Journal of Statistical Planning and
Inference 78 (1999) 191–204

journal of
statistical planning
and inference

A stochastic process applied to sequential parametric analysis of censored survival data

A. Maul[a], A.A. Bartolucci[a,*], K.P. Singh[b]

[a] *Département Statistique et Traitement Informatique des Données, Institut Universitaire de Technologie, Université de Metz, 57045 Metz, France*

[b] *Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Room 101, Bishop Building, 900 19th Street South, Birmingham, AL 35294-2030, USA*

Abstract

A stochastic process that allows sequential parametric estimation of the hazard function is presented. The analysis of censored survival data is based on a discrete time definition of the hazard which is expressed as a logistic function of a number of time-dependent covariates. The method adequately handles large sets of data with many tied failure times and high rates of type I censored values. A procedure available to estimate the relative risk parameter characterizing two groups of individuals over a specific period of time is also given. Likelihood methods are used in estimating the parameters of the model and making inference about the survivor function, especially beyond the value of censoring. The method is illustrated by an example concerning the induction period between infection with the AIDS virus and the onset of clinical AIDS. The effects of censoring on the inference analysis of the survivor function corresponding to several groups of individuals are examined and discussed. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Survival analysis; Censored data; Logistic regression; Time-dependent covariates; Latency period; AIDS

1. Introduction

A major objective of biomedical investigations is to assess the effects of a number of covariates on a time-related response variable such as an incubation period or a survival time. This can be done by expressing the hazard as a function of a number of explanatory variables that may depend on time. These covariates are used for

[*] Corresponding author. Tel.: +1 205 934 4905; fax: +1 205 975 2540; e-mail: abartolucci@biostat.soph.uab.edu

PII: S0378-3758(98)00213-4

characterizing the individuals or the different groups to which they belong. Such situations are encountered, namely when

(i) Studying the induction period of a new unknown spreading disease or similarly the latency period between documented exposure to an environmental agent and subsequent disease diagnosis.

(ii) Performing clinical trials. Considering the possible implications for public health intervention and prevention policy in both of the previous medical problems it is desirable that the outcome of the study be stated within the shortest possible period of time. To maintain the duration of such a follow-up study within an acceptable time limit it is often necessary to take stock of the situation at a given prespecified date by noting the state of each individual whether he/she has failed. More generally, reviewing the situation at successive chronological terms may generate data sets containing a high number of censored values or tied failure times. On the other hand, inferences about the parameters of interest obviously improve as the available information which is accumulated over time increases. Thus, the statistical problem consists in making inferences about a stochastic process of exposure and subsequent failure for which realizations are subject to right censoring in chronologic time.

The purpose of the present paper is to present a parametric approach that is adapted for estimating and comparing the induction distributions of several groups of individuals by taking into account all the information available at successive expiration dates. The method presented here is much more efficient in reaching a compromise between the previous antagonistic time constraints than the current parametric or nonparametric statistical procedures in the sense that

(i) Usual parametric methods (Kalbfleisch and Prentice, 1980) may not be satisfactory due to a lack of generality since most of the existing methods are characterized by the stringency of the underlying hypotheses and/or the subsequent narrowness of the areas in which they can be applied.

(ii) Nonparametric methods, e.g., the product-limit method developed by Kaplan and Meier (1958), do not allow to define the survivor function beyond the last observed value if it is censored and an estimate of the mean survival time is then unavailable.

(iii) The widely used semi-parametric proportional hazards regression model (Cox, 1972) or other continuous models become rather inadequate to deal with large data sets comprising many censored and/or tied failure times.

The difficulties arising from the use of a continuous model for analyzing such data sets have been discussed by Lawless (1982). A number of models have therefore been developed to perform survival analysis in discrete time (Cox, 1972; Lawless, 1982; Kalbfleisch and Prentice, 1973; Prentice and Gloeckler, 1978). However, the results of asymptotic maximum likelihood inference on the parameters of interest may be influenced by the way of grouping the failure times (Prentice and Gloeckler, 1978).

The method used in this paper takes explicit account of the discrete nature of the data since it is based on a discrete time expression of the hazard. It is a generalization of the discrete time logistic model given by Lawless (1982) to include time-dependent regressor variables. Clearly, the hazard function associated with each individual of the sample is considered a time series that is expressed as a logistic function involving a number of time-dependent covariates. The further developed technique is shown to have an appeal in terms of conceptual simplicity and it accommodates the possibility of large sets of grouped survival data with high rates of censored values.

The different aspects and the usefulness of the method for the assessment and the comparison of the induction distributions characterizing several groups of individuals are illustrated by means of a simple numerical example (Lagakos et al., 1988) which is from the AIDS literature. The example is concerned with the latency period of several groups of individuals each of whom contracted AIDS as a result of being infected from a contaminated blood transfusion.

2. The model

The hazard p_{it} corresponding to an individual i with regressor variable $\boldsymbol{x}_{it}$ at time t (i.e. the probability of failure at time t provided that the individual was still at risk at time t) is modelled as follows:

Let $\{Z_t^i\}$ $(i=1,\dots,n)$ be a collection of independent time series. Each series has binomially distributed random variables with probability distribution defined as

$$\begin{aligned}&P(Z_t^i=1)=p_{it}=1/(1+\exp(\boldsymbol{x}_{it}\cdot\boldsymbol{\beta})),\\&P(Z_t^i=0)=1-p_{it}=1/(1+\exp(-\boldsymbol{x}_{it}\cdot\boldsymbol{\beta})),\quad (i=1,\dots,n;\ t=1,2,\dots)\end{aligned}\tag{1}$$

where $\boldsymbol{x}_{it}=(x_{it}^1,x_{it}^2,\dots,x_{it}^p)$ is a vector of explanatory variables, describing patient etiology, clinical stage of disease etc, which is associated with the ith individual at time t where $\boldsymbol{\beta}=(\beta_1,\dots,\beta_p)$ is a vector of unknown parameters. The various covariables represented by the vector $\boldsymbol{x}_{it}$ may be discrete, continuous, time-fixed or time-dependent. It is interesting to point out that if $\boldsymbol{x}_{it}\cdot\boldsymbol{\beta}$ is great enough, the continuous proportional hazards model given by Cox (1972) may be obtained as a special case of Model (1) given by Maul (1994).

Let Y_i $(i=1,\dots,n)$ be the random variable associated with the failure time corresponding to the ith individual, i.e. the value of t when $Z_t^i=1$ for the first time. It is assumed that $P(Y_i=+\infty)=0$, $(i=1,\dots,n)$. The method used to assess the dependence of the hazard function associated with an individual on the p explanatory variables in Model (1) will be referred to as discrete time logistic Bernoulli regression (DTLBR).

3. Statistical methods

The present paper is concerned with the DTLBR method which will be used to

(i) Estimate $\boldsymbol{\beta}$ and thereby determining the hazard function.

(ii) Estimate the survivor function and the expected life, i.e. $E[Y]$, corresponding to a given profile, $\{\boldsymbol{x}_t\}$, of the effects of the explanatory variables $x^1,x^2,\dots,x^p$.

(iii) Compare the hazard rate between two populations or profiles. This will enable us to extend the concept of relative risk by integrating the mean value of the hazard ratio over an interval of time. Usually, the relative risk is *instantaneous,* that is examined at a given time. Here we suggest to generalize the concept of relative risk by considering it on an interval of time. This is useful in the case of a time-dependent hazard function (cf. Section 3.3).

These objectives are achieved on the basis of a sequential process of censoring which involves discrete type I right-censored observation times. Such data may be obtained in a wide range of biomedical investigations by reviewing the situation at regularly spaced dates during the follow-up study. Other useful aspects of the DTLBR method are given by Maul (1994), namely estimating quantiles, testing the equality of two or more distributions and assessing the adequacy of Model (1) to describe the data set examined.

3.1. Estimation of the parameters and hazard function

Let $\boldsymbol{y}=(y_1,\dots,y_n)$ be the observed failure times in a sample of size n. If Y_i is censored on the right, the observed survival time of the ith unit will be denoted by y_i^{c}. This means that the ith individual still was in the study without failing at time y_i^{c}.

We have

$$P(Y_i=y_i)=\left[\prod_{t=1}^{y_i-1}(1-p_{it})\right]p_{iy_i} \tag{2a}$$

and

$$P(Y_i>y_i^{\mathrm{c}})=\prod_{t=1}^{y_i^{\mathrm{c}}}(1-p_{it}) \quad (i=1,\dots,n,\ y_i \text{ or } y_i^{\mathrm{c}}=1,\dots,). \tag{2b}$$

Note that all the duration variables and values in Eqs. (2a) and (2b) are integers. If we assume that the last k ordered observations are censored on the right at $y_{n-k+1}^{\mathrm{c}},\dots,y_n^{\mathrm{c}}$, and that the censoring and failure mechanisms are independent then the maximum likelihood (ML) estimates of the parameters in Model (1) are obtained by maximizing the log-likelihood function of the sample which is given by

$$L(\boldsymbol{y}|\boldsymbol{\beta})=\prod_{i=1}^{n-k}\left\{\left[\prod_{t=1}^{y_i-1}(1-p_{it})\right]p_{iy_i}\right\}\prod_{i=n-k+1}^{n}\left\{\prod_{t=1}^{y_i^{\mathrm{c}}}(1-p_{it})\right\}. \tag{3}$$

This is done by solving the set of p equations,

$$\frac{\partial \ln L}{\partial \beta_j}=0,\ \text{ i.e. } \sum_{i=1}^{n}\left\{\sum_{t=1}^{y_i^*}x_{it}^{j}p_{it}\right\}=\sum_{i=1}^{n-k}x_{iy_i}^{j}\quad (j=1,\dots,p), \tag{4}$$

where y_i^* is for y_i^{c} or y_i according as Y_i has been censored on the right at y_i^{c} or not, respectively.

The foregoing results are presented by using matrix notation. Let X be the matrix of explanatory variables of order $\{\sum_{i=1}^{n} y_i^* \times p\}$ with $\boldsymbol{x}_{it}$ as the tth row associated with the

ith individual. The successive individuals being in the same order as the observations, that is $y_1,\ldots,y_{n-k},\ y^{\mathrm{c}}_{n-k+1},\ldots,y^{\mathrm{c}}_n$, provided that the last k ordered observations have been censored on the right. Let $\boldsymbol{O}$ be the column vector of length $\sum_{i=1}^n y_i^*$ with all its elements equal to 0 with the exception of the last element corresponding to a non-censored value y_i $(i = 1,\ldots,n-k)$ which is equal to 1. Let $D_{(\beta)}$ be a diagonal matrix of order $\{\sum_{i=1}^n y_i^* x \sum_{i=1}^n y_i^*\}$ with its successive diagonal elements equal to $p_{i1}, p_{i2},\ldots,p_{iy_i^*}$ $(i=1,\ldots,n)$ that is $p_{i1},\ p_{i2},\ldots,p_{i(y_i-1)},\ p_{iyi}$ $(i = 1,\ldots,n-k)$ and $p_{i1}, p_{i2},\ldots,p_{i(y_i^{\mathrm{c}}-1)}, p^{\mathrm{c}}_{iy_i}$ $(i = n-k+1,\ldots,n)$ provided that the last k ordered observations have been censored on the right. Then it is easy to show that the system of Equation (4) is

$$X' \cdot D_{(\boldsymbol{\beta})} \cdot \mathbf{1} = X' \cdot \boldsymbol{O} \tag{5}$$

where X^t is the transpose of the matrix X and $\mathbf{1}$ is a column vector of length $\sum_{i-1}^n y_i^*$ with all its elements equal to 1. The maximum likelihood Eqs. (4) and (5) can be solved by Newton–Raphson iteration (Bard, 1974) which requires the evaluation of the Fisher information matrix I. The element in the rth $(r=1,\ldots,p)$ row and sth $(s=1,\ldots,p)$ column of the observed Fisher information matrix, $I(\hat{\boldsymbol{\beta}})$, evaluated at $\hat{\boldsymbol{\beta}}$, is given as

$$I(r,s) = \sum_{i=1}^{n}\left[\sum_{t=1}^{y_i^*} x_{it}^r \boldsymbol{x}_{it}^s \frac{\exp(\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i\cdot\hat{\boldsymbol{\beta}})^2}\right]. \tag{6}$$

If $\hat{\boldsymbol{\beta}}_{(l)}$ is a column vector representing the solution at stage l in the iteration process, which is performed to solve Eq. (5), then the solution $\hat{\boldsymbol{\beta}}_{(l+1)}$ at iteration $(l+1)$ is

$$\hat{\boldsymbol{\beta}}_{(l+1)} = \hat{\boldsymbol{\beta}}_{(l)} - I^{-1}_{(\hat{\boldsymbol{\beta}}_l)} \cdot (X' \cdot D_{(\hat{\boldsymbol{\beta}}_l)} \cdot 1 - X' \cdot \boldsymbol{O}), \tag{7}$$

where $I^{-1}_{(\hat{\boldsymbol{\beta}}_l)}$ is the inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\beta}}_{(l)}$. This process is iterated until convergence.

Assuming n is sufficiently large, $\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}}$ has approximately a normal distribution with mean $\boldsymbol{x}_{it}\cdot\boldsymbol{\beta}$ and standard deviation $\sqrt{Var[\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}}]}$. We abbreviate this as

$$\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{x}_{it}\cdot\boldsymbol{\beta};\sqrt{Var[\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}}]}\right) \quad (i = 1,\ldots,n; t = 1,\ldots,y_i^*). \tag{8}$$

An estimate of $Var[\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}}]$ is given as

$$\sum_{r=1}^{p}(x_{it}^r)^2 I^{-1}_{(r,r)} + 2\sum_{r<s} x_{it}^r x_{it}^s I^{-1}_{(r,s)} \quad (i=1,\ldots,n; t=1,\ldots,). \tag{9}$$

Thus, it is easy to show that an estimate and associated confidence limits at level α for the hazard function p_{it} evaluated at $\boldsymbol{x}_{it}$ are given as

$$\hat{p}_{it} = 1/(1+\exp(x_{it}\cdot\hat{\boldsymbol{\beta}}))$$

and

$$1 \Big/ \left(1+\exp\left(\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}} \pm u_{(1-\alpha/2)}\sqrt{Var[\boldsymbol{x}_{it}\cdot\hat{\boldsymbol{\beta}}]}\right)\right), \quad \text{respectively, } (i=1,\dots,n;\ t=1,2,\dots), \tag{10}$$

where $u_{(1-\alpha/2)}$ is obtained from the table of the standard normal distribution.

3.2. Estimation of the survivor function and the expected life

The survivor function associated with the set of vectors $x_t^i=\{\boldsymbol{x}_{iu},\ u=1,\dots,t\}$, that is, the probability for the ith individual to be still at risk at time t under the conditions specified by x_t^i, is defined as

$$S_i(0)=1,$$

$$S_i(t)=P(Y_i>t)=\prod_{u=1}^{t}(1-p_{iu}) \quad (t=1,2,\dots,). \tag{11}$$

Thus, the ML estimate of $S_i(t)$ given by

$$\hat{S}_i(t)=\prod_{u=1}^{t}(1-\hat{p}_{iu}). \tag{12}$$

The expectancy of Y_i $(i=1,\dots,n)$ is given as

$$E[Y_i]=\sum_{t=1}^{\infty} tp(Y_i=t)=\sum_{t=0}^{\infty} S_i(t),$$

provided that this series converges. Consequently, $E[Y_i]$ can be estimated by $\sum_{t=0}^{\infty}\hat{S}_i(t)$.

Using large sample approximations, it can be shown that $\log \hat{S}_i(t)$ has an asymptotic normal distribution with mean $\log S_i(t)$. If the covariates are time-fixed, i.e. $p_{it}=p_i$ $(t=1,2,\dots)$ the asymptotic distribution for the $\ln(\hat{S}(t))$ is

$$\ln(\hat{S}_i(t)) \simeq N\left(\ln(S_i(t));\ t\cdot\hat{p}_i\cdot\sqrt{Var[\boldsymbol{x}_i\cdot\hat{\boldsymbol{\beta}}]}\right). \tag{13}$$

Another special case of interest arises when one of the covariates in Model (1) is time, e.g. $x_{it}^1=t$ $(i=1,\dots,n)$ with β_1 representing a monotonic trend in the time-hazard relationship.

If we assume that exp $(-\boldsymbol{x}_t\cdot\boldsymbol{\beta})$ is small in such a way that the ratio $\ln(1-p_{t+1})/\ln(1-p_t)$ can be approximated by $\exp(-\beta_1)$, then it becomes easy to show that

$$\ln(\hat{S}(t)) \simeq \ln(1-\hat{p}_1)\cdot[(1-\exp(-\hat{\beta}_1 t))/(1-\exp(-\hat{\beta}_1))] \tag{14a}$$

and

$$Var[\ln(\hat{S}(t)] \simeq [(1-\exp(-\hat{\beta}_1 t))/(1-\exp(-\hat{\beta}_1))]^2\cdot(\hat{p}_1)^2.Var[\boldsymbol{x}_1\cdot\hat{\boldsymbol{\beta}}]. \tag{14b}$$

The approximate results in Eqs. (14a) and (14b) are valid under the condition that the probability of failure p_t is less than 0.5 and provided that $\exp(-\beta_1)$ is close to unity.

Note that these conditions are met in most of the practical situations. These results can be used for determining confidence limits for the survivor function at time t, assuming a simple monotonic dependance on time of the hazard function. Moreover, it is of interest to note that the standard deviation in Eq. (13) can be obtained from Eq. (14b) as $\hat{\beta}_1$ goes to zero.

3.3. Estimation of the relative risk

The instantaneous relative risk characterizing two populations (denoted by the subscripts 0 and 1) at time t is expressed as the ratio $p_1(t)/p_0(t)$. However, in the case of a time-dependent hazard function it is preferable to define a mean relative risk R_T calculated for a prespecified period of time T as follows:

$$R_T = \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} p_1(t)/p_0(t). \tag{15}$$

Note that if the hazard function is independent on time, $\hat{R}_T$ reduces to $(1+\exp(\boldsymbol{x}_0 \cdot \hat{\boldsymbol{\beta}}))/(1+\exp(\boldsymbol{x}_1 \cdot \hat{\boldsymbol{\beta}}))$ or even (if p_0 and p_1 are small) to $\exp((\boldsymbol{x}_0 - \boldsymbol{x}_1) \cdot \hat{\beta})$ which is the result obtained for the proportional hazards regression model (3). Furthermore, if p may be considered a continuous function of t it is convenient to compute R_T as the mean value of the hazard ratio integrated over the interval $[t_0, t_0+T]$, that is

$$R_T = \frac{1}{T} \int_{t_0}^{t_0+T} \frac{p_1(t)}{p_0(t)} \, \mathrm{d}t.$$

In the case of a simple monotonic dependence on time of the hazard, i.e. $x_{it}^1 = t$ $(t = 1, \ldots, n)$, after simplification it can be shown that R_T can be estimated by

$$\begin{aligned} \hat{R}_T = 1 + {} & \frac{\exp((\boldsymbol{x}_0 - \boldsymbol{x}_1) \cdot \hat{\boldsymbol{\beta}}) - 1}{\beta_1 T} \\ & \times \ln[1 + \exp(\hat{\beta}_1(t_0+T) + \hat{\beta}_2 x_1^2 + \cdots + \hat{\beta}_p x^p))/ \\ & (1 + \exp(\hat{\beta}_1 t_0 + \hat{\beta}_2 x_1^2 + \cdots + \hat{\beta}_p x_1^p))] \end{aligned} \tag{16}$$

for any two given sets of values of the explanatory variables $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$.

4. Application to aids data

The data set (Lagakos et al., 1988) which is presented in Table 1 gives the infection time, (I), and failure time (i.e. onset of clinical AIDS), (F), for 258 adults and 37 children who were infected by contaminated blood transfusions and developed AIDS by 10 June 1986. All dates have been expressed in 3-month time interval units from 1 April 1978 onwards. Thus, an event occurring between 1 January 1985 and 30 March 1985 will be recorded at 27. These data are used to illustrate the statistical

Table 1
Infection time, I, and failure time, F, for 258 adults and 37 children (^{a}) with transfusion-related AIDS. Numbers in parentheses denote multiplicities (adapted from Lagakos et al., 1988)

	F																	
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
0						(1)												
1														(1)				
2																		
3									(2)									(1)
4							(1)					$(1)^a$	(1)		(1)	(1)		
5							(1)	(1)		(1)				(1)				
6	$(1)^a$		(1)				(1)					(1)		(1)				(1)
7				(1)	(1)									(2)				
8			(1)			(1)				(1)		(1)	(1)	(1)	(2)	(2)		(1)
9							(1) $(1)^a$										(1)	
10					(4)	(1)	(1)	(1)	(2)			(3)						
11	$(1)^a$	(1)	(1)				(1)		(2)	(1)		(1)			(2)		(2)	(5)
12					(1)	(1)					(1)	(1)	(1)	(2)	(3)		(3)	(1)
13				(1)		(1)	(2)			(1)	(2)		(2)			(1)	(1)	
14			$(1)^a$		(1)				(2)	(1)	(2)	(1)	(1)	(2)		(2)	(1)	(2)
15				$(1)^a$	$(1)^a$	(1)		(2)	(1)			(1) $(1)^a$	(3) $(1)^a$		$(1)^a$		(1)	(2) $(1)^a$
16						(1) $(1)^a$		(2)		(1)	(1)		(1)		(1)		(2)	(1)
17								(1)	(2)	$(1)^a$	(3)	(1)	(3)		(1)		(2)	
18								(1)		(3)	(1)		(2)	(4)	(4)	(3)	(2) $(1)^a$	
19									(1) $(1)^a$		(3)	(2)	(1)	(1)		(1)	(2) $(1)^a$	(6)
20								(1) $(1)^a$	$(1)^a$		$(1)^a$	(2)	(1)	(1)	(3)	(2) $(1)^a$		(3)
21								(2) $(1)^a$		(3)	(2) $(1)^a$	(2)	(4) $(1)^a$			(2)	(2)	(1)
22										$(1)^a$		(3)	(2)	$(1)^a$	(1)	(1)	(2)	(1) $(1)^a$
23										(1)		(1)	(1)		(3)		(2) $(1)^a$	(1)
24												(1) $(1)^a$	(3)	(3)	(2) (1)	(2)	(3)	(1)
25													$(1)^a$	(1)	(1)	(1) $(1)^a$		(2)
26													(2)		(1) $(1)^a$	(1)	(1)	(1)
27															$(1)^a$	(3) $(1)^a$	(1)	(3)
28																	(1) $(1)^a$	
29																(1) $(1)^a$		

methods outlined in Sections 2 and 3. In particular, the stochastic process of infection and disease are used to

(i) Estimate the hazard and survivor functions corresponding to each of the two groups of individuals considered. This is done for a given censoring value, that is, assuming the situation has been reviewed at a specific date.

(ii) Assess the effects on the estimated hazard and survivor functions of the accumulated information within both groups by setting the reviewing date at regularly spaced chronologic times.

It must be emphasized that the set of data examined undoubtedly induces biased estimates for the induction times since all the individuals have been involved in the study conditionally on having contracted clinical AIDS by June 30, 1986. This means that the data examined here are truncated (i.e. only individuals with diagnosed AIDS are in the sample). Nevertheless, this feature is ignored in the further statistical analysis. A substantive analysis of these data should therefore be done cautiously.

However, notwithstanding its limitations (Lui et al., 1986; Medley et al., 1987) the structure of the data presented in Table 1 is particularly convenient to illustrate the efficiency and aptitude of the DTLBR approach for making inferences about the hazard rate of a process with right censoring.

The hazard function at time t, that is, the probability of contracting AIDS during the tth interval of time, is modelled as

$$p_t = 1/(1+\exp(\beta_0+\beta_1\varepsilon+\beta_2 t+\beta_3 t^2)), \tag{17}$$

where ε is an indicator variable with values -1 and $+1$ according as the individual considered is a child or an adult.

Table 2 presents the results of the asymptotic likelihood inference analysis for the regression parameters as given in Model (17). The estimated limit of the survivor function as $t \to \infty$ and the estimated medians corresponding to both the adults and the children are also given in Table 2. All these values were calculated for different successive six-month time interval spaced censoring dates ranging from 31 March 1983 until 30 June 1986 and thus covering a proportion of censored values going from 91 down to 0%, respectively.

The analysis started by testing the significance of the regression coefficients in Model (17) which were observed at the successive censoring dates. Most of the estimated values for $\beta_{0'}$, β_1 and β_2 were significantly different from zero at the 0.1% level, in terms of individual statistical significance. This indicates strong evidence that

(i) The hazard may not be considered the same among the two groups (i.e. adults vs. children); the instantaneous failure rate to contract AIDS at a given time after infection is higher for children than for adults. It is interesting to note that such a statement could already have been made at time 19 (i.e. 31 March 1983).

(ii) The assumption of a time-dependent expression of the hazard is reasonable. In this regard, testing the hypothesis H_0: $\beta_3 \leqslant 0$ is of special interest since a positive value of β_3 (β_2 being negative) may thus represent a non-monotonic trend in the time-hazard relationship. Moreover, a positive β_3 means that the hazard $p_{t'}$ goes to zero as

Table 2
Analysis of the hazard function for successive 6-month intervals censoring dates ranging from 31 March 1983 until 30 June 1986 and characteristics of the estimated survivor functions (adults vs. children) assuming Model (17) is used for the hazard function

Date of censoring[a]	19	21	23	25	27	29	31	33
Proportion of censored values (%)	91	85	78	69	57	41	22	0
Estimated effects (standard error)								
β_0	6.49***	5.51***	4.42***	3.95***	3.73***	3.35***	3.01***	3.00***
	(1.39)	(0.79)	(0.48)	(0.37)	(0.30)	(0.24)	(0.21)	(0.20)
β_1	1.24**	0.98***	0.76***	0.65***	0.56***	0.43**	0.46**	0.44***
	(0.34)	(0.24)	(0.19)	(0.16)	(0.13)	(0.12)	(0.10)	(0.10)
β_2	−0.78*	−0.55**	−0.30**	−0.22**	−0.23***	−0.16***	−0.14***	−0.16***
	(0.42)	(0.21)	(0.12)	(0.09)	(0.07)	(0.05)	(0.04)	(0.04)
β_3	0.044	0.024*	0.010	0.007	0.008*	0.004	0.003	0.002
	(0.029)	(0.012)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$\lim_{t\to\infty} \hat{S}(t)$								
Adults	0.898	0.696	0.447	0.271	0.221	0.038	0.005	0.000
Children	0.297	0.091	0.030	0.011	0.013	0.001	0.000	0.000
Median								
Adults	/	/	21.64	17.03	14.24	12.28	10.84	9.72
Children	8.89	8.23	8.38	8.04	7.40	7.25	5.95	5.68

*Value is significant at the 5% level.
**Value is significant at the 1% level.
***Value is significant at the 0.1% level.
[a]Note that the time is expressed in 3 month intervals beginning 1 April 1978.

$t \to \infty$. Clearly, this indicates the possibility for an individual of becoming safe from contracting the disease provided that he has not failed before a sufficiently long period of time. Thus, assuming the complete model is used for the instantaneous failure rate, the estimated proportion of individuals that is expected to avoid contracting the disease is also presented in Table 2 for each group and the various censoring thresholds.

However, since none of the individual tests on β_3 yielded a significant value at the 1% level of probability ($p > 0.05$ for six out of the eight values tested), it seems preferable that Model (17) should be reduced to a three-parameter expression including β_0, β_1, and β_2 only for modelling the hazard function on the basis of the data set examined.

Thus, the previous analysis was reconsidered by using the reduced model for the hazard which is hence taken to be of the form

$$p_t = 1/(1 + \exp(\beta_0 + \beta_1 \varepsilon + \beta_2 t)) \tag{18}$$

Table 3 presents the ML estimates and standard errors for the regression parameters as given in Model (18) which have been obtained for the 6-month spaced censoring dates. The different estimated expectancies and medians of the induction times and the relative risk as calculated by (16) for both groups of individuals are also given in Table 3.

Table 3
An analysis of the hazard function for successive censoring dates and characteristics of the estimated survivor functions (adults vs children) assuming the reduced Model (18) is used for the hazard function

Date of[a] censoring	19	21	23	25	27	29	31	33
Proportion of censored values (%)	91	85	78	69	57	41	22	0
Estimated effects (standard error)								
β_0	4.69	4.23	3.83	3.53	3.22	3.05	2.79	2.79
	(0.59)	(0.38)	(0.28)	(0.23)	(0.19)	(0.16)	(0.14)	(0.13)
β_1	1.18	0.92	0.73	0.62	0.53	0.42	0.45	0.43
	(0.33)	(0.23)	(0.19)	(0.16)	(0.13)	(0.12)	(0.10)	(0.10)
β_2	−0.122	−0.132	−0.109	−0.094	−0.079	−0.080	−0.079	−0.105
	(0.076)	(0.043)	(0.030)	(0.023)	(0.018)	(0.014)	(0.012)	(0.010)
$\hat{E}[Y]$								
Adults	27.36	20.85	19.11	17.68	16.00	13.79	12.29	10.73
Children	11.57	10.15	9.99	9.60	8.98	8.65	7.26	6.73
Median								
Adults	27.95	21.10	19.05	17.22	15.08	12.67	11.02	9.76
Children	10.87	9.39	8.98	8.38	7.53	7.21	5.80	5.50
Relative risk								
R_{20}	9.64	5.52	3.83	3.12	2.62	2.11	2.19	2.04

[a] Note that the time is expressed in 3-month intervals beginning 1 April 1978.

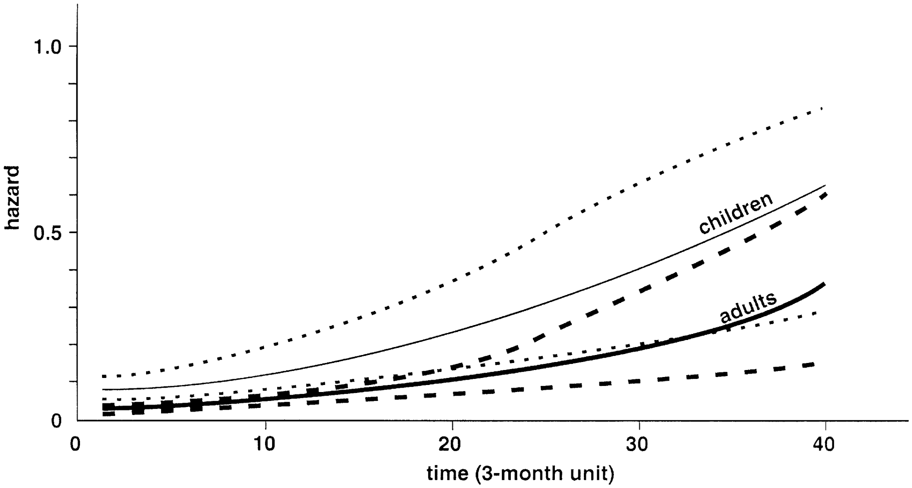

Fig. 1. Estimated hazard functions (solid lines) and 95% confidence bands (dotted lines) for children and adults, assuming the date of censoring was settled by 31 March 1985.

The differences in the induction dynamics of AIDS between adults and children are illustrated in Figs. 1 and 2 which show, respectively, the estimated hazard function and survivor function corresponding to each of the two groups by using (18) as a model for the instantaneous failure rate. The different curves are given with related

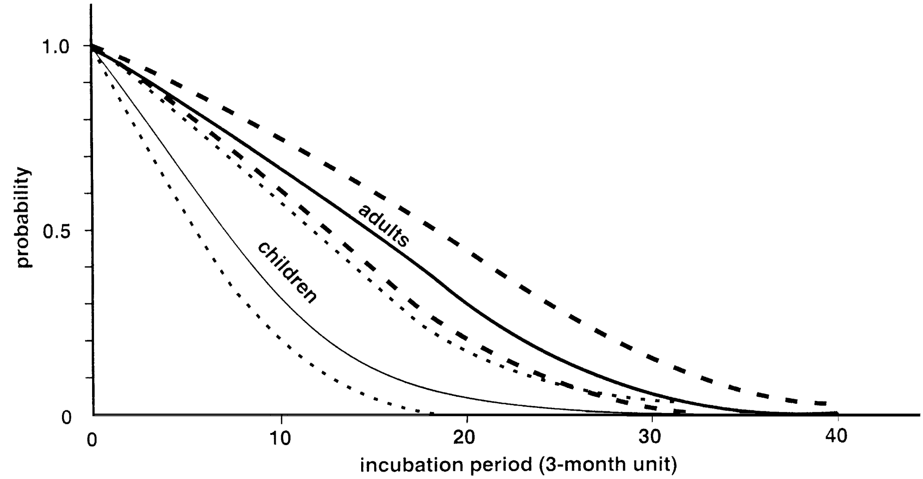

Fig. 2. Survival function estimates (solid lines) and 95% confidence bands (dotted lines) for children and adults, assuming the date of censoring is 31 March 1985.

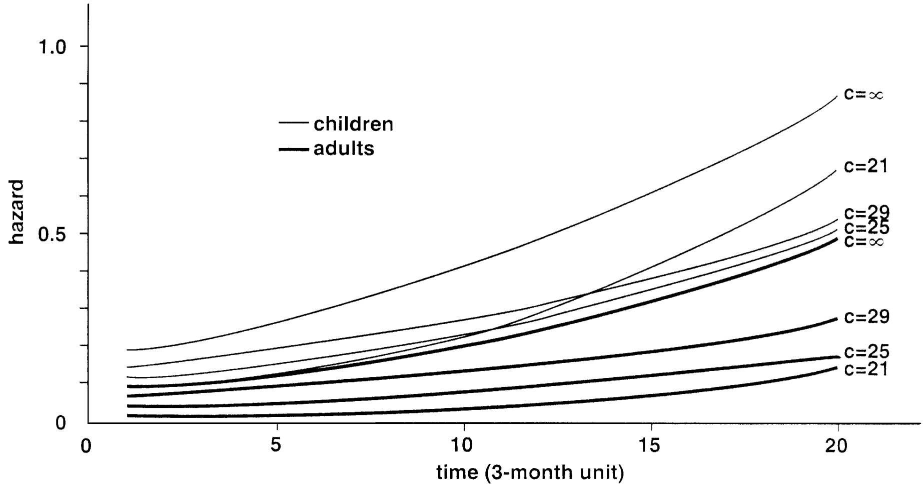

Fig. 3. Hazard function estimates for adults and children assuming different values (c) of the date of censoring.

95% confidence bands assuming the date of censoring was fixed on 31 March 1985. Note that all the graphs should be discontinuous since the functions considered are defined for discrete values of the time only, but the hazard and survivor functions have been interpolated between observations for convenience in plotting and reading. The effects on the estimated curves as a result of accumulating information by deferring the censoring threshold in chronologic time are shown in Fig. 3 for the hazard function and in Fig. 4 for the survivor function. The curves in Fig. 3 represent the hazard functions

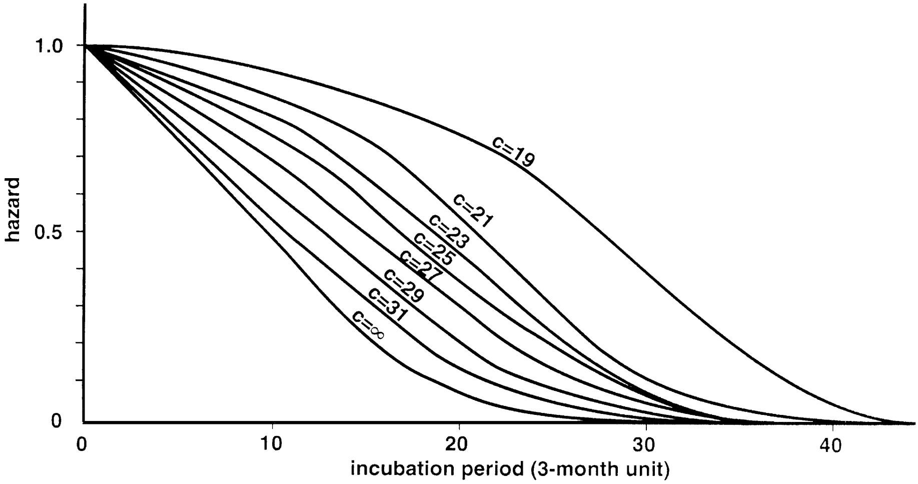

Fig. 4. Survivor functions (adults only) estimated for different 6-month intervals censoring dates (c) ranging from 31 March 1983 until 31 March 1986.

which have been estimated both for adults and children by assuming three different one-year spaced dates of censoring. The estimated survivor curves corresponding to several 6-month spaced censoring periods are plotted in Fig. 4 for adults only.

From Table 3 and all the previous figures, it becomes clear that:

(i) The instantaneous failure rate is higher for children than adults. This result is in full agreement with the conclusion stated by Lagakos et al. (1988). However, it must be emphasized that the difference in the induction times between the two groups examined is shown to be significant from $t = 19$ (i.e. 31 March 1983) onwards using the DTLBR method.

(ii) The hazard as estimated from the present data set is shown to be an increasing function of time.

(iii) The estimated survivor functions become more and more depreciative as the amount of information available increases as a result of postponing the censoring threshold in time. This can also be observed numerically by means of the estimated induction times expectations and medians as shown in Table 3.

Nevertheless, a substantive interpretation of the last two statements is difficult in the sense that the results obtained are likely to be induced by the peculiar structure of the data set examined. Moreover, one must be aware that extrapolation of the model beyond the data in order to estimate lifetime parameters may be of dubious reliability since it relies on specific parametric assumptions.

5. Concluding remarks

The approach proposed in this paper provides a particularly convenient and useful way of making inferences about the hazard rate of a process with right censoring. The

DTLBR method is applicable to a wide range of biomedical investigations, namely:

(i) the estimation of the mean latency period of a disease in order, for example, to reach a better understanding of the features which may influence the mechanism of spreading and/or the survival time after the date of diagnosis.

(ii) the comparison of the relative efficiency of treatments with respect to longevity when performing therapeutic trials.

Its interest lies in both the generality of the statistical model including concomitant information and the possibility of making inferential analysis on patient survival data with high numbers of censored values or tied failure times. Furthermore, the DTLBR method undoubtedly will find its highest interest within the framework of a stepwise assessment, that is, by following a sequential process of reviewing the situation when carrying out a follow-up study.

References

Bard, J., 1974. Non-Linear Parameter Estimation. Academic Press, New York.

Cox, D.R., 1972. Regression models and life-tables (with discussion). J. Roy. Statist. Soc. Ser. B 34, 187–220.

Kalbfleisch, J.D., Prentice, R.L., 1973. Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267–278.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457–481.

Lagakos, S.W., Barraj, L.M., DeGruttola, V., 1988. Nonparametric analysis of truncated survival data, with application to AIDS. Biometrika 75, 515–523.

Lawless, J.F., 1982. Statistical Models and Methods for Lifetime Data. Wiley, New York.

Lui, K.J., Lawrence, D.N., Morgan, W.M., Peterman, T.A., Haverkos, H.H., Bregman, D.J., 1986. A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. Proc. Natl. Acad. Sci. 83, 2913–2917.

Maul, A., 1994. A discrete time logistic regression model for analyzing censored survival data. Environmetrics 5, 145–157.

Medley, G.F., Anderson, R.M., Cox, D.R., Billard, L., 1987. Incubation period of AIDS in patients infected via blood transfusion. Nature 328, 719–721.

Prentice, R.L., Gloeckler, L.A., 1978. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57–67.