Different Approaches for Modeling Grouped Survival Data: A Mango Tree Study

Suely Ruiz GIOLO, Enrico Antonio COLOSIMO, and Clarice Garcia Borges DEMÉTRIO

Interval-censored survival data, in which the event of interest is not observed exactly but is only known to occur within some time interval, occur very frequently. In some situations, event times might be censored into different, possibly overlapping intervals of variable widths; however, in other situations, information is available for all units at the same observed visit time. In the latter cases, interval-censored data are termed grouped survival data. Here we present alternative approaches for analyzing interval-censored data. We illustrate these techniques using a survival data set involving mango tree lifetimes. This study is an example of grouped survival data.

Key Words: Accelerated failure time; Additive hazards; Cox model; Discrete model; Imputation; Interval censoring.

1. INTRODUCTION

In some studies, survival response can be interval-censored, such that the event of interest is not observed exactly but is only known to occur within some time intervals that may overlap and vary in length. Often, interval-censored data are analyzed using imputation methods, in which each event time interval is replaced by a single value and the analysis is performed as though this value were the exact time event. Certainly, imputation has some attractive features; for example, it allows standard analysis for continuous time-to-event data. Midpoint imputation is one of the most widely used methods; however, its statistical properties depend strongly on the width of the intervals. Law and Brookmeyer (1992), for instance, noted that using midpoint imputation to estimate the regression parameter in a proportional hazards model might result in a biased estimate if the intervals are wide and varied. Moreover, the standard error of the estimator is underestimated, because midpoint

Suely Ruiz Giolo is Professor, Department of Statistics, Federal University of Parana, 81531-990 Curitiba, Parana, Brazil (E-mail: *giolo@ufpr.br*). Enrico Antonio Colosimo is Professor, Department of Statistics, Federal University of Minas Gerais, 31270-901 Belo Horizonte, Minas Gerais, Brazil (E-mail: *enricoc@est.ufmg.br*). Clarice Garcia Borges Demétrio is Professor, Department of Exact Science, University of Sao Paulo, 13418-900 Piracicaba, Sao Paulo, Brazil (E-mail: *clarice@carpa.ciagri.usp.br*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 14, Number 2, Pages 154–169
DOI: 10.1198/jabes.2009.0010

imputation assumes that the failure times are exactly known, when in fact they are not (Kim 2003). Rücker and Messerer (1988), Odell, Anderson, and D'Agostino (1992), and Dorey, Little, and Schenker (1993) also noted that applying methods for standard time-to-event data on the imputed times can lead to biased and misleading results. Other imputation procedures have been discussed by Pan (1999, 2000) and Hsu et al. (2007), among others.

Other methods for the analysis of interval-censored data besides the imputation approach have been proposed. Analogs of the classical nonparametric estimator of Kaplan and Meier (1958) were proposed by Peto (1973) and Turnbull (1976) to estimate the survival function for interval-censored data; however, these estimators do not allow for covariates. Recently, Sen and Banerjee (2007) proposed confidence sets for the survival function. Parametric and semiparametric regression models, such as the proportional hazards model (Finkelstein 1986) and the accelerated failure time (AFT) model (Rabinowitz, Tsiatis, and Aragon 1995), also have been proposed for interval-censored data, but these require some specialized methods for estimation. Some of these methods are based on the EM algorithm, considering that interval-censored data can be viewed as incomplete data (Goggins et al. 1998; Betensky et al. 1999; Goetghebeur and Ryan 2000). Other methods are based on a Bayesian approach, which has become increasingly popular (see, e.g., Sinha, Chen, and Ghosh 1999). Recently, Komárek, Lesaffre, and Hilton (2005) suggested and implemented in an R library a maximum likelihood-based approach for the AFT model that exploits penalized smoothing of the baseline density of the error distribution. This method provides estimates of both regression parameters and baseline density without making any strong parametric assumptions. Lessafre, Komárek, and Declerck (2005) has provided an overview of methods for interval-censored data, along with an example using the penalized AFT model in an oral health study. In particular, for grouped data, in which event times can be grouped into mutually exclusive intervals, methods based on discrete models are usually recommended (Lawless 2002).

The aims of this article are (a) to describe different methodologies for handling interval-censored data (particularly grouped survival data), presenting models not often used to analyze this sort of data, such as the Aalen additive hazards model; (b) to compare imputation and interval-censored approaches; and (c) to highlight the differences among different fitted models in analyzing a grouped survival data set involving mango tree lifetimes. The article is organized as follows. Section 2 presents the mango tree data set. Section 3 describes methods for analyzing interval-censored data, and Section 4 gives results and comparisons from the analysis of the data set. Section 5 presents some concluding remarks.

2. THE DATA SET

The data set that we consider was first used by Colosimo, Chalita, and Demétrio (2000) to illustrate a score test statistic proposed to discriminate between two discrete models for grouped survival data. An experiment was conducted in a completely randomized block design with five blocks and six treatments in a 6×7 factorial design involving six different scions (Extrema, Oliveira, Pahiri, Imperial, Carlota, and Bourbon) grafted on seven different rootstocks (Espada, Extrema, Oliveira, Carlota, Bourbon, Coco, and Pahiri), for

Table 1. Mango tree survival data.

Visit year	At risk	Dead	Alive	Visit year	At risk	Dead	Alive
1973	210	12	198	1986	156	13	143
1974	198	8	190	1987	143	16	127
1975	190	1	189	1988	127	28	99
1981	189	8	181	1989	99	10	89
1983	181	2	179	1990	89	27	62
1985	179	23	156	1992	62	6	56

a total of 210 experimental units. The aim of the experiment was to identify the scion–rootstock combination most resistant to a disease of the mango tree, *seca*, caused by the *Ceratocystis fimbriata* fungus. The experimental study began in 1971; the site was visited 12 times, in 1973, 1974, 1975, 1981, 1983, 1985–1990, and 1992, with the condition of each experimental unit (alive or dead) recorded at each visit. The data are summarized in Table 1.

In this study, our interest lies in determining the lifetimes of the six scions and seven rootstocks and in identifying any interaction effect. These lifetimes are not exactly known, however. For mango trees observed to be dead in 1973, for instance, all we know is that the event occurred at some time between 1971 and 1973; thus these trees survived between 0 and 2 years. Similarly, mango trees that were alive in 1973 but observed to be dead in 1974 survived between 2 and 3 years, and so on. In addition, data are available for all units in every visit, and because many mango trees die in the same time interval, many ties are observed, which can be grouped into disjoint intervals. These types of data, known as grouped survival data, represent a particular type of so-called "interval-censored" data.

3. METHODS FOR INTERVAL-CENSORED DATA

For the mango tree data, let t_i denote the event time of interest for the ith tree ($i = 1, \ldots, n$), which we do not observe directly. It is only known to belong to an interval denoted by lower and upper time points, (ℓ_i, u_i). In addition, let $\mathbf{x}_i$ be a vector of covariates for the ith tree. Both left and right censoring can be expressed as special cases of interval censoring where the observed time interval is $(0, u_i)$ for left-censored observations and (ℓ_i, ∞) for right-censored observations. Alternative approaches with different properties can be used to analyze these types of data, as shown in Table 2.

3.1 BASIC METHODS

One of the first methods for estimating the survival function for continuous time-to-event data was proposed by Kaplan and Meier (1958). Although this method is not adequate for interval-censored data, it can be used when an imputation procedure is being considered. A key component of any imputation method for the analysis of interval-censored survival data is the selection of a value, t_i, for each interval-censored observation from an appropriate distribution that satisfies $\ell_i < t_i < u_i$. For example, considering the imputed

Table 2. Properties associated with alternative approaches to analyzing interval-censored survival data.

Property	Turnbull	Cox	Parametric AFT	Discrete	Aalen
Allows covariates		✓	✓	✓	✓
Allows time-dependent covariates		✓		✓	✓
Nonparametric type fit	✓	✓			✓
Allows overlapping intervals	✓	✓	✓		
Allows time-varying effects		✓			✓

times t_i $(i = 1, \ldots, n)$ as the midpoint of the intervals (ℓ_i, u_i), the Kaplan–Meier estimator for the survival function, $S(t) = 1 - F(t)$, is given by

$$\widehat{S}(t) = \prod_{j\,:\,t_j \le t} \left(1 - \frac{d_j}{n_j}\right), \tag{3.1}$$

where $t_1 < t_2 < \cdots < t_k$ are the k ordered failure times, d_j is the number of failures at t_j, and n_j is the number of individuals at risk on time t_j, $j = 1, \ldots, k$.

An analogous estimator for interval-censored data that has no closed form and is based on an iterative procedure was proposed by Turnbull (1976). Because the failure times for these data are not observed directly, Turnbull suggested considering a grid of times $0 = \tau_0 < \tau_1 < \cdots < \tau_m$, which includes all of the points ℓ_i and u_i for $i = 1, \ldots, n$. For the ith individual, a weight α_{ij} $(j = 1, \ldots, m)$ also is defined; this is an indicator of whether the event occurring in the interval (ℓ_i, u_i) could have occurred at τ_j as well. Taking these weights and the initial estimate of $S(\tau_j)$, Turnbull's algorithm estimates both the number of events, d_j, occurring at τ_j and the number, n_j, of individuals at risk at τ_j. The estimates, d_j and n_j, are then used in (3.1) to update the estimate of the survival function, $S(\tau_j)$. This procedure is repeated until the survival function stabilizes. A detailed description of the steps of Turnbull's algorithm have been provided by Klein and Moeschberger (2003). Code implemented in the R statistical environmental for this procedure has been provided by Giolo (2004).

The disadvantage of these two estimators is that they do not allow the use of covariates. In what follows, we explore a semiparametric model (Cox), two parametric AFT models (Weibull and log-logistic), two discrete models (Cox and logistic), and the Aalen additive hazards model. For all of these models, excluding the discrete models, we consider an imputation method as well.

3.2 COX PROPORTIONAL HAZARDS MODEL

The Cox proportional hazards model specifies that the hazard and survival functions for the ith individual with a given covariate vector $\mathbf{x}_i$ are expressed as

$$\lambda(t; \mathbf{x}_i) = \lambda_0(t) \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i) \tag{3.2}$$

and

$$S(t; \mathbf{x}_i) = \exp\left\{ - \int_0^t \lambda_0(v) \exp(\boldsymbol{\beta}' \mathbf{x}_i)\, dv \right\} = \{S_0(t)\}^{\exp(\boldsymbol{\beta}' \mathbf{x}_i)},$$

where $\lambda_0(\cdot)$, called the baseline hazard function, is assumed to be unknown, $S_0(\cdot) = 1 - F_0(\cdot)$ is the corresponding unknown baseline survival function, and $\boldsymbol{\beta}$ is the regression coefficient vector. In this model the hazard functions for different values of $\mathbf{x}$ are assumed to be proportional, so that the regression coefficient β_m $(m = 1, \ldots, p)$ describes the change in the hazard on a logarithmic scale for a change in the corresponding covariate x_m of one unit, while all other covariates are kept fixed.

The likelihood contribution for the ith individual can be expressed as the difference of the survivorship functions evaluated at the observed lower and upper time points, that is, $\{S(\ell_i; \mathbf{x}_i) - S(u_i; \mathbf{x}_i)\}$. Under the assumption of proportional hazards, $S(\ell_i; \mathbf{x}_i)$ is equal to $\{S_0(\ell_i)\}^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}$. The likelihood function is then proportional to

$$L = \prod_{i=1}^{n} \{S(\ell_i; \mathbf{x}_i) - S(u_i; \mathbf{x}_i)\} = \prod_{i=1}^{n} \left[\{S_0(\ell_i)\}^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)} - \{S_0(u_i)\}^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right]. \tag{3.3}$$

From a nonparametric standpoint, the likelihood function (3.3) is maximized by a discrete distribution for S with mass points at a subset formed by the visit times. This set of times induces a discrete distribution for S_0, because no probability mass is associated with the follow-up times outside of this set. Finkelstein (1986) proposed a modified Newton–Raphson method for determining the maximum likelihood estimators. With this approach, the number of unknown parameters could be very large for continuous time models without grouping. This could create some numerical problems, especially for inverting the Hessian matrix at each iteration of the method. Pan (1999a) overcame this drawback by proposing an extension of the iterative convex minorant (ICM) algorithm, which uses only the diagonal elements of the Hessian matrix. Confidence intervals for the estimates are obtained using a bootstrap resampling method. The R library intcox (Henschel, Heiss, and Mansmann 2004) can be used to fit this model using the ICM algorithm.

When using an imputation method to analyze the data is desired, the standard Cox model for continuous survival data can be used. The partial likelihood suggested by Cox (1975) then can serve as an estimation method for the proportional hazards model. This method simply estimates the regression coefficients $\boldsymbol{\beta}$, allowing the baseline hazard function $\lambda_0(t)$ to be a nuisance parameter.

For observed or imputed times, an extension of the Cox model (3.2) can be used to accommodate time-varying regression coefficients. This is expressed as

$$\lambda(t; \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\boldsymbol{\beta}'(t)\mathbf{x}_i(t)\}, \tag{3.4}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))'$ is the vector of time-varying regression coefficients. For situations in which time-varying effects are not needed for all covariates, an important version of the model (3.4) studied by Martinussen and Scheike (2006) is

$$\lambda(t; \mathbf{x}_i(t), \mathbf{z}_i(t)) = \lambda_0(t) \exp\{\boldsymbol{\beta}'(t)\mathbf{x}_i(t) + \boldsymbol{\gamma}'\mathbf{z}_i(t)\}, \tag{3.5}$$

where $\boldsymbol{\beta}(t)$ is a q-dimensional vector of time-varying regression coefficients and $\boldsymbol{\gamma}$ is an r-dimensional vector of regression coefficients with $q + r = p$. According to Martinussen and Scheike (2006), these two models, particularly model (3.4), may be difficult to fit for small- to medium-sized data sets and for data sets with many ties.

3.3 PARAMETRIC AFT MODEL

Let T be the random variable time to the event of interest. The standard way to describe an AFT model is

$$\log(T) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} + \sigma\nu, \tag{3.6}$$

where the random variable ν has a density from the location-scale family and σ is a scale parameter. Then we have

$$T = \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x})\exp(\sigma\nu).$$

Several distributions can be assumed for ν. Here we assume extreme-value and logistic distributions, implying that T has Weibull and log-logistic distributions, respectively. For the Weibull AFT model, the survival function for the ith individual with a given vector of covariates $\mathbf{x}_i$ is given as

$$S(t;\mathbf{x}_i) = \exp\{-t^{\gamma}\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)\},$$

where $\gamma = 1/\sigma$ is the Weibull shape parameter and β_0 is the constant term. A nice feature of the Weibull distribution is that it is the only distribution that can be formulated as either an AFT model or a proportional hazards model. In the case where T has a log-logistic distribution, the survival function is expressed as

$$S(t;\mathbf{x}_i) = \frac{1}{1+\exp[\{\log(t) - (\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)\}/s]},$$

where $s = 1/\sigma$ is the shape parameter.

Under the AFT model, maximum likelihood estimates can be obtained using a Newton–Raphson optimization procedure for the corresponding likelihood function, $L = \prod_{i=1}^{n}\{S(\ell_i;\mathbf{x}_i) - S(u_i;\mathbf{x}_i)\}$. Using the convention that $\{S(\ell_i;\mathbf{x}_i) - S(u_i;\mathbf{x}_i)\} = f(\ell_i;\mathbf{x}_i)$ if $\ell_i = u_i$, we can accommodate exact failure times.

Note that the covariate effects in the hazard function for the Cox proportional hazards model are given by (3.2), whereas those for the AFT model are

$$\lambda(t;\mathbf{x}_i) = \lambda_0\big(t\exp\{-(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)\}\big)\exp\{-(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)\},$$

where $\lambda_0(\cdot)$ is the baseline hazard function. Thus, in contrast to the proportional hazards formulation, in which the covariates just cause the instantaneous hazard value to be multiplied up or down, the AFT model makes the covariates act directly on the time scale by accelerating or decelerating the hazard curve. Lesaffre, Komárek, and Declerck (2005) have provided a nice comparison of the AFT and Cox proportional hazards models. An important feature of the AFT model is that it does not have a nonparametric formulation, because an estimation method for estimating the baseline hazard nonparametrically is not available in this model. When considering an imputation method for analyzing the interval-censored data using the AFT model (3.6), the maximum likelihood for continuous time-to-event data can be used as an estimation method.

3.4 DISCRETE MODELS FOR GROUPED DATA

The models proposed herein can be used in a general framework characterized by interval-censored data. But in some situations, information is available for all units at the same observed visit time. In this case, interval-censored data are termed grouped survival data. Certain statistical methods have been designed especially for this particular case.

Suppose that the event times are grouped into k intervals, $I_j = [a_{j-1}, a_j)$, $j = 1, \ldots, k$, where $0 = a_0 < a_1 < \cdots < a_k = \infty$, and assume that all censoring is done at the end of the intervals. In addition, let R_j be the risk set at time a_{j-1}, and let δ_{ij} be 1 if the lifetime of subject i ends within I_j and 0 otherwise. Assuming that $p_j(\mathbf{x}_i) = P(t_i \leq a_j \mid t_i \geq a_{j-1}; \mathbf{x}_i)$, the probability of the ith subject's death in I_j conditional on being alive at a_{j-1}, and the covariate values $\mathbf{x}_i$, the likelihood function is then given by

$$\prod_{j=1}^{k} \prod_{i \in R_j} \{p_j(\mathbf{x}_i)\}^{\delta_{ij}} \{1 - p_j(\mathbf{x}_i)\}^{(1-\delta_{ij})}, \tag{3.7}$$

which is the likelihood function from a Bernoulli distribution with response δ_{ij} and probability of success $p_j(\mathbf{x}_i)$. The regression structure represented by the probability $p_j(\mathbf{x}_i)$ could be modeled by adopting a proportional hazards model (Cox 1972) or a proportional odds model (Collett 1991). The proportional hazards approach takes $p_j(\mathbf{x}_i)$ of the form

$$p_j(\mathbf{x}_i) = 1 - \gamma_j^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}, \tag{3.8}$$

where $\gamma_j = S_0(a_j)/S_0(a_{j-1})$, $j = 1, \ldots, k$, and $S_0(\cdot)$ is the baseline survival function. In contrast, the proportional odds approach takes $p_j(\mathbf{x}_i)$ as

$$p_j(\mathbf{x}_i) = 1 - \{1 + \gamma_j \exp(\boldsymbol{\beta}'\mathbf{x}_i)\}^{-1}, \tag{3.9}$$

where $\gamma_j = p_j(0)/\{1 - p_j(0)\}$, $j = 1, \ldots, k$. Plugging the equations (3.8) and (3.9) into the likelihood function (3.7) gives us proportional hazards and proportional odds models for grouped survival data. Model (3.8) can be linearized by using a complementary log-log transformation (i.e., $\log[-\log\{1 - p_j(\mathbf{x}_i)\}]$), and model (3.9) can be linearized by using a logistic transformation, such as $\log[p_j(\mathbf{x}_i)/\{1 - p_j(\mathbf{x}_i)\}]$. Therefore, these models can be fitted using standard methods for modeling binary data.

3.5 ADDITIVE HAZARDS MODEL

The nonparametric Aalen additive hazards model is an alternative approach to handling interval-censored data that can be used when an imputation method is considered. This model is based on assuming that the covariates act in an additive manner, instead of multiplicatively, on an unknown baseline hazard rate. As in the multiplicative hazards model, in the additive model of Aalen (1989) we have an event time with a distribution depending on a vector of (possibly time-dependent) covariates, $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$. Thus, for individual i, the conditional hazard rate at time t, given $\mathbf{x}_i(t)$, is a linear combination expressed as

$$\lambda(t \mid \mathbf{x}_i(t)) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) x_{ik}(t), \tag{3.10}$$

where the regression risk coefficients $\beta_k(t)$, $k = 1, \ldots, p$, are unknown functions that are to be estimated and are allowed to be functions of time, so that the effect of a covariate may vary over time. Because directly estimating $\beta_k(t)$, $k = 0, 1, \ldots, p$, is difficult, estimating the cumulative risk function, defined as

$$B_k(t) = \int_0^t \beta_k(u)\, du, \qquad k = 0, 1, \ldots, p,$$

is suggested. A least squares technique is used to find the estimates of $B_k(t)$ and the standard errors of these functions. Defining a $n \times (p+1)$ design matrix, $\mathbf{X}(t)$, with the ith row given by $\mathbf{X}_i(t) = (1, \mathbf{x}_i(t)')$ if the ith individual is a member of the risk set at time t or with $\mathbf{X}_i(t)$ containing only 0's if the ith individual is not in the risk set at time t, the least squares estimate of the vector $\mathbf{B}(t) = (B_0(t), B_1(t), \ldots, B_p(t))'$ is

$$\widehat{\mathbf{B}}(t) = \sum_{t_i \le t} \mathbf{Z}(t_i)\mathbf{I}(t_i), \tag{3.11}$$

where $t_1 < t_2 < \cdots$ are the ordered event times, $\mathbf{I}(t_i)$ is a $n \times 1$ vector with the ith element equal to 1 if the subject i dies at time t and 0 otherwise, and $\mathbf{Z}(t_i)$ is defined by $\{\mathbf{X}'(t_i)\mathbf{X}(t_i)\}^{-1}\mathbf{X}'(t_i)$. The estimator (3.11) exists only up to the time at which $\mathbf{X}'(t_i)\mathbf{X}(t_i)$ becomes singular, denoted here by τ. The covariance matrix estimator of $\widehat{\mathbf{B}}(t)$ is

$$\widehat{\text{var}}(\widehat{\mathbf{B}}(t)) = \sum_{t_i \le t} \mathbf{Z}(t_i)\mathbf{I}^D(t_i)\mathbf{Z}'(t_i),$$

where $\mathbf{I}^D(t_i)$ is a diagonal matrix with diagonal elements equal to $\mathbf{I}(t_i)$.

To test the hypothesis of no regression effects for one or more covariates, Aalen (1993) described a test in which a $(p+1) \times (p+1)$ matrix of weights $\mathbf{W}(t)$ with diagonal elements $W_0(t), W_1(t), \ldots, W_p(t)$ is needed. This matrix puts weights on the observed follow-up time similar to the weighted least squares procedure. Possible candidates are $W_k(t)$ equal to the number of individuals at risk at time t and $W_k(t) =$ constant for $k = 0, 1, \ldots, p$. (For more information, see Aalen 1993 or Lee and Weissfeld 1998.) The test statistic vector $\mathbf{U} = (U_0, U_1, \ldots, U_p)'$ is then expressed as

$$\mathbf{U} = \sum_{t_i \le \tau} \mathbf{W}(t_i)\mathbf{Z}(t_i)\mathbf{I}(t_i),$$

and the covariance matrix of $\mathbf{U}$ is given by

$$\mathbf{V} = \sum_{t_i \le \tau} \mathbf{W}(t_i)\{\mathbf{Z}(t_i)\mathbf{I}^D(t_i)\mathbf{Z}'(t_i)\}'\mathbf{W}(t_i).$$

Then the hypothesis $\text{H}_0 : \beta_k(t) = 0$ for all $t \le \tau$ and all $k \in \mathbf{k}$, where $\mathbf{k}$ is a subset of $\{0, 1, \ldots, p\}$, can be tested using $Q = \mathbf{U}_\mathbf{k}'\mathbf{V}_\mathbf{k}^{-1}\mathbf{U}_\mathbf{k}$, where $\mathbf{U}_\mathbf{k}$ is the subvector of $\mathbf{U}$ corresponding to elements in $\mathbf{k}$ and $\mathbf{V}_\mathbf{k}$ the corresponding subcovariance matrix. Under H_0, Q has a chi-squared distribution.

The cumulative regression functions, $\widehat{B}_k(t)$, and their respective 95% confidence intervals obtained by

$$\widehat{B}_k(t) \pm 1.96\sqrt{\widehat{\text{var}}\{\widehat{B}_k(t)\}}, \qquad k = 0, 1, \ldots, p,$$

when plotted against t, provide a useful tool for evaluating the cumulative excess of risk over time.

There is no special estimation procedure for the Aalen model that includes interval-censored data. But because of this model's nonparametric structure, the same results are obtained for any imputation method used in the analysis of grouped survival data, because the intervals are disjoint and the estimation method is based only on the failure time ranks.

4. RESULTS AND DISCUSSION

In this section we use the models presented in Section 3 to analyze the mango tree data set described in Section 2. The full model formulation for $\boldsymbol{\beta}'\mathbf{x}$ was blocks + scions + stocks + scions ∗ stocks, where scions ∗ stocks represents the interaction term between scions and rootstocks. In addition, the age effect was included for the discrete models. Only the results for scions and blocks are shown, because the other factors (stocks and interaction) were not significant in any of the models. Survival function estimates comparing significant scion differences are presented for each model. Results considering the data as interval-censored are given in Section 4.1, and those for the midpoint imputation are given in Section 4.2. The R statistical environment (R Development Core Team 2008) was used for all analyses.

4.1 ESTIMATES CONSIDERING THE DATA AS INTERVAL-CENSORED

Table 3 gives parameter estimates and the p-values of the Wald-type tests (Cox and Hinkley 1974) for the Cox proportional hazards model and two parametric AFT models (Weibull and log-logistic), fitted by considering the data to be interval-censored. For these models, Figures 1 and 2(a) present the estimated survival probability for each scion. These curves are based on the first block, but the curves for all blocks are similar.

Table 3. Parameter estimates of the Cox, Weibull, and log-logistic AFT models fitted by considering the mango tree data set as interval-censored, and the corresponding bootstrap confidence intervals for the Cox model parameters.

	Cox		Weibull AFT		log logistic AFT	
Parameter	Estimate	95% bootstrap confidence intervals	Estimate	p-value	Estimate	p-value
β_0			3.027	<0.001	2.955	<0.001
β_1: Block 2	−0.023	(−0.618, 0.567)	−0.012	0.924	−0.100	0.517
β_2: Block 3	0.011	(−0.493, 0.549)	−0.021	0.865	−0.014	0.924
β_3: Block 4	0.566	(0.135, 1.046)	−0.237	0.039	−0.219	0.126
β_4: Block 5	0.568	(0.090, 1.155)	−0.245	0.036	−0.289	0.051
β_5: Scion 2: Oliveira	−0.621	(−0.873, −0.736)	0.218	0.097	0.204	0.204
β_6: Scion 3: Pahiri	−0.070	(−0.628, 0.522)	−0.030	0.811	−0.227	0.199
β_7: Scion 4: Imperial	−0.319	(−0.831, 0.267)	0.113	0.375	0.097	0.535
β_8: Scion 5: Carlota	−0.423	(−1.098, 0.266)	0.154	0.238	0.074	0.646
β_9: Scion 6: Bourbon	0.694	(0.302, 1.193)	−0.201	0.083	−0.151	0.306
$\log(1/\gamma)$			−0.797	<0.001		
$\log(s)$					−0.918	<0.001

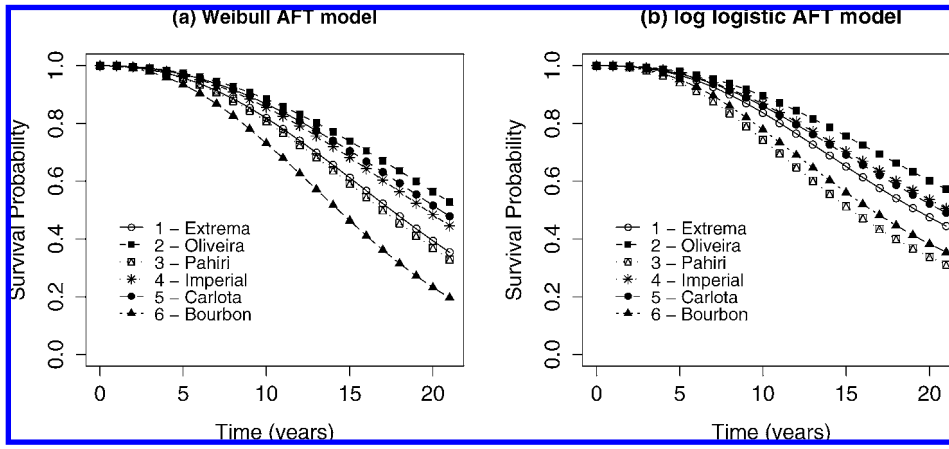

Figure 1. Survival probabilities estimated from the (a) Weibull AFT model and (b) log-logistic AFT model considering the mango tree data as interval-censored.

As shown in Table 3, estimated fitted values were consistent for all models, indicating that Extrema differs from Oliveira and Bourbon. According to Figures 1 and 2(a), Oliveira was the most disease-resistant scion variety studied, exhibiting the highest probability of survival over time. Based on the Cox and Weibull AFT models, Bourbon was the most susceptible scion variety, exhibiting the lowest survival probabilities. The log-logistic AFT model indicated that Pahiri and Bourbon were the most susceptible varieties, and the other scion varieties were of intermediate susceptibility.

Table 4 gives the parameter estimates for the two fitted discrete models, along with age-effect estimates. The results of block and scion effects are in agreement with those presented in Table 3. Figures 3(a) and (b) show survival probabilities estimated from discrete proportional hazards and discrete proportional odds models in the context of grouped data. From these graphs, we can see that both models yielded similar results. Conclusions

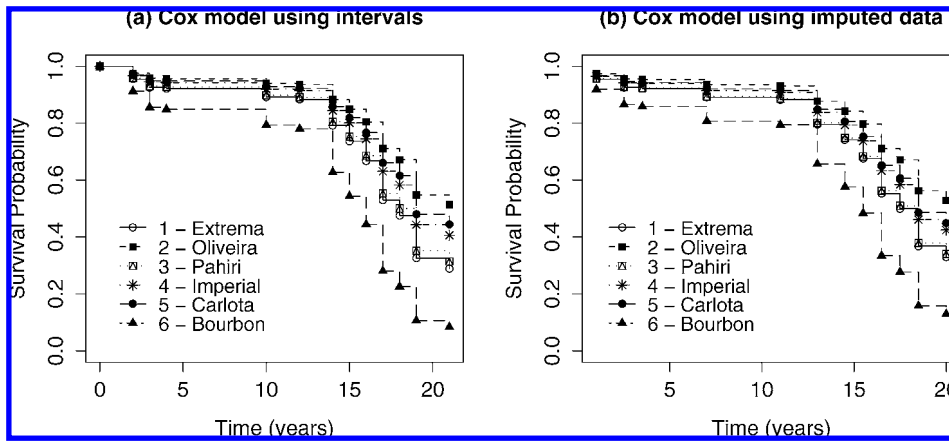

Figure 2. Survival probabilities estimated from the Cox proportional hazards model considering (a) the mango tree data as interval-censored and (b) the midpoint imputation.

Table 4. Results obtained from the discrete models fitted by considering the mango tree data as grouped data (disjoint intervals) with $\gamma_r^* = \log(\gamma_r)$ representing the rth age effect.

	Proportional odds model		Proportional hazards model	
Parameter	Estimate	p-value	Estimate	p-value
γ_1^*	−3.089	<0.001	−3.086	<0.001
γ_2^*	−3.471	<0.001	−3.449	<0.001
γ_3^*	−5.563	<0.001	−5.508	<0.001
γ_4^*	−3.427	<0.001	−3.401	<0.001
γ_5^*	−4.806	<0.001	−4.751	<0.001
γ_6^*	−2.171	<0.001	−2.201	<0.001
γ_7^*	−2.609	<0.001	−2.630	<0.001
γ_8^*	−2.276	<0.001	−2.280	<0.001
γ_9^*	−1.349	<0.001	−1.433	<0.001
γ_{10}^*	−2.186	<0.001	−2.217	<0.001
γ_{11}^*	−0.782	0.029	−0.916	0.004
γ_{12}^*	−2.093	<0.001	−2.104	<0.001
β_1: Block 2	−0.001	0.999	−0.025	0.927
β_2: Block 3	0.013	0.965	0.012	0.965
β_3: Block 4	0.615	0.028	0.576	0.023
β_4: Block 5	0.630	0.026	0.577	0.025
β_5: Scion 2: Oliveira	−0.653	0.040	−0.629	0.030
β_6: Scion 3: Pahiri	−0.033	0.914	−0.072	0.796
β_7: Scion 4: Imperial	−0.304	0.324	−0.324	0.251
β_8: Scion 5: Carlota	−0.384	0.227	−0.432	0.141
β_9: Scion 6: Bourbon	0.768	0.009	0.711	0.006

regarding the most resistant and most susceptible scion varieties are the same as those obtained from the Cox proportional hazards and Weibull AFT models.

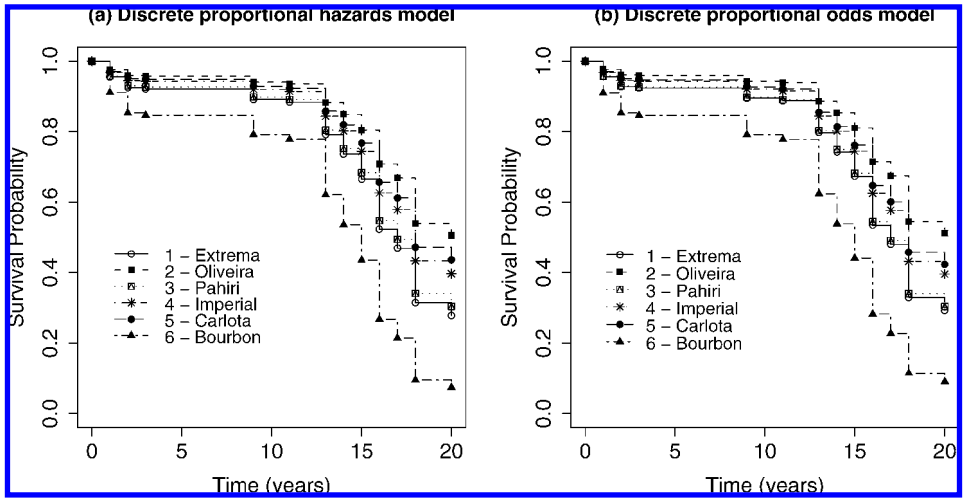

Figure 3. Survival probabilities estimated from the discrete proportional hazards model (a) and the discrete proportional odds model (b) for the mango tree data set.

We could not check the adequacy of the fitted models, because, as far as we know, no statistical techniques are available for this purpose.

4.2 ESTIMATES CONSIDERING THE MIDPOINT IMPUTATION

Table 5 gives parameter estimates for the same models as in Table 3 as well as for the Aalen model (3.10), now considering the midpoint imputation in the analysis of the mango tree data. In general, the results are very similar, especially when the corresponding values are compared in Tables 3 and 5. The survival curve estimates for Cox proportional hazards and both parametric AFT models in Table 5 are shown in Figures 2(b) and 4.

To compare the six scions of the mango tree data set, we obtained the estimated cumulative rates and their respective 95% confidence intervals from the Aalen additive hazards model, as shown in Figure 5. Figure 5(a) shows that the risk of the Extrema variety increased over time, becoming more accentuated from 14 years on. Figure 5(b) compares the cumulative excess of risk for Oliveira and Extrema. Apparently there was no excess risk from 0 to 14 years, because the slopes are close to zero for this period. After that, the risk decreased, demonstrating that Oliveira is more resistant than Extrema. Figure 5(c) shows a slightly increased risk for Pahiri compared with Extrema. Figures 5(d) and (e) show no excess risk for either Imperial and Carlota compared with Extrema. Finally, the excess risk for Bourbon compared with Extrema increased from 14 years on. Thus, as affirmed by the Aalen additive model, Oliveira is the most resistant scion variety, and Bourbon is the most susceptible scion variety.

To evaluate the adequacy of the fitted models, we obtained the Schoenfeld residuals (Therneau and Grambsch 2000) for the Cox model (3.2). No violation was detected in the residuals (not shown), and thus the model can be considered adequate for the imputed data set. The residual plots for the Weibull and log-logistic AFT models presented in Figures 4(b) and (d) show evidence that these models are inadequate, because no straight

Table 5. Results from Cox, Weibull, and log-logistic AFT and Aalen models fitted by considering the midpoint imputation for analyzing the mango tree data set.

	Cox		Weibull AFT		log logistic AFT		Aalen (at $\tau = 18.5$)	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
β_0			3.032	<0.001	2.961	<0.001	1.166	<0.001
β_1: Block 2	−0.002	0.990	−0.013	0.916	−0.106	0.513	−0.116	0.964
β_2: Block 3	0.012	0.960	−0.022	0.863	−0.015	0.920	−0.042	0.821
β_3: Block 4	0.503	0.048	−0.245	0.041	−0.229	0.128	0.606	0.079
β_4: Block 5	0.514	0.046	−0.253	0.037	−0.301	0.052	0.494	0.111
β_5: Scion 2: Oliveira	−0.551	0.060	0.224	0.102	0.210	0.213	−0.607	0.038
β_6: Scion 3: Pahiri	−0.028	0.920	−0.034	0.797	−0.241	0.194	−0.238	0.790
β_7: Scion 4: Imperial	−0.256	0.360	0.116	0.382	0.100	0.543	−0.469	0.220
β_8: Scion 5: Carlota	−0.325	0.260	0.158	0.245	0.077	0.646	−0.563	0.238
β_9: Scion 6: Bourbon	0.614	0.018	−0.204	0.092	−0.150	0.330	1.419	0.015
$\log(1/\gamma)$			−0.758	<0.001				
$\log(s)$					−0.866	<0.001		

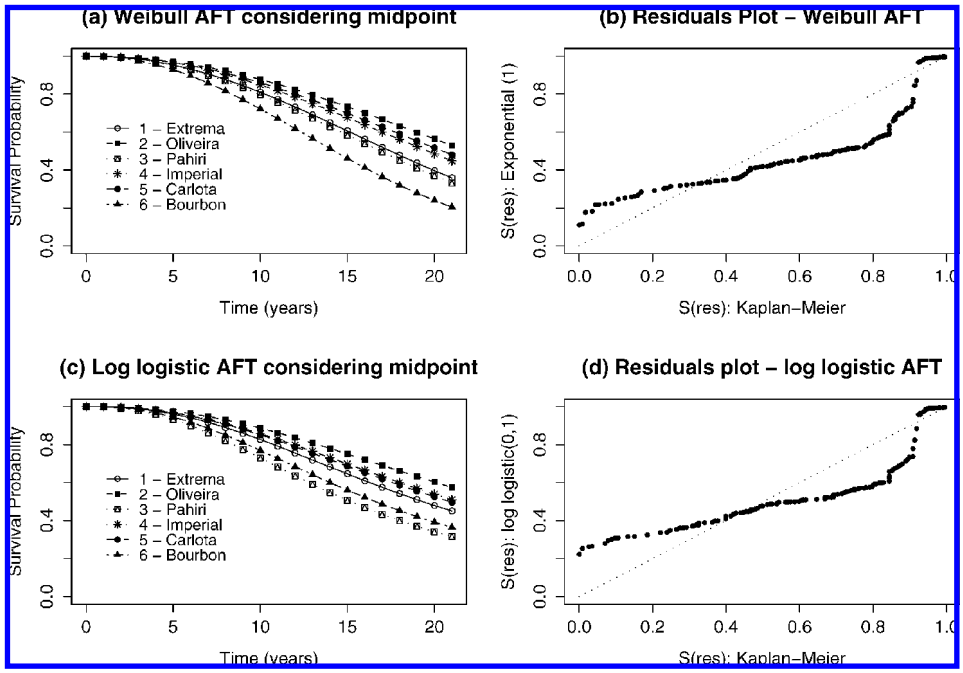

Figure 4. Survival probabilities estimated from the Weibull AFT model (a) and log-logistic AFT model (c) considering the midpoint imputation, and residual plots for the Weibull AFT model (b) and the log-logistic AFT model (d).

lines can be seen. A residual analysis (not shown) of the Aalen additive model (3.10) provided evidence of this model's suitabilty for analyzing the mango tree data. Klein and Moeschberger (2003) have provided details of this residual analysis.

In general, the estimated survival curves obtained from both discrete models in the context of grouped data were very similar to those obtained from the Cox proportional hazards model. These results were not observed for the two parametric AFT models when the data were considered either as interval-censored or imputed. The parametric survival curves for these models decreased more rapidly than those from the other models. Therefore, Cox proportional hazards (3.2), Aalen additive hazards (3.10), and both discrete models (3.8) and (3.9) are adequate for analyzing the mango tree data set. Based on our data, we conclude that, independent of the block, Oliveira was the most resistant scion variety to the disease under study and Bourbon was the most susceptible variety. All other scion varieties exhibited intermediate resistance.

5. CONCLUDING REMARKS

Table 6 summarizes the four classes of models considered in this article for analyzing a grouped survival data set. From this table, it can be seen that the Aalen model can be used when an imputation method is considered and the discrete models are formulated only for grouped survival data.

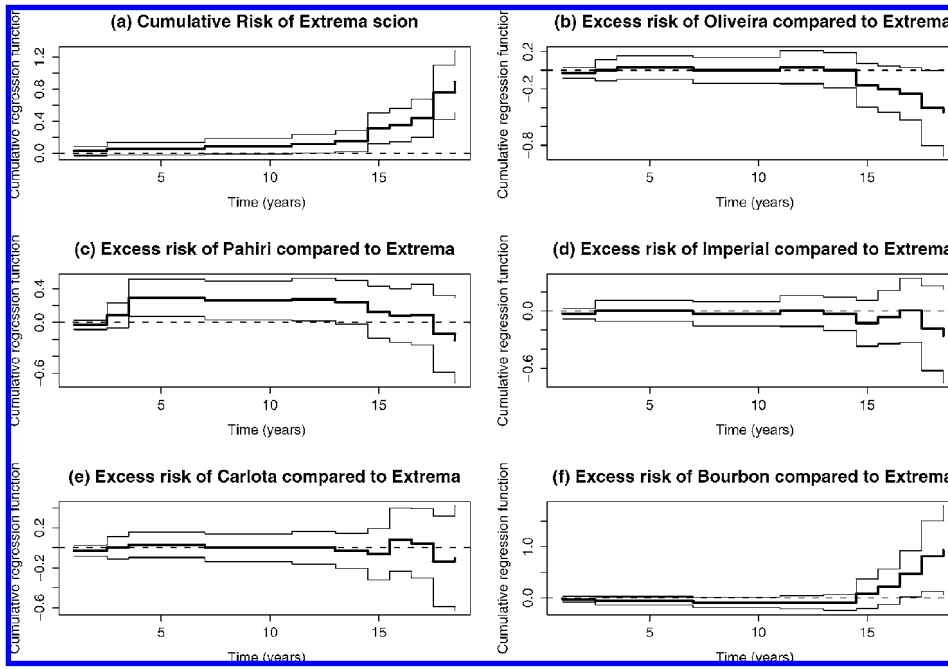

Figure 5. Estimated cumulative regression functions for $t \leq 18.5$ years (bold lines) and respective 95% confidence intervals (thin lines) obtained by fitting the Aalen additive hazards model for the mango tree data set using the midpoint imputation.

Although imputation seems to be a common practice among statistical users, some caution is needed. Imputation might result in biased estimates, especially if the intervals are wide and a large proportion of ties fall into just a few intervals. Some features of imputation make it attractive, however. For instance, many software programs and techniques for evaluating model adequacy are available. For the mango tree data, the results obtained when the midpoint imputation was used were very similar to those obtained when the data were considered to be grouped, indicating that the midpoint imputation worked well. Imputation likely can be justified because the intervals were not too wide for these data, as shown in Table 1.

An important issue related to the parametric AFT models considered in this article is their adequacy. These models are completely parametric and thus are subject to biased estimates if their assumptions are not satisfied. The Cox proportional hazards model is more robust, but still requires verification of the proportional hazards assumption. The

Table 6. Models and approaches considered in the mango tree data.

Approach	Cox	Parametric AFT	Discrete	Aalen
Imputation method	✓	✓		✓
Grouped data	✓	✓	✓	

results obtained from the Cox and AFT models differed for the mango tree data analyzed here. Graphical residual analysis performed for these models in the imputed data approach demonstrated that the AFT models did not provide reasonable fits to the data.

The discrete models considered in this article can be used only when the data set is grouped. Both of these models gave very similar results for the mango tree data set. Their results also were similar to those obtained by fitting the Cox model, considering the data as either grouped or imputed. Finally, the additive hazards model, which is not often used for analyzing grouped data, appears to be an interesting alternative for handling these types of data. The main advantage of this model is that the unknown risk coefficients are allowed to be functions of time, so that the effect of a covariate may vary over time. This feature can be seen for the mango tree data set in Figure 5. Because of this model's nonparametric nature, estimates from it are robust to the choice of imputation values in the presence of grouped data.

ACKNOWLEDGMENTS

This work was partially supported by the Brazilian CAPES Foundation (BEX 0298/01-8), as part of the first author's PhD thesis at the Depto de Ciências Exatas, ESALQ/USP, Brazil, and by grants from the CNPq to the second and third authors. The authors thank Dr. Salim Simão (*in memoriam*) for providing the mango tree data, Dr. Marco Rodriguez from the University of Quebec at Trois-Rivieres for revising the English, and the associate editor for providing many useful suggestions that improved the article considerably.

[Received February 2007. Revised May 2008.]

REFERENCES

Aalen, O. O. (1989), "A Linear Regression Model for the Analysis of Lifetimes," *Statistics in Medicine*, 8, 907–925.

——— (1993), "Further Results on the Non-Parametric Linear Regression Model in Survival Analysis," *Statistics in Medicine*, 12, 1569–1588.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (1999), "Local EM Estimation of the Hazard Function for Interval-Censored Data," *Biometrics*, 55, 238–245.

Collett, D. (1991), *Modelling Binary Data*, New York: Chapman & Hall.

Colosimo, E. A., Chalita, L. V. A. S., and Demétrio, C. G. B. (2000), "Tests of Proportional Hazards and Proportional Odds Models for Grouped Survival Data," *Biometrics*, 56, 1233–1240.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Dorey, F. J., Little, R. J., and Schenker, N. (1993), "Multiple Imputation for Threshold-Crossing Data With Interval Censoring," *Statistics in Medicine*, 12, 1589–1603.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845–854.

Giolo, S. R. (2004), "Turnbull's Nonparametric Estimator for Interval-Censored Data: An R Code," Technical Report 2004/01-C. Available at *www.est.ufpr.br/rt.*

Goetghebeur, E., and Ryan, L. (2000), "Semiparametric Regression Analysis of Interval-Censored Data," *Biometrics*, 56, 1139–1144.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998), "A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data Under the Cox Proportional Hazards Model," *Biometrics*, 54, 1498–1507.

Henschel, V., Heiss, C., and Mansmann, U. (2004), "The Intcox Package," available at *http://cran.r-project.org/web/packages/intcox/index.html*.

Hsu, C.-H., Taylor, J. M. G., Murray S., and Commenges, D. (2007), "Multiple Imputation for Interval Censored Data With Auxiliary Variables," *Statistics in Medicine*, 26, 769–781.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Kim, J. (2003), "Maximum Likelihood Estimation for the Proportional Hazards Model With Partly Interval-Censored Data," *Journal of the Royal Statistical Society*, Ser. B, 65, 489–502.

Klein, J. P., and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.), New York: Springer.

Komárek, A., Lesaffre, E., and Hilton, J. F. (2005), "Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution," *Journal of Computational & Graphical Statistics*, 14 (3), 726–745.

Law, G., and Brookmeyer, R. (1992), "Effects of Midpoint Imputation on the Analysis of Doubly Censored Data," *Statistics in Medicine*, 11, 1569–1578.

Lawless, J. F. (2002), *Statistical Models and Methods for Lifetime Data* (2nd ed.), New York: Wiley.

Lee, E. T., and Weissfeld, L. A. (1998), "Assessment of Covariates Effects in Aalen's Additive Hazard Model," *Statistics in Medicine*, 17, 983–998.

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 (6), 539–552.

Martinussen, T., and Scheike, T. H. (2006), *Dynamic Regression Models for Survival Data*, New York: Springer.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992), "Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull-Based Accelerated Failure Time Model," *Biometrics*, 48, 951–959.

Pan, W. (1999), "A Comparison of Some Two-Sample Tests With Interval Censored Data," *Journal of Nonparametric Statistics*, 12, 133–146.

——— (1999a), "Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data," *Journal of Computational & Graphical Statistics*, 8 (1), 109–120.

——— (2000), "A Multiple Imputation Approach to Cox Regression With Interval Censored Data," *Biometrics*, 5, 192–203.

Peto, R. (1973), "Experimental Survival Curves for Interval Censored Data," *Applied Statistics*, 22, 86–91.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at *http://www.R-project.org*.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995), "Regression With Interval-Censored Data," *Biometrika*, 82, 501–513.

Rücker, G., and Messerer, D. (1988), "Remission Duration: An Example of Interval-Censored Observations," *Statistics in Medicine*, 7, 1139–1145.

Sen, B., and Banerjee, M. (2007), "A Pseudolikelihood Method for Analyzing Interval-Censored Data," *Biometrika*, 94, 71–86.

Sinha, D., Chen, M.-H., and Ghosh, S. K. (1999), "Bayesian Analysis and Model Selection for Interval-Censored Data," *Biometrics*, 55, 585–590.

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

Turnbull, B. W. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society*, Ser. B, 38, 290–295.