Statistical Modelling 2009; **9**(**4**): 259–297

Tutorial on methods for interval-censored data and their implementation in R

Guadalupe Gómez[1], M Luz Calle[2], Ramon Oller[3] and Klaus Langohr[1,4]
[1]Departament d'Estadística i I.O., Universitat Politècnica de Catalunya, Spain
[2]Departament de Biologia de Sistemes, Universitat de Vic, Vic, Spain
[3]Departament d'Economia, Matemàtica i Informàtica, Universitat de Vic, Vic, Spain
[4]Programa de Recerca en Neuropsicofarmacologia, Institut Municipal d'Investigació Mèdica, Spain

Abstract: Interval censoring is encountered in many practical situations when the event of interest cannot be observed and it is only known to have occurred within a time window. The theory for the analysis of interval-censored data has been developed over the past three decades and several reviews have been written. However, it is still a common practice in medical and reliability studies to simplify the interval censoring structure of the data into a more standard right censoring situation by, for instance, imputing the midpoint of the censoring interval. The availability of software for right censoring might well be the main reason for this simplifying practice. In contrast, several methods have been developed to deal with interval-censored data and the corresponding algorithms to make the procedures feasible are scattered across the statistical software or remain behind the personal computers of many researchers. The purpose of this tutorial is to present, in a pedagogical and unified manner, the methodology and the available software for analyzing interval-censored data. The paper covers frequentist non-parametric, parametric and semiparametric estimating approaches, non-parametric tests for comparing survival curves and a section on simulation of interval-censored data. The methods and the software are described using the data from a dental study.

Key words: interval censoring; simulation; software R; survival analysis

Received November 2008; **revised** February 2009; **accepted** February 2009

1 Introduction

Most statistical methods developed for the analysis of survival data assume that the event which defines the start of the survival is known and allows the event $\mathcal{E}$

Address for correspondence: Guadalupe Gómez, Departament d'Estadística i I.O., Universitat Politècnica de Catalunya, Barcelona, Spain. E-mail: lupe.gomez@upc.edu

 10.1177/1471082X0900900402

that determines failure and hence the survival time, to be right censored. In many situations, however, the event of interest $\mathcal{E}$ cannot be observed and it is only known to have occurred within two times, say L and R. In this set-up, we say that the time T to $\mathcal{E}$ is interval censored.

Interval-censored data is an actively researched area which can be encountered nowadays in a large number of situations. More than 600 papers (according to the Web of Science's search on 17 October 2008) have been published since Peto's pioneer paper in 1973 analyzing girls' sexual maturity development. Among those, more than 150 have appeared in the last 3 years and although most of them may be classified under the area of Statistics and Probability, a great number belong to many other areas within medical studies such as Oncology, Immunology, Infectious Diseases, Transplantation, as well as in Mathematical and Computational Biology, Computer Science, Electrical Engineering, Environmental Sciences, Veterinary Sciences, Food Science, among others. These papers have focused on the different types of interval-censored mechanisms, in deriving theoretical properties for the survival function or hazard function estimators, in dealing with regression problems where either the response or one covariate is interval censored, and mainly in addressing scientific questions where one of the features in the data is their incompleteness due to an interval censoring mechanism.

The theory for the analysis of interval-censored data has been developed over the past three decades and several reviews have been written. The first two reviews written by Huang and Wellner (1997) and Lindsey and Ryan (1998) have been a keystone but are outdated by many of the interval-censored methods. Two recent reviews are by Gómez *et al.* (2004) and Lesaffre *et al.* (2005) who attempt to collect and unify methods of analysis for lifetime data when part of the data is interval censored. The former emphasizes the nature and the consequences of the censoring mechanism and states the assumptions on the inspection process so that the survival function is estimable, whereas the latter places emphasis on the use of accelerated failure time models. However, both of them need to be updated to include the major contributions on software of the last 4 years. The recent comprehensive book by Sun (2006) is intended for a technical audience with a thorough background in statistics and in computational methods. Sun's book addresses statistical issues and describes statistical methods for the analysis of singly and doubly interval-censored survival data arising from AIDS, cancer and other disease studies, though it lacks more detailed and hands-on examples with indications on how to use the available software to analyze interval-censored data. More details and extensions of what our paper covers can be found in the first six chapters of Sun's book.

The purpose of this tutorial is to present, in a pedagogical and unified manner, the methodology to analyze univariate interval-censored lifetime data together with the available software to do so. As already mentioned, the literature on interval-censored data is so extensive that a comprehensive description of all the methodologies is not feasible. Instead we have picked a number of methods based on the availability of free software for applying them to real data. The paper is organized as follows. This section is completed with the presentation of several motivating examples and the

dental study that will be used throughout the paper to illustrate the methodologies (Sections 1.1 and 1.2); the description of different formulations of interval censoring (Section 1.3); the notation is introduced in Section 1.4 together with a discussion of the non-informative and identifiability assumptions and, finally, Section 1.5 is devoted to the available software for interval censoring. Section 2 is devoted to non-parametric methods for interval censoring, describing in detail the expectation–maximization (EM) self-consistency algorithm and the software for the EM and EM–iterative convex minorant (ICM) algorithms. Section 3 presents a general framework of hypothesis testing for comparing survival curves extending the Fleming and Harrington family of tests. Section 4 includes parametric and semiparametric regression models. We discuss the log-linear representation of parametric survival models and the available methods for analyzing interval-censored data under a semiparametric proportional hazards model. The alternative approach based on a semiparametric accelerated failure time model is addressed by Komárek and Lesaffre (2009) in the paper following this one in this same volume. We have included a section on simulation (Section 5) where we discuss how to generate non-informative interval-censored data. Section 6 overviews, not exhaustively, several issues not considered in this paper. The paper closes with two appendices covering the available libraries in the R package to analyze interval-censored data and with a succinct review on commercial statistical software which can deal with interval-censored data.

1.1 Motivating illustrations

We have chosen five papers to exemplify different interval censoring mechanisms. One of the first applied papers to deal with the interval censoring problem was by De Gruttola and Lagakos (1989) who estimated the chronological time to HIV infection among haemophiliacs receiving contaminated blood factor between 1978 and 1988. The data were collected at the Hospitals Kremlin Bicêtres and Coeur des Yvelines in France where blood samples were periodically collected and stored and retrospectively tested to determine a time interval during which the infection occurred. The infection was only known to be between the times specified by the last negative and the first positive assessment, yielding, hence, interval-censored observations.

Interval-censored data are quite usual in longitudinal studies where subjects are not monitored continuously but scheduled to be inspected at certain times. In these cases, the time to the event of interest is observed within consecutive visits. One such study is described in Leung *et al.* (1997) when analyzing the time from employment to the development of asthmatic symptoms (wheezing and dyspnoea) among aluminum potroom workers from a respiratory health study in the Nordic countries between 1986 and 1989. The study endpoint here is only determined by periodical health examinations and hence is interval censored.

In a retrospective study aimed to estimate the elapsed time to HIV infection since entering the intravenous drug users risk group, Gómez *et al.* (2000) and Langohr *et al.* (2004) analyze the data on 306 intravenous drug users who were admitted to the

Hospital Germans Trias i Pujol in Badalona, Spain, for a desintoxication treatment between 1987 and 1995. The time to HIV infection since the start of injecting drugs is interval censored since for some patients we only have one date for the latest negative test and another in which she/he is HIV-positive, and these two dates define the interval where the HIV infection has occurred. For others, we have only one test indicating whether the patient is HIV-negative or HIV-positive and in those cases, a semi-interval of seroconversion can be defined.

Interval censoring is found, as well, in non-medical studies, for instance in a demographic panel survey on social, geographical and wealth mobility in the 19th and 20th centuries in France where the use of retrospective surveys and population registry data permits to estimate the distribution of migrations or job changes over time. Courgeau and Najim (1996) analyze these data taking into account their interval-censored nature due to the fact that residential or occupational mobility is only known between two censuses or family events.

In shelf-life studies, the probability of consumers accepting or rejecting a product beyond a certain storage time is of interest. Hough *et al.* (2003) use the storage times of the product together with whether the consumer accepts or rejects the product at this time to define censoring intervals based on which the distribution function for the time to rejection is estimated.

1.2 The Signal Tandmobiel® study

Throughout the tutorial, we will use the data from the Signal Tandmobiel® study. This is a longitudinal prospective oral health study conducted in Flanders (Belgium) from 1996 to 2001. A cohort of 4430 randomly sampled schoolchildren (2297 boys and 2133 girls) who attended the first year of the primary school at the beginning of the study was annually dental examined by one of 16 trained dentists. The original data set consists of at most six dental observations for each child including time of tooth emergence, caries experience and data on dietary and oral hygiene habits. For details of the study design and research methods see Vanobbergen *et al.* (2000). From a methodological viewpoint, this data set has already been presented in Bogaerts and Lesaffre (2004) and in Komárek and Lesaffre (2006, 2007, 2008), among others.

In this paper, we will restrict the analysis to the emergence time distribution of the permanent upper left first premolars (teeth 24 in European dental notation). We define T_{24} as the age (in years) of emergence of permanent tooth 24. Baseline information includes also `gender`: 0 = boy (52%), 1 = girl (48%) and `dmf` which stands for the status of the primary predecessor of this tooth: 0 (57%) if the primary predecessor was sound, 1 (43%) if it was decayed, missing due to caries or filled. For the purpose of the illustration, we have excluded 44 cases for which the variable `dmf` is missing. The age of emergence of permanent tooth 24 is both interval censored (2775, 63%) as well as right censored (1611, 37%), being the length of the intervals between 0.4 and 4.5 years with a mean time of 0.95 year. The distribution of the censoring times looks alike between boys and girls and whether the primary predecessor

of the permanent first premolar was sound or decayed. As suggested in Komárek and Lesaffre (2009), since permanent teeth do not emerge before the age of 5, the origin time for all analyses is set at 5 years and $T_i - 5$ is taken instead of T_i throughout.

We will describe the times to emergence by means of the non-parametric maximum likelihood estimator (NPMLE) of the distribution function in Section 2.4 and we will investigate whether the rate of emergence of these teeth is similar for boys and girls in Section 3.3. The question on the impact of gender and dmf on the emergence time of T_{24} will be answered both via a parametric regression analysis in Section 4.1.2 and under the assumptions of a proportional hazards model in Section 4.2.1.

The same data set is being analyzed in Komárek and Lesaffre (2009), who extend the analysis to include remaining permanent first premolars (teeth 14, 34, 44). In their paper, multivariate analyses by means of flexible accelerated failure time models are considered.

1.3 Interval censoring types

Different censoring mechanisms give rise to interval-censored data of varying nature and the methods and the theoretical developments behind these are also different and not necessarily interchangeable. Basically, we refer to interval-censored data to those situations where instead of observing the actual value of a random variable T, we only observe a window $(L, R]$ where T has occurred. The practical use of closed or half-open intervals depends on the way the observations were collected. The continuous nature of the variables would induce us to think that such a precision is not important, but in fact different interpretations lead to different likelihood functions, which in turn imply different non-parametric estimates (Ng, 2002) and the theoretical challenges posed by these types of censoring are of a very distinct nature. Turnbull's (1976) pioneer paper considered closed intervals. The advantage of closed intervals is that they incorporate both uncensored exact observations (when $L = R$) as well as group data which might appear when the inspection times are measured in a rather rough scale (for instance in years). However, since half-open intervals are more common and appear in situations where the individuals are inspected intermittently, we will mainly focus on the inference and analysis based on the observation of half-open intervals as in $(L, R]$. The methods explained in this tutorial could be easily adapted to closed intervals.

Since several types of interval-censored data can be seen in practice and their corresponding censoring mechanism is formalized in different ways, we next list the most relevant cases.

Case I interval-censored data or current status data: T is only known to be larger or smaller than an observed monitoring time, C. In this case, the study subject is observed only once producing either a left- or a right-censored observation. This type of data is encountered, for instance, in animal tumourigenicity experiments where animals die or are sacrificed at predetermined time intervals and are examined for the presence or absence of a tumour. If the tumours are irreversible, the observed

death times (natural and sacrifices) provide left- and right-censored observations on the time until tumour onset; that is, if the animal dies at time t and if the tumour is present we have the interval $(0, t]$, while if the tumour is absent we have the interval $(t, \infty]$ (Gómez and Van Ryzin, 1992).

Case II interval-censored data: In experiments with two monitoring times, U and V with $U < V$, the survival time of interest T is only known to be before the first monitoring time ($T \leq U$), between the two monitoring times ($U < T \leq V$), or after the second monitoring time ($T > V$). A situation that might be encountered sometimes is a mixed one where the observed intervals are complemented by some exact observation. Yu *et al.*, (2000) generalize the case II model, so that exact observations are as well allowed.

Case K interval-censored data: In longitudinal studies with periodic follow-up and K monitoring times $M_1, M_2, \cdots, M_K$, the event of interest is only observed between two consecutive inspecting times M_l, M_{l+1} and the observed data reduce to the interval $(M_l, M_{l+1}]$. This censoring scheme corresponds to a natural extension of case I and case II mechanisms and is discussed and extended in Schick and Yu (2000). These authors generalize the model so that the number of monitoring times K is random.

1.4 Notation, observable data, non-informativity and identifiability assumptions

We assume that the positive random variable T, the time until the event of interest $\mathcal{E}$, is governed by a right continuous distribution function $W(t) = \text{Prob}(T \leq t)$ with survival function $S(t)$ and support $\mathcal{S}_W = \{t \geq 0 : dW(t) > 0\}$.

A model for interval-censored data is described by the joint distribution $F_{L,R,T}$ between the random variable T and the observables (L, R), with range $\{(l, r, t) : 0 \leq l < t \leq r < \infty\}$, that is, under the constraint that $\text{Prob}(T \in (L, R]) = 1$. The marginal laws are denoted by $dW(t)$ and $dF_{L,R}(l, r)$, the latter represents the contribution to the likelihood of an individual with observed interval $(l, r]$ and is the proper basis for inference.

The main assumption we use throughout all the paper is that censoring occurs non-informatively in the following sense: the conditional distribution of T given L and R satisfies:

$$dF_{T|L,R}(t|l, r) = \frac{dW(t)}{P(T \in (l, r])} \mathbb{1}_{\{t \in (l,r]\}}(t), \tag{1.1}$$

that is, the only information provided by the censoring interval $(l, r]$ about the survival time t is that the interval contains t (Self and Grossman, 1986). The intuition behind the non-informative assumption is clarified by the equivalent notion of coarsening at random (Heitjan and Rubin, 1991) which establishes that given any two specific values t and t' of T consistent with the observables, they always provide the same information, that is, on the set $\{(l, r) : t \in (l, r] \text{ and } t' \in (l, r]\}$,

$dF_{L,R|T}(l,r|t) = dF_{L,R|T}(l,r|t')$. It is also worth noting that an alternative non-informative condition stating that the observables (l,r) are not influenced by the specific value of T in $(l,r]$, that is,

$$dF_{L,R|T}(l,r|t) = \frac{dF_{L,R}(l,r)}{P(T \in (l,r])} \mathbb{1}_{\{t \in (l,r]\}}(l,r), \tag{1.2}$$

is also equivalent to (1.1). The non-informative assumption is relevant and not always fulfilled. We refer the reader to Oller *et al.* (2004) for a sound explanation with illustrative examples of this concept.

The likelihood function is of paramount importance in making inferences, thus its appropriateness is worth a discussion. In a study with n subjects, their potential times to $\mathcal{E}$, say $T_1, T_2, \ldots, T_n$, are not observed and the observable data set is then $\mathcal{D} = \{(l_1, r_1], \ldots, (l_n, r_n]\}$. The contribution to the likelihood of the ith individual with observed interval $(l_i, r_i]$ is given by $dF_{L,R}(l_i, r_i) = \text{Prob}(L \in dl_i, R \in dr_i, T \in (l_i, r_i])$, hence the overall likelihood is given by

$$L_O\left(S(\cdot)|\mathcal{D}\right) = \prod_{i=1}^{n} dF_{L,R}(l_i, r_i) = \prod_{i=1}^{n} \text{Prob}(L \in dl_i,\ R \in dr_i,\ l_i < T \leq r_i).$$

However, the interval censoring problem has been generally treated via the so-called simplified likelihood function defined as

$$\mathcal{L}\left(S(\cdot)|\mathcal{D}\right) = \prod_{i=1}^{n} \int_{\{t:t \in (l_i, r_i]\}} dW(t) = \prod_{i=1}^{n} [S(l_i) - S(r_i)]. \tag{1.3}$$

This likelihood considers the observed intervals as being fixed in advance and ignores their randomness. There are two situations where the censoring process can be ignored and the simplified likelihood be used for inferences: the censorship model holds either the non-informative assumption or the constant-sum assumption (see Oller *et al.*, 2004). The non-informative assumption implies that the censoring process does not affect the survival process. For instance, in longitudinal studies with periodic follow-up, this assumption holds when the monitoring times are independent of T. The constant-sum property is a slightly weaker assumption and in this situation the censoring process may affect survival but does not alter the results of non-parametric inference.

Another important issue concerns the identifiability of the model. Given a censorship model $F_{T,L,R}$, we say that W is non-identifiable when there exists another censorship model having different lifetime distribution but sharing the same lifetime support $\mathcal{S}_W$ and the same distribution for the observables $F_{L,R}$. Oller *et al.* (2007) give a constructive way of obtaining censorship models with W being non-identifiable. The authors also show that to ensure complete identifiability of W, the constant-sum assumption is necessary but not sufficient. Under the constant-sum assumption, there are specific situations in which it is possible to ensure complete identifiability, for instance, when uncensored data are allowed for the whole support of the lifetime

variable, i.e., when $dF_{L,R}(t,t) > 0$ for any $t \in \mathcal{S}_W$. This identifiability assumption is rather mild and it is typically satisfied in applications with right-censored data. We could assure the identifiability of W, for instance, in those situations where each individual is inspected a countable number of times by means of an inspection process independent of T if the support of L or R covers $\mathcal{S}_W = (0, +\infty)$.

1.5 Software for interval censoring

Most of the available statistical software support survival analysis methods for right-censored data. However, few of them are ready to incorporate interval-censored data and those which are have different capabilities. As far as we know, SAS (SAS Institute Inc.) and STATA (StataCorp LP) incorporate functions to estimate the survival function non-parametrically and to run parametric regression analysis. Neither of these is ready to incorporate semiparametric and flexible accelerated failure time analysis. By contrast, the commercial statistical software SPSS (SPSS Inc.), in its present version 17, does not offer functions to deal with interval-censored data. We address the reader to Appendix B for a description of the capabilities of these commercial software packages.

The commercial software S-PLUS (TIBCO Software Inc.) and, especially, the free software R from the R Development Core Team (2008) are nowadays the most complete sources for survival analysis with interval-censored data. In recent years, the R software has become more and more popular among statisticians not only because of its free availability but also because it offers the same statistical procedures as any other commercial software. Many statisticians worldwide contribute continuously with specific packages to extend R's possibilities for statistical analyses. In Appendix A, different R packages to analyze interval-censored data are presented and several of the capabilities of S-PLUS are briefly described (see Appendix B.1).

For the reasons mentioned above, we have chosen to present and illustrate our analyses with R. These illustrations can be found at the end of each of the following sections and the reader is encouraged to reproduce them. To do so, one needs R and the data set (`tooth24.RData`) which can be downloaded from the following website: `http://www-eio.upc.es/grass/`. Knowledge on R is helpful but not indispensable and if it is not available yet on the computer, it can be installed following the instructions on the web pages of the R Project for Statistical Computing: `http://www.r-project.org/`. Once R is installed successfully, the data set can be opened by double clicking on `tooth24.RData` and the instructions can be copied from our illustrations.

2 Non-parametric estimation of the survival function

The most basic approach for analyzing interval-censored survival data is the non-parametric estimation of the survival function. Without the need of any modelling assumption, the estimated curves can be easily interpreted in a similar way as the

Kaplan–Meier curves for right censoring. This is usually the first analysis performed on a survival time with interval censoring and can be the basis for further parametric or semiparametric analysis.

In a study of n individuals with interval-censored observations and non-informative censoring, as defined in (1.1), inferences can be based on the above simplified likelihood function (see (1.3)), i.e.,

$$\mathcal{L}\left(S(\cdot)|\mathcal{D}\right) = \prod_{i=1}^{n} \text{Prob}\{l_i < T_i \leq r_i\} = \prod_{i=1}^{n} \left[S(l_i) - S(r_i)\right],$$

where $\mathcal{D} = \{(l_i, r_i], i = 1, \ldots, n\}$.

The goal is to find a monotonically decreasing function $\hat{S}_n(t)$ which maximizes the likelihood function $\mathcal{L}\left(S(\cdot)|\mathcal{D}\right)$. There are different algorithms for obtaining the NPMLE of the survival function under interval censoring which are described in Sections 2.1 and 2.2.

2.1 Self-consistency algorithm

One of the most popular methods to obtain a non-parametric estimator for the survival function under interval censoring is the use of self-consistency or Turnbull's algorithm (Turnbull, 1976). Turnbull extended the ideas of Peto (1973) to a more general problem, the analysis of arbitrarily grouped, censored and truncated data and proposed to obtain the non-parametric estimator of the survival function through the self-consistency equations.

The search of the NPMLE of the survival function under interval censoring requires the definition of a set of intervals, the so-called Turnbull intervals, denoted by $\mathcal{I} = \{(q_1, p_1], (q_2, p_2], \ldots, (q_m, p_m]\}$. These intervals are obtained from the set of all left and right interval endpoints in such a way that q_j is a left endpoint, p_j is a right endpoint and there is no other left or right endpoint between q_j and p_j. Turnbull proved that a maximum likelihood estimator of the survival function under interval censoring concentrates its mass on this set of intervals. Specifically, he stated that the search of the non-parametric MLE of S should be performed within the class of survival curves which are constant outside the set of Turnbull intervals and that the estimated survival curve is unspecified within each $(q_j, p_j]$.

From these results and denoting by $w_j = \text{Prob}\{q_j < T \leq p_j\} = S(q_j) - S(p_j)$ the weight of the jth Turnbull's interval, $j = 1, \ldots, m$, maximization of $\mathcal{L}\left(S(\cdot)|\mathcal{D}\right)$ for obtaining the NPMLE of the survival function reduces to maximization of the following likelihood function:

$$L_T(w_1, \ldots, w_m) = \prod_{i=1}^{n} \Big(\sum_{j=1}^{m} \alpha_j^i w_j \Big), \tag{2.1}$$

where $\alpha_j^i = \mathbb{1}\{(q_j, p_j] \subseteq (L_i, R_i]\}$ indicates whether or not the interval $(q_j, p_j]$ is contained in $(l_i, r_i]$ and the parameters are subject to the constraints $w_j \geq 0$ and $\sum_{j=1}^m w_j = 1$.

The NPMLE for the survival function is a decreasing step function with gaps inbetween corresponding to the unidentifiability of the function within each Turnbull's interval $(q_j, p_j]$. Specifically, the NPMLE for $S(t)$, based on $\mathcal{D}$, is given by

$$\hat{S}_n(t) = \begin{cases} 1, & \text{if} \quad t \leq q_1, \\ 1 - (\hat{w_1} + \cdots + \hat{w_j}), & \text{if} \quad p_j \leq t \leq q_{j+1}, \quad 1 \leq j \leq m-1, \\ 0, & \text{if} \quad t \geq p_m, \end{cases}$$

and is not specified within $(q_j, p_j]$, for $1 \leq j \leq m$. We illustrate the construction of Turnbull intervals and the likelihood function in the following example.

Example: Given a set of six individuals with censoring intervals $\{(l_i, r_i], 1 \leq i \leq 6\} = \{(4, 7], (3, 5], (0, 2], (1, 4], (6, 9], (8, 10]\}$, the corresponding Turnbull intervals are given by $\mathcal{I} = \{(q_1, p_1] = (1, 2], (q_2, p_2] = (3, 4], (q_3, p_3] = (4, 5], (q_4, p_4] = (6, 7], (q_5, p_5] = (8, 9]\}$. The likelihood corresponding to this data set is given by

$$\begin{aligned} L_T(w_1, w_2, w_3, w_4, w_5) &= \prod_{i=1}^6 \left[S(r_j) - S(l_j)\right] = \prod_{i=1}^6 \left(\sum_{j=1}^5 \alpha_j^i w_j\right) \\ &= (w_3 + w_4)(w_2 + w_3)(w_1)(w_1 + w_2)(w_4 + w_5)(w_5), \end{aligned}$$

and the point $(\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4, \hat{w}_5) = (1/3, 0, 1/3, 0, 1/3)$ is the maximizing solution.

Turnbull's approach for maximizing the likelihood $L_T(w_1, \ldots, w_m)$ is based on the solution of the self-consistent equations and is a special case of the EM algorithm. A self-consistent estimator of $(w_1, \ldots, w_m)$ can be obtained as the solution of the following simultaneous equations

$$\hat{w}_j = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_j^i}{\sum_{l=1}^m \alpha_l^i \hat{w}_l} \, \hat{w}_j, \quad 1 \leq j \leq m,$$

where the dependence on the sample size n of the weights $\hat{w}_j$ has been omitted for notational simplicity. Turnbull stated the equivalence between the self-consistent estimator and the NPMLE. Specifically, he proved that (i) any NPMLE $(\hat{w}_1, \ldots, \hat{w}_m)$ satisfies the self-consistent equations and (ii) conversely, the solution $(\hat{w}_1, \ldots, \hat{w}_m)$ of the self-consistent equations is the NPMLE of $(w_1, \ldots, w_m)$ provided that the Kuhn–Tucker conditions are fulfilled. The beauty of Turnbull's EM algorithm resides in its simplicity and intuitive behaviour; however, its convergence can be very slow for relatively large sample sizes. More efficient algorithms, such as the ICM proposed by Groeneboom and Wellner (1992) and the EM-ICM developed by Wellner and Zahn (1997), have been proposed as alternatives and are explained in Section 2.2. Another reformulation of the ICM is proposed by Pan (1999) and it is succinctly covered in 4.2.

2.2 ICM and EM–ICM algorithms

Groeneboom and Wellner (1992) proposed more efficient optimization techniques to obtain the NPMLE of the distribution function under interval censoring using an ICM algorithm which is a special case of the generalized gradient projection optimization method. This method combines a Newton–Raphson scheme with isotonic least squares regression that guarantees that the new estimate at each step of the iterative process is a proper distribution function. Wellner and Zhan (1997) proposed a hybrid algorithm that combines the self-consistency algorithm and the ICM algorithm, named EM–ICM. Each step in the EM–ICM algorithm performs the ICM algorithm and uses the current ICM estimate to obtain a new EM estimate. Both ICM and EM–ICM converge in fewer iterations than the EM algorithm, but this advantage does not always result in shorter computational times. The main advantage of ICM and EM–ICM is that their global convergence is guaranteed while the solution of the self-consistency algorithm is not necessarily the NPMLE.

2.3 Asymptotic behaviour of the NPMLE

The study of the asymptotic properties of the NPMLE of the survival function under interval censoring is much more complicated than under exact and right censoring due, on one hand, to the 'poor' amount of information that an interval could entail; second, because often the number of parameters to be estimated increases with sample size and hence standard likelihood theory fails, and third, because the NPMLE cannot be expressed in terms of a counting process and martingale theory cannot be applied. Groeneboom and Wellner (1992), under the assumptions that $S(t)$ is continuous and that the support of $S(t)$ is contained in the union of the supports of L and R, prove the uniform strong consistency of the NPMLE and show that the estimator of $S(t)$ obtained at the first step of the ICM algorithm converges to $S(t)$ at the $\sqrt{n \log n}$ rate and that its asymptotic distribution is not normal. The reader is addressed to Groeneboom and Wellner (1992), and the following papers of these authors, for a thorough discussion.

In what follows, we summarize the required main assumptions to assure uniform consistency and $\sqrt{n}$ convergence rate together with an asymptotical normal distribution. In the restricted, but very common case II situation, in which the inspection times are discrete random variables but the theoretical survival function $S(t)$ is continuous—this includes those studies with a periodic longitudinal follow-up and with a fixed number of scheduled visits—Yu *et al.* (1998) prove that:

(*i*) $\hat{S}_n(t)$ is strongly consistent at each point in the set $\mathcal{A}$ of all possible values of L and R. If $\mathcal{A}$ is finite or dense in $[0, \infty)$, then $\hat{S}_n(t)$ is uniformly strongly consistent, i.e., $\lim_{n\to\infty} \sup_{t\in\mathcal{A}} |\hat{S}_n(t) - S(t)| = 0$, almost surely.

(*ii*) If $\mathcal{A} = \{a_1, a_2, \cdots, a_m\}$ contains finitely many elements and $0 < S(b) < S(a) < 1$ for all $a, b \in \mathcal{A}$ such that $a < b$, then

$$\sqrt{n}\{\hat{S}_n(a_1) - S(a_1), \hat{S}_n(a_2) - S(a_2), \cdots, \hat{S}_n(a_m) - S(a_m)\}$$

is asymptotically normal with mean 0 and explicit asymptotic variance which can be consistently estimated by the observed information matrix.

(*iii*) Special consideration has to be given to the estimates of the parameters w_j in equation (2.1) since they could approach the boundary of the parameter space, that is, the value $\hat{w}_j$ could be close to zero. In this situation, the inverse of the matrix $I(\hat{w}_1, \ldots, \hat{w}_m)$ where $I(\hat{w}_1, \ldots, \hat{w}_m) = (-\partial^2 \log L_T(w_1, \ldots, w_m)/\partial(\hat{w}_1, \ldots, \hat{w}_m)\partial(\hat{w}_1, \ldots, \hat{w}_m)')$, cannot be obtained.

Whenever we have enough exact failure times together with censoring intervals, $S(t)$ is continuous and strictly decreasing and $\mathcal{A}$ contains finitely many elements, Huang (1999) proves that $\sqrt{n}\{\hat{S}_n(t_0) - S(t_0)\}$ converges to a Gaussian process in $\mathcal{D}[0, \infty)$ (the class of bounded right continuous functions with left limits on $[0, \infty)$, equipped with the supremum norm) with mean 0 and a variance which achieves the information lower bound for the estimation of $S(t)$. In general, in order that the NPMLE $\hat{S}_n(t)$ be uniformly consistent, the theoretical distribution function $S(t)$ has to be continuous and its support has to be contained in the support of the inspection times.

2.4 Illustration: Signal Tandmobiel® data

In this section, we illustrate the use of several functions of the R package `Icens` (see also Appendix A.1) to compute and plot the NPMLE for the survival and the distribution function based on the dental interval-censored data set presented in Section 1.2. Recall that we are interested in the emergence times of permanent tooth 24 in Belgian school children and whether they depend on sex and the status of the primary predecessor of that tooth. Since this tooth does not emerge before age 5, we set this age as the origin time.

Let us first have a look at our dataset stored in the dataframe `tooth24` which contains the data of a total of 4386 children. That is, for the following illustrations, we take into account only those children with complete information in all the variables considered.

```
> str(tooth24)
'data.frame':   4386 obs. of  5 variables:
 $ id   : num  1 2 3 4 5 6 7 8 9 10 ...
 $ left :Class 'labelled'  atomic [1:4386] 2.7 2.4 4.5 5.9 4.1 ...
  .. ..-attr(*, "label")= chr "Lower limit of tooth emergence"
  .. ..-attr(*, "units")= chr "Years since age 5"
 $ right:Class 'labelled'  atomic [1:4386] 3.5 3.4 5.5 999 5 4.5 ...
```

```
  .. ..-attr(*, "label")= chr "Upper limit of tooth emergence"
  .. ..-attr(*, "units")= chr "Years since age 5"
 $ sex  :Class 'labelled'  atomic [1:4386] 1 0 1 1 1 0 0 1 1 1 ...
  .. ..-attr(*, "label")= chr "Gender"
  .. ..-attr(*, "levels")= chr [1:2] "Boys" "Girls"
 $ dmf  :Class 'labelled'  atomic [1:4386] 1 1 0 0 1 1 1 1 1 0 ...
  .. ..-attr(*, "label")= chr "Status of primary predecessor"
  .. ..-attr(*, "levels")= chr [1:2] "Sound" "Decayed, missing,
 +                                                    or filled"
 -attr(*, "comment")= chr "Data on tooth 24 of 4386 children in
 +                                                      Flanders"
```

As the variable labels indicate, the variables `left` and `right` form the observed intervals which contain the unkown times of tooth emergence. Note that in the case of the 1611 right-censored observations, the value of `right` is set to an arbitrary value (999) beyond the maximum value of the variables `left` and `right`. The reason is that the functions of package `Icens` require real values for both the lower and upper limits of the interval-censored survival time. As mentioned in Section 1.2, 52% of the children are boys and the proportion of a sound primary predecessor is 57%.

In order to use the functions of package `Icens`, the package needs to be installed once in the computer and then has to be loaded at the beginning of every R session (see also Appendix A):

```
> install.packages("Icens")   # only necessary if not installed yet
> library(Icens)
```

The application of function `MLEintvl` yields Turnbull intervals within which the distribution function is a constant function. In the following, we show part of the output obtained, where 999, actually, stands for infinity. Note that `MLEintvl` is applied within function `with()` which permits to identify variables `left` and `right` of dataframe `tooth24`.

```
> with(tooth24,MLEintvl(cbind(left,right)))
      [,1]  [,2]
 [1,]  2.5   2.6
 [2,]  2.6   2.7
 [3,]  2.7   2.8
 [4,]  2.8   2.9
 [5,]  2.9   3.0
  :     :     :
[48,]  7.2   7.3
[49,]  7.3   7.4
[50,]  7.4 999.0
```

These intervals are useful to construct a matrix with the values of the NPMLE obtained by any of the available algorithms. To do so, we store the intervals in object `TBintvls`, apply both the EM–ICM and the EM algorithm to our data and assign their results to objects `EMICMest` and `EMest`, respectively:

```
> TBintvls<-with(tooth24,MLEintvl(cbind(left,right)))
> EMICMest<-with(tooth24,EMICM(cbind(left,right)))
> EMest<-with(tooth24,EM(cbind(left,right)))
Warning message:
In EM(cbind(left, right)) : EM may have failed to converge
```

Both functions return a list with several details of the NPMLE: whereas `EMICM` stores the NPMLE of the distribution function in the element `sigma`, the function `EM` stores the estimates of the probabilities $w_j = \text{Prob}(q_j < T \leq p_j)$ in the element `pf`. Hence, to obtain $\hat{W}_n(t) = 1 - \hat{S}_n(t)$ (see below the commands to plot the estimator of the survival function and an illustration in Figure 1), we have to apply the R function `cumsum` which calculates the cumulative sums for each element of a vector. Doing so we can see that both algorithms yield similar values despite the warning message which appeared applying function `EM`. To achieve convergence with that function, one can change, for example, the default value of 500 iterations with option `maxiter`.

```
> cbind(TBintvls,EMICM=round(EMICMest$sigma,4),
 +                EM=round(cumsum(EMest$pf),4))
                   EMICM      EM
 [1,]  2.5   2.6  0.0050  0.0050
 [2,]  2.6   2.7  0.0108  0.0108
 [3,]  2.7   2.8  0.0118  0.0118
 [4,]  2.8   2.9  0.0130  0.0130
 [5,]  2.9   3.0  0.0130  0.0130
  :     :     :    :       :
[29,]  5.3   5.4  0.4904  0.4901
[30,]  5.4   5.5  0.5121  0.5122
  :     :     :    :       :
[48,]  7.2   7.3  0.9453  0.9440
[49,]  7.3   7.4  0.9595  0.9610
[50,]  7.4 999.0  1.0000  1.0000
```

According to the output, the probability that tooth 24 has emerged by age 5 + 3 = 8 years is equal to 0.0130 and the estimated median time of tooth emergence is a value between 5 + 5.4 = 10.4 and 5 + 5.5 = 10.5 years. We can also see that the probability for tooth emergence before age 5 + 7.4 = 12.4 years is equal to 0.9595.

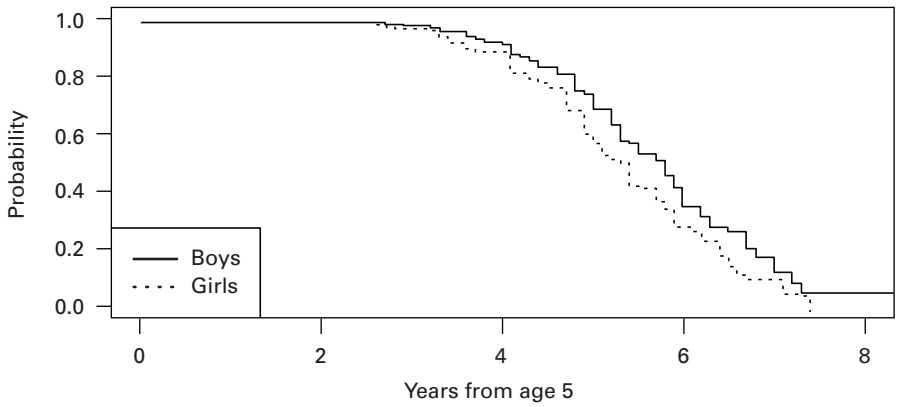

Figure 1 Survival function of emergence times of permanent tooth 24

To get a first idea of whether tooth 24 emerges earlier in boys or girls, we represent the NPMLE of the survival function $S(t)$ graphically for both sexes. Therefore, we apply function `EMICM` to both subsets and draw the resulting NPMLE with function `plot.icsurv`. Therein, we use the option `type="lc"` to assign the probability w_j of every Turnbull's interval $(q_j, p_j]$ to the left endpoint q_j and the option `surv=TRUE` to draw the survival function (instead of the distribution function). The resulting survival functions are defined for every $t > 0$. The result is shown in Figure 1 which shows earlier emergence times in girls compared to boys. In Section 4.1.2, we will check, by means of an accelerated failure time model, whether these differences do also hold when adjusting for variable dmf.

The commands to draw Figure 1 are the following:

```
> EMICMboys<-with(subset(tooth24,sex==0),EMICM(cbind(left,right)))
> EMICMgirls<-with(subset(tooth24,sex==1),EMICM(cbind(left,right)))
> x11(width=10,height=6)
> par(font=2,font.lab=2,font.axis=2,lwd=3,las=1,cex=1.5)
> plot.icsurv(EMICMboys,type="lc",surv=TRUE,xlim=c(0,8),main="",
+             xlab="Years from age 5")
> plot.icsurv(EMICMgirls,type="lc",surv=TRUE,lty=3,new=F)
> legend('bottomleft',c('Boys','Girls'),lty=c(1,3),lwd=3)
```

Note that function `x11()` opens a new graphical window; several graphical parameters are defined within function `par()`. We remark as well that while executing the previous command `plot.icsurv(EMICMgirls,···)` a straight line appeared in the top right corner of the plot. In Figure 1, we have, subsequently, deleted that line.

3 Comparison of survival curves

Some test statistics for comparing curves under interval censoring, which extend known test statistics for right-censored data, have been proposed; see Sun (2006) for a comprehensive exposition. They can be divided into rank-based and survival-based comparison procedures. Rank-based tests rely on the integrated weighted difference between the estimated hazard functions in the different groups and include extensions of the log-rank and the Wilcoxon tests and, more generally, of any test statistic in the class of weighted log-rank tests (Gómez and Oller, 2008; Huang *et al.*, 2008). These statistics are appropriate to detect ordered hazards alternatives and fail to detect crossing hazards. On the other hand, survival-based procedures rely on the integrated weighted differences between the estimated survival functions in the groups and are as well appropriate to detect ordered survivals but inappropriate to detect crossing survivals (Fang *et al.*, 2002; Lim and Sun, 2003; Yuen *et al.*, 2006).

Deriving the asymptotic behaviour of test statistics based on interval-censored data is more difficult than for right censoring because the counting process theory does not apply. This difficulty is faced up either with resampling methods or by standard inference methods which require additional assumptions for the censoring mechanism. Resampling methods include permutation, multiple imputation and bootstrap procedures. Fay and Shih (1998) and Gómez and Oller (2008) consider permutation tests, Pan (2000) and Huang *et al.* (2008) use multiple imputation, while Yuen *et al.* (2006) approach the inference problem via bootstrap methods. Methods using standard likelihood theory, as it has been already discussed in Section 2.3, require that the inspection times are discrete random variables and that the estimated parameters are not on the boundary of the parameter space (Fay, 1996, 1999; Gómez and Oller, 2008).

In this section, we focus on the new family of tests developed by Gómez and Oller (2008). This family includes the log-rank and the Wilcoxon–Peto test proposed by Peto and Peto (1972) and extends the unquestionable most popular Fleming and Harrington family of tests for right-censored data. We use a permutation approach, valid for either discrete or continuous data, instead of the Fisher's information-based approach.

3.1 Weighted log-rank tests

The k-sample comparison problem corresponds to testing the null hypothesis of identical survival functions between groups against the alternative of some groups having a different survival curve. That is, we test the hypothesis $H_0 : S_1 = \cdots = S_k$ against $H_a : S_l \neq S_{l'}$ for some $l \neq l'$, where $S_1, \ldots, S_k$ are the survival functions of T under each one of k groups $G_1, \ldots, G_k$. Let $N_1, \ldots, N_k$ be the sample sizes in each group ($n = N_1 + \cdots + N_k$) and let $\alpha_i^{(l)}$ be the indicator function ($i = 1, \ldots, n$ and $l = 1, \ldots, k$) which is equal to 1 if the individual i belongs to group G_l and 0 otherwise.

Let $\hat{S}(t)$ be Turnbull's estimator for the survival function based on the pooled sample for the k groups. We now define an estimate of the survival function for the ith individual, $\hat{S}^i(t)$, as the pooled survival $\hat{S}(t)$ truncated at the ith observed interval, as follows:

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

where $P_{\hat{S}}$ denotes the probability measure of T given by the pooled survival function $\hat{S}(t)$ and $l \vee r$ stands for the maximum value between l and r. These individual estimators satisfy $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}^i(t)$.

The average of these estimates for individuals in group G_l is an estimate of the survival function for each group and is denoted by

$$\hat{S}_{(l)}(t) = \frac{1}{N_l} \sum_{i=1}^{n} \alpha_i^{(l)} \hat{S}^i(t). \tag{3.1}$$

Recall that $\hat{S}(t)$ is unspecified in each Turnbull's interval $(q_1, p_1], \ldots, (q_m, p_m]$. In the sequel of this section, we identify each Turnbull's interval $(q_j, p_j]$ with the right endpoint p_j and assign the probability $P_{\hat{S}}\big((q_j, p_j]\big)$ to p_j. We also define $p_0 = 0$.

Let $d_j = n\big[\hat{S}(p_{j-1}) - \hat{S}(p_j)\big]$ be the estimated total number of events at p_j and $n_j = n\,\hat{S}(p_{j-1})$ be the estimated total number at risk just prior to p_j. Let $d_{lj} = N_l\big[\hat{S}_{(l)}(p_{j-1}) - \hat{S}_{(l)}(p_j)\big]$ and $n_{lj} = N_l\,\hat{S}_{(l)}(p_{j-1})$ be the corresponding estimates in group l. Then, as defined in Fay (1999), a weighted log-rank test statistic for interval-censored data takes the form

$$\boldsymbol{U} = \big(U_1, \ldots, U_k\big)', \tag{3.2}$$

where

$$U_l = \sum_{j=1}^{m} U_{lj} = \sum_{j=1}^{m} v_j \left[d_{lj} - \frac{n_{lj}}{n_j} d_j\right]$$

and v_j is a weighting function. As with the Fleming–Harrington family for right-censored data, the statistic $\boldsymbol{U}$ can be seen as a weighted sum of the differences between observed and expected deaths, $\boldsymbol{U} = \sum_{j=1}^{m} v_j \left[\boldsymbol{O}_j - \boldsymbol{E}_j\right]$, where $\boldsymbol{O}_j = \big(d_{1j}, \ldots, d_{kj}\big)$, $\boldsymbol{E}_j = \big(\frac{n_{1j}}{n_j} d_j, \ldots, \frac{n_{kj}}{n_j} d_j\big)$ and $d_j = \sum_{l=1}^{k} d_{lj}$ and $n_j = \sum_{l=1}^{k} n_{lj}$.

Several weighted log-rank statistics can be derived as score test statistics in discrete interval-censored data models. Under a discrete linear transformation model, Gómez and Oller (2008) obtain an extension of the $G^{\rho,\lambda}$ family originally proposed in

Fleming and Harrington (1991) for right-censored data. The weighting function is

$$v_j^{\rho,\lambda} = \hat{S}(p_{j-1})\,\frac{B(1-\hat{S}(p_j);\lambda+1,\rho) - B(1-\hat{S}(p_{j-1});\lambda+1,\rho)}{\hat{S}(p_{j-1}) - \hat{S}(p_j)},$$

where $B(t,a,b) = \int_0^t x^{a-1}(1-x)^{b-1}\,\boldsymbol{dx}$ is an incomplete beta function. When $\rho \to 0$ and $\lambda = 0$, or $\rho = 1$ and $\lambda = 0$, the test statistics reduce, respectively, to the extension of the log-rank and Wilcoxon test statistics given by Peto and Peto (1972). Since $v_j^{\rho,\lambda} \to \left(\hat{S}(p_{j-1})\right)^{\rho}\left(1-\hat{S}(p_{j-1})\right)^{\lambda}$ as $\frac{\hat{S}(p_{j-1})}{\hat{S}(p_j)} \to 1$, these weights have a similar interpretation as the weights in the original family. A straightforward manipulation gives

$$U_l = \sum_{j=1}^{m} v_j\, n_{lj}\, \left[\widehat{h}_{(l)}(p_j) - \widehat{h}(p_j)\right],$$

where $\widehat{h}_{(l)}(p_j) = \frac{d_{lj}}{n_{lj}}$ and $\widehat{h}(p_j) = \frac{d_j}{n_j}$ are estimates of the hazard functions under the lth group and the pooled sample, respectively. This result shows that $\boldsymbol{U}$ is an integrated weighted difference between estimated hazard functions in the groups. When $\lambda = 0$, early hazard differences are emphasized stronger as ρ increases. When $\rho = 0$, late hazard differences are emphasized stronger as λ increases. Finally, when $\rho = \lambda$, hazard differences around the overall median are emphasized stronger as ρ and λ increase.

3.2 Asymptotic permutation distribution

To determine an asymptotic distribution of the weighted log-rank statistic, the most straightforward way of doing so is to use a permutation approach. A permutation test applies for discrete as well as continuous interval-censored data. The main assumption which needs a permutation test is that the underlying censoring distribution is non-informative and identical across groups.

For the permutation approach, it is more convenient to write $\boldsymbol{U}$ in a linear form with a term for each individual. As shown in Gómez and Oller (2008), the $G^{\rho,\lambda}$ family of test statistics can be written as

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

where

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

and $\mathbf{z}_i = \left(\alpha_i^{(1)}, \alpha_i^{(2)}, \ldots, \alpha_i^{(k)}\right)'$ is a k-vector of group indicators associated with the ith observation. Note that this linear form of the statistic $\boldsymbol{U}$ only depends on the observed intervals $(l_i, r_i]$ and does not depend on the Turnbull's intervals $(q_j, p_j]$. This means that not specifying $\hat{S}(t)$ inside the Turnbull intervals does not have consequences in the computation of $\boldsymbol{U}$. With this form we do not need to assign the probability $P_{\hat{S}}\left((q_j, p_j]\right)$ to the right endpoint p_j.

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. This 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 allows a normal approximation with expectation $E(\boldsymbol{U}) = n\bar{c}\bar{\mathbf{z}}'$ and variance–covariance matrix $V_0 = \frac{1}{n-1}(\sum_{i=1}^n c_i^2 - n\bar{c}^2)(\sum_{i=1}^n (\mathbf{z}_i\mathbf{z}_i' - \bar{\mathbf{z}}\bar{\mathbf{z}}'))$. In our situation, $\bar{c} = 0$ and, consequently, $E(\boldsymbol{U}) = \mathbf{0}$. The permutation test is then based on the Mahalanobis distance $\boldsymbol{U}'V_0^-\boldsymbol{U}$, where V_0^- is the generalized inverse of V_0. We would reject the null hypothesis if $\boldsymbol{U}'V_0^-\boldsymbol{U}$ is an extreme value from a χ^2_{k-1} distribution.

3.3 Illustration: Signal Tandmobiel® data

As far as we know, there is no R package available which implements k-sample methods with interval-censored data. In this section, we describe in detail how to implement the methodology presented above. We use functions of the R package and the package `Icens` as well as the output of the function `PGM`.

The implementation of the weighted log-rank test statistics is illustrated analyzing the differences between boys (`sex==0`) and girls (`sex==1`) with respect to the emergence time of permanent tooth 24.

First, we show how to compute the survival functions $\hat{S}_{(1)}(t)$ and $\hat{S}_{(2)}(t)$ for boys and girls, respectively. To do that, we estimate the overall survival function by means of the function `PGM`.

```
> library(Icens)
> attach(tooth24)
> svf <- PGM(cbind(left,right))
```

The returned object `svf` contains several useful components. The component `svf$sigma` is the NPMLE of the cumulative distribution function $W(t)$. The component `svf$clmat` is a matrix with $m = 50$ rows and $n = 4386$ columns, which contains the indicators $\alpha_j^i = \mathbb{1}\{(q_j, p_j] \subseteq (L_i, R_i]\}$ $(i = 1, \ldots, 4386,\ j = 1, \ldots, 50)$ stating whether or not Turnbull's interval $(q_j, p_j]$ is contained in the observed interval $(l_i, r_i]$. Turnbull intervals are given in the matrix `svf$intmap` and their probabilities

$\hat{S}(q_j) - \hat{S}(p_j)$ in the vector `svf$pf`. From `svf$clmat` and `svf$pf`, we compute the probabilities $P_{\hat{S}^i}((q_j, p_j])$,

```
> p <- svf$pf*svf$clmat
> ptrunc <- t(t(p)/colSums(p))
```

The survival functions $\hat{S}_{(1)}(t)$ and $\hat{S}_{(2)}(t)$ given in equation (3.1) are then obtained as follows:

```
> pboys<-rowMeans(ptrunc[,sex==0])
> svfboys <- 1-cumsum(pboys)

> pgirls<-rowMeans(ptrunc[,sex==1])
> svfgirls <- 1-cumsum(pgirls)
```

A plot of these two survival functions can be generated by a generic function for plotting R objects. The survival plot for the emergence time of permanent tooth 24 for boys and girls is given in Figure 1.

Table 1 gives an outline of the required calculations in the weighted log-rank test statistic formulation given in (3.2). We consider the parameters ρ and λ equal to $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$. We only give the component U_1 corresponding to boys because in a two-sample problem $U_2 = -U_1$. The results show that: (i) when $(\rho, \lambda) = (0, 0)$ the differences $\boldsymbol{O}_j - \boldsymbol{E}_j$ are weighted similarly (see column $v_j^{0,0}$); (ii) when $(\rho, \lambda) = (1, 0)$ the weights decrease; (iii) when $(\rho, \lambda) = (0, 1)$ the weights increase; (iv) when $(\rho, \lambda) = (1, 1)$ the weights increase up to the interval $(5.4, 5.5]$ which contains the median and decrease afterwards.

Table 1 Calculation of the weighted log-rank statistic for Signal Tandmobiel® data

q_j	p_j	d_j	n_j	d_{1j}	n_{1j}	$v_j^{0,0}$	$v_j^{1,0}$	$v_j^{0,1}$	$v_j^{1,1}$	$U_{1j}^{0,0}$	$U_{1j}^{1,0}$	$U_{1j}^{0,1}$	$U_{1j}^{1,1}$
2.5	2.6	18.9	4386.0	8.0	2278.0	1.0	1.0	0.0	0.0	−1.8	−1.8	−0.0	−0.0
2.6	2.7	20.5	4367.1	9.5	2270.0	1.0	1.0	0.0	0.0	−1.1	−1.1	−0.0	−0.0
.....													
5.3	5.4	148.5	2244.8	76.0	1370.2	1.0	0.6	0.5	0.3	−7.5	−4.0	−3.4	−1.9
5.4	5.5	133.6	2296.3	69.4	1294.2	1.0	0.5	0.5	0.3	−6.0	−3.1	−3.0	−1.5
5.5	5.6	108.3	2162.7	56.2	1224.8	1.0	0.5	0.5	0.3	−5.3	−2.5	−2.7	−1.3
.....													
7.2	7.3	105.9	332.5	64.7	203.9	1.2	0.1	1.1	0.1	−0.4	−0.0	−0.4	−0.0
7.3	7.4	109.2	226.5	66.5	139.2	1.4	0.1	1.3	0.0	−0.8	−0.0	−0.7	−0.0
7.4	∞	117.3	117.3	72.7	72.7		0.0		0.0	0.0	0.0	0.0	0.0
Total:		4386		2278						−210.9	−136.3	−74.6	−35.5

The values d_j, n_j, d_{1j} and n_{1j} are computed as follows:

```
> d <- svf$pf*length(left)
> n <- (1-svf$sigma+svf$pf)*length(left)
> dboys <- pboys*sum(sex==0)
> nboys <- (svfboys+pboys)*sum(sex==0)
```

Regarding the computation of the weighting functions, the code below gives $v_j^{0,0}$,

```
> rho <- 0
> lambda <- 0
> v00 <- pbeta(svf$sigma,lambda+1,rho+10^-15)-
         pbeta(svf$sigma-svf$pf,lambda+1,rho+10^-15)
> v00 <- (1-svf$sigma+svf$pf)*v00/svf$pf
> v00[1] <- pbeta(svf$sigma[1],lambda+1,rho+10^-15)/svf$pf[1]
> v00 <- v00*gamma(lambda+1)*gamma(rho+10^-15)/gamma(lambda+rho+1)
> v00[is.na(v00)] <- 0
```

When a Turnbull's interval $(q_j, p_j]$ has null probability mass, the corresponding $v_j^{0,0}$ gives a missing value which has to be changed to 0. Another problem which appears is that the last weight $v_m^{0,0}$ is equal to ∞. In practice, we can ignore this weight because in this case the observed and expected number of deaths coincide, $\boldsymbol{O}_m = \boldsymbol{E}_m$.

Now, we give the basic code to implement the permutation approach. First, we compute the pooled survival function evaluated at each observed interval, that is, $\hat{S}(r_i)$ and $\hat{S}(l_i)$ for $i = 1, \ldots, 4386$.

```
> svfright <- 1-svf$sigma[max.col(t(svf$clmat),ties.method="last")]
> svfright[right>=svf$intmap[2,length(svf$pf)]] <- 0
> svfleft <- svfright+colSums(p)
> svfleft[left<=svf$intmap[1,1]] <- 1
```

The code to compute the score values c_i with $\rho = 0$ and $\lambda = 0$ is shown below:

```
> rho <- 0
> lambda <- 0
> ci <- svfright*pbeta(1-svfright,lambda+1,rho+10^-15)-
        svfleft*pbeta(1-svfleft,lambda+1,rho+10^-15)
> ci <- ci*gamma(lambda+1)*gamma(rho+10^-15)/gamma(lambda+rho+1)
> ci <- ci/(svfleft-svfright)
```

The test statistic $\boldsymbol{U} = \sum_{i=1}^{n} \mathbf{z}_i c_i$ and the Mahalanobis distance $\boldsymbol{U}' V_0^{-} \boldsymbol{U}$ can be computed as follows:

```
> library(MASS)
> zi <- 1*cbind(sex==0,sex==1)
> u <- t(ci)%*%zi
> m <- t(length(left)*mean(ci)*colMeans(zi))
> V <- var(ci)*((t(zi)%*%zi)-length(left)*
       (t(t(colMeans(zi)))%*%t(colMeans(zi))))
> distM <- c((u-m)%*%ginv(V)%*%t(u-m))
```

Although the theoretical value of $E(\boldsymbol{U})$ is 0, we have to compute it since due to numerical precision the value obtained is not exactly zero. Note that we have to load the package `MASS` to compute the generalized inverse V_0^-. The values of the Mahalanobis distances for different choices of the parameters ((0, 0), (1, 0), (0, 1) and (1, 1)) are, respectively, 67.2, 70.4, 36.6 and 50.3, much larger than 6.63, the 0.99 quantile of the χ^2 distribution with one degree of freedom. Thus, although the curves for girls and boys are fairly close, we have found statistically significant differences between the survival curves of boys and girls for any of the weighting functions.

4 Regression models

4.1 Parametric regression models

The parametric approach for analyzing interval-censored data is computationally straightforward. A variety of parametric models can be used (Lindsey, 1998) to obtain smooth representations of both the hazard and the survival functions. Maximum likelihood methods can then be applied to provide useful and meaningful parameter-based quantities.

Under the non-informative censoring assumption (see Section 1.1), standard likelihood inference and usual large sample properties apply. Hence, for a study of n individuals, if censoring occurs non-informatively, inferences can be based on the likelihood function

$$L(\theta|\mathcal{D}) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} \int_{l_i}^{r_i} w(u;\theta, Z_i)\, \boldsymbol{du} = \prod_{i=1}^{n} \left[S(l_i;\theta, Z_i) - S(r_i;\theta, Z_i)\right],$$

where $\theta = (\theta_1, \ldots, \theta_p)$ is the vector of the unknown model parameters, Z_i is the covariate vector of subject i and $S(t;\theta, Z_i)$ and $w(t;\theta, Z_i)$ are the survival and density functions, respectively. The expression of both functions depends on the error term distribution of the model chosen for the analyses (see Section 4.1.1).

The maximum likelihood estimator, $\hat{\theta}_n$, of the unknown parameter vector θ can be obtained as the solution of the score equation $U(\theta) = 0$ by using any numerical algorithm such as the Newton–Raphson algorithm, where $U(\theta) = \sum_{i=1}^{n} \partial L_i(\theta)/\partial\theta$.

Under some regularity conditions, the obtained estimator is unique and consistent. The asymptotic distribution of $\hat{\theta}_n$ can be approximated by a multivariate normal distribution with mean θ and covariance matrix $I^{-1}(\theta)$, where $I(\theta)$ is the observed information matrix, namely $I(\theta) = -\sum_{i=1}^{n} \partial^2 L_i(\theta)/\partial\theta\partial\theta'$.

The parametric approach is appealing because of its simplicity but its disadvantage is that all the inferences depend upon the assumption of a model which is difficult to assess based on an interval-censored sample, with the risk of deriving inconsistent estimators for the parameters of interest leading to inaccurate conclusions. Ren (2003) proposes a goodness-of-fit method, called the leveraged bootstrap and Calle and Gómez (2008) propose a sampling-based chi-squared test.

4.1.1 Log-linear representation of parametric survival models

Most commonly used survival models can be expressed as a log-linear model, which is equivalent to the accelerated failure time model. The advantage of this representation is that most standard statistical packages such as SAS, S-PLUS or R are able to fit log-linear models in the presence of interval-censored data. The general expression is in terms of the natural logarithm of the survival time T:

$$Y = \log(T) = \mu + \boldsymbol{\beta}' Z_i + \sigma E, \tag{4.1}$$

where E is the error term distribution. Common choices for T are the Weibull, the log-logistic and the log-normal model and in these cases the error E follows an extreme value, a logistic and a normal distribution, respectively. The acceleration factor, $\exp(\boldsymbol{\beta}' Z)$, in the log-linear models is such that the pth quantile $t_p(z)$ for the population with $Z = z$ is proportional to the pth quantile $t_p(0)$ for the baseline population, i.e., $t_p(z) = t_p(0)\exp(\boldsymbol{\beta}' z)$. If we choose the Weibull model, we are picking the only log-linear model that can as well be interpreted as a proportional hazards model (see Section 4.2) and therefore its parameters can be interpreted as relative risks. For instance, when comparing a subject with covariate vector $Z = z$ to another with $Z = 0$ the relative risk is $RR = \exp(-\frac{\boldsymbol{\beta}' z}{\sigma})$. By contrast, if we choose the log-logistic model, we are in fact under a proportional odds model. With this model, the term $\exp(-\frac{\boldsymbol{\beta}' z}{\sigma})$ can be interpreted as the relative odds of a subject with covariate vector $Z = z$ compared to another with $Z = 0$. Recently, Sparling *et al.* (2006) proposed a general family of parametric regression models for interval-censored survival data that accommodates both fixed and time-dependent covariates.

4.1.2 Illustration: Signal Tandmobiel® data

Herein, we illustrate the use of function `survreg` of the package `survival` (see also Appendix A) to analyze the differences between boys and girls with respect to the emergence time of permanent tooth 24. In order to do so, we first have to load

the package `survival`:

```
> library(survival)
```

The use of function `survreg` requires to create a so-called `Surv` object, which combines all those vectors containing information on the survival times and its censoring status. With 'pure' interval-censored data (not containing exact observations), a vector indicating the censoring status is not necessary. Hence, with our data, the `Surv` object can be defined as follows:

```
> sur24<-with(tooth24,Surv(left,right,type="interval2"))
> dim(sur24)
[1] 4386    3
> sur24[1:5]
[1] [2.7, 3.5] [2.4, 3.4] [4.5, 5.5] [5.9, 999.0] [4.1, 5.0]
```

The object `sur24` contains the observed intervals of emergence times of permanent tooth 24 of 4386 children, five of which are shown above. Recall that we are modelling $T_i - 5$ instead of T_i. Hence, the values [2.7, 3.5] in `sur24` correspond to 7.7 and 8.5 years of age. The value 999 indicates a right-censored observation, that is, tooth 24 of child 4 had not emerged yet by its last dental examination at the age of $5 + 5.9 = 10.9$ years.

Note that unlike function Icens (in Section 2.4), function `Surv` allows right-censored observations to be defined as such. If one wants to do so, a vector indicating the type of censoring has to be defined being 0 the indicator for right- and 3 the indicator for interval-censored data. In the sequel, a new censoring variable (`cens`) is added to dataframe `tooth24` and then included in function `Surv`:

```
> tooth24$cens<-with(tooth24,ifelse(right==999,0,3))
> sur24b<-with(tooth24,Surv(left,right,cens,type="interval"))
> dim(sur24b)
[1] 4386    3
> sur24b[1:5]
[1] [2.7, 3.5] [2.4, 3.4] [4.5, 5.5] 5.9+       [4.1, 5.0]
```

As it can be seen, there is now a slight difference in presenting right-censored data.

In order to fit model (4.1) including both sex and dmf (see Section 1.2) as covariates, we may use either one of the `Surv` objects. For the model, we choose a Weibull distribution, store the model fit in object `weimod` and have a close look at the estimated parameters.

```
> weimod<-survreg(sur24~sex+dmf,data=tooth24,dist="weibull")
> summary(weimod)
```

```
Call:
survreg(formula = sur24 ~ sex + dmf, data = tooth24)
              Value Std. Error       z        p
(Intercept)  1.8439    0.00627  294.11 0.00e+00
sex         -0.0607    0.00723   -8.40 4.61e-17
dmf         -0.0633    0.00724   -8.74 2.32e-18
Log(scale)  -1.6796    0.01587 -105.81 0.00e+00

Scale= 0.186

Weibull distribution
Loglik(model)= -5523.9   Loglik(intercept only)= -5597
        Chisq= 146.23 on 2 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 9
n= 4386
```

According to the results, both variables are highly significant. The negative sign of both parameter estimates indicates, on average, shorter times until tooth emergence for the categories with label 1, that is, girls and children with a decayed, filled or missing (due to caries) primary predecessor of permanent tooth 24. We can interpret the estimated parameters in terms of the relative risk. For example, comparing girls and boys of the same age and with the same value of variable dmf, the instant risk of tooth emergence in girls is $RR = \exp\{-(-0.0607/0.186)\} = 1.386$ times the instant risk in boys. The corresponding acceleration factor is $\exp(-0.0607) = 0.941$. That is, the median time (from age 5) until tooth emergence in girls is 0.941 times the median time in boys.

If we would like to fit the model under different parametric assumptions, we could change option `dist`, for example, to `dist="loglogistic"` or `dist="lognormal"`.

As explained in Appendix A, the generic function `predict` permits model-based prediction, such as the estimation of the distribution's quantiles. For example, the following R code can be used to estimate the median times until tooth emergence in each of the four groups considered in the model:

```
> (new<-data.frame(sex=rep(0:1,2),dmf=rep(0:1,each=2)))
  sex dmf
1   0   0
2   1   0
3   0   1
4   1   1

> Median<-round(predict(weimod,newdata=new,type='quantile',p=0.5),
  2)+5
> data.frame(new,Median)
```

```
  sex dmf  Median
1   0   0   10.90
2   1   0   10.56
3   0   1   10.54
4   1   1   10.22
```

We see that the estimated median emergence time of tooth 24 is lowest (10.22 years) in girls with a decayed, filled or missing (due to caries) primary predecessor of that tooth and highest among boys with dmf=0.

Function `predict` may also be used to draw the model-based distribution functions in each group. This can be accomplished with the following R commands, which produce Figure 2.

```
> pred<-predict(weimod,newdata=new,type='quantile',p=c(0:999/1000))
> x11(width=10,height=6)
> par(font=2,font.lab=2,font.axis=2,las=1,lwd=3,cex=1.5)
> plot(pred[1,],c(0:999/1000),type="l",xlab = "Years from age 5",
+      ylab = "Probability")
> lines(pred[2,],c(0:999/1000),type="l",col=2)
> lines(pred[3,],c(0:999/1000),type="l",lty=2)
> lines(pred[4,],c(0:999/1000),type="l",col=2,lty=2)
> legend("topleft",c('Boy: dmf=0','Girl: dmf=0','Boy: dmf=1',
+        'Girl: dmf=1'), text.width=strwidth('Boy: dmf=0 '),
+        lty=c(1,1,2,2),col=c(1,2,1,2),lwd=3)
```

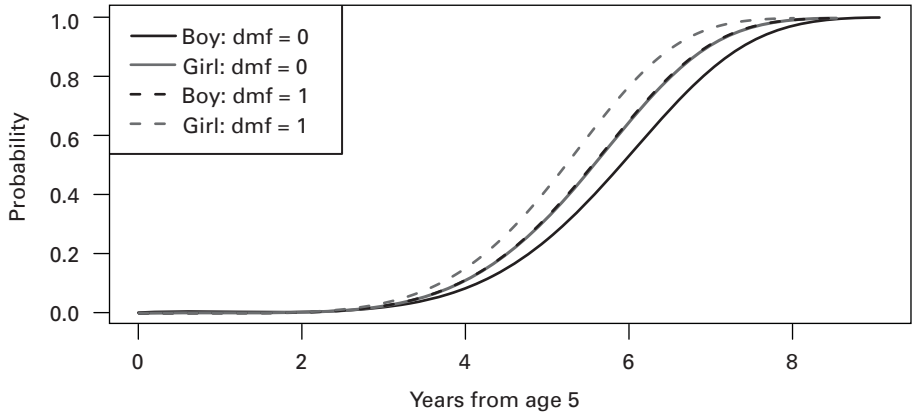

Figure 2 Model-based distribution functions of emergence times of permanent tooth 24 under the Weibull model

4.2 Cox proportional hazards model

The Cox proportional hazards model is the most widely used method for the analysis of right-censored survival data and it is available in all the major statistical software packages. The model itself is semiparametric in nature though general, in the sense that it establishes the relationship between the survival times and the covariates by means of an unspecified baseline hazard function and the exponential of a linear form of the covariates. The statistical inference, being based on the observables, has to acknowledge the different types of censoring. The theory underlying the proportional hazards model for right censoring is well established, not only for the main basic model but also for its many extensions and as said, is widely implemented. The theory for the interval censoring case is, however, not yet at the same theoretical level, with some unresolved issues and lacking an unified approach, partly because many methods rely on the ranks of the exact and right-censored observations and these cannot be identified for interval-censored data.

Most of the first methods to cope with the proportional hazards models are based on the EM algorithm and an approximate likelihood function (Finkelstein, 1986 and Goetghebeur and Ryan, 2000, among others). Other authors deal with the problem via multiple imputation of the unobserved survival times (Satten, 1996; Goggins *et al.*, 1998; Pan, 2000) while Kooperberg and Clarkson (1997) and Betensky *et al.* (2002) propose a non-parametric smoothing of the baseline hazard via regression splines. See Lesaffre *et al.* (2005) for an extensive review of several approaches to fit a proportional hazards model when data are interval censored. Lastly, it is worth mentioning a very recent paper written by Zhang and Davidian (2008) proposing a general framework for semiparametric regression analysis of different patterns of censoring data including the proportional hazards model and interval-censored data. This method is, so far, only available as a SAS macro.

In an attempt to provide in this tutorial an implemented and friendly use of the proportional hazards model for interval-censored data, surprisingly, we have found only one method available. It is the one proposed by Pan (1999) which is implemented in the R package `intcox` developed by Henschel *et al.* (2007). Pan's approach (1999) reformulates the ICM algorithm proposed by Groeneboom and Wellner (1992) as a generalized gradient projection method and it is a reasonably fast algorithm allowing the use of a large sample size. For the purpose of this tutorial, we will only give the details of Pan's (1999) approach and its implementation in R given in the package `intcox`. The `intcox` package, however, is not complete since it fails to directly provide standard errors for the estimated regression parameters, proposing bootstrap to do so.

The proportional hazards model is most often stated in terms of the hazard function of the random variable T as follows:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta'\mathbf{Z}\} = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p\},$$

where $\mathbf{Z}$ stands for the covariate vector and $\lambda_0(t)$ for the unknown baseline hazard function and where we see that the relative hazard $\exp\{\beta'\mathbf{Z}\}$ acts as a factor on the

hazard function. The simplified likelihood function (see expression 1.3) based on the observed data $\mathcal{D} = \{((l_i, r_i], \mathbf{Z_i}), i = 1, \cdots, n\}$ becomes

$$L(S_0, \beta|\mathcal{D}) = \prod_{i=1}^{n} \left\{ S_0(l_i)^{\exp\{\beta'\mathbf{Z_i}\}} - S_0(r_i)^{\exp\{\beta'\mathbf{Z_i}\}} \right\}, \tag{4.2}$$

where $S_0(t)$ corresponds to the baseline survival function.

Pan obtains the non-parametric maximum likelihood estimate of β along with that of the baseline survival function $S_0(t)$ by assuming that $S_0(t)$ is piecewise constant and extending the ICM algorithm. Pan's algorithm parameterizes $L(S_0, \beta|\mathcal{D})$ in terms of the cumulative hazard function $\Lambda_0(t)$ because it is more convenient since there is no upper bound on $\Lambda_0(t)$ and hence no constraints are needed. Simulation studies have shown a slight positive bias in the estimated regression coefficients.

4.2.1 Illustration: Signal Tandmobiel® data

We want now to examine the emergence times of permanent tooth 24 with the proportional hazards model using the same covariates as before. Hence, as a first step, we have to load the package `intcox` (see also Appendix A.2):

```
> install.packages("intcox")   # only necessary if not installed
  yet
> library(intcox)
```

The authors of package `intcox` recommend to identify right-censored data by a missing value (`NA`), that is why we first define a new variable, `rightNA`, with the same right endpoints as vector `right`, but with `NA` instead of 999. After that step, we fit the Cox model with both covariates and obtain the estimates shown below:

```
> tooth24$rightNA<-with(tooth24,ifelse(right==999,NA,right))
> coxmod<-intcox(Surv(left,rightNA,type="interval2")~sex+dmf,
  tooth24)
> coxmod
Call:
intcox(formula = Surv(left, rightNA, type = "interval2") ~ sex +
    dmf, data = tooth24)

     coef exp(coef) se(coef)  z  p
sex 0.288      1.33       NA NA NA
dmf 0.316      1.37       NA NA NA
```

Compared with the log-linear model, the interpretation of the sign of the parameters in the Cox model is just the other way around: both positive signs imply a higher risk and hence, earlier times of tooth emergence among girls and children with a decayed, filled or missing (due to caries) primary predecessor of permanent tooth 24,

respectively. The magnitude of the corresponding relative risks is given by $\exp(\hat{\beta})$. For example, according to this model, comparing girls and boys of the same age and with the same value of variable dmf, the relative risk is exp(0.288) = 1.33, similar to the result of the log-linear model (1.386) assuming a Weibull distribution.

As said in Section 4.2, the estimation of the model parameters is not complete: standard errors are not computed and therefore no values for Z and p are provided. The same happens, for example, with the values of the likelihood ratio test; see Henschel *et al.* (2007) for the reasons. Hence, applying the function `summary` does not really provide more information on the model fit:

```
> summary(coxmod)
Call:
intcox(formula = Surv(left, rightNA, type = "interval2") ~ sex +
    dmf, data = tooth24)

  n= 4386
     coef exp(coef) se(coef)  z  p
sex 0.288      1.33       NA NA NA
dmf 0.316      1.37       NA NA NA

    exp(coef) exp(-coef) lower .95 upper .95
sex      1.33      0.750        NA        NA
dmf      1.37      0.729        NA        NA

Rsquare= NA   (max possible= 0.913 )
Likelihood ratio test= NA  on 2 df,   p=NA
Wald test            = NA  on 2 df,   p=NA
Score (logrank) test = NA  on 2 df,   p=NA
```

As suggested by the package's authors, we computed bootstrap intervals to check whether the association between emergence times and both covariates is significant at a 5% significance level. With 1000 replications, we obtained the following 95% confidence intervals for both parameters: [0.21, 0.367] (sex) and [0.236, 0.391] (dmf) which confirm a significant association between these variables and emergence times of permanent tooth 24. An estimation of the baseline hazard function λ_0 can be obtained by means of the command `coxmod$lamda0`.

5 Simulating interval-censored data

Simulating data is an important part of research and, often, a relevant part in exploring the small and moderate behaviour of the estimators as well as a natural way of comparing statistical procedures under different scenarios. We address here how to generate interval-censored data so that the simulated data are non-informative with

respect to the main time variable of interest in the sense described in (1.2). In the paper by Lawless and Babineau (2006), the reader can find a thorough discussion of how to generate interval-censored data.

Let T be a failure time random variable following a specific distribution $W(t)$. We want to generate censoring intervals $(L, R]$ from $F_{L,R}(l, r)$ such that censoring occurs non-informatively, that is, the conditional distribution of L and R given T satisfies equation (1.2). For a given sample size n of potential times (T_i, L_i, R_i), $(i = 1, \cdots, n)$, we start generating $T_1, \cdots, T_n$ from $W(t)$ following standard procedures. The methods below describe three different ways of generating $(L_1, R_1), \cdots, (L_n, R_n)$.

(*i*) The censoring mechanism of T could mimic a longitudinal study in which there is a periodical follow-up with scheduled visits, taking into account that patients might miss some of their appointments. We assume that there are M potential inspection times a_j $(j = 0, 1, \cdots, M)$, for instance $a_j = j$. The probability that patients attend each of these scheduled visits is p. For an individual i, the observed censoring interval $(L_i, R_i]$ is constructed by defining R_i as the first visit at which the event of interest is observed and L_i as the previous visit. That is, $L_i = \max\{a_j : a_j < T_i; \delta_j^i = 1\}$ and $R_i = \min\{a_j : a_j \geq T_i; \delta_j^i = 1\}$, where δ_j^i is the indicator of whether the visit at time a_j occurred $(\delta_j^i = 1)$ or was missed $(\delta_j^i = 0)$. Different values of p lead to different lengths of intervals, for instance, $p = 0.3$ would imply that 70% of the visits are missed, which would lead to wide intervals of observation for T. In Calle and Gómez (2005), M was taken to be 10 and the distribution of T was taken to be a discrete exponential with values $1, 2, \cdots, 10$ and defined in the following way: $T = [T^*] + 1$ for $T^* < 10$ and $T = 10$ for $T^* > 10$, in which T^* has an exponential distribution of mean 8 and $[t]$ denotes the greatest integer less than or equal to t.

(*ii*) Another way of mimicking a longitudinal study, with periodical follow-up and scheduled visits, is following Schick and Yu's model (2000). In this case and for every individual i, consider a set of examination times $\{Y_{ai}, a = 1, \ldots, \tau_i\}$ which are the sum of independent and identically distributed inter-follow-up times, $Y_{ai} = \sum_{b=1}^{a-1} \xi_{bi}$. 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 observed intervals are defined by $L_i = \max\{Y_{ai} : Y_{ai} < T_i\}$ and $R_i = \min\{Y_{ai} : Y_{ai} \geq T_i\}$. The parameters $E(\xi_{bi}) = \mu$ and τ provide a control of the length of the observed intervals and the percentage of right-censored observations, respectively.

(*iii*) It can be shown that the naive way of simulating intervals by defining $L_i = T_i - U_i^{(1)}$ and $R_i = T_i + U_i^{(2)}$ where $U^{(1)}$ and $U^{(2)}$ are independent continuous variables with uniform distribution in the interval $(0, c)$ does not satisfy the non-informativity condition (1.2). One way to go around this method is by constructing $L_i^* = \max(T_i - U_i^{(1)}, T_i + U_i^{(2)} - c)$ and $R_i^* = \min(T_i + U_i^{(2)}, T_i - U_i^{(1)} + c)$

which can be shown that satisfies the non-informative condition. Zhang (2009) is using this approach, with $c = 1$, in the paper following this in this same volume.

6 Discussion

The aim of the present paper was to present the fundamentals for interval-censored data in a unified manner together with the available software to perform analyses with such data. Although we did not try to do an exhaustive review, some further topics that could not be covered in this paper but could be relevant for the analysis of particular data sets are briefly discussed next and some references are given.

We can encounter in practice other interval-censored mechanisms such as those yielding current status data, doubly censored data or panel count data which require specific methodology and software. The reader is addressed to Sun (2006) who discusses *ad hoc* methods in Chapters 5, 8 and 9, respectively.

Interval censoring is not exclusive of a response endpoint in a regression model but may also be found when a time variable is used as an explanatory covariate. For instance, in the randomized clinical trial ACTG359 for HIV-infected patients, delays in initiating treatment led to concerns that patients who had failed indinavir several months previously, might be different from those who had just recently failed. The effect of the waiting time from indinavir failure to enrollment—which is interval censored because viral load levels are monitored periodically on the log viral load levels at the time of enrollment was of great interest (Gómez *et al.*, 2003). Other instances are described in Goggins *et al.* (1999) and Tian and Lagakos (2006) when estimating the effect of a binary time-varying covariate on failure times when the change time of the covariate is interval censored, or in Langohr *et al.* (2004) who modelled the latency time of AIDS in injection drug users as a function of time from first injection drug use to (interval-censored) HIV infection.

There has been a large number of contributions dealing with the interval-censored problem from a Bayesian viewpoint. See Sinha *et al.* (1999), Gómez *et al.* (2000), Gómez *et al.* (2004), among others.

Interval censoring is an active area of both methodological and applied research and the unresolved issues are still plenty. In particular, important topics such as goodness-of-fit tests and regression diagnostics as well as residuals in regression analysis are not well developed yet. Some useful references are Ren (2003), Topp and Gómez (2004), Lawless and Babineau (2006) and Calle and Gómez (2008).

Concerning hypothesis testing, we have presented a family of tests for testing k identical survival functions under the assumption that the underlying censoring distributions are identical across groups. It will be of practical interest to develop tests for situations when this assumption does not hold, which, to the best of our knowledge, have not been addressed yet.

From a practical point of view, the effort of gathering together all the available routines for interval censoring in this tutorial has made clear that quite a lot remains to be done before interval-censored data can be routinely analyzed in all their extent.

In many situations, even if there are methods derived, the computational aspects are not developed or are not easily available. We hope that this tutorial will facilitate the use of the different functions currently available for interval censoring in R. We have chosen this software for the illustrations and as a connecting thread because it provides a wide range of functions for interval-censored data and, as a free and open source software, it is under constant development by many statisticians worldwide. Hence, any substantial methodological development on interval-censored data will surely find its way quickly to R. In order that the reader will have the possibility to repeat our analyses with the same data. The data used in this tutorial, as well as the functions for the k-sample problem, can be downloaded from the following website: `http://www-eio.upc.es/grass/`.

Acknowledgements

This work is partially supported by grant MTM2008–06747–C02-00 from the Ministerio de Ciencia y Tecnología (Spain). Part of this paper has been written under the support of Grant AI24643 from National Institute of Allergy and Infectious Diseases and Grant 050831 from Marató de TV3 Foundation (Spain). We are indebted to the GRASS group for fruitful discussions. We are grateful to the Signal Tandmobiel® project for providing the background of the problem and the data used in the analysis. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel®project comprises the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

Appendices

Appendix A: R packages for interval-censored data

R's main package for survival analysis is the package `survival` (Therneau and Lumley, 2009). It comprises most of the functions for standard methods in survival analysis such as the Kaplan–Meier estimator, several non-parametric tests, as well as the accelerated failure time model and the proportional hazards model. To work with the package's functions, one has to load the package at the beginning of an R session:

```
> library(survival)
```

Functions in package `survival` do mainly work with right censored, but not with interval-censored data. An exception is function `survreg` which accomplishes the accelerated failure time model. This function allows for several parametric choices such as the exponential, the Weibull, the log-logistic or the log-normal distribution. Several generic R functions are available for `survreg`. For instance, in order to obtain details on the model fit we can invoke the function `summary`; to compute different types of model residuals we have `residuals` or to make predictions based on the model we can use `predict`. Their use is shown in Section 4.1.2.

However, if one wants to compute Turnbull's estimator or to fit the proportional hazards model, the functions of the R packages `Icens` and `intcox`, respectively, are required. In contrast with package `survival`, both packages need to be installed once on the computer. This can easily be done executing the following command and choosing one of the available CRAN mirrors from the list which appears after execution:

```
> install.packages(c("Icens","intcox"))
```

As with other packages, if one wants to use their functions, one has to load them at the beginning of any R session by means of the command `library`. In the sequel, both packages as well as some others offering functions for interval-censored data will briefly be presented. More information on packages developed for survival analysis may be found on the web page 'CRAN Task View: Survival Analysis': `http://cran.r-project.org/web/views/Survival.html`.

A.1 Package Icens

The R package `Icens` of Gentleman and Vandal (2008) (can be downloaded from `http://cran.r-project.org/web/packages/Icens/index.html`) provides different algorithms to compute the NPMLE of the distribution function under interval censoring including the self-consistency algorithm (EM) and the EM-ICM algorithm. The most relevant functions are:

EM: Implementation of the EM algorithm developed by Turnbull (1976).

EMICM: Implementation of the hybrid EM–ICM estimator of the distribution function proposed by Wellner and Zahn (1997).

PGM: Estimation of the NPMLE obtained by using projected gradient methods. It is a particular case of the methods described in Wu (1978).

A graphical representation of the NPMLE can be obtained by applying function `plot.icsurv` to any of these functions. Another useful function is `MLEintvl` which computes Turnbull intervals for a given set of interval-censored data. It is important to note that all these functions require the interval-censored data to be

stored into a two-column matrix, one column containing the lower, the other the upper limit of the censoring interval. Confidence intervals of the NPMLE are not provided. Its use is illustrated in Section 2.4.

A.2 Package intcox

Package `intcox` developed by Henschel *et al.* (2007) implements Pan's (1999) approach (can be downloaded from `http://cran.r-project.org/web/packages/intcox/vignettes/intcox.pdf`). It consists, basically, of one function with the same name which fits the proportional hazards model with interval-censored data. It provides the estimation of the model parameters as well as the estimation of the baseline hazard function. As with function `survreg`, its main argument is a `Surv` object (see Section 4.1.2). However, standard errors of the model parameters are not provided and the authors suggest the use of bootstrap intervals instead. We illustrate that in Section 4.2.1.

A.3 Packages smoothSurv and bayesSurv

For a detailed overview of Komárek's R package `smoothSurv` and `bayesSurv`, developed for the use of an accelerated failure time model with flexible error distributions, we refer the reader to the paper by Komárek and Lesaffre (2009). The last version of `smoothSurv` available on CRAN is 0.3-12 and the author's web is `http://www.karlin.mff.cuni.cz/~komarek/`.

A.4 Packages Epi and Design

The R packages `Epi` of Carstensen *et al.* (2008) and Harrell, (2008) are not specific for survival analysis but do comprise some interesting and nice features. The former is mainly thought for demographical and epidemiological analyses. Among many other functions, it also contains a function called `Icens` which accomplishes the fit of a regression model to interval-censored data assuming 'a piecewise constant baseline rate in intervals specified by the argument breaks, and for the covariates either a multiplicative relative risk function (default) or an additive excess risk function'.

On the other hand, `Design` is a package for 'Regression modeling, testing, estimation, validation, graphics, prediction and typesetting by storing enhanced model design attributes in the fit'. For example, it is possible to fit the accelerated failure time model with interval-censored data using function `psm`. It furnishes basically the same results as function `survreg`, but has the nice feature that model-based estimates for survival probabilities can then be obtained for given times using function `survest`.

Appendix B: Analyses of interval-censored data with other software

Herein, we briefly present some functions offered by other statistical software packages for the analyses of interval-censored data, in particular the commercial software packages S-PLUS (TIBCO Software Inc.; `http://www.insightful.com/products/splus/default.asp`), SAS (SAS Institute Inc.; `http://www.sas.com`) and STATA (StataCorp LP; `http://www.stata.com`). By contrast, as mentioned previously in Section 1.5, the commercial software SPSS (SPSS Inc.; `http://www.spss.com`), in its current version 17.0, can only handle right-censored data.

B.1 S-PLUS

Since S-PLUS uses the same programming language as R, namely S, most of R's functions are also available under S-PLUS and work in the same way. However, when it comes to interval-censored data, its functions are different from R. To compute Turnbull's estimator of the survival function, one has to use a function with a somewhat misleading name: `kaplanMeier`. Its main argument, the observed intervals, is not defined by function `Surv` as in R, but by a function called `censor` which works in the same way. As equivalent to the R function `survreg`, which fits the accelerated failure time model, two different functions may be used in S-PLUS: `survReg` and `censorReg`. Both work with interval-censored data, however, the latter function has a wider range of options. Among others, a threshold value may be specified and the generic function `plot` may be applied to produce several figures which can be helpful to judge the goodness-of-fit of the model. As far as we know, the proportional hazards model with interval-censored data is not available under S-PLUS.

B.2 SAS

The basic SAS functions for survival analysis are `PROC LIFETEST`, `PROC LIFEREG` and `PROC PHREG` for non-parametric, parametric and semi parametric analyses, respectively. Whereas the accelerated failure time model can be fitted with interval-censored data using `PROC LIFEREG`, the other two procedures do not handle these kind of data. Turnbull's estimator, however, can be computed in the presence of interval-censored data using the ICE macro. As mentioned in Section 4.2, there is a SAS macro written by Zhang and Davidian (2008) which enables the fit of the proportional hazards model. An alternative to `PROC LIFEREG` is `PROC RELIABILITY`, which allows to '. . . construct probability plots and fitted life distributions with left-, right- and interval-censored lifetime data' as well as to 'fit regression models, including accelerated life test models, to combinations of left-, right- and interval-censored data' (SAS Help and Documentation). For further information, see also Allison (1997) or Cantor (2003).

B.3 STATA

In STATA, once the data have been declared survival time data by means of the function `stset`, one can apply different survival analysis functions for different kinds of analyses. The basic functions are `sts` for the non-parametric estimation of the survival function, `streg` for the fit of a parametric survival model and `stcox` which fits the proportional hazards model. As far as we know, there are no specific STATA modules for Turnbull's estimator and the Cox model with interval-censored data, but, others exist to fit a parametric survival model. On the one hand, STATA module INTCENS (Griffin, 2005) performs interval-censored survival analysis:

> This program fits various distributions by maximum likelihood to non-negative data which can be left-, right- or interval-censored or point data. The supported distributions are exponential, Weibull, Gompertz, log-logistic, log-normal, 2 and 3 parameter gamma, inverse Gaussian and an extension of the inverse Gaussian which is the time to reach a certain point for a Wiener process with random drift.

On the other hand, there is the module STPM which fits flexible parametric models for survival time data. According to Royston (2001), this module 'fits spline-based distributional models to right-, left- or interval-censored survival data'. It supports different link functions.

References

Allison P (1997) *Survival analysis using SAS. A practical guide*. Cary, NC: SAS Institute Inc.

Betensky R, Lindsey J, Ryan L and Wand W (2002) A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, **21**, 263–75.

Bogaerts K and Lesaffre E (2004) A new, fast algorithm to find the regions of possible support for bivariate interval-censored data. *Journal of Computational and Graphical Statistics*, **13**, 330–40.

Calle ML and Gómez G (2005) Semiparametric hierarchical method for a regression model with an interval-censored covariate. *Australian and New Zealand Journal of Statistics*, **47**, 351–64.

Calle ML and Gómez G (2008) A sampling based chi-squared test for interval–censored data in statistical models and methods for biomedical and technical systems. In Vonta F, Nikulin MS, Limnios N and Huber-Carol C eds. *Statistics for industry and technology*. Springer: Birkhäuser, 303–14.

Cantor A (2003) *SAS. Survival analysis techniques for medical research*, 2nd edition. Cary, NC: SAS Institute Inc.

Carstensen B, Plummer M, Laara E, Hills M, *et al.* (2008) *Epi: a package for statistical analysis in epidemiology. R package version 1.0.8*. Available at: `http://www.pubhealth.ku.dk/~bxc/Epi/`

Courgeau D and Najim J (1996) Interval-censored event history analysis. *Population: An English Selection*, **8**, 191–298.

De Gruttola V and Lagakos SW (1989) Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1–11.

Fang H, Sun J and Lee MT (2002) Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica*, **12**, 1073–83.

Fay MP (1996) Rank invariant tests for interval-censored data under the grouped continuous model. *Biometrics*, **52**, 811–22.

Fay MP (1999) Comparing several score tests for interval-censored data. *Statistics in Medicine*, **18**, 273–85.

Fay MP and Shih JH (1998) Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, **93**, 387–96.

Finkelstein DM (1986) A proportional hazards models for interval-censored failure time data. *Biometrics*, **42**, 845–54.

Fleming TR and Harrington DP (1991) *Counting processes and survival analysis*. New York: Wiley.

Gentleman R and Vandal A (2008) *Icens: NPMLE for censored and truncated data. R package version 1.2.0.* Available at `http://cran.r-project.org/web/packages/Icens/index.html`

Gómez G, Espinal A and Lagakos SW (2003) Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, **22**, 409–25.

Griffin J (2005) *INTCENS: Stata module to perform interval-censored survival analysis*. Statistical Software Components, Boston College Department of Economics. Available at `http://ideas.repec.org/c/boc/bocode/s453501.html`

Goggins W, Finkelstein D, Schoenfeld D and Zaslavsky A (1998) A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, **54**, 1498–507.

Goggins W, Finkelstein D and Zaslavsky A (1999) Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. *Biometrics*, **55**, 445–51.

Gómez G, Calle ML, Muga R and Egea JM (2000) Estimation of the risk of HIV infection as a function of the length of intravenous drug use. A nonparametric Bayesian approach. *Statistics in Medicine*, **19**, 2641–56.

Gómez G, Calle ML and Oller R (2004) Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers*, **45**, 139–73.

Gómez G and Oller R (2008) *A new class of rank tests for interval-censored data*. Harvard University Biostatistics Working Paper Series, Working Paper 93. Available at `http://www.bepress.com/harvardbiostat/paper93`

Gómez G and Van Ryzin J (1992) Estimation of the subsurvival function for time-to-tumor in survival/sacrifice experiments. *Statistics and Probability Letters*, **13**, 5–13.

Goetghebeur E and Ryan L (2000) Semiparametric regression analysis of interval-censored data. *Biometrics*, **56**, 1139–44.

Groeneboom P and Wellner JA (1992) *Information bounds and nonparametric maximum likelihood estimation*. Basel: Birkhäuser Verlag.

Harrell F Jr (2008) *Design: Design Package. R package version 2.1-2*. Available at `http://biostat.mc.vanderbilt.edu/s/Design,http://biostat.mc.vanderbilt.edu/rms`

Heitjan DF and Rubin DB (1991) Ignorability and coarse data. *The Annals of Statistics*, **19**, 2244–53.

Henschel V, Heiss C and Mansmann U (2007) intcox: Compendium to apply the iterative convex minorant algorithm to interval censored event data. Available at `http://cran.r-project.org/web/packages/intcox/vignettes/intcox.pdf`

Hough G, Langohr K, Gómez G and Curia A (2003) Survival analysis applied to sensory

shelf-life of foods. *Journal of Food Science*, **68**, 359–62.
Huang J (1999) Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, **9**, 501–19.
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**, 3217–26.
Huang J and Wellner JA (1997) Interval censored survival data: a review of recent progress. Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis. Springer-Verlag, New York, 123–169.
Komárek A and Lesaffre E (2006) Bayesian semi-parametric accelerated failure time model for paired doubly-interval–censored data. *Statistical Modelling*, **6**, 3–22.
Komárek A and Lesaffre E (2007) Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *Statistica Sinica*, **17**, 549–69.
Komárek A and Lesaffre E (2008) Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, **103**, 523–33.
Komárek A and Lesaffre E (2009) The regression analysis of correlated interval-censored data: illustration using accelerated failure time models with flexible distributional assumptions. *Statistical Modelling*.
Kooperberg C and Clarkson DB (1997) Hazard regression with interval-censored data. *Biometrics*, **53**, 1485–94.
Langohr K, Gómez G and Muga R (2004) A parametric survival model with an interval-censored covariate. *Statistics in Medicine*, **23**, 3159–75.
Lawless J and Babineau D (2006) Models for interval censoring and simulation-based inference for lifetime distributions. *Biometrika*, **93**, 671–86.
Lesaffre E, Komárek A and Declerck D (2005) An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, **14**, 539–52.
Leung K, Elashoff ME and Afiffi AA (1997) Censoring issues in survival analysis. *Annual Review of Public Health*, **18**, 83–104.
Lim HJ and Sun J (2003) Nonparametric tests for interval-censored failure time data. *Biometrical Journal*, **45**, 263–76.
Lindsey JK (1998) A study of interval censoring in parametric regression models. *Lifetime Data Analysis* **4**, 329–54.
Lindsey JC and Ryan LM (1998) Tutorial in biostatistics methods for interval censored data. *Statistics in Medicine*, **17**, 219–238.
Ng MP (2002) A modification of Peto's nonparametric estimation of survival curves for interval-censored data. *Biometrics*, **58**, 439–42.
Oller R, Gómez G and Calle ML (2004) Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, **32**, 315–26.
Oller R, Gómez G and Calle ML (2007) Interval censoring: identifiability and the constant- sum property. *Biomerika*, **94**, 61–70.
Pan W (1999) Extending the iterative convex minorant algorithm to the Cox model for interval censored data. *Journal of Computational and Graphical Statistics*, **8**, 109–20.
Pan W (2000) A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, **19**, 1–11.
Peto R (1973) Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society, Series C*, **22**, 86–91.
Peto R and Peto J (1972) Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, **135**, 185–207.

R Development Core Team (2008) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at `http://www.R-project.org`.

Ren JJ (2003) Goodness of fit test with interval censored data. *Scandinavian Journal of Statistics*, **30**, 211–26.

Royston P (2001) *STPM: Stata module to fit flexible parametric models for survival-time data*. Statistical Software Components, Boston College Department of Economics. Available at `http://ideas.repec.org/c/boc/bocode/s418605.html`

Satten G (1996) Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, **83**, 355–70.

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

Self SG and Grossman EA (1986) Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*, **42**, 521–30.

Sinha D, Chen MH and Ghosh SK (1999) Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, **55**, 585–90.

Sparling YH, Younes N, Lachin JM and Bautista OM (2006) Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, **7**, 599–614.

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

Therneau Terry, original Report by Thomas Lumley (2009) Survival: Survival analysis, including penalised likelihood. R package version 235–37.

Tian L and Lagakos S (2006) Analysis of a partially observed binary covariate process and a censored failure time in the presence of truncation and competing risks. *Biometrics*, **62**, 821–28.

Topp R and Gómez G (2004) Residual analysis in linear regression models with an interval-censored covariate. *Statistics in Medicine*, **23**, 3377–91.

Turnbull B (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290–95.

Vanobbergen J, Martens L, Lesaffre E and Declerck D (2000) The Signal Tandmobiel® project — a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, **2**, 87–96.

Wellner JA and Zahn Y (1997) A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Society*, **92**, 945–59.

Wu C (1978) Some algorithmic aspects of the theory of optimal designs. *Annals of Statistics*, **6**, 1286–301.

Yu Q, Li L and Wong GYC (2000) On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, **27**, 35–44.

Yu Q, Schick A, Li L and Wong GYC (1998) Asymptotic properties of the GMLE with case 2 interval-censored data. *Statistics & Probability Letters*, **37**, 223–28.

Yuen KC, Shi J and Zhu L (2006) A k-sample test with interval censored data. *Biometrika*, **93**, 315–28.

Zhang M and Davidian M (2008) 'Smooth' semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, **64**, 567–76.

Zhang Z (2009) A class of transformed regression models for interval censoring. *Statistical Modelling*, **9**.