STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:263–275 (DOI: 10.1002/sim.993)

A local likelihood proportional hazards model for interval censored data

Rebecca A. Betensky[1,2,*,†], Jane C. Lindsey[2], Louise M. Ryan[2,3] and M. P. Wand[2]

[1]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*
[2]*Harvard School of Public Health, Boston, U.S.A.*
[3]*Dana Farber Cancer Institute, Boston, U.S.A.*

SUMMARY

We discuss the use of local likelihood methods to fit proportional hazards regression models to right and interval censored data. The assumed model allows for an arbitrary, smoothed baseline hazard on which a vector of covariates operates in a proportional manner, and thus produces an interpretable baseline hazard function along with estimates of global covariate effects. For estimation, we extend the modified EM algorithm suggested by Betensky, Lindsey, Ryan and Wand. We illustrate the method with data on times to deterioration of breast cosmeses and HIV-1 infection rates among haemophiliacs. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: local likelihood methods; interval censored data; proportional hazards, regression model

1. INTRODUCTION

Recent years have seen an increasing interest in statistical methods for studying the relationship between covariates and time to event data. Although many different approaches can be used, the proportional hazards (PH) regression model [1] is one of the most popular. Suppose T represents time to the event of interest and $\mathbf{z}$ a covariate vector that we wish to relate to the distribution of T. Then, the proportional hazards model implies

$$\lambda(t;\mathbf{Z}) = \lambda(t)\mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}} \tag{1}$$

where $\lambda(t)$ is the baseline hazard function at time t. The PH model is generally fit using partial likelihood methods which allow the baseline hazard to remain unspecified. Thus, the global regression parameter, $\boldsymbol{\beta}$, is assumed to be of primary interest. Sometimes, however, the

* Correspondence to: R. A. Betensky, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
† E-mail: betensky@hsph.harvard.edu

Contract/grant sponsor: NIH; contract/grant numbers: CA75791 and CA48061
Contract/grant sponsor: NIAID; contract/grant number: 1U01 4I 41110.

Copyright © 2002 John Wiley & Sons, Ltd.
Received July 2000
Accepted April 2001

shape of the hazard function is also of interest. For right censored data, the Nelson–Aalen hazard estimator [2] can be used, although, being a step function, it is not always informative. The estimate can also be highly variable in regions of sparse data. To avoid these problems, Gray [3] applies a kernel smoother to the estimated hazards. In contrast to this *post hoc* approach, Whittemore and Keller [4] and Rosenberg *et al.* [5] model the baseline hazard function using splines and simultaneously estimate its parameters and the regression parameter. Wu and Tuma [6] describe a model in which the baseline hazard function is estimated using local likelihood methods, although they do not describe how estimation occurs and they do not provide any examples of this particular model. Here we propose an iterative procedure to obtain a local likelihood estimate of the baseline hazard function in conjunction with an estimate of the global regression parameter for right censored data.

Interval censored data arise frequently in practice, for example, in a clinical trial in which prespecified visit times are required of participants and status of disease is assessed only at these times. Several authors [7–9] have discussed non-parametric estimation of survival curves. Smoothing methods can be particularly useful for the analysis of interval censored survival data, for which the non-parametric hazard estimators are even less informative than they are for right censored data. Both local likelihood estimation and splines have been used to estimate the hazard function based on interval censored data [10–12].

Parametric regression approaches for interval censored data (for example, exponential, Weibull, log-normal, log-logistic and gamma) are straightforward and programs are available in standard statistical software packages [13]. Another computationally appealing approach is to use flexible, 'loosely parametric' models such as a piecewise exponential model [13–17]. Satten *et al.* [18] assume a parametric form for the baseline hazard, but just for the purpose of imputing failure times, whose rankings are then used in a marginal likelihood approach. Estimation of the proportional hazards model with unspecified baseline hazard function has been addressed by several authors. Finkelstein [19] discusses the two-group case, proposing a generalization of the logrank test. A computational difficulty is that the regression coefficients need to be constrained so that estimated discrete hazards lie between 0 and 1. Huang [20], Satten [21], Sun [22], Alioum and Commenges [23] and Goggins *et al.* [24] also discuss this problem. As for right censored data, these non-parametric methods do not provide a useful estimate of the baseline hazard function. We extend our local likelihood estimation procedure for interval censored without covariates to the proportional hazards model for the case of covariates, thereby obtaining a smooth baseline hazard and a global covariate vector.

Global regression parameters, as produced by the proportional hazards model, are useful in situations in which a simple summary of the data is desired, such as treatment effect. In some applications, however, the proportional hazards assumption may be questionable and it is of interest to allow for an arbitrary smooth covariate function, $s(\boldsymbol{Z})$, rather than $\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})$. Several authors have applied smoothing methods to such non-parametric hazard models for right censored data. These include splines [25, 26], kernels [27] and local likelihood [6, 28, 29]. Our methods for interval censored data presently do not apply to these situations; appropriate extensions are possible [30], although are likely to be computationally challenging.

In Section 2 we describe the technique of local likelihood and its application to fitting regression models for right censored survival data. In Section 3 we discuss the extension to interval censored data. We apply the method to data from Kroner *et al.* [31] and Finkelstein [19] in Section 4 and conclude with a discussion of future research directions in Section 5.

Copyright © 2002 John Wiley & Sons, Ltd.

2. LOCAL LIKELIHOOD ESTIMATION FOR RIGHT CENSORED DATA

2.1. Background and global PH model

Local likelihood estimation was popularized by Tibshirani and Hastie [28] as a means of allowing covariates to enter likelihood-based regression models via an unspecified smooth function. To illustrate this approach, let $(z_1, y_1), \ldots, (z_n, y_n)$ be n independent realizations of covariates Z and responses Y. Suppose that conditional on $Z = z$, Y has density $f_Y(y, \beta(z))$. Instead of maximizing the full likelihood

$$L = \prod_{i=1}^{n} f(y_i, \beta)$$

with the covariate effect characterized by a strong modelling assumption, this approach assumes that $\beta = s(z_i)$ where $s(.)$ is an unspecified smooth function. The local likelihood estimate of $s(z_i)$ is simply a polynomial approximation to $s(.)$ at z_i. The parameters that characterize the shape of the function in the kth local likelihood are estimated by maximizing

$$L_k = \prod_{i: z_i \in N_k} f(y_i, \beta_k z_i)$$

where N_k is a neighbourhood of the z_i's. An alternative representation is the weighted likelihood

$$L_k = \prod_{i=1}^{n} f(y_i, \beta_k z_i)^{w_{ik}}$$

where w_{ik} is an indicator of $z_i \in N_k$.

While local likelihood methods have become popular in the generalized linear models setting, they have been less widely used for proportional hazards survival data. One of the complications in the survival setting is the presence of the infinite dimensional baseline hazard function. In this paper, we focus on the use of local likelihood to smooth the baseline hazard in a proportional hazards regression model. That is, we fit model (1) with a fully parameterized 'global' covariate function and a locally parameterized baseline hazard function, $\lambda(t)$.

Similar ideas are discussed by Wu and Tuma [6, 32] who propose the following hazard model:

$$\lambda(t; v, z) = \exp(\beta v) \exp\{s_0(t) + s_1(t) z\}$$

Thus, they assume proportional hazards for the covariate v, but the covariate z can have non-proportional effects. They propose using local likelihood to estimate the s functions (note that $\exp(s_0(t))$ serves the role of the baseline hazard function). In their paper, however, they consider only the case of no covariates (neither v nor z) and do not describe how estimation can be accomplished more generally. They use the local Fisher information to obtain pointwise confidence intervals. They have no way of performing global tests of group effects. Loader (1998, documentation for Locfit package available at the Internet site `cm.bell-labs.com/stat/project/locfit`) proposes a similar hazard model:

$$\lambda(t; v, z) = \exp(a_0 + a_1 t + a_2 z + a_3 t^2 + a_4 z t + a_5 z^2)$$

Copyright © 2002 John Wiley & Sons, Ltd.

This is more general than Wu and Tuma's [6] model in that it does not assume linearity in the covariate z. Instead, it assumes that the relationship between survival time t and the covariate z can be described as a local plane. However, the method has the same limitation as Wu and Tuma in that it does not allow for a global covariate, such as might be used to summarize group differences. Loader [29] uses a kernel function to define the neighbourhoods and to weight the observations in the neighbourhoods. In theory, one could calculate pointwise confidence intervals for the hazard function for a given t and z.

Let $T_1, \ldots, T_n$ be a set of observed failure or censoring times. Setting up a local likelihood approach requires consideration of the full log-likelihood:

$$\ell = \sum_{i=1}^{n} \left\{ \delta_i \ln \lambda(T_i) + \delta_i \boldsymbol{\beta}^{\mathsf{T}} \mathbf{z}_i - \int_0^{T_i} \mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{z}_i} \lambda(u)\, \mathrm{d}u \right\} \tag{2}$$

where $\lambda(.)$ is the hazard function, δ_i indicates whether or not the ith individual failed, $\boldsymbol{Z}$ represents the $n \times q$ matrix of covariate values, $\mathbf{z}_i$ represents the $q \times 1$ covariate vector for individual i, and $\boldsymbol{\beta}$ a $q \times 1$ vector of coefficients. Let $\boldsymbol{\delta}$ be the $n \times 1$ vector of censoring indicators, $\boldsymbol{\lambda}$ be the $n \times 1$ vector of individual baseline hazards evaluated at the observed failure or censoring times and $\boldsymbol{\Lambda}$ be the $n \times 1$ vector of cumulative hazards, also evaluated at the observed failure or censoring times. In other words, the jth element of $\boldsymbol{\lambda}$ is $\lambda(T_i)$, while the jth element of $\boldsymbol{\Lambda}$ is $\int_0^{T_i} \lambda(u)\, \mathrm{d}u$. Finally, for a vector $\mathbf{v}$, let $\log(\mathbf{v})$ be the vector obtained by applying the logarithm to each entry of $\mathbf{v}$; $\mathrm{e}^{\mathbf{v}}$ is defined similarly. Then we may rewrite (2) in this matrix notation as

$$\ell = \boldsymbol{\delta}^{\mathsf{T}} \log \boldsymbol{\lambda} + \boldsymbol{\delta}^{\mathsf{T}} \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{\Lambda}^{\mathsf{T}} \mathrm{e}^{\boldsymbol{Z}\boldsymbol{\beta}}$$

The local likelihood estimate of the hazard function at the point t begins with a local approximation to $\ln\{\lambda(t)\}$ in the smoothing window with bandwidth $h(t)$ around t:

$$\ln\{\lambda(s)\} \approx \alpha_{0t} + \alpha_{1t}(s-t) + \cdots + \alpha_{pt}(s-t)^p \quad \text{for } |s-t| \leqslant h$$

For given $\boldsymbol{\beta}$, and using a generic kernel function K, the contribution of the ith individual to the local log-likelihood at the point t is equal to

$$\begin{aligned} &\delta_i \boldsymbol{\beta}^{\mathsf{T}} \mathbf{z}_i + \delta_i K\left(\frac{T_i - t}{h}\right)\{\alpha_{0t} + \alpha_{1t}(v-t) + \cdots + \alpha_{pt}(v-t)^p\} \\ &\quad - \int_0^{T_i} \mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{z}_i} \mathrm{e}^{\alpha_{0t} + \alpha_{1t}(u-t) + \cdots + \alpha_{pt}(u-t)^p} K\left(\frac{u-t}{h}\right) \mathrm{d}u \end{aligned} \tag{3}$$

Note that this is a more general case of the Wu and Tuma [6, 32] model, since they use a rectangular kernel with overlapping regions and only the constant term α_{0t} to model the hazards.

2.2. Estimation

Generalized estimating equations provide a flexible and powerful approach for estimation [33]. Although not necessary for estimation with right censored data, we set up here a general framework which will facilitate extensions to the interval censored case.

Copyright © 2002 John Wiley & Sons, Ltd.

First define a grid, $\mathscr{G}$, of M points at which the hazard function will be estimated. Note that M should be chosen so that integrals involving the hazard function can be accurately approximated using quadrature based on these points. At a grid point $t \in \mathscr{G}$, the $p+1$ score equations (from a pth order polynomial approximation to the hazard function) are given by:

$$\sum_{i=1}^{n}\left\{\delta_i K\left(\frac{T_i-t}{h}\right)(T_i-t)^d - \int_0^{T_i} \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} z_i}(u-t)^d \mathrm{e}^{\{\alpha_{0t}+\cdots+\alpha_{pt}(u-t)^p\}} K\left(\frac{u-t}{h}\right)\mathrm{d}u\right\}=0$$

for $d=0,\ldots,p$. When the baseline hazard is known, the regression coefficient $\boldsymbol{\beta}$ can be estimated by solving the q score equations given by

$$\sum_{i=1}^{n}\left\{\delta_i Z_{ij} - Z_{ij}\mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} z_i}\int_0^{T_i}\lambda(u)\,\mathrm{d}u\right\}=0$$

for $j=1,\ldots,q$.

In Section 3.1, we describe in detail an algorithm for solving these systems of estimating equations which iterates between estimation of the underlying hazards, assuming the $\boldsymbol{\beta}$ are known, and estimation of the $\boldsymbol{\beta}$, assuming the hazards are known. Alternatively, the equations could be solved using a Newton–Raphson algorithm, although this approach would involve computing complicated derivatives, and thus is not as appealing as the iterative alternating algorithm.

For the model with $p=0$, corresponding to the constant hazards model, the equations reduce to

$$\sum_{i=1}^{n}\left\{\delta_i K\left(\frac{T_i-t}{h}\right) - \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} z_i+\alpha_{0t}}\mathscr{K}\left(\frac{u-t}{h}\right)\right\}=0$$

$$\sum_{i=1}^{n}\left\{\delta_i Z_{ij} - Z_{ij}\mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} z_i+\alpha_{0t}}\Lambda\left(\frac{u-t}{h}\right)\right\}=0$$

where $\mathscr{K}(s)$ is the integral from 0 to s of the kernel function K. The estimates for the hazard $\hat{\lambda}(t)=\exp(\hat{\alpha}_{0t})$ have the form

$$\frac{\sum_{i=1}^{n}\delta_i K(\frac{T_i-t}{h})}{\sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} z_i}\mathscr{K}(\frac{T_i-t}{h})}=\frac{\boldsymbol{\delta}^\mathsf{T}\mathbf{K}}{\mathrm{e}^{\mathbf{Z}\boldsymbol{\beta}}\boldsymbol{\mathscr{K}}}$$

where $\mathbf{K}$ and $\boldsymbol{\mathscr{K}}$ are the $n\times 1$ vectors of the kernel function, K, and the cumulative kernel function, $\mathscr{K}$, evaluated at $(T_i-t)/h$. The solution has an appealing analogy to the estimator of the baseline hazard rate from a piecewise exponential model [13]; the numerator corresponds to a weighted number of failures and the denominator to a weighted time at risk in the region defined by the kernel.

3. LOCAL LIKELIHOOD ESTIMATION FOR INTERVAL CENSORED DATA

In some applications only intervals of time, (L_i,R_i), $i=1,\ldots,n$, which are known to contain the failure times T_i, are observed. In this section, we extend the methods described in Section 2 for right censored data to obtain smooth hazard and covariate estimates for data where some

Copyright © 2002 John Wiley & Sons, Ltd.

failure times are observed ($L_i = R_i$), some are right censored ($R_i = \infty$), and some are interval censored. The log-likelihood contributions of interval censored observations are of the form

$$\ln\left\{\int_{L_i}^{R_i} \lambda(s)\mathrm{e}^{-\int_0^s \lambda(u)\,\mathrm{d}u}\,\mathrm{d}s\right\}$$

Betensky *et al.* [10] describe a procedure for estimating the hazard function with no covariates using a local EM algorithm that applies to estimating equations [34]. We extend their methods by alternating between the algorithm they describe for estimating the baseline hazards, assuming the covariate effects are known, and the simpler estimation of the covariate effects, assuming the baseline hazard function is known.

We use the example of a locally linear fit for illustration.

Step 0 (*initialization step*). Choose the kernel K, the grid of points for estimation, $\mathscr{G}$, and the bandwidth h. For observed failures, set $\delta_i = 1$ and $T_i = L_i$, for right censored observations, set $\delta_i = 0$ and $T_i = L_i$, and for interval censored observations set $\delta_i = 1$ and $T_i = (L_i + R_i)/2$, the midpoint of (L_i, R_i). Estimate the initial $\boldsymbol{\beta}$ by fitting a standard proportional hazards model. Set $\hat{\lambda}(t)$ to be the local linear estimator of $\lambda(t)$ based on the (T_i, δ_i), as described in Section 2.

Step 1. Estimate λ, assuming $\boldsymbol{\beta}$ is fixed and known:

Step 1.1. For each $t \in \mathscr{G}$, set $\hat{S}(t) = \exp\{-\int_0^t \hat{\lambda}(u)\,\mathrm{d}u\}$ and $\hat{f}(t) = \hat{\lambda}(t)\hat{S}(t)$, where the integral is approximated using the trapezoidal rule.

Step 1.2. Let $\hat{S}_i(t) = \hat{S}(t)^{\exp(\boldsymbol{\beta}^{\mathsf{T}} z_i)}$ and $\hat{f}_i(t) = \hat{\lambda}(t)\exp(\boldsymbol{\beta}^{\mathsf{T}} z_i)\hat{S}_i(t)$, for $i = 1, \ldots, n$. For each $t \in \mathscr{G}$, reset $\hat{\lambda}(t) = \mathrm{e}^{\hat{\alpha}_{0t}}$ where

$$\begin{bmatrix}\hat{\alpha}_{0t}\\ \hat{\alpha}_{1t}\end{bmatrix} = \underset{(\alpha_{0t},\alpha_{1t})^{\mathsf{T}}}{\operatorname{argmax}}\left[\sum_{i=1}^n E_{\hat{f}_i}\ell(T_i,t)|L_j \leqslant T_j \leqslant R_j, 1 \leqslant j \leqslant n\right]$$
$$= \underset{(\alpha_{0t},\alpha_{1t})^{\mathsf{T}}}{\operatorname{argmax}} \sum_{i=1}^n \left[\frac{\int_{L_i}^{R_i} \ell_i(v,t)\hat{f}_i(v)\,\mathrm{d}v}{\int_{L_i}^{R_i} \hat{f}_i(v)\,\mathrm{d}v}\right]$$

and $\ell_i(v,t)$ is the contribution to the local likelihood at t of a failure at v, obtained from the summand in (3):

$$\ell_i(v,t) = K\left(\frac{v-t}{h}\right)\{\alpha_{0t} + \alpha_{1t}(v-t)\} + \boldsymbol{\beta}^{\mathsf{T}} z_i$$
$$- \mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}} z_i}\int_0^v \mathrm{e}^{\alpha_{0t}+\alpha_{1t}(u-t)}K\left(\frac{u-t}{h}\right)\mathrm{d}u$$

The notation '$\operatorname{argmax}_x g(x)$' denotes the value of x that maximizes g, and $E_{\hat{f}}$ refers to the expectation calculated under the distribution $\hat{f}$. The integrals are computed using the trapezoidal rule, and the Newton–Raphson algorithm is used to solve for $\hat{\alpha}_{0t}$ and $\hat{\alpha}_{1t}$.

Step 1.3. Repeat steps 1.1 and 1.2 until convergence.

Copyright © 2002 John Wiley & Sons, Ltd.

Step 2. Find the maximum likelihood estimates of the $\boldsymbol{\beta}$ assuming the baseline hazard is known. That is, solve

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \ln\{\hat{S}(L_i)^{\exp(\boldsymbol{\beta}^{\mathsf{T}} z_i)} - \hat{S}(R_i)^{\exp(\boldsymbol{\beta}^{\mathsf{T}} z_i)}\} \right.$$

using a Newton-Raphson algorithm, for example.

Step 3. Iterate steps 1 and 2 until convergence.

In the examples that we describe in Section 4, this algorithm was quite fast; it required five iterations at about 45 seconds per iteration. Modifications of the algorithm are needed to accommodate right censored and exactly observed observations. With right censored observations, the contribution to the local log-likelihood is simply expressed as

$$-\mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}} z_i} \int_0^{T_i} \mathrm{e}^{\alpha_{0t}+\alpha_{1t}(u-t)} K\left(\frac{u-t}{h}\right) \mathrm{d}u \tag{4}$$

where T_i is the time of right censoring. Thus, the algorithm is easily modified to accommodate right censored observations by substituting (4) for the term

$$\frac{\int_{L_i}^{R_i} \ell(v,t) \hat{f}(v)\,\mathrm{d}v}{\int_{L_i}^{R_i} \hat{f}(v)\,\mathrm{d}v} \tag{5}$$

in step 1.2. For exact data or even when $[L_i, R_i]$ is very narrow it is recommended [10] that one replace (5) by the numerically more stable approximation $\ell((L_i+R_i)/2, t)$. This expression follows from application of l'Hopital's rule as $L_i \to R_i$.

4. DATA EXAMPLES

4.1. Breast cosmesis data

Our first example involves a reanalysis of data provided by Finkelstein [19] from a study that followed breast cancer patients to determine the long-term effects of therapy on the occurrence of cosmetic deterioration. Of the 94 women in the study, 46 received radiotherapy alone and 48 received radiotherapy plus chemotherapy. Because they were assessed only at their clinic visits, the time of deterioration is known only to have occurred between successive visits (usually about 6 months apart) and so they are interval censored. By the end of the study, 56 women had experienced cosmetic deterioration and thus were interval censored, and 38 had not shown evidence of deterioration and hence were right censored.

Figure 1 shows the estimated hazard function for each group of women based on the assumption of proportionality. This estimate is based on a binned implementation of a Gaussian kernel with 201 grid points. The bandwidth was taken to be locally varying according to a simple nearest neighbour rule that ensured that 40 per cent of the data were in the fitting window. The resulting estimate of β is 1.053, and a bootstrap standard error based on 1000 bootstrap samples is 0.270, suggesting that hazard for breast cosmesis deterioration is significantly higher when radiotherapy is accompanied by adjuvant chemotherapy than when it is given alone. A kernel density estimate of the bootstrap distribution of $\hat{\beta}$ is displayed in

Copyright © 2002 John Wiley & Sons, Ltd.

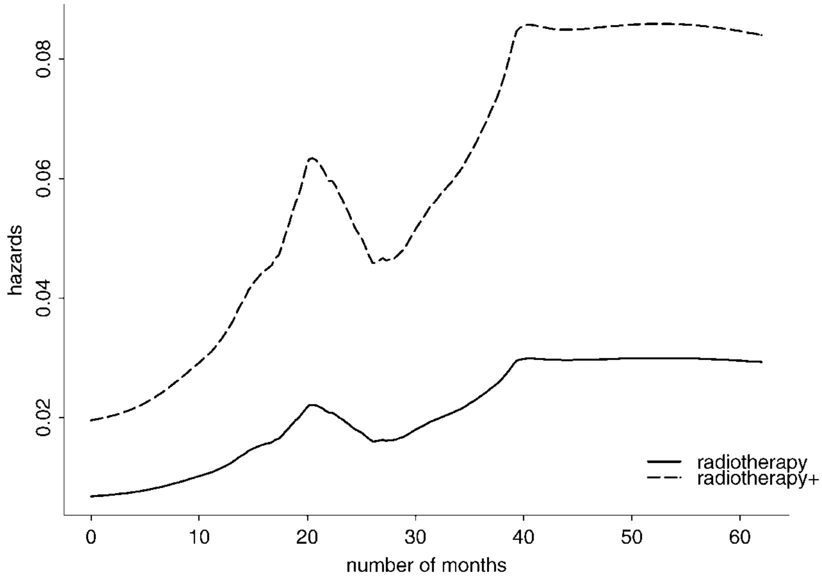

Figure 1. Estimated hazards for the breast cosmesis data.

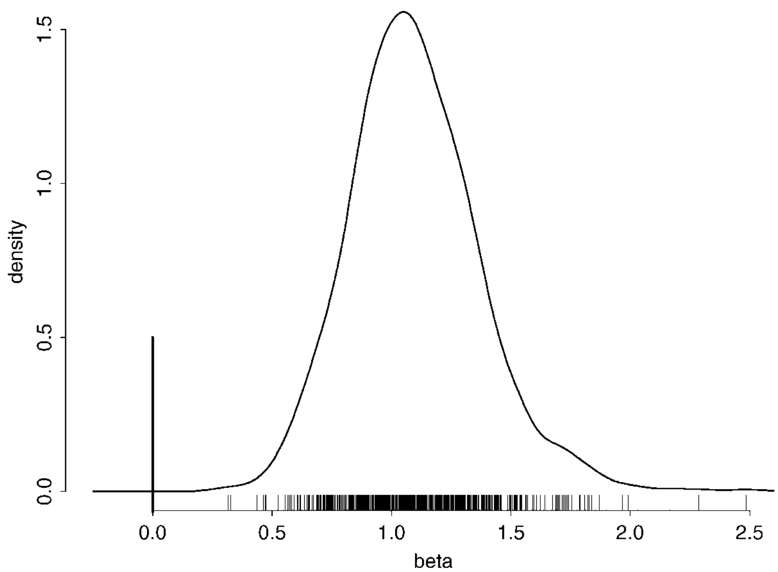

Figure 2. Sample of size 1000, with kernel density estimate, from bootstrap distribution of $\hat{\beta}$ for the breast cosmesis data.

Figure 2. For comparison, Table I lists the parameter estimates and standard errors obtained from a variety of different approaches. Figure 3 displays the local likelihood estimates of the baseline hazard function assuming each of the values of β given in Table I (that is, the

Copyright © 2002 John Wiley & Sons, Ltd.

Table I. Treatment effect for breast cosmesis data.

Model	Estimate	Standard error
Finkelstein [19]	0.791	0.288
Exponential	0.742	0.277
Satten [21]	0.890	0.297
Satten *et al.* [18]	0.878	0.294
Goggins *et al.* [24]	1.450	0.371
Piecewise (eight intervals)	0.930	0.287
Local likelihood	1.053	0.270

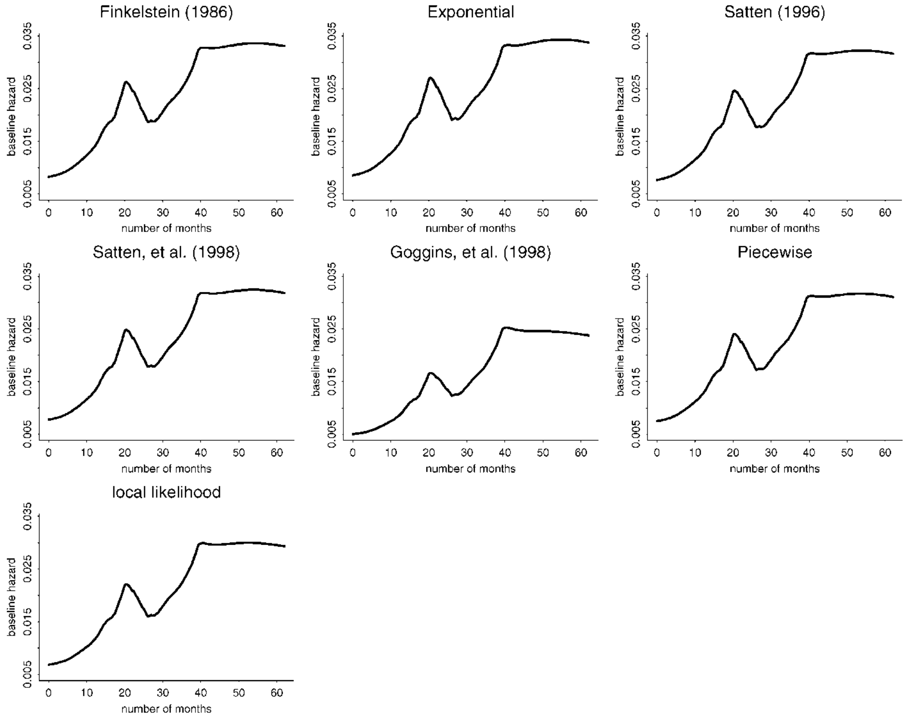

Figure 3. Estimated baseline hazards for the breast cosmesis data.

solutions to step 1 of our algorithm). Not surprisingly, given the similarity of the estimates in Table I, these baseline hazard functions all look similar, except for that based on the β of Goggins *et al.* [24], which is lower and flatter than the others. There are smaller differences among the other curves in their heights; the local likelihood estimator gets as large as 0.30, whereas the exponential estimator gets as large as 0.35.

It is of interest to compare the hazard functions in Figure 1 with the hazard function derived based on ignoring treatment displayed in Figure 2 of Betensky *et al.* [10]. The shapes of the

Copyright © 2002 John Wiley & Sons, Ltd.

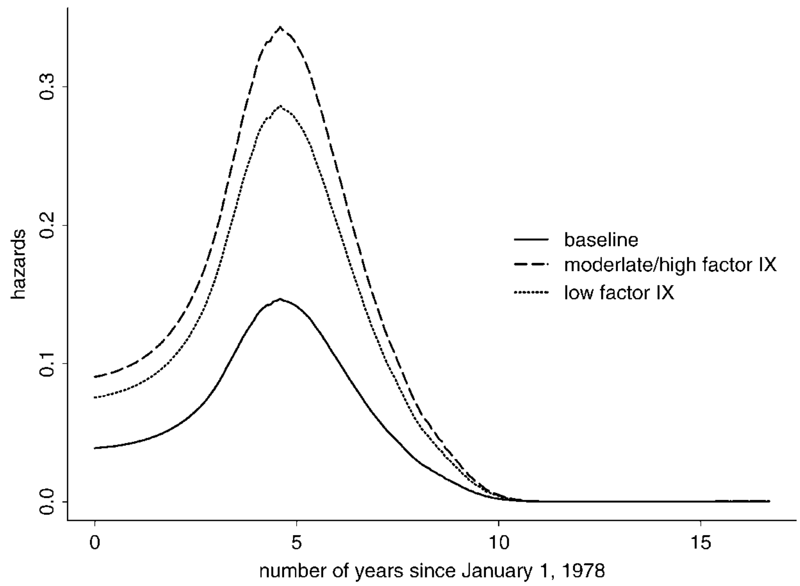

Figure 4. Estimated hazards for the haemophilia data.

hazard functions are similar, suggesting, via straightforward calculations, that the baseline hazard function is small relative to the treatment effect, exp(β), which indeed we see to be the case.

4.2. Haemophilia data

People with haemophilia were at high risk of infection with HIV-1 during the 1980s due to their need for infusions of factor VIII or factor IX concentrate, products made from the plasma of thousands of donors. HIV-1 infection incidence was studied using data from a 16-centre cohort of haemophiliacs in the United States and Europe [12, 31]. In five of these centres, patients were enrolled without regard to HIV-1 antibody status. Positive samples were collected between early 1978 until early 1987. Any samples before 1 January 1978 were considered to be HIV-1 negative as the earliest positive tests for the disease occurred in the spring of that year. Here we assess the effect of dose of factor IX concentrate on the hazard for HIV-1 infection, without regard for the dose of factor VII concentrate. Our data set contained information on 681 individuals with dose information and is updated from the data set discussed in the previous papers. Data are interval censored because individuals in the study were only tested periodically for infection. Information was also available on several different characteristics of the individuals under study.

Figure 4 displays the estimated hazard functions for individuals who received no factor IX concentrate ('baseline'), individuals who received a low dose of factor IX concentrate (<20000 U), and individuals who received a high dose of factor IX concentrate ($\geqslant 20000$ U). The kernel smoothing parameters were the same as for the analysis of the

Copyright © 2002 John Wiley & Sons, Ltd.

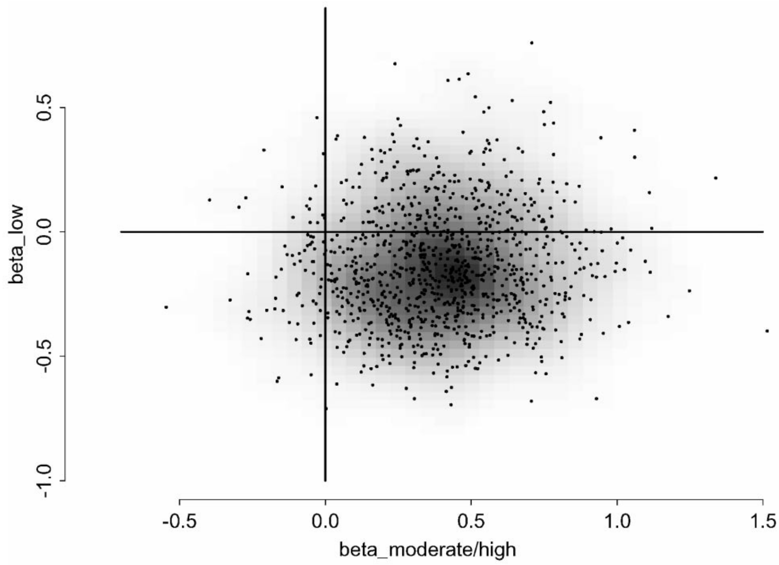

Figure 5. Sample of size 1000 from bootstrap distribution of $(\hat{\beta}_{\text{moderate/high}}, \hat{\beta}_{\text{low}})$ for the haemophilia data. The grey levels correspond to a kernel density estimate.

breast cosmesis data described in Section 4.1. The estimates and bootstrap standard deviations are $\hat{\beta}_{\text{moderate/high}} = 0.850$ (0.271) and $\hat{\beta}_{\text{low}} = 0.669$ (0.234). As expected, those individuals who received the highest dose of factor IX concentrate had the highest hazard for infection with HIV-1. Figure 5 displays a kernel density estimate of the bivariate bootstrap distribution of $(\hat{\beta}_{\text{moderate/high}}, \hat{\beta}_{\text{low}})$.

5. DISCUSSION

We have proposed the use of local likelihood methods for the proportional hazards regression analysis of right and interval censored data. Specifically, we assume a parametric, global covariate function, but make only minimal assumptions on the baseline hazard function. Our method produces estimates of the global covariate vector, as well as a smooth baseline hazard. Estimation of smooth baseline hazard functions increases the interpretability and understanding of the failure process, beyond the contribution of the regression coefficients which simply convey the relative impacts of the covariates. This is particularly useful in the presence of multiple covariates, which our approach can accommodate, when estimating hazard functions separately by covariate groups is not feasible due to sparse data. We use the bootstrap to obtain variance estimates of the covariate effects in our examples. Further research on the possibility of analytic approximations to the variance is warranted, with the downside of requiring complicated calculations. Fruitful approaches

Copyright © 2002 John Wiley & Sons, Ltd.

are likely to be based on the framework of estimating equations or on a profile likelihood function.

ACKNOWLEDGEMENTS

This research was supported by NIH grants CA75791 and CA48061 and NIAID contract 1U01 4I 41110. We thank Philip Rosenberg for use of the hemophilia data.

REFERENCES

1. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
2. Nelson WB. Theory and applications of hazard plotting for censored failure data. *Technometrics* 1972; **14**: 945–965.
3. Gray RJ. Some diagnostic methods for Cox regression models through hazard smoothing. *Biometrics* 1990; **46**:93–102.
4. Whittemore AS, Keller JB. Survival estimation using splines. *Biometrics* 1986; **42**:495–506.
5. Rosenberg PS, Goedert JJ, Biggar RJ. Effect of age at seroconversion on the natural AIDS incubation distribution. *AIDS* 1994; **8**:803–810.
6. Wu LL, Tuma NB. Local hazard models. *Sociological Methods* 1990; **20**:141–180.
7. Peto R. Experimental survival curves for interval-censored data. *Applied Statistics* 1973; **22**:86–91.
8. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, *Series B* 1976; **38**:290–295.
9. Gentleman R. Geyer CJ. Maximum likelihood for interval censored data: consistency and computation. *Biometrika* 1994; **81**:618–623.
10. Betensky RA, Lindsey JC, Ryan LM, Wand MP. Local EM estimation of the hazard function for interval-censored data. *Biometrics* 1999; **55**:238–245.
11. Kooperberg C, Stone CJ. Logspline density estimation for censored data. *Journal of Computation and Graphical Statistics* 1992; **1**:301–328.
12. Rosenberg PS. Hazard function estimation using B-splines. *Biometrics* 1995; **51**:874–887.
13. Lindsey JC, Ryan LM. Tutorial in Biostatistics. Methods for interval-censored data. *Statistics in Medicine* 1998; **17**:219–238.
14. Farrington CP. Interval censored survival data. *Statistics in Medicine* 1996; **15**:283–292.
15. Borgan Ø, Liestol K, Ebbesen P. Efficiencies of experimental designs for an illness-death model. *Biometrics* 1984; **40**:627–638.
16. Chiang Y, Hardy RJ, Hawkins CM, Kapadia AS. An illness-death process with time-dependent covariates. *Biometrics* 1989; **45**:669–681.
17. Brookmeyer R, Goedert JJ. Censoring in an epidemic with an application to Hemophilia-associated AIDS. *Biometrics* 1989; **45**:325–335.
18. Satten G, Datta S, Williamson JM. Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* 1998; **93**:318–327.
19. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986; **42**: 845–854.
20. Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* 1996; **24**:540–568.
21. Satten G. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* 1996; **83**:355–370.
22. Sun J. A non-parametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* 1996; **15**:1387–1395.
23. Alioum A, Commenges D. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics* 1996; **52**:512–524.
24. Goggins WB, Finkelstein DM, Schoenfeld DA, Zaslavsky AM. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* 1998; **54**:1498–1507.
25. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**:941–949.
26. Kooperberg C, Clarkson DB. Hazard regression with interval-censored data. *Biometrics* 1996; **53**:1485–1494.
27. Gray RJ. Hazard rate regression using ordinary nonparametric regression smoothers. *Journal of Computational and Graphical Statistics* 1996; **5**:190–207.

Copyright © 2002 John Wiley & Sons, Ltd.

28. Tibshirani R, Hastie T. Local likelihood estimation. *Journal of the American Statistical Association* 1987; **82**:559–567.
29. Loader CR. Local likelihood density estimation. *Annals of Statistics* 1996; **24**:1602–1618.
30. Fan J, Gijbels I, King M. Local likelihood and local partial likelihood in hazard regression. *Annals of Statistics* 1997; **25**:1661–1690.
31. Kroner BL, Rosenberg PS, Aledort LM, Alvord WG, Goedert JJ. HIV-1 infection incidence among persons with hemophilia in the United States and Western Europe, 1978–1990. *Journal of AIDS* 1994; **7**:279–286.
32. Wu LL, Tuma NB. Assessing bias and fit of global and local hazards models. *Sociological Methods in Research* 1991; **19**:354–387.
33. Liang KY, Zeger SL. Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* 1995; 158–173.
34. Elashoff M, Ryan LM. Using the ES algorithm to adjust for missing data in estimating equations. *Journal of Computational and Graphical Statistics* 2001 (tentatively accepted).

Copyright © 2002 John Wiley & Sons, Ltd.