol{\beta} = \mathbf{0}$. *Then,*

(a) *Finding the nonparametric maximum likelihood estimator of the survival function from (1) is equivalent to obtaining the maximum likelihood estimator of the nuisance parameters because* $\hat{S}(t) = \left[S(t \mid \mathbf{z}_i'\boldsymbol{\beta}, \boldsymbol{\theta}) \right]_{\beta=0,\theta=\hat{\theta}_0}$.

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

(b) *The efficient score statistic* $\left[\frac{\partial \log(Lik(\boldsymbol{\beta},\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}} \right]_{\beta=0,\theta=\hat{\theta}_0}$ *is given by (7) and (8) with* $\lambda = 0$.

The proof of this result is omitted because it is analogous to Fay (1996, 1999). We remark 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 (7) and

$$c_i = \frac{S_\epsilon'\left(S_\epsilon^{-1}(\hat{S}(l_i))\right) - S_\epsilon'\left(S_\epsilon^{-1}(\hat{S}(r_i))\right)}{\hat{S}(l_i) - \hat{S}(r_i)} \tag{11}$$

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 directly from (11) when we consider the survival function $S_\epsilon(t) = [1 + \rho \exp(t)]^{-1/\rho}$.

Proposition 2 uses the "non standard" formulation $g(T_i) = -\mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i$ instead of $g(T_i) = \mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i$ because the former, together with $S_\epsilon(t) = [1 + \rho \exp(t)]^{-1/\rho}$ and $g(t) = \log t$, includes the well-known proportional hazards model (when $\rho \to 0$) and the proportional odds model (when $\rho = 1$). See Kalbfleisch & Prentice (2002) for a thorough discussion of the equivalence between linear transformation models and regression models.

Another interesting issue is whether or not the $G^{\rho,\lambda}$ family when $\lambda \neq 0$ is a class of score statistics for discrete data under the linear transformation model. From Equations (8) and (11), it follows that the survival function of the error term S_ϵ would have to be a solution of the following differential equation,

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

which, except for the special case $\lambda = 0$, does not have, in general, an analytic solution.

The asymptotic behaviour of a score vector $\boldsymbol{U}$ under discrete interval-censored data follows from maximum likelihood theory. Since in this case the number of parameters is finite, under the null hypothesis H_0 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. The explicit formula for V is not presented here but it can be provided upon request.

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. The next section presents the main aspects of this approach.

5. PERMUTATION DISTRIBUTION

In comparison with the likelihood method given above, the permutation approach is straightforward and 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 be asymptotically efficient. The main assumption to apply a permutation test is that the underlying censoring process has to be identical across groups. Although this restriction is not necessary in the previous likelihood approach, it is also necessary with other methods such as the multiple imputation methods proposed in Zhao & Sun (2004) and Huang, Lee, & Yu (2008) or the asymptotic method proposed in Sun, Zhao, & Zhao (2005).

The permutation approach can be applied easily to the linear form of the score statistic $\boldsymbol{U}$ given by (7). 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 (Sen, 2006) can be applied yielding a normal approximation with permutation expectation $E(\boldsymbol{U}) = n\bar{c}\bar{\mathbf{z}}$ and variance–covariance matrix $V_0 = \frac{1}{n-1}\left(\sum_{i=1}^{n} c_i^2 - n\bar{c}^2\right)\left(\sum_{i=1}^{n}(\mathbf{z}_i\mathbf{z}_i' - \bar{\mathbf{z}}\bar{\mathbf{z}}')\right)$, where $\bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i$ and $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i$ are the sample means.

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, so 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 $\bar{c}^{(j)} = \frac{1}{n^{(j)}}\sum_{i=1}^{n} c_i\alpha_i^{(j)}$. The statistic $\boldsymbol{U}'V_0^-\boldsymbol{U}$ follows a χ^2_{k-1} distribution, for n large enough, and the null hypothesis would be rejected for large values of $\boldsymbol{U}'V_0^-\boldsymbol{U}$. In the sequel we consider either discrete or continuous data and we use the normal approximation of the permutation distribution of $\boldsymbol{U}$.

The permutation approach is, in fact, a conditional approach since the distribution of the test statistic is computed conditioned 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 & 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 behaviour of the permutation distribution of $\boldsymbol{U}$ with an unconditional distribution, for instance the likelihood distribution, deserves further attention. With case II interval-censored data, that is, the particular situation where the observable data are determined by two inspection times, Sun, Zhao, & Zhao (2005) give an asymptotic unconditional distribution of $\boldsymbol{U}$. 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 $\boldsymbol{U}$ given by the permutation approach is asymptotically equivalent to the unconditional distribution given by Sun, Zhao, & Zhao (2005).

6. SIMULATION STUDY

A large simulation study has been carried out to assess the performance of the $G^{\rho,\lambda}$ family of tests given in Section 3 (from now on "OG") and to validate, in terms of their power, the interpretation of the weight function given by (5). Throughout, we use the normal approximation of the permutation distribution presented in Section 5. By means of the simulation study we also compare OG with other existing extensions of the $G^{\rho,\lambda}$ family. We have computed the power of the class of test statistics given by Sun, Zhao, & Zhao (2005) and by Sun (2006). We first summarize these two classes of tests.

The class of tests given in Sun, Zhao, & Zhao (2005) (from now on "S05") can be written either as a linear form in terms of scores, as in Equation (7), or as a weighted log-rank test (6) with weight function (10) derived from

$$\gamma(t) = \log(t)t^{\rho}(1-t)^{\lambda}. \tag{13}$$

The Sun, Zhao, & Zhao (2005) class reduces to the log-rank test proposed in Peto & Peto (1972) for $\rho = 0$ and $\lambda = 0$. For values $\rho \neq 0$ or $\lambda \neq 0$, S05's family differs from the OG family and does not include the Wilcoxon–Peto test statistic. Furthermore, since for any pair of values (ρ, λ) such as $\rho \neq 0, \lambda = 0$ or $\rho = \lambda \neq 0$ the function (13) is not necessarily nondecreasing, the corresponding weights could be negative and hence, the interpretation of them would be troublesome. The results given below for S05 are based on the normal approximation of the permutation distribution which, as we discuss in Section 5, is equivalent to the asymptotic distribution proposed in Sun, Zhao, & Zhao (2005).

Chapter 4 of Sun (2006) proposes a generalization of the weighted log-rank test statistic. This approach is a resampling method based on M imputed right-censored samples. From the original interval-censored sample, M right-censored samples are generated in the following way: first, for each interval-censored observation, an exact failure time is generated at random from the estimated survival function $\hat{S}_i(t)$ given in (2); second, the right observations are preserved. The methodology consists in computing the Fleming and Harrington statistic $\boldsymbol{U}^m$ and the martingale variance V^m for every imputed sample of right-censored data ($m = 1, \dots, M$). In our simulation study we are using the test statistic $\bar{\boldsymbol{U}}' V^- \bar{\boldsymbol{U}}$, where $\bar{\boldsymbol{U}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{U}^m$ and $V = \frac{1}{M}\sum_{m=1}^{M} V^m + \left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\boldsymbol{U}^m - \bar{\boldsymbol{U}}\right)\left(\boldsymbol{U}^m - \bar{\boldsymbol{U}}\right)'$. Under the null hypothesis, $\bar{\boldsymbol{U}}' V^- \bar{\boldsymbol{U}}$ (from now on "S06") is approximately χ^2 with $k-1$ degrees of freedom. The simulation results reported below for this proposal are based on $M = 20$. We obtained similar results for larger values of M. We note that S06 slightly differs from the statistic considered in Sun (2006) but this is a minor change done for practical purposes, and the statistic considered here is common in multiple imputation approaches (Pan, 2000).

The censoring mechanism for T has been simulated mimicking a longitudinal study where there is a periodical follow-up with scheduled visits, see Schick & Yu's model (2000). Specifically, for an individual i, we consider a set of examination times $\{Y_{ai}, a = 1, \dots, \tau_i\}$ which are the sum of the 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 $\tau_i = \sup\{a \geq 1 : \sum_{b=1}^{a-1} \xi_{bi} \leq \tau\}$ where τ represents the length of the study. The parameters μ and τ control 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 $\tau = 14$. We did compare the three methods (OG, S05, and S06) for different values of τ and their behaviour was similar to the results reported below for $\tau = 14$. All the scenarios are based on 1,000 replications. The resulting average percentage of right-censored observations is about 30% for those scenarios in Subsection 6.1 and 20% for those in Section 6.2.

We simulated a large number of scenarios where the null hypothesis was true and in all cases the nominal significance level $\alpha = 0.05$ was roughly reached by the statistics OG and S05. The empirical significance level was near 0.05 for scenarios with small sample sizes (for instance scenarios of two or three groups and sample size 50 for each group) and for large percentages of right-censored data (from 30% to 50%). However, in most of these scenarios the empirical significance level of the statistic S06 was clearly below the nominal level of 0.05. We do not present the results here but they can be provided upon request.

6.1. Accelerated Failure Time Models

Since the OG family is a class of score statistics only for discrete data, in this section we study whether the OG statistics perform as efficient score tests when both the failure and the censoring times are continuous. We have simulated T from an accelerated failure time (AFT) model with error distribution $S_\epsilon(t)$ holding Equation (12). The survival function for each group is given by

$$S^{(1)}(t) = S_\epsilon\left(\frac{\log(t) - \log(\alpha_1)}{\beta}\right) \text{ and } S^{(2)}(t) = S_\epsilon\left(\frac{\log(t) - \log(\alpha_2)}{\beta}\right)$$

where $\frac{S'_\epsilon(S_\epsilon^{-1}(t)))}{t} = -B(1-t; \lambda+1, \rho)$ and $S_\epsilon(0) = \frac{1}{2}$. This last condition ensures that in all the scenarios the two groups have medians equal to α_1 and α_2, respectively. We have generated 10 different scenarios for several values of (ρ, λ) which we describe next.

- Scenario 1 is a proportional hazards model with $\rho = 0, \lambda = 0$ ($\alpha_1 = 5.8, \alpha_2 = 8.8$ and $\beta = 0.9$).
- In Scenarios 2, 3, and 4 we have considered $\lambda = 0$ and $\rho = 1, 2, 3$ ($\alpha_1 = 5.5, 5.7, 5.7, \alpha_2 = 8.5, 8.5, 8.3$ and $\beta = 0.6, 0.4, 0.3$, respectively); we remark that Scenario 2 is a proportional odds model.
- In Scenarios 5, 6, and 7 we have considered $\rho = \lambda = 1, 2, 3$ ($\alpha_1 = 5.2, 5.5, 5.1$, $\alpha_2 = 8.4, 8.5, 8.3$ and $\beta = 0.2, 0.04, 0.01$, respectively).
- Finally, in Scenarios 8, 9, and 10 we have considered $\rho = 0$ and $\lambda = 1, 2, 3$ ($\alpha_1 = 9.1, 9.3, 9.4$, $\alpha_2 = 10.4, 10.2, 10.1$ and $\beta = 0.14, 0.06, 0.03$, respectively).

Table 1 gives the empirical powers for several parameterizations of the three proposals OG, S05, and S06. We have considered two groups with sample sizes $n^{(1)} = n^{(2)} = 100$.

Table 1 shows that, indeed, our proposal (OG) performs as expected in accordance with an efficient score test statistic. That is, the highest powers are achieved in the diagonal of the table (when the parameterization (ρ, λ) of the test statistics coincides with the parameterization (ρ, λ) of the AFT model). When we compare OG with S06 (Sun, 2006) and S05 (Sun, Zhao, & Zhao, 2005), the statistic OG achieves the highest power in all scenarios except in Scenario 8. Test statistics OG and S05 behave similarly in Scenarios 8, 9, and 10 when $\rho = 0$ and λ is equal to 1, 2, and 3, respectively. This behaviour is confirmed in other simulation results not presented here.

We also observe that the power of OG decreases when the test parameters and the AFT model parameters move away. For instance, under a proportional hazards model (Scenario 1), the power of the test statistics decreases as ρ or λ increases and the weight function is emphasizing early, middle or late hazards differences. Under a proportional odds model (Scenario 2) the hazard functions show early differences and 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.

6.2. Piecewise Exponential Models

In this section the goal is to validate the interpretation of the weight function in the OG class and to introduce a guideline to choose adequate parameters. We have simulated scenarios where the two hazard functions for each group differ in three different ways: An *Early times situation* where the hazards differ at early times, a *Middle times situation* for hazards differing at times around the median and a *Late times situation* for late time hazard differences. In order to reproduce these situations, T has been simulated from a piecewise constant hazard function as follows: for a fixed set of points $0 = x_0 < x_1 < \cdots < x_b < x_{b+1} = +\infty$, the hazard function for group $G^{(j)}$ ($j = 1, 2$) is constant within the interval $[x_{a-1}, x_a)$, that is, $h^{(j)}(t) = h_a^{(j)}$ when $x_{a-1} \le t < x_a$ ($a = 1, \ldots, b+1$). In all the scenarios we have applied a parameter configuration such that the median of the pooled sample (Me) is 5, that is, $S(5) = \frac{1}{2}$ where $S(t) = \frac{1}{2}S^{(1)}(t) + \frac{1}{2}S^{(2)}(t)$. In the sequel, the jth decile of the pooled survival function is denoted by D_j. We have considered sample sizes $n^{(1)} = n^{(2)} = 150$.

For the *Early times situation* we take $b = 1$ and consider hazard differences ($h_1^{(1)} \neq h_1^{(2)}$) until the point x_1 and equal hazards ($h_2^{(1)} = h_2^{(2)}$) thereafter. For the *Middle times situation* we take $b = 2$ and consider only differences between x_1 and x_2 ($h_2^{(1)} \neq h_2^{(2)}$ and $h_1^{(1)} = h_1^{(2)} = h_3^{(1)} = h_3^{(2)}$). We distinguish between symmetric and asymmetric middle scenarios. We remark that the degree of symmetry does not depend on the distances $x_1 - Me$ and $Me - x_2$ ($Me = 5$), but it depends on the distances $S(x_1) - 0.5$ and $0.5 - S(x_2)$. For the *Late times situation* we take $b = 1$ and consider

TABLE 1: Empirical powers for the three (ρ, λ)-families under accelerated failure time models: OG stands for the class presented in Section 3, S05 stands for the class in Sun, Zhao, & Zhao (2005) and S06 stands for the class in Sun (2006).

Scenarios			(ρ, λ)									
No.	(ρ, λ)	Method	(0, 0)	(1, 0)	(2, 0)	(3, 0)	(1, 1)	(2, 2)	(3, 3)	(0, 1)	(0, 2)	(0, 3)
1	(0, 0)	OG	0.842	0.779	0.674	0.589	0.805	0.746	0.710	0.725	0.602	0.494
		S05	0.842	0.541	0.233	0.125	0.707	0.505	0.351	0.753	0.656	0.547
		S06	0.835	0.746	0.614	0.513	0.761	0.696	0.654	0.710	0.603	0.503
2	(1, 0)	OG	0.762	0.829	0.820	0.777	0.660	0.612	0.571	0.470	0.266	0.172
		S05	0.762	0.760	0.501	0.341	0.743	0.643	0.513	0.519	0.317	0.211
		S06	0.756	0.794	0.746	0.686	0.599	0.525	0.473	0.438	0.262	0.170
3	(2, 0)	OG	0.714	0.829	0.838	0.825	0.530	0.447	0.401	0.290	0.156	0.100
		S05	0.714	0.812	0.676	0.504	0.707	0.605	0.526	0.346	0.175	0.124
		S06	0.708	0.786	0.775	0.737	0.452	0.355	0.296	0.261	0.147	0.096
4	(3, 0)	OG	0.624	0.808	0.851	0.866	0.401	0.312	0.274	0.209	0.097	0.077
		S05	0.624	0.851	0.785	0.668	0.646	0.551	0.466	0.247	0.126	0.080
		S06	0.619	0.772	0.793	0.769	0.323	0.231	0.180	0.189	0.094	0.073
5	(1, 1)	OG	0.808	0.705	0.549	0.420	0.835	0.833	0.821	0.774	0.644	0.523
		S05	0.808	0.377	0.095	0.040	0.741	0.572	0.427	0.803	0.705	0.593
		S06	0.806	0.671	0.480	0.328	0.815	0.802	0.772	0.761	0.642	0.523
6	(2, 2)	OG	0.761	0.619	0.429	0.305	0.816	0.812	0.806	0.731	0.568	0.445
		S05	0.761	0.276	0.059	0.054	0.705	0.551	0.427	0.765	0.645	0.531
		S06	0.750	0.584	0.358	0.216	0.781	0.777	0.763	0.720	0.581	0.452
7	(3, 3)	OG	0.712	0.563	0.383	0.258	0.797	0.821	0.827	0.706	0.571	0.425
		S05	0.712	0.227	0.058	0.063	0.688	0.554	0.432	0.737	0.638	0.522
		S06	0.714	0.529	0.309	0.164	0.775	0.784	0.780	0.698	0.570	0.432
8	(0, 1)	OG	0.759	0.404	0.209	0.130	0.770	0.746	0.718	0.873	0.855	0.791
		S05	0.759	0.097	0.049	0.068	0.367	0.187	0.135	0.867	0.877	0.847
		S06	0.672	0.350	0.174	0.103	0.700	0.659	0.607	0.823	0.793	0.734
9	(0, 2)	OG	0.588	0.237	0.106	0.083	0.602	0.573	0.546	0.824	0.855	0.826
		S05	0.588	0.058	0.058	0.067	0.198	0.107	0.068	0.804	0.854	0.853
		S06	0.499	0.193	0.087	0.060	0.512	0.492	0.461	0.750	0.797	0.778
10	(0, 3)	OG	0.418	0.160	0.082	0.064	0.454	0.430	0.398	0.760	0.827	0.833
		S05	0.418	0.064	0.076	0.090	0.126	0.069	0.056	0.699	0.802	0.830
		S06	0.335	0.131	0.070	0.049	0.388	0.359	0.345	0.646	0.759	0.792

The scenarios are based on the parameters (ρ, λ) of the error distribution. The results are based on two groups with sample sizes $n^{(1)} = n^{(2)} = 100$ and 1,000 replications.

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

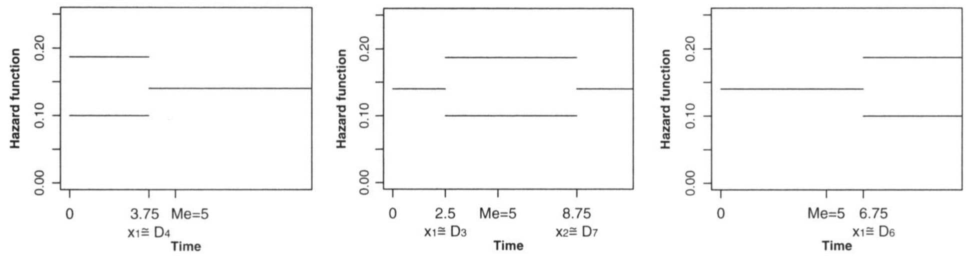

FIGURE 1: Early, middle and late hazard differences in the two sample piecewise exponential models. The values x_1 and x_2 denote the cut points. D_j is the jth decile of the pooled survival function $S(t) = \frac{1}{2}S^{(1)}(t) + \frac{1}{2}S^{(2)}(t)$ ($Me = D_5 = 5$).

equal hazards ($h_1^{(1)} = h_1^{(2)}$) until x_1 and hazard differences ($h_2^{(1)} \neq h_2^{(2)}$) thereafter. Figure 1 gives the hazards configuration in these situations.

The simulation study has allowed us to learn about the behaviour of the OG family under different alternatives. Table 2 presents some of the results. We now summarize and highlight the main findings:

- In scenarios with early hazard differences, the highest powers of the OG statistic are reached with $\lambda = 0$. When $x_1 < Me$ (*Pure early* scenario), the highest power of OG is achieved with $\rho = 2$. Whenever $x_1 > Me$, that is, the early pattern is mixed with middle and late hazard differences (*Mixed early* scenario), the highest power of OG is achieved with ρ equal to 1 and the power decreases as ρ increases.
- In scenarios with middle hazard differences, the highest powers of the OG statistic are reached with the parameters ρ and λ being equal. In general, the power of OG is higher for larger values of $\rho = \lambda$ and a value of ρ close to 3 is enough to reach a substantial improvement of the power.
- In scenarios with late hazard differences, the highest powers of the OG statistic are reached with $\rho = 0$. When $x_1 > Me$ (*Pure late* scenario), the highest power of OG is achieved with $\lambda = 3$. When the late pattern is mixed with middle and early hazard differences (*Mixed late* scenario), the highest power of OG is achieved with λ equal to 1 and the power decreases as λ increases.

In view of these findings, we recommend to use the family of statistics OG with parameterizations $(\rho, 0)$, (ρ, ρ) or $(0, \lambda)$ for testing against alternatives which set out early, middle or late hazard differences, respectively.

Finally, we have compared the behaviour of the three existing statistics OG, S05, and S06. Note that in Sun, Zhao, & Zhao (2005) the authors do not discuss the behaviour of their test statistics in terms of the hazard differences. Anyhow, for early hazard differences the statistic OG performs noticeably better than S05 when $\rho \neq 0$ and $\lambda = 0$ (we observe that in 9 of the 10 cases the power of OG is higher). Likewise, for middle hazard differences the statistic OG also performs noticeably better that S05 when $\rho = \lambda \neq 0$ (we observe that in 13 of the 15 cases the power of OG is higher). For late hazard differences, the statistic S05 performs slightly better than OG when $\rho = 0$ and $\lambda \neq 0$ (we observe that in 6 of the 10 cases the power of S05 is higher). Our simulation study reveals, as well, that, as compared to S06, OG has higher power for early and middle scenarios and similar power for late scenarios.

7. 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

TABLE 2: Empirical powers for the three (ρ, λ)-families under piecewise exponential models: OG stands for the class presented in Section 3, S05 stands for the class in Sun, Zhao, & Zhao (2005) and S06 stands for the class in Sun (2006).

		(ρ, λ)							
Early scenarios	Method	(0, 0)	(1, 0)	(2, 0)	(3, 0)	(4, 0)	(5, 0)	(1, 1)	(0, 1)
Pure early	OG	0.631	0.805	0.835	0.819	0.812	0.800	0.456	0.210
$x_1 = 3.75 \approx D_4$	S05	0.631	0.829	0.751	0.577	0.401	0.281	0.696	0.258
	S06	0.629	0.744	0.754	0.724	0.681	0.619	0.351	0.183
Mixed early	OG	0.933	0.965	0.962	0.942	0.915	0.884	0.898	0.664
$x_1 = 6.75 \approx D_6$	S05	0.933	0.943	0.705	0.393	0.233	0.142	0.962	0.742
	S06	0.929	0.954	0.931	0.876	0.801	0.723	0.859	0.637
Middle scenarios	Method	(0, 0)	(1, 1)	(2, 2)	(3, 3)	(4, 4)	(5, 5)	(1, 0)	(0, 1)
Left asymmetric	OG	0.461	0.557	0.600	0.626	0.638	0.643	0.393	0.432
$x_1 = 2.5 \approx D_3$	S05	0.461	0.574	0.556	0.516	0.485	0.438	0.203	0.465
$x_2 = 6.75 \approx D_6$	S06	0.440	0.512	0.539	0.533	0.524	0.526	0.332	0.411
Symmetric	OG	0.669	0.802	0.830	0.852	0.859	0.855	0.525	0.676
$x_1 = 2.5 \approx D_3$	S05	0.669	0.731	0.639	0.526	0.426	0.341	0.203	0.717
$x_2 = 8.75 \approx D_7$	S06	0.654	0.783	0.803	0.801	0.797	0.787	0.458	0.670
Right asymmetric	OG	0.420	0.522	0.580	0.604	0.622	0.628	0.278	0.500
$x_1 = 3.75 \approx D_4$	S05	0.420	0.382	0.272	0.197	0.150	0.126	0.092	0.515
$x_2 = 8.75 \approx D_7$	S06	0.391	0.505	0.536	0.546	0.539	0.539	0.217	0.496
Late scenarios	Method	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(1, 1)	(1, 0)
Pure late	OG	0.238	0.440	0.534	0.561	0.553	0.541	0.242	0.098
$x_1 = 6.75 \approx D_6$	S05	0.238	0.403	0.497	0.548	0.565	0.565	0.071	0.045
	S06	0.213	0.427	0.525	0.549	0.539	0.507	0.230	0.072
Mixed late	OG	0.662	0.838	0.830	0.795	0.746	0.690	0.756	0.355
$x_1 = 3.75 \approx D_4$	S05	0.662	0.832	0.848	0.814	0.787	0.744	0.406	0.072
	S06	0.634	0.837	0.815	0.780	0.730	0.667	0.740	0.302

Early scenarios are based on $b = 1$, $h_1^{(1)} = 0.1$, $h_1^{(2)} = 0.1866$ and $h_2^{(1)} = h_2^{(2)} = 0.1352, 0.1341$ respectively; middle scenarios are based on $b = 2$, $h_2^{(1)} = 0.1$, $h_2^{(2)} = 0.1866$ and $h_1^{(1)} = h_1^{(2)} = h_3^{(1)} = h_3^{(2)} = 0.1363, 0.1363, 0.1375$ respectively; late scenarios are based on $b = 1$, $h_2^{(1)} = 0.1$, $h_2^{(2)} = 0.1866$ and $h_1^{(1)} = h_1^{(2)} = 0.1386, 0.1375$ respectively. The results are based on two groups with sample sizes $n^{(1)} = n^{(2)} = 150$ and 1000 replications.

[a] Remark: D_j is the jth decile of the pooled survival function $S(t) = \frac{1}{2}S^{(1)}(t) + \frac{1}{2}S^{(2)}(t)$.

and November 1997. The details of 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 1981, Period 2 (P2) includes IDU patients who started the addiction between

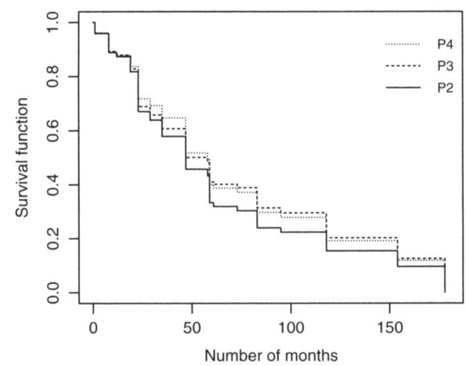

FIGURE 2: Elapsed time to seroconversion since starting intravenous drug use for men entering at risk either in calendar period P2 ($n^{(1)} = 300$), P3 ($n^{(2)} = 240$), or P4 ($n^{(3)} = 73$).

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 in 1992 or later. 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 focus on men and compare the elapsed time to seroconversion for periods P2, P3, and P4. The estimates of the survival functions given in Equation (4) have been applied to these data and the plot is shown in Figure 2. The different calendar periods represent different stages in the individuals' knowledge of the HIV epidemic as well as different health policies. The effect of the patients' behaviour as a result of the health policies is not expected to be seen at earlier times after the initiation of intravenous drugs but rather later on. For this reason we are particularly interested in exploring the differences after 3 or 4 years, hence in detecting middle hazard differences. To this end we use the statistic $\boldsymbol{U}$ and we choose parameters $\rho = 3$ and $\lambda = 3$. We obtain $\boldsymbol{U} = (0.22, -0.19, -0.03)$ with P-value 0.022. We remark here that the statistic S05 with the same parameters $\rho = 3$ and $\lambda = 3$ would yield a P-value of 0.163 and would not have shown the marked differences. Also a log-rank statistic, with P-value 0.089, would not have shown the differences.

The second analysis takes into account the age at which patients started to use drugs, since it is very likely that this could be a risk factor for HIV infection. We centre this analysis in period P3. The median and mean age for starting IDU in this period is 20 and 20.8 years, respectively. According to these measures, 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. Due to sexual risk habits, the younger group is expected to show differences from the older group at earlier times after the initiation of intravenous drugs. For this reason we choose parameters $\rho = 1$ and $\lambda = 0$ which emphasize early hazard differences. We obtain $\boldsymbol{U} = (8.36, -8.36)$ and a statistically significant difference with P-value 0.043. In this case the statistic S05 with parameters $\rho = 1$ and $\lambda = 0$ and the log-rank would yield a P-value equal to 0.035 and 0.061, respectively. The statistic S05 would therefore identify differences between the two groups while the log-rank would not.

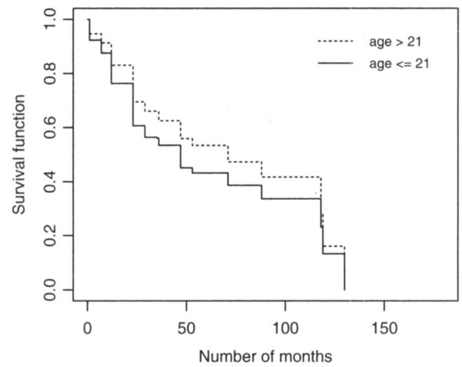

FIGURE 3: Elapsed time to seroconversion since starting intravenous drug use for individuals younger than 21 ($n^{(1)} = 192$) and individuals older than 21 ($n^{(2)} = 114$) entering at risk in calendar period P3.

8. DISCUSSION

This paper proposes a new class of test statistics for interval-censored data which extends the $G^{\rho,\lambda}$ family given in Fleming & Harrington (1991) for right-censored data. The proposed weighted log-rank test corresponds to the efficient score test if the data are discrete and performs in accordance with an efficient score test statistic for continuous data. The weight functions for the proposed family can be chosen to emphasize either early, middle or late hazards differences in an equivalent way as the weights of the Fleming and Harrington family do for right-censored data. Furthermore, a strategy is developed to choose the parameters ρ and λ of the family for each different situation.

In this paper we handle right-censored observations as interval-censored observations with right endpoints being equal to $+\infty$. Other methodologies like Sun (2006)'s family distinguish between interval-censored and right-censored observations and define a statistic which can be accommodated by the weighted log-rank statistic in the case of right-censored data. The question here is whether there is any benefit in treating differently interval and right-censored data. Our simulation study does not show this possible benefit.

Finally, the simulation study demonstrates that our proposed family of test statistics has a better power behaviour than the families of Sun, Zhao, & Zhao (2005) and Sun (2006), specially when there are early and middle hazard differences.

ACKNOWLEDGEMENTS

The authors are grateful to the GRASS group for their fruitful discussions. This research was partially supported by Grant MTM2008-06747-C02-00 from the Ministerio de Ciencia e Innovación. Part of this paper has been written under the support of Grant AI24643 from the National Institute of Allergy and Infectious Diseases. We thank the reviewers of this paper for their useful comments and suggestions resulting in an improvement of the manuscript.

BIBLIOGRAPHY

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. & 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. & Harringon, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.

Gómez, G., Calle, M. L., Egea, J. M., & 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.

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

Gómez, G., Calle, M. L., Oller, R., & Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9, 259–297.

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

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

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed., Wiley, New York.

Lawless, J. F. (2004). A note on interval-censored lifetime data and the constant-sum condition of Oller, Gómez, & Calle (2004). *The Canadian Journal of Statistics*, 32, 327–331.

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

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

Pan, W. (2000). A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, 19, 1–11.

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

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

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

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

Sen, P. K. (2006). Permutational Central Limit Theorems. *Encyclopedia of Statistical Sciences*, Wiley, New York.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

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

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

APPENDIX

Proof of Proposition 1. We prove only the result given in *(b)*. The remaining results are almost immediate.

Replacing the weight function $w(t) = \hat{S}(t-)\frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)}$ in Equation (9) gives

$$\begin{aligned}
c_i &= \int_0^{+\infty} \frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)} \left[\hat{S}_i(t-)\mathrm{d}\hat{S}(t) - \hat{S}(t-)\mathrm{d}\hat{S}_i(t)\right] = \int_0^{+\infty} \frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)} \left[\hat{S}_i(t-)\hat{S}(t) - \hat{S}(t-)\hat{S}_i(t)\right] \\
&= \int_0^{+\infty} \frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)} \left[\hat{S}_i(t-)\hat{S}(t) - \hat{S}_i(t)\hat{S}(t) + \hat{S}_i(t)\hat{S}(t) - \hat{S}(t-)\hat{S}_i(t)\right] \\
&= \int_0^{+\infty} \frac{\mathrm{d}\gamma(\hat{S}(t))}{\mathrm{d}\hat{S}(t)} \left[\hat{S}_i(t)\mathrm{d}\hat{S}(t) - \hat{S}(t)\mathrm{d}\hat{S}_i(t)\right].
\end{aligned}$$

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

$$\begin{aligned}
c_i &= \int_0^{l_i} \mathrm{d}\gamma(\hat{S}(t)) + \int_{l_i}^{r_i} \frac{\hat{S}(t) - \hat{S}(r_i)}{\hat{S}(l_i) - \hat{S}(r_i)} \mathrm{d}\gamma(\hat{S}(t)) - \int_{l_i}^{r_i} \frac{\hat{S}(t)}{\hat{S}(l_i) - \hat{S}(r_i)}\, \mathrm{d}\gamma(\hat{S}(t)) \\
&= \gamma(\hat{S}(l_i)) - \gamma(1) - \frac{\hat{S}(r_i)\{\gamma(\hat{S}(r_i)) - \gamma(\hat{S}(l_i))\}}{\hat{S}(l_i) - \hat{S}(r_i)} = \frac{\hat{S}(l_i)\gamma(\hat{S}(l_i)) - \hat{S}(r_i)\gamma(\hat{S}(r_i))}{\hat{S}(l_i) - \hat{S}(r_i)}.
\end{aligned}$$

This completes the proof of *(b)*. ■

Received 03 February 2011
Accepted 03 March 2012