SMOOTHING GROUPED BIVARIATE DATA TO OBTAIN THE INCUBATION PERIOD DISTRIBUTION OF AIDS

JEREMY M. G. TAYLOR AND YUN CHON

Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90024, U.S.A.

SUMMARY

We use a penalized likelihood approach to obtain a smooth estimate of a bivariate distribution from grouped data where each observation consists of a region in a plane. The purpose of the analysis is to estimate the incubation period distribution of AIDS from the Multicenter AIDS Cohort Study, a prevalent cohort of homosexual men. In this article we illustrate the usefulness of the penalized likelihood approach. We also discuss the use of a cross-validation and a Bayesian scheme to choose the smoothing parameters and bootstrap samples to assess uncertainty.

INTRODUCTION

This paper describes an analysis of a specific AIDS-related data set, and follows a similar analysis of an older published version of this data set.[1] We develop in more depth the statistical issues in the analysis. The scientific problem is estimation of the incubation period distribution of AIDS from a cohort study of homosexual men recruited in Los Angeles in 1984–5. The incubation period is the time interval from infection with the AIDS virus (HIV) to the onset of clinical symptoms (AIDS). Its distribution is important both as a summary of the natural history of the disease and for its utility in predicting the future course of the epidemic. It has been shown[2] that to estimate the AIDS incubation period with data from a cohort study, one must model jointly both the incubation period and the date of HIV infection. Because of the nature of the study, however, for most subjects the exact values of these two variables are unknown but there is some information concerning their possible values. In statistical terms, the problem is that of estimating the joint bivariate distribution of two random variables when the observed data are grouped, that is each observation consists of a region in the plane. In the estimation scheme, we make minimal assumptions concerning the bivariate distribution and use a penalized likelihood approach to obtain smooth marginal distributions. The methods used also incorporate truncation in the sampling scheme and we discuss how we can introduce covariates that influence the joint distribution.

Previous work in this area using related methodology to estimate the incubation period distribution of AIDS has been performed by others, using both parametric models[3–5] and semi-parametric and non-parametric approaches.[1,2,6,7]

STATISTICAL DESCRIPTION OF THE PROBLEM

Because the methodology applies to situations other than AIDS, we describe it first in general terms. There is a sample of n subjects; the observation on subject i consists of a known region B_i in

CCC 0277-6715/94/090969-13

Received August 1992
Revised June 1993

the bivariate positive quadrant $\mathbb{R}^+ \times \mathbb{R}^+$. That is, there is an unknown specific value $(x_i, y_i) \in \mathbb{R}^+ \times \mathbb{R}^+$ for each subject, but this value has not been observed. All that we know is that (x_i, y_i) lies within a known region B_i. In addition, covariates Z_i, which could depend on (x, y) exist for each subject. Also, the samples is truncated in the sense that there is a truncated region T_i such that if (x_i, y_i) had been in region T_i, then subject i is excluded from the sample, with no knowledge of his existence. By definition, B_i and T_i are disjoint. The aim is to estimate the joint distribution of (x, y) given Z; denote the corresponding density indexed by parameter θ by

$$m(x, y; Z, \theta), \quad \text{where} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} m(x, y; Z, \theta)\, \mathrm{d}x\mathrm{d}y = 1.$$

The likelihood of the observations is

$$L = \prod_{i=1}^{n} L_i(m) = \prod_{i=1}^{n} \left\{ \frac{\int_{B_i} m(x, y; Z_i, \theta)\mathrm{d}x\mathrm{d}y}{1 - \int_{T_i} m(x, y; Z_i, \theta)\mathrm{d}x\mathrm{d}y} \right\}$$

We need the denominator in this expression to account for the truncation in the sampling scheme. One approach is to specify m parametrically, for example, with a bivariate log-normal distribution, to maximize L and base inference on the likelihood surface. In this article, we use a more non-parametric approach by making only weak assumptions concerning m, and we ensure that the estimate of m is smooth by maximizing a penalized likelihood.[6,8] In particular, we maximize

$$\log L - P(m, \lambda),$$

where $P(m, \lambda)$ is a penalty function which is large if the density m is 'rough' and small if m is 'smooth'. $P(m, \lambda)$ is a non-negative function for which $P(m, \mathbf{0}) = 0$. Penalized likelihood balances agreement between the data and the model, as measured by large $\log L$, against smoothness of the estimator, as measured by small $P(m, \lambda)$. The vector λ controls the degree of smoothness, in the sense that larger values of each coordinate of λ will increase the smoothness of the estimated bivariate distribution of m. In these problems, one usually estimates the value of λ separately from the parameters of m.

DESCRIPTION OF THE DATA

The data are from the 1637 homosexual men who enrolled in the Los Angeles portion of the Multicenter AIDS Cohort Study.[9,10] All participants were AIDS-free at the time of their enrolment between April 1984 and February 1985. Follow-up visits were scheduled every 6 months during which blood was drawn for HIV antibody testing. AIDS dianosis information is obtained from the participants, their friends, their doctors, disease registries, death certificates and newspaper obituaries.

There is a potential for bias in these procedures in the sense that lost participants are likely to reappear later only if they develop AIDS. To counter this, we assumed that any drop-out prior to July 1987 who, at the time of the final gathering of the data (January 1991), had not yet been reported to have AIDS, did not have AIDS before July 1987.

As the definition of AIDS changed in September 1987, we excluded AIDS diagnoses that were only applicable after this date to ensure homogeneity of the endpoint. Some subjects in the study had stopped attending the scheduled 6 month visits but continued to be followed by telephone or mail.

The bivariate ($\mathbb{R}^+ \times \mathbb{R}^+$) region fundamental to these data is (date of HIV infection) $\times$ (incubation period), denoted by (x, y). Each subject has a true value of (x, y), which is not known exactly but is known to lie in a region B_i. Let $t = x + y$ be the date of AIDS. In our analysis we take $x = 0$ as 1 January 1979, and the closing date as 31 December 1989. We could know both x and t exactly, or they could be right censored or interval censored. The exact position and shape of B_i differ for each subject. We can classify each of the 1637 participants into one of eight possible shapes for B_i. Each subject has a number of important dates that determine the boundary for B_i. These dates are $I_0, I_1, I_2, A_1, A_2, A_3$ and A_4, where I_0 is the date of enrolment, I_1 is the date of the last HIV negative test, I_2 is the date of the first HIV positive test, A_1 is the date of AIDS diagnosis, A_2 and A_3 bound the date of AIDS when it is interval censored and A_4 is the last follow-up date at which we know AIDS had not been diagonsed. The boundaries that define the eight shapes for B_i are (1) $x < I_0$, $t > A_4$; (2) $x < I_0$, $t = A_1$; (3) $x < I_0$, $A_2 < t \leqslant A_3$; (4) $I_1 < x \leqslant I_2$, $t > A_4$; (5) $I_1 < x \leqslant I_2$, $t = A_1$; (6) $I_1 < x \leqslant I_2$, $A_2 < t \leqslant A_3$; (7) $x > I_1$, $t > I_1$; (8) $x > I_1$, $t > A_4$. Participants in region 1 were HIV seroprevalent at enrolment and have not developed AIDS; regions 2 and 3 are seroprevalent with AIDS; regions 4, 5 and 6 are seroconverters with or without AIDS; and regions 7 and 8 are seronegative. The numbers of subjects who possess each of the eight different shapes are 545, 255, 9, 93, 12, 0, 589, 134, respectively. There are 914 subjects known seropositive some time during their follow-up, and a total of 276 AIDS cases, but only 12 people for whom we know both the date of infection and the incubation period with reasonable accuracy. The truncation region T_i is $\{(x, y): x + y < I_0\}$.

BIVARIATE CONTINGENCY TABLE MODEL

We adopt a semi-parametric approach for the estimation of m. The data are discretized into 6 month units, converting the $\mathbb{R}^+ \times \mathbb{R}^+$ bivariate space into a 23×23 contingency table. Let j denote the index of date of infection, where $j = 1$ indicates $x \in [1$ January 79, 30 June 79$]$ and $j = 23$ indicates $x > 31$ December 89. Similarly $k = 1$ indicates $y \in (0, 0{\cdot}5]$ and $k = 23$ indicates $y > 11$ years.

The contingency table model for the ith subject in cell (j, k) is

$$P_{jk}(Z_i) = f_j\, g_{k|j}(Z_i), \quad \text{where } \sum_{j=1}^{23} f_j = 1 \text{ and } \sum_{k=1}^{23} g_{k|j}(Z_i) = 1 \text{ for all } j, i,$$

where f_j is the probability distribution for date of infection, and $g_{k|j}(Z_i)$ is the probability distribution for the incubation period given that the date of infection is j and the covariates are Z_i. With this discrete formulation, the integrals in the likelihood become sums in an obvious manner.

Motivated by theoretical models for the growth of the HIV epidemic,[11] we assume $f_j = c e^{j-4}/(1 + e^{j-4})$ for $j = 1, \ldots, 6$, where c is a parameter to estimate from the data. This parameterization is also useful to reduce instability problems in the estimation procedure partly because we have reduced the number of parameters by 5. The enrolment of the cohort occurs between April 1984 ($j = 11$) and February 1985 ($j = 13$), so there is little information in the data to assist in the estimation of f_j for $j < 10$. We chose to bridge part of this gap by making the above mild, yet flexible and reasonable, parametric assumption, rather than to rely completely on the smoothness induced by the penalty function to compensate for the lack of information. As part of a sensitivity analysis, we considered two other parametric forms for the early f_j values, namely (i) $f_j = c e^{j-4}/(1 + e^{j-4})$, $j = 1, \ldots, 4$ and (ii) $f_j = c e^{j} - 1$, $j = 1, \ldots, 6$.

We parameterize the incubation period distribution g in terms of the hazard h, that is

$$g_{k|j}(Z_i) = h_{k|j}(Z_i) \prod_{l=1}^{k-1} (1 - h_{l|j}(Z_i)), \quad k = 1, \ldots, 22.$$

In this paper, we will focus on the independence model and do not consider covariates; thus $g_k|_j(Z_i)=g_k$ and $h_{k|j}(Z_i)=h_k$. One could incorporate covariates through a log-linear model[7]

$$h_{k|j}(Z_i)=h^{\circ}_{k|j}\exp(\beta Z_i),$$

where $h^{\circ}_{k|j}$ is a baseline hazard, or through the model[12, 13]

$$g_{k|j}(Z_i)=\left[1-(Q_{k|j}(Z_i))^{\exp(\beta Z_i)}\right]\left[\prod_{l=1}^{k-1} Q_{l|j}(Z_i)\right]^{\exp(\beta Z_i)},$$

where $Q_{k|j}(Z_i)=1-h_{k|j}(Z_i)$. Both these models become the standard proportional hazards model in continuous time. The possible covariates of interest in this study are age, treatment variables and other demographic and genetic factors.

In the independence model, there are 39 parameters, so a procedure such as maximum likelihood will give unstable estimates. To force smoothness into the estimates of f and h we used a penalized likelihood approach. The penalty function we used was

$$\frac{\lambda_1}{2}J_1(f)+\frac{\lambda_2}{2}J_2(h)=\frac{\lambda_1}{2}\sum_{j=5}^{20}(f_j+f_{j+2}-2f_{j+1})^2+\frac{\lambda_2}{2}\sum_{k=0}^{20}(h_k+h_{k+2}-2h_{k+1})^2.$$

Note that the second sum contains the term h_0, defined as zero because it is well known that the hazard of developing AIDS is low in the first 2 years after infection. Setting $h_0=0$ also assists in computational instability problems. Also note that the first sum does not include f_{23}, which is the catch-all category for HIV infection after December 1989. With the above penalty function, very large values of λ_1 and λ_2 would force the estimates of f and h to be linear.

We performed all computations using a Fortran program on an IBM 3090 with calls to IMSL program DBCONF for maximization of the penalized likelihood. The speed and convergence of the algorithm were unreliable when we used poor starting values or of the values of λ_1 and λ_2 were small.

CHOOSING THE SMOOTHING PARAMETERS

A simple method of determining the appropriate amount of smoothness is to choose the values of λ that give a 'reasonable' amount of smoothness to the solutions. That is, we wish enough smoothness to eliminate irregularities from the estimate of m, but not so much smoothness that the penalty rather than the data dominates the solution. This method, although subjective, is frequently satisfactory. Related to this approach one can graph the likelihood component of the penalized likelihood as a function of λ_1 and λ_2, and look for an elbow or inflection points in these graphs.

Another method of choosing a suitable value for λ is cross-validation. Let $\hat{m}(\lambda, i)$ denote the maximum penalized likelihood estimate of m when we delete observation i. Then, regarding the likelihood as a distance measure, we choose λ that maximizes $\Pi_{i=1}^{n} L_i(\hat{m}(\lambda, i))$. In our specific application, this is computationally too expensive. Instead, we perform 20-fold cross-validation in which, instead of omitting a single observation at a time, we omit 1/20 of the data for each refitting of the model.

Maximum penalized likelihood (MPL) has a Bayesian interpretation in which we view $\exp(-P(m, \lambda))$ as proportional to, a possibly improper, prior for m; then the maximum penalized likelihood estimate (MPLE) is the posterior mode estimate. A third approach to the choice of a suitable value of λ is by exploiting this Bayesian interpretation and using ABIC, the Bayesian information criterion type A.[14] In our application, we can formulate the problem such that

m is a vector of parameters describing the joint density of (x, y), and $c(\lambda)\exp(-P(m, \lambda))$ is the prior distribution for m. Then viewing λ as parameters, we can regard the expression $\exp\{\log L - P(m, \lambda) + \log c(\lambda)\}$ as proportional to a joint posterior distribution for m and λ. Then we can choose λ to maximize the marginal posterior distribution having integrated out m.[14] There are some non-trivial problems with this approach, in particular that the prior for m is improper and we use a Laplace approximation to estimate the marginal posterior distribution. In our application let $L(c, f, h)$ denote the likelihood; let

$$\Pi(f|\lambda_1) = k_1\lambda_1^8 \exp\left\{\frac{-\lambda_1}{2}\sum_{j=5}^{20}(f_j + f_{j+2} - 2f_{j+2})^2\right\}$$

denote the improper Gaussian prior for f; and let

$$\Pi(h|\lambda_2) = k_2\lambda_2^{21/2}\ \exp\left\{\frac{\lambda_2}{2}\sum_{k=0}^{20}(h_k + h_{k+2} - 2h_{k+1})^2\right\}$$

denote the improper Gaussian prior of h. Note that we can think of $\Pi(f|\lambda_2)$ as 16 independent $N(0, 1/\lambda_1)$ priors for $(f_j + f_{j+2} - 2f_{j+1})$, $j = 5, \ldots, 20$; and $\Pi(h|\lambda_2)$ as 21 independent $N(0, 1/\lambda_2)$ priors for $(h_k + h_{k+2} - 2h_{k+1})$, $k = 0, \ldots, 20$. For convenience of notation let $\theta = (c, f, h)$, $\lambda = (\lambda_1, \lambda_2)$, $L(\theta) = L(c, f, h)$ and $\Pi(\theta|\lambda) = \Pi(f|\lambda_1)\Pi(h|\lambda_2)$. The MPLE, $\hat{\theta} = \hat{\theta}(\lambda)$, maximizes $T(\theta, \lambda) = \log(L(\theta)\Pi(\theta|\lambda))$ for fixed. λ. Assume that $T(\theta, \lambda)$ is approximately quadratic with respect to θ, that is that the posterior distribution is approximately normal; then

$$T(\theta, \lambda) \simeq T(\hat{\theta}, \lambda) - \tfrac{1}{2}(\theta - \hat{\theta})'\ H(\theta - \hat{\theta}), \quad \text{where } H = H(\hat{\theta}, \lambda) = \frac{-\mathrm{d}^2}{\mathrm{d}\theta^2}T(\theta, \lambda)|_{\hat{\theta}}.$$

Then the method of choosing λ is to maximize the marginal distribution with respect to λ. That is, choose the λ that maximizes

$$\int L(\theta)\,\Pi(\theta|\lambda)\mathrm{d}\theta \simeq \mathrm{e}^{T(\hat{\theta}, \lambda)}\int \mathrm{e}^{-\frac{1}{2}(\theta - \hat{\theta})' H(\theta - \hat{\theta})}\mathrm{d}\theta = \mathrm{e}^{T(\hat{\theta}, \lambda)}(2\pi)^{k/2}|H|^{-1/2},$$

where $k = \dim(\theta) = 39$. Minus twice the logarithm of this quantity is called ABIC by Ogata and Katsura.[14] Evaluation of this criterion requires numerical calculation of the 39 × 39 hessian matrix for each choice of λ.

CONFIDENCE INTERVALS

We investigated two different approaches for constructing approximate confidence intervals for quantities of interest: a bootstrap scheme and a method that exploits the Bayesian interpretation of penalized likelihood estimation. The bootstrap confidence intervals are based on 100 bootstrap samples of the 1637 subjects. We considered three methods for constructing confidence intervals: the simple percentile method; the bias-corrected percentile method; and a scheme in which the 95 per cent confidence intervals for a quantity of interest Q are of the form $(v^{-1}(\overline{v(Q)} - 2\mathrm{SD}(v(Q)))$, $v^{-1}(\overline{v(Q)} + 2\mathrm{SD}(v(Q))))$, where v is a suitably chosen transformation to make the bootstrap distribution approximately symmetric. As most of the quantities of interest are probabilities, we used the logit transformation of v. In practice, the three confidence intervals were similar for nearly all quantities. The bias-corrected percentile method results are presented here.

We can also obtain confidence intervals using the Bayesian interpretation described above. In particular the posterior distribution for $\theta = (c, f, h)$ is approximately $N(\hat{\theta}, H^{-1})$.

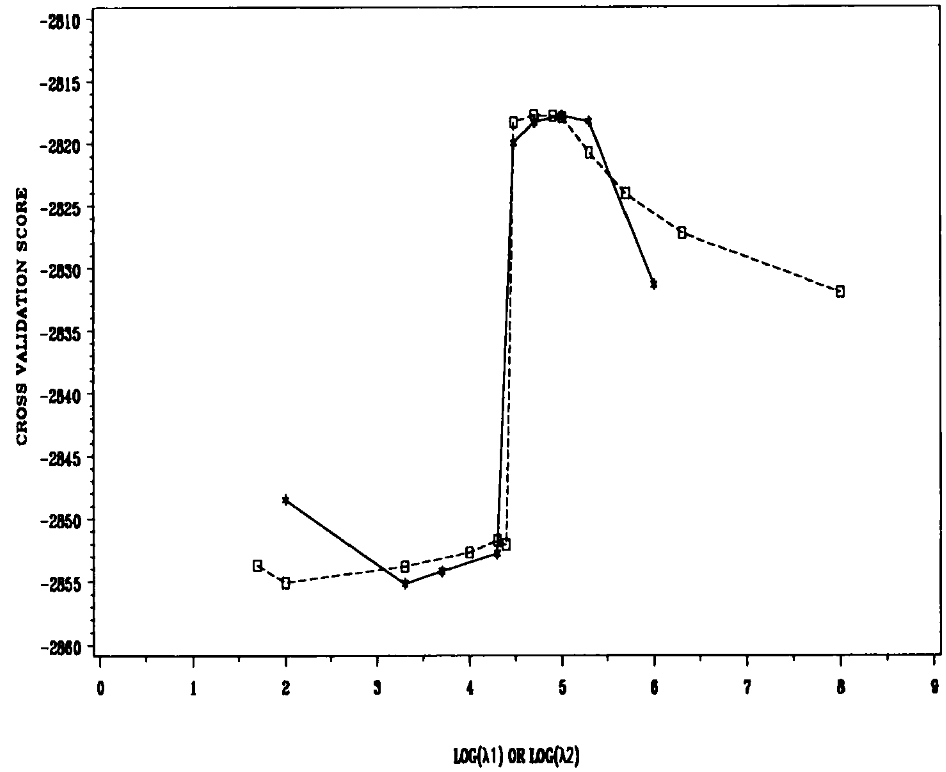

Figure 1. Cross-validation score versus $\log_{10}(\lambda_1)$ for $\lambda_2 = 5 \times 10^4$(solid line); cross-validation score versus $\log_{10}(\lambda_2)$ for $\lambda_1 = 10^5$(dashed line)

RESULTS OF DATA ANALYSIS

Figure 1 illustrates the cross-validation score for a range of values of λ_1 and λ_2. From these graphs and others (not shown) in which we considered different combinations of λ_1 and λ_2, it appears that the best choices of λ_1 and λ_2 are $\lambda_1^* = 10^5$ and $\lambda_2^* = 5 \times 10^4$, respectively. Figures 2 and 3 show graphs of the estimates of f and h for a variety of values of λ_1 and λ_2; it appears from these graphs that λ_1^* and λ_2^* are reasonable choices. The estimates of the infection density f are insensitive to the choice of λ_2 for λ_2 in the range 10^3 to 10^8. Similarly the estimates of the incubation hazard h are insensitive to the choice of λ_1, for λ_1 in the range 10^2 to 10^7.

Notice that the estimate of h is nearly linear for incubation times greater than 8 years. At these long follow-up times there is little information in the data so the estimate has been driven by the penalty term part of the estimation procedure.

Figure 4 shows the value of ABIC, where

$$-2\text{ABIC} = \log L(\hat{f}(\lambda), \hat{h}(\lambda)) - \frac{\lambda_1}{2} J_1(\hat{f}(\lambda)) - \frac{\lambda_2}{2} J_2(\hat{h}(\lambda)) + 8\log\lambda_1 + 10{\cdot}5\log\lambda_2$$
$$-\tfrac{1}{2}\log(|H|) + 19{\cdot}5\log(2\pi),$$

where $L(f, h)$ is the likelihood evaluated at f and $h, \hat{f}(\lambda)$ and $\hat{h}(\lambda)$ are the MPL estimates for a fixed value of λ, and H is the hessian of the penalized likelihood. Note that ABIC does appear to give information about the right choice of $\lambda_1 (\lambda_1 = 2 \times 10^4)$ which roughly agrees with that from

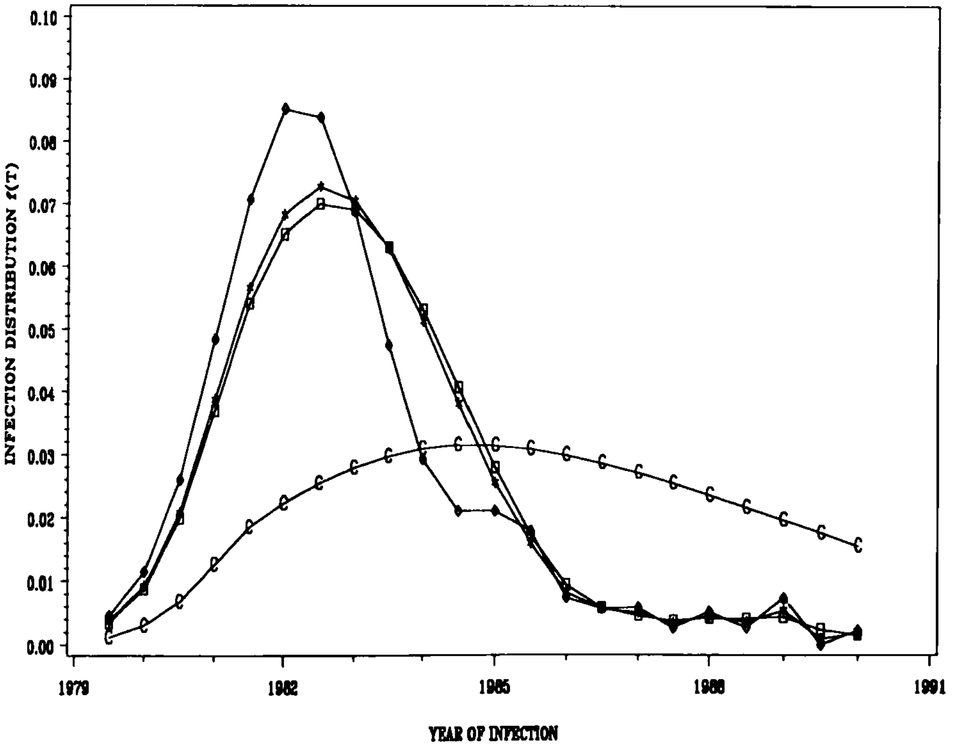

Figure 2. Estimates of the infection distribution f for four choices of λ_1: 4×10^2(◇), 2×10^4(*), 10^5(□), 10^8(c). $\lambda_2 = 5 \times 10^4$ in all cases

Figures 1, 2 and 3, but it gives little information about λ_2 as the graph is flat with respect to λ_2.

Figure 5 illustrates the likelihood component of the penalized likelihood function for different choices of λ_1 and λ_2. The graph does suggest that a reasonable value of λ_1 is in the range 10^4–10^5, consistent with the other methods of choosing λ_1. The λ_2 graph, however, again indicates that the data are of little help in providing a good choice of λ_2.

Figures 6 and 7 show the cumulative distribution of infection times and the incubation period with pointwise 95 per cent bootstrap confidence intervals. These confidence intervals do not reflect the uncertainty associated with the estimation procedure for λ. Figure 8 illustrates this and shows the 95 per cent confidence intervals for the estimate of h using λ_1^* and λ_2^*, and also the estimate of h for two other choices of λ_2. Note that the contribution to the total uncertainty associated with estimating λ is not negligible. A computationally intensive solution to this problem is to include in the bootstrap the estimation of λ as well.

The confidence intervals based on the Bayesian interpretation of penalized likelihood were about 10–50 per cent narrower than the bootstrap confidence intervals. The Bayesian-based confidence intervals were less satisfactory because many of the terms in the hessian were dominated by the penalty part of the penalized likelihood. This was particularly true for the hazard estimates h_j for j greater than 10. Thus the confidence intervals associated with these parameters were driven by the choice of λ_2 rather than by the data. This is an example where, in a Bayesian sense, the improper prior assists considerably in obtaining a good point estimate, but the resulting posterior distribution is less useful because of its dependence on λ_2.

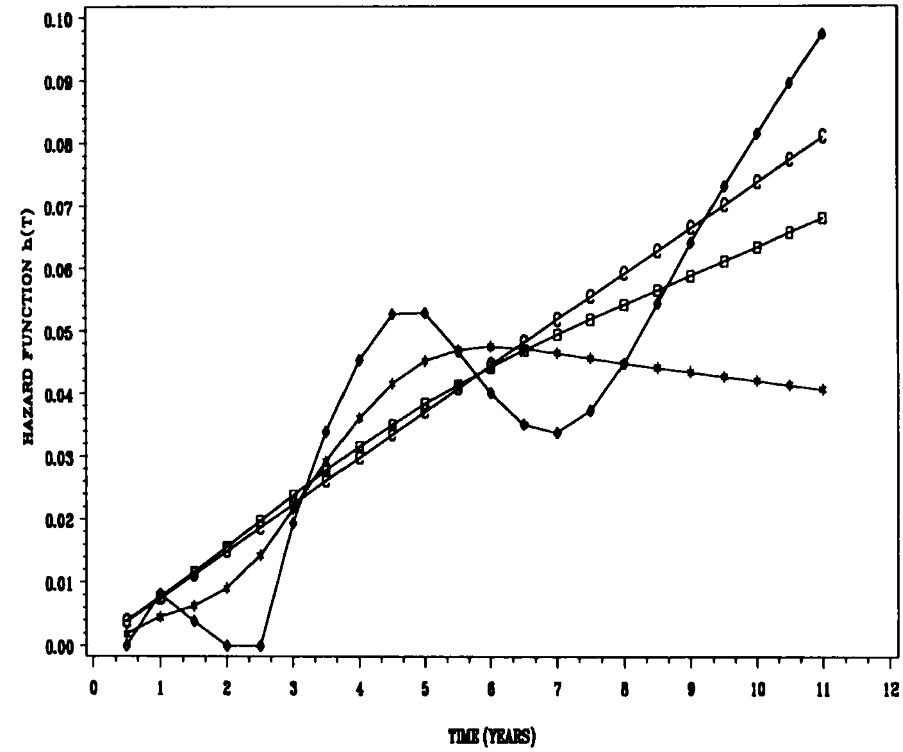

Figure 3. Estimates of the incubation distribution hazard h for four choices of λ_2: 2×10^3 (◇), 5×10^4(*), 2×10^6(□), 10^8(c). $\lambda_1 = 10^5$ in all cases

Bootstrapping the 1637 subjects introduces into the total uncertainty study aspects that are perhaps not relevant to inference about f and h. Bootstrapping cases incorporate variability associated with the design of the study, such as when participants enrolled and when they choose to miss a clinic visit. One could use a stratified bootstrap approach to correct for this overestimation of the uncertainty. We attempted this by resampling observation from within four strata defined by the date of enrolment; this caused a slight but inconsequential reduction in the bootstrap uncertainty.

Figures 9 and 10 show the estimates of the infection density f and the incubation period hazard h for the three choices of the parametric form of f_j for small j. In both figures, and particularly in Figure 10, there is negligible difference between the results, indicating that the assumptions within this parametric part of the model have a negligible effect on the results.

DISCUSSION

The main difficulty with our approach is the computational effort required. The convergence was slow for all values of λ_1 and λ_2, although worse for small λ_1 and λ_2; the speed was substantially improved when we used good starting values. For small λ_1 and λ_2 there were multiple local maxima of the likelihood. Other authors[6] extending the work of Turnbull[15] have used the EM algorithm to reduce the instability of the numerical problems.

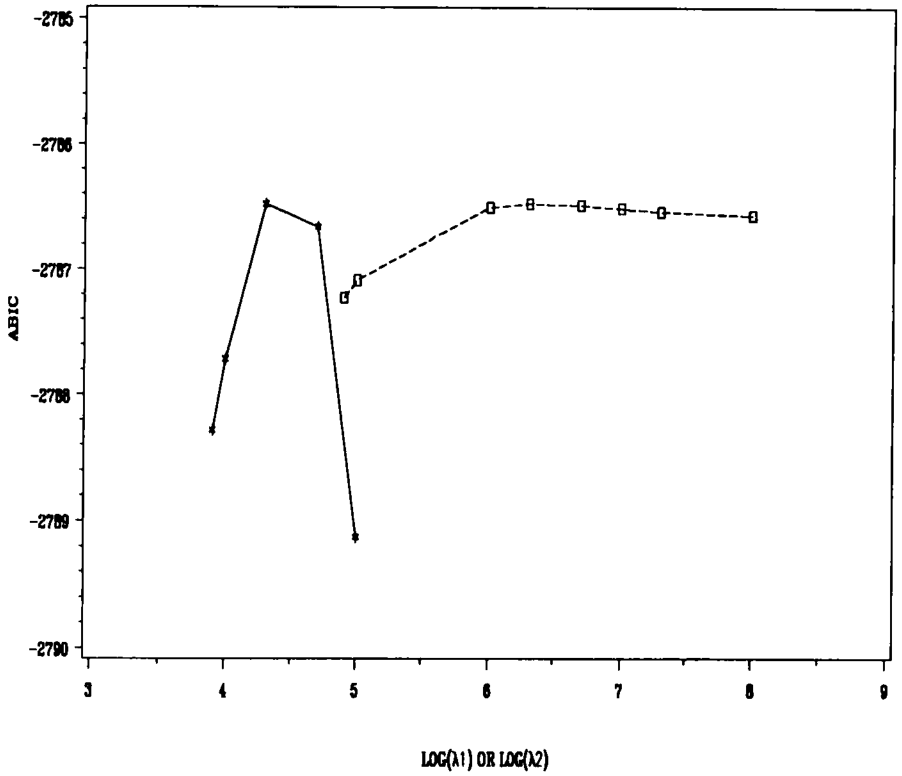

Figure 4. ABIC versus $\log_{10}(\lambda_1)$ for $\lambda_2 = 2 \times 10^6$ (solid line); ABIC versus $\log_{10}(\lambda_2)$ for $\lambda_1 = 2 \times 10^4$ (dashed line)

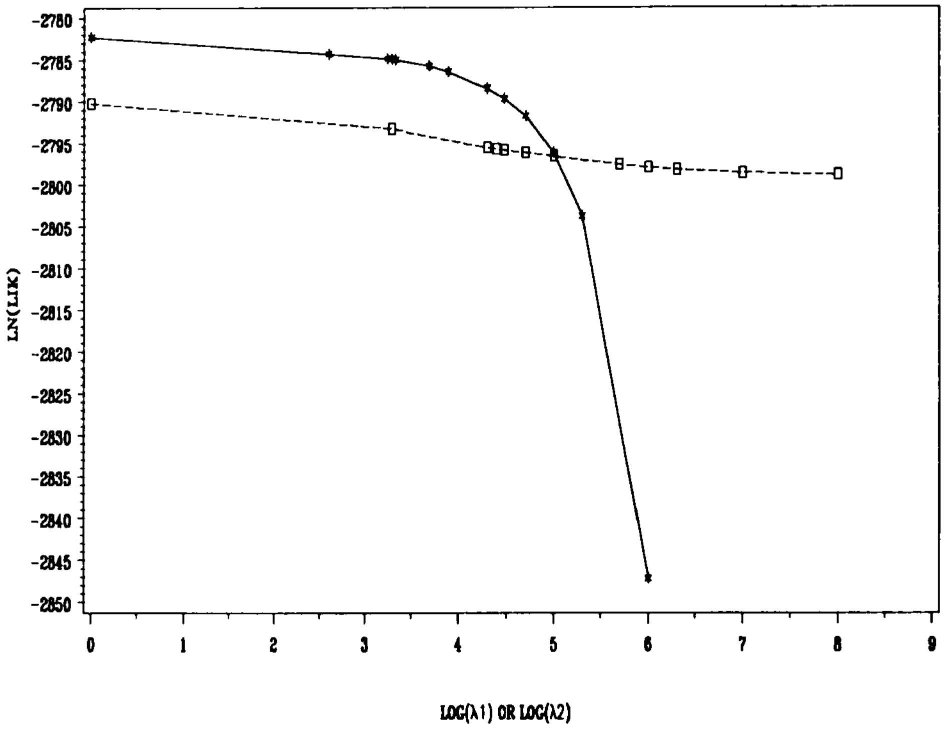

Figure 5. Likelihood component of penalized likelihood (log L) versus $\log_{10}(\lambda_1)$ for $\lambda_2 = 5 \times 10^4$ (solid line); log L versus $\log_{10}(\lambda_2)$ for $\lambda_1 = 10^5$ (dashed line)

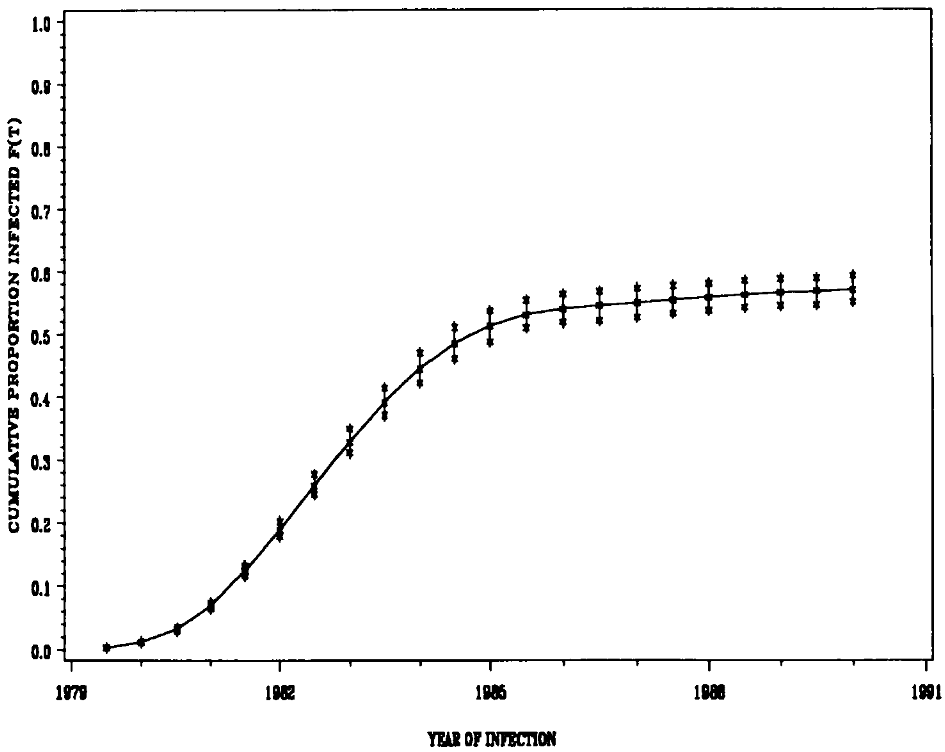

Figure 6. Cumulative distribution of infection distribution with 95 per cent bootstrap confidence intervals: $\lambda_1 = 10^5$, $\lambda_2 = 5 \times 10^4$

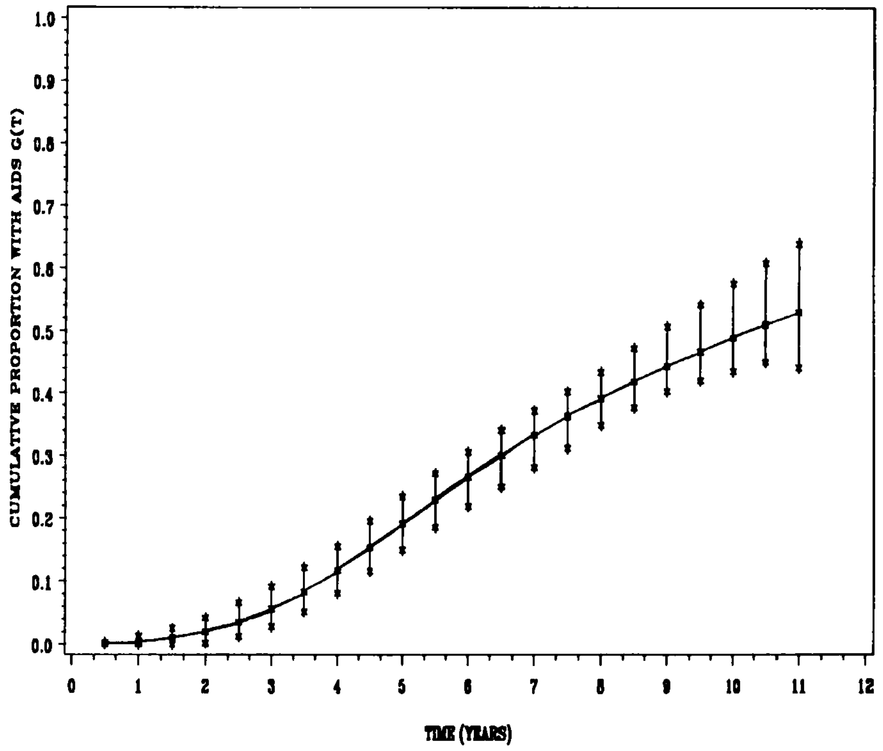

Figure 7. Incubation period distribution with 95 per cent bootstrap confidence intervals: $\lambda_1 = 10^5$, $\lambda_2 = 5 \times 10^4$

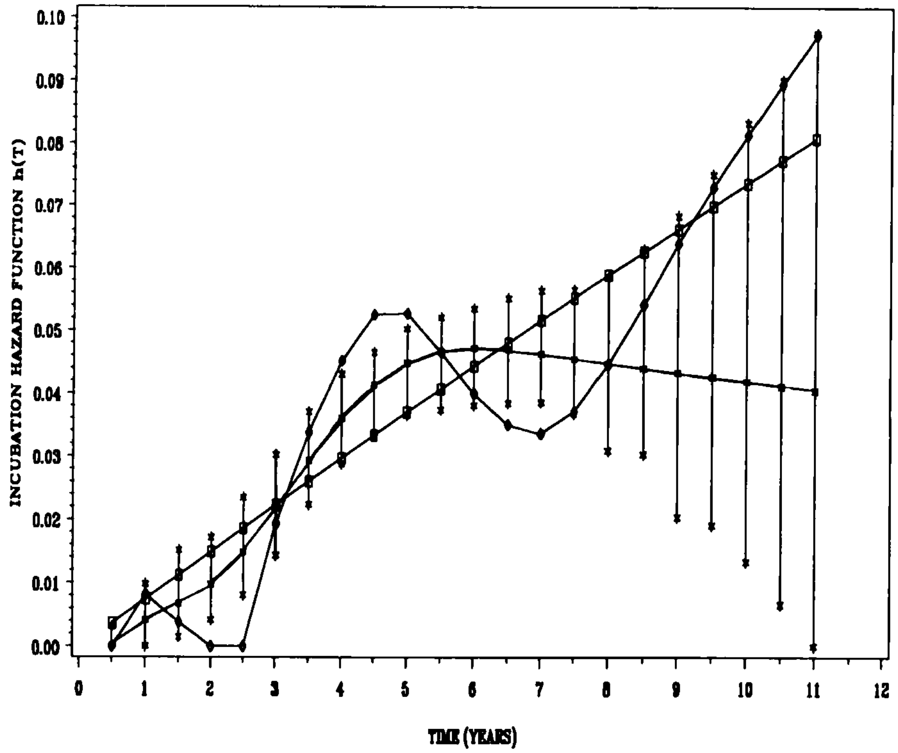

Figure 8. Uncertainty in estimate of the incubation period hazard: $\lambda_2=5\times10^4$ and 95 per cent confidence intervals (*) compared with $\lambda_2=10^8$ (□) and $\lambda_2=2\times10^3$ (□). $\lambda_1=10^5$ in all cases

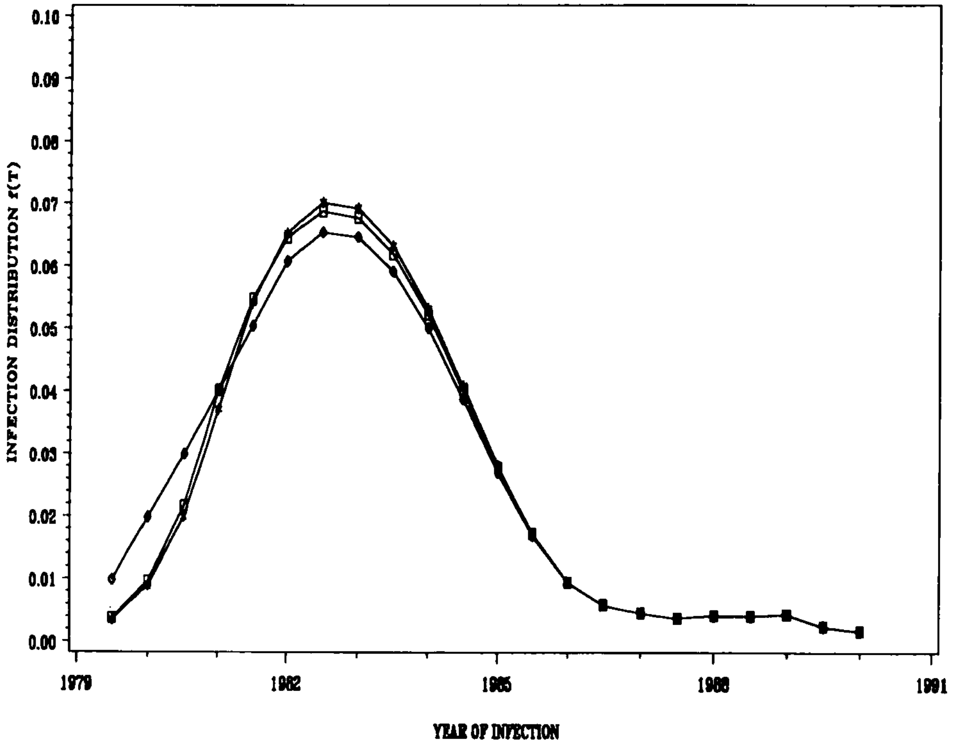

Figure 9. Influence of choice of parametric form for f_j $(j\leqslant6)$ on estimate of infection distribution: $f_j=c\ \exp(j)-1$, $j=1,\ldots,6$ (◇); $f_j=c\exp(j-4)/(1+\exp(j-4)), j=1,\ldots,6$ (*); $f_j=c\exp(j-4)/(1+\exp(j-4)), j=1,\ldots,4$ (□). $\lambda_1=10^5$, $\lambda_2=5\times10^4$

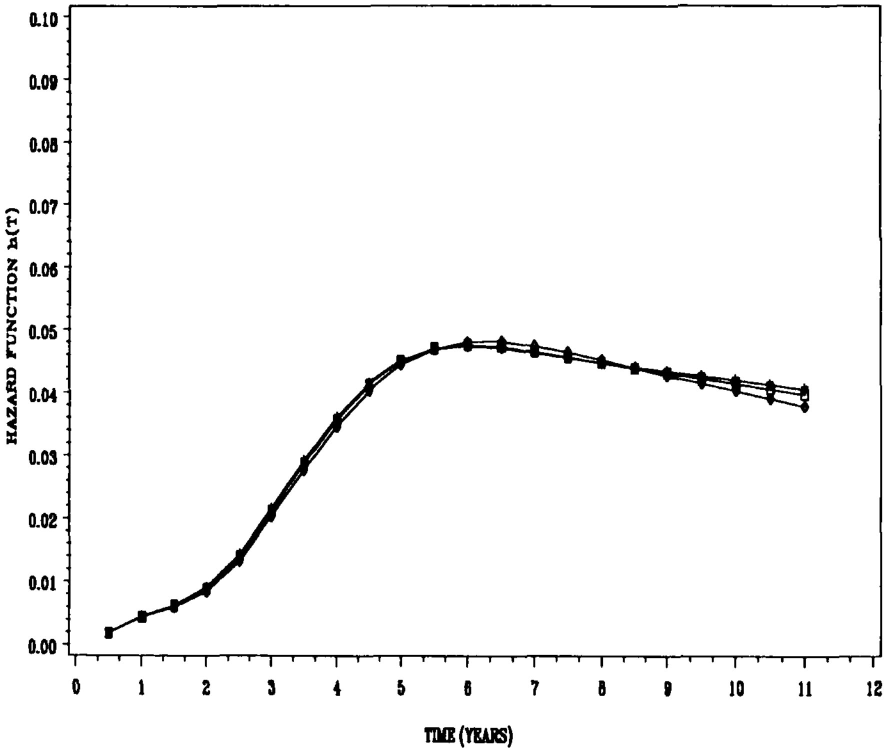

Figure 10. Influence of choice of parametric form for f_j ($j \leqslant 6$) on estimate of incubation hazard: $f_j = c\ \exp(j) - 1$, $j = 1, \ldots, 6$(◇); $f_j = c\exp(j-4)/(1+\exp(j-4)), j = 1, \ldots, 6$ (*); $f_j = c\exp(j-4)/(1+\exp(j-4)), j = 1, \ldots, 4$ (□). $\lambda_1 = 10^5$, $\lambda_2 = 5 \times 10^4$

One validation of the numerical results is that they agree with other epidemiologic data. In particular, the infection curves in Figure 2 are consistent with what is known about the spread of the epidemic.[16] The results in Figures 3 and 7 are similar to previous estimate (Reference 1, and references therein). From the model, we can estimate the number truncated from the sample as

$$\sum_{i=1}^{1637} \frac{Q_i}{1-Q_i}, \quad \text{where } Q_i = \sum_{(jk) \in T_i} \sum P_{jk}(Z_i).$$

The estimated number varied between 37 and 47 depending upon the choice of smoothing parameters. This range is epidemiologically reasonable given the incidence of AIDS cases in Los Angeles, although the relevance of this is questionable as the cohort was not a population-based sample. An alternative procedure for estimating the size of this 'unseen' sample[17] gave a value of 45 for this cohort.

In summary we believe that the penalized likelihood approach adopted in this paper is a reasonable, although computationally intensive, method of extracting good estimates of the incubation period distribution from this prevalent cohort study.

ACKNOWLEDGEMENTS

This work was partially supported by contracts and grants A1-29196 and AI-72631 from the U.S. Public Health Service.

REFERENCES

1. Taylor, J. M. G., Kuo, J.-M. and Detels, R. 'Is the incubation period of AIDS lengthening?', *Journal of AIDS*, **4**, 69–75 (1991).
2. deGruttola, V. and Lagakos, S. W. 'Analysis of doubly-censored survival data, with application to AIDS', *Biometrics*, **45**, 1–11 (1989).
3. Brookmeyer, R. and Goedert, J. J. 'Censoring in an epidemic with an application to hemophilia-associated AIDS', *Biometrics*, **45**, 325–353 (1989).
4. Kuo, J.-M., Taylor, J. M. G. and Detels, R. 'Estimating the AIDS incubation period from a prevalent cohort', *American Journal of Epidemiology*, **133**, 1050–1057 (1991).
5. Darby, S. C., Doll, R. and Thakrar, B. 'Time from infection with HIV to onset of AIDS in patients with haemophilia in the U.K.', *Statistics in Medicince*, **9**, 681–689 (1990).
6. Bacchetti, P. and Jewell, N. P. 'Non-parametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times', *Biometrics*, **47**, 947–960 (1991).
7. Taylor, J. M. G. and Chon, Y. 'Semi-parametric estimation of the incubation period of AIDS', in Dietz, K., Farewell, V. and Jewells, N. P. (eds.), *Statistical Methodology for Study of the AIDS Epidemic*, Birkhauser, Boston, 1992.
8. Green, P. J. 'Penalized likelihood for general semi-parametric regression models', *International Statistical Review*, **55**, 245–259 (1987).
9. Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. Jr. 'The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants', *American Journal of Epidemiology*, **126**, 310–318 (1987).
10. Detels, R., English, P. A., Giorgi, J. V., Visscher, B. R., Fahey, J. L., Taylor, J. M. G, Dudley, J. P., Nishanian, P., Munoz, A., Phair, J. P., Polk, B. F. and Rinaldo, C. R. 'Patterns of CD4$^+$ cell changes after HIV-1 infection indicate the existence of a codeterminant of AIDS', *Journal of AIDS*, **1**, 390–395 (1988).
11. Isham, V. 'Mathematical modelling of the transmission dynamics of HIV infection and AIDS: a review', *Journal of the Royal Statistical Society, Series A*, **151**, 5–49 (1988).
12. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
13. Kim, M. Y., deGruttola, V. G. and Lagakos, S. W. 'Analyzing doubly censored data with covariates, with application to AIDS', *Biometrics*, **49**, 13–22 (1993).
14. Ogata, Y. and Katsura, K. 'Likelihood analysis of spatial inhomogeneity for marked point patterns', *Annals of the Institute of Statistical Mathematics*, **40**, 29–39 (1988).
15. Turnbull, B. 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society, Series B*, **38**, 290–295 (1976).
16. Bacchetti, P. 'Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns', *Journal of the American Statistical Association*, **89**, 1002–1008 (1990).
17. Hoover, D. R., Munoz, A., Carey, V., Odaka, N., Taylor, J. M. G., Chmiel, J. S., Armstrong, J. and Vermund, S. H. 'The unseen sample in cohort studies: estimation of its size and effect', *Statistics in Medicine*, **10**, 1993–2003 (1991).