STATISTICS IN MEDICINE, VOL. 10, 1527-1542 (1991)

A METHOD OF NON-PARAMETRIC BACK-PROJECTION AND ITS APPLICATION TO AIDS DATA

NIELS G. BECKER AND LYNDSEY F. WATSON

Department of Statistics, La Trobe University, Bundoora VIC 3083, Australia

AND

JOHN B. CARLIN

Department of Community Medicine, The University of Melbourne, Parkville VIC 3052, Australia

SUMMARY

The method of back-projection has been used to estimate the unobserved past incidence of infection with the human immunodeficiency virus (HIV) and to obtain projections of future AIDS incidence. Here a new approach to back-projection, which avoids parametric assumptions about the form of the HIV infection intensity, is described. This approach gives the data greater opportunity to determine the shape of the estimated intensity function. The method is based on a modification of an EM algorithm for maximum likelihood estimation that incorporates smoothing of the estimated parameters. It is easy to implement on a computer because the computations are based on explicit formulae. The method is illustrated with applications to AIDS data from Australia, U.S.A. and Japanese haemophiliacs.

INTRODUCTION

Methods used to predict future trends in the incidence of acquired immunodeficiency syndrome (AIDS), tend to be classified into three types.[1,2] These methods are similar in the sense that they all fit some function of calendar time to AIDS incidence data, but differ in the degree to which the mechanisms that generate the data are incorporated into the model.

At one extreme, a simple exponential, polynomial or other mathematically convenient curve is used and no attempt is made to incorporate knowledge about the way the data are generated.[3,4] Predictions are made by extrapolation from the fitted curve. It is well known that such extrapolations must be treated with care. The prevalence of HIV infection cannot be estimated with this approach.

At the other extreme are transmission models that attempt to describe the exact nature of the spread of the disease.[5] The AIDS incidence data alone are not adequate to estimate the parameters of these models with any degree of precision and much other data, concerned with sizes of risk groups, types of sexual practices, and so on, are needed.

The method of back-projection fits between these extremes. It incorporates only elementary assumptions about the way the data are generated and so requires only limited additional information. The data used in applying the method are the numbers of incident cases of AIDS by month, or whatever time period is chosen for analysis. The only additional information required is the distribution of the time from infection to clinical diagnosis of AIDS. The model is

0277-6715/91/101527-16$08.00

Received May 1990
Revised February 1991

sufficiently detailed to allow estimation of the unobserved HIV infection incidence curve up to the present, as well as forming a basis for short-term projections of future AIDS incidence. It is because of the long incubation period of the disease that back-projection lends itself to short-term forecasting.

In practice, it is difficult to estimate past incidence of HIV infection precisely, because of the variable period between infection with the virus and affliction with AIDS (even leaving aside the difficulty of estimating the distribution of this incubation period). Many quite different forms for the HIV incidence curve may be consistent with the observed data on AIDS incidence.[6,7] Previous approaches to back-projection have assumed particular functional forms for the HIV incidence, either in the form of smooth curves[7-9] (characterized by up to four parameters to give richness of HIV epidemic form), or by step functions,[8,10] with up to four discrete steps at chosen time points. Another approach[11] has been to assume a parametric form for AIDS incidence itself.

Here we propose an alternative approach to resolving the lack of identifiability of the HIV incidence curve. Instead of assuming a parametric family of curves, we impose a smoothness restriction on a non-parametric form. Non-parametric maximum likelihood estimation is easily implemented for this problem using the EM algorithm. We propose a modification that incorporates smoothing at each step of the iterative algorithm. The basic method has been proposed previously for applications in stereology and emission tomography.[12] The approach ensures that all estimated values of HIV incidence will be non-negative and computations are easy to implement on a computer because they are based on explicit formulae.

Smoothed non-parametric back-projection seems particularly suited to situations where AIDS cases arise from a series of epidemics out of phase with each other, because parametric models tend to have limited potential for representing more elaborate features of the infection intensity.

The method is illustrated with applications to AIDS data from Australia, U.S.A. and Japanese haemophiliacs.

METHOD

For the method of back-projection we formulate a model from knowledge that AIDS is the result of infection with HIV followed by an incubation period of variable duration. The incubation period is the time from infection to clinical AIDS.

Assumptions, notation and the likelihood function

We choose a time point prior to the introduction of the virus into the community as the time origin. In general, we choose a month as the unit of time and formulate the discussion in terms of discrete time, because incidence data for AIDS are usually reported as monthly counts. However, the development applies to any choice of discrete time interval, and in one of our examples we analyse quarterly data. A discrete time formulation involves approximation, but this has little impact on conclusions relative to other assumptions.

Infections are assumed to arise according to some random process. Let N_t denote the number of individuals infected during month t. The number of AIDS cases diagnosed in month t is denoted by Y_t, $t = 1, 2, \ldots, T$, where T is the month beyond which no reliable AIDS incidence data are available. Let f_d be the probability that the duration of the incubation period is d months, $d = 0, 1, 2, \ldots$. Under the assumption that the distribution of the incubation period is the same irrespective of when the individual is infected, we have

$$E[Y_t | N_1, N_2, \ldots, N_t] = \sum_{i=1}^{t} N_i f_{t-i}.$$

Then the mean number of cases of clinical AIDS in month t will be

$$\mu_t = \sum_{i=1}^{t} \lambda_i f_{t-i} \tag{1}$$

where $\mu_t = E[Y_t]$ and $\lambda_i = E[N_i]$.

The f_d are assumed known in the method of back-projection. In the present context this means that there must be sufficient data available from sources such as transfusion-associated AIDS[13] and cohort studies[14] to enable accurate estimation of the f_d. In practice it is important to perform sensitivity analyses to allow for considerable uncertainties in current knowledge of the incubation distribution.

Next, it is common to specify a parametric family of curves for either the μ_t or the λ_t. The most common approach[2,7-10] is to postulate a parametric form $\lambda_t = g(t;\ \theta)$ and then obtain a maximum likelihood estimate of θ. On the other hand, Isham[11] fits a parametric form in μ_t to the observed Y_t, $t = 1, 2, \ldots, T$. After substituting the fitted μ_t she estimates the λ_t by solving the equations (1) for λ_t. The fact that solutions tend to include some negative λ_t values is unsatisfactory, even though they do not seem to have much effect on prediction.

Here we do not specify a parametric form for either μ_t or λ_t. Instead, we obtain estimates of the λ_t that are closely related to non-parametric maximum likelihood estimates under the assumption that $N_1, N_2, \ldots, N_T$ are independent Poisson variates. This assumption implies that $Y_1, Y_2, \ldots, Y_T$ are independent Poisson variates. Corresponding to the observed monthly AIDS cases $y_1, y_2, \ldots, y_T$ we then have the likelihood function

$$\prod_{t=1}^{T} \left(\sum_{i=1}^{t} \lambda_i f_{t-i} \right)^{y_t} \exp\left(- \sum_{i=1}^{t} \lambda_i f_{t-i} \right),$$

so that the data are linear Poisson. This observation suggests that the model may be fitted by standard statistical software such as GLIM,[15] but there are difficulties with this in practice. Without specifying a parametric form for the λ_t, there are too many parameters relative to the number of data points, y_t. (Rosenberg and Gail[8] avoid this problem by fitting models based on assuming λ_t constant over extended periods.) Furthermore, unique maximum likelihood estimates do not necessarily exist and GLIM may well produce unsatisfactory negative estimates for some λ_t's. On the other hand, maximization of the likelihood function with respect to the λ_t via the EM algorithm[16] always leads to non-negative estimates and can also be implemented conveniently. However, this does not remove the difficulty that for most datasets non-parametric maximum likelihood estimates of the λ_t's are unstable, for example exhibiting consecutive values that vary greatly. This is not consistent with strong prior knowledge that the infection intensity should be a smooth curve. We should of course allow for variation in the infection intensity because the HIV epidemic can consist of a number of separate epidemics which are out of phase, but haphazard jumps in the infection intensity are implausible. One way of getting non-parametric estimates to lie on a smooth curve is to incorporate a smoothing step into the iterative EM algorithm that is used to maximize the likelihood.[12]

The EM algorithm

The EM algorithm is a technique for obtaining maximum likelihood estimates in situations where only incomplete data are observed but where it is possible to define a set of 'complete data' for which closed-form, or at least very straightforward, maximum likelihood estimates exist. Here we take the 'complete data' to include information on the time of infection of each AIDS case. That is, we suppose observation of N_{td} ($d = 0, 1, \ldots, T-t;\ t = 1, 2, \ldots, T$), where N_{td} is the number of individuals infected in month t with an incubation period of duration d months. Since the N_{td} are independent Poisson variates with means $\lambda_t f_d$ under the present assumptions, estimation of the λ_t would then be very easy indeed. In reality, it is only the $Y_t = \sum_{d=0}^{t-1} N_{t-d,d}$ that are observable.

Following the recipe for the EM algorithm given by Dempster *et al.*[16] we know that the likelihood function is increased at each step when we use the iterative equation

$$\lambda_t^{\text{new}} = \frac{\lambda_t^{\text{old}}}{F_{T-t}} \sum_{d=0}^{T-t} \frac{Y_{t+d} f_d}{\sum_{i=1}^{t+d} \lambda_i^{\text{old}} f_{t+d-i}}, \tag{2}$$

where $F_{T-t} = \sum_{d=0}^{T-t} f_d$. This formula combines both the E step and the M step of the EM algorithm. Equation (2) is a slight modification of equation (2.13) of Vardi *et al.*,[17] to suit the present context.

It is possible to give an intuitive explanation for the iterative formula (2). The maximum likelihood estimate of λ_t would be $\sum_{d=0}^{T-t} N_{td}/F_{T-t}$ if all the N_{td} were observed. As only the Y_t are observed we replace the N_{td} by

$$E(N_{td} \mid Y_1, Y_2, \ldots, Y_T) = Y_{t+d} \frac{\lambda_t^{\text{old}} f_d}{\sum_{i=1}^{t+d} \lambda_i^{\text{old}} f_{t+d-i}}.$$

This leads to equation (2).

Incorporating a smoothing step

The smoothing step is incorporated after each application of equation (2). In other words, we let

$$\phi_t^{\text{new}} = \frac{\lambda_t^{\text{old}}}{F_{T-t}} \sum_{d=0}^{T-t} \frac{Y_{t+d} f_d}{\sum_{i=1}^{t+d} \lambda_i^{\text{old}} f_{t+d-i}}, \tag{3a}$$

and then let

$$\lambda_t^{\text{new}} = \sum_{i=0}^{k} w_i \phi_{t+i-k/2}^{\text{new}}. \tag{3b}$$

Now λ_t^{new} is a weighted average of the new parameter values which the E and M steps produce near t, when applied to old parameter values. The value of k determines the 'window width' for the weighted average and should be an even integer. We must choose a value for k and the weights w_i (i = 0, 1, . . . , k). In our applications, we have chosen the same symmetric binomial weights as used by Silverman *et al.*,[12] namely

$$w_i = \binom{k}{i} \bigg/ 2^k \qquad i = 0, 1, \ldots, k. \tag{4}$$

Other kernel functions may of course be used and in the Discussion we include some comments on initial sensitivity analyses performed using double exponential, quadratic, and uniform kernels.

Convergence

Each iteration of the EM algorithm is known to increase the likelihood, and convergence of the algorithm is guaranteed since the log-likelihood considered here is concave. In practice, convergence of the EM algorithm to a maximum likelihood estimate requires very many iterations in the present type of application. Furthermore, as we have seen, there will not necessarily be a unique maximum likelihood estimate. The EM algorithm converges to one of the configurations maximizing the likelihood, dependent on the choice of the $\lambda_t^{(0)}$, the values of λ_t used to start the iterations described by (2).

Our experience with the applications considered here is that convergence to the same configuration occurs irrespective of the starting configuration $\lambda_t^{(0)}$, both with and without the smoothing step added to the algorithm.

When the smoothing step is included, there are no compelling reasons for using a convergence criterion based on the value of the likelihood because the likelihood is no longer being maximized. It seems natural to use a convergence criterion based on the values of the parameters of interest. We cannot expect the individual λ_t, with values of t near T, to be estimated with great precision. Therefore, we choose a time T', where $T' \leqslant T$, and focus on the mean number infected up to this time. More specifically, we choose T' and a small positive ε and stop iteration when

$$\frac{|\sum_{t=1}^{T'} \lambda_t^{\text{new}} - \sum_{t=1}^{T'} \lambda_t^{\text{old}}|}{\sum_{t=1}^{T'} \lambda_t^{\text{old}}} < \varepsilon. \tag{5}$$

Values such at $T' = T - 24$ and $\varepsilon = 10^{-4}$ seem appropriate in the present context. We have used the convergence criterion (5), but other convergence criteria based on the parameter values, such as

$$\sum_{t=1}^{T'} \frac{|\lambda_t^{\text{new}} - \lambda_t^{\text{old}}|}{\lambda_t^{\text{old}}} < \varepsilon,$$

deserve consideration. Note that the only purpose of introducing T' into the criterion is to reduce the amount of computation. Any additional computation resulting from the unstable nature of the estimates $\hat{\lambda}_t$ near T seems unnecessary in view of the imprecision in these estimates.

Adding the smoothing step reduces the number of iterations required for convergence.[12] Nevertheless convergence may still require several hundred iterations. The simple nature of the computations fortunately means that convergence will occur within a few seconds of computer time.

APPLICATIONS

Our main application is to Australian AIDS data, but applications to AIDS data from U.S.A. and Japanese haemophiliacs are considered briefly for the purpose of comparing the present approach to those considered by Rosenberg and Gail[8] and Tango.[2] These three applications will indicate the performance of this method of back-projection for a wide range of datasets. The number of AIDS cases in the U.S. dataset is very large, while the Japanese example involves only a total of 45 cases. The Australian case counts lie between these extremes and provide an example more typical of AIDS counts found in a number of other countries.

AIDS incidence in Australia

Smoothed non-parametric back-projection is first applied to the Australian AIDS incidence data shown in Table I. The first Australian case of AIDS was diagnosed in December 1982, while the earliest Australian HIV-positive case found from tests of stored blood samples is dated 9 October 1980. On the basis of these observations we choose the time origin to be the start of 1979, so that $t = 1$ corresponds to January 1979. Table I gives cases up to October 1989 ($t = 130$), as reported to the National Centre in HIV Epidemiology and Clinical Research[18] by the end of March 1990.

Table I. Australian AIDS incidence data

Year						Months							Total
	J	F	M	A	M	J	J	A	S	O	N	D	
1982	0	0	0	0	0	0	0	0	0	0	0	1	1
1983	0	0	0	1	0	0	0	1	1	0	1	2	6
1984	0	0	1	2	0	2	3	6	7	6	5	11	43
1985	11	11	6	9	19	8	9	4	11	10	9	11	118
1986	15	14	13	14	18	18	16	23	22	30	25	13	221
1987	30	25	31	17	46	34	24	27	37	29	44	26	370
1988	39	42	27	28	34	41	49	45	41	53	58	40	497
1989	53	47	32	25	39	42	38	48	48	50			422
													1678

More recent counts of clinical AIDS are considered unreliable, because of reporting delays, and therefore $T = 130$.

For the incubation period of AIDS we postulate the Weibull distribution. In our discrete time formulation this is specified by

$$F_d = 1 - \exp[-\beta(d+1)^{\gamma}] \qquad d = 0, 1, 2, \ldots .$$

There are few Australian data available for estimating the distribution of the incubation period of AIDS. Accordingly we are guided in the choice of incubation distribution by data from the U.S.A.. Following Rosenberg and Gail[8] we set $\gamma = 2{\cdot}516$ on the basis of estimates obtained by Brookmeyer and Goedert[19] from data on haemophiliacs. Several values are chosen for the parameter β, because of uncertainty in our knowledge of the incubation distribution. More specifically, we choose the five values of β corresponding to median incubation periods of 6, 8, 10, 12 and 14 years.

For each of five incubation period distributions we used the iteration (3), with window width $k = 8$, until convergence to obtain estimates $\hat{\lambda}_t$, $t = 1, 2, \ldots$. The resulting five graphs are shown in Figure 1(*a*).

The degree of smoothing is determined by the window width k. We tried the values 2, 4, 6, 8, 10 and 12 for k, and judged $k = 8$ to give the appropriate degree of smoothing. The non-parametric maximum likelihood estimate (no smoothing), if allowed to converge using $\varepsilon = 10^{-6}$, has tall sharp spikes and is clearly implausible, although providing a good fit to the data. As the degree of smoothing is increased, there tends to be a decrease in the goodness of fit. This should be expected because we decrease the flexibility in the choice of estimates for the $\lambda_1, \lambda_2, \ldots, \lambda_T$, as we increase the amount of smoothing. Sensitivity of the estimated time of peak HIV intensity, and of short-term forecasts of AIDS cases, to the degree of smoothing, was slight. For example, with the binomial weight function, the estimated peak in HIV incidence moved from October 1983 for $k = 0$ to December 1983 for $k = 12$. The total number of AIDS cases forecast for the three-year period November 1989 to October 1992 increased from 2170 to 2270 for the same change in k, while the goodness of fit (defined as the sum of squares of standardised residuals; see Table II) decreased by about 8 per cent.

Concern that the time origin should be placed at an earlier calendar time is allayed by the observation that, in each of the graphs, $\hat{\lambda}_t$ is essentially zero for the first few months.

For most of the estimated curves $\hat{\lambda}_t$ is essentially zero over the last few months, which seems too optimistic. On the other hand, the estimated curve based on a median incubation period of six

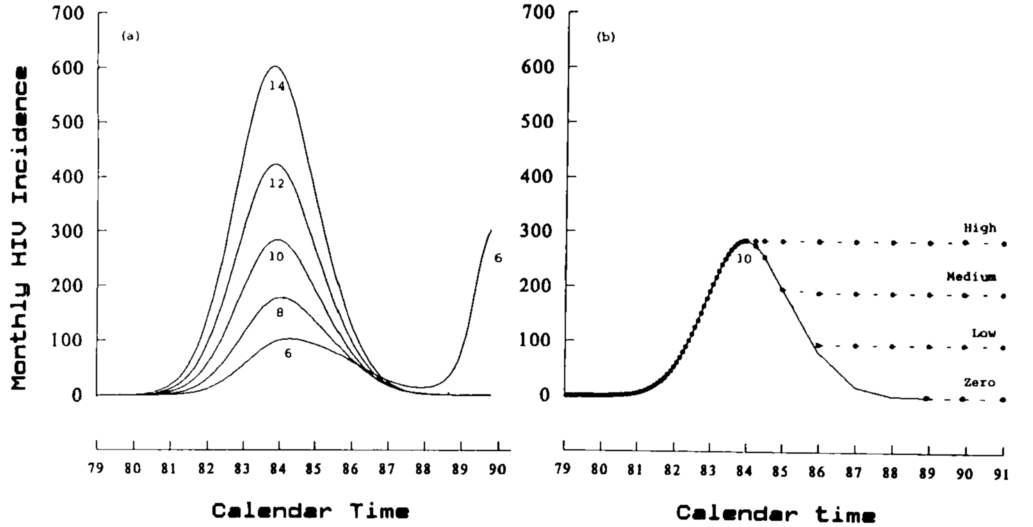

Figure 1. Estimated HIV incidence curves for Australia using $T' = 106$ (October 1987) and $\varepsilon = 10^{-4}$ in convergence criterion (5). The curves are smoothed estimates with window $k = 8$ (months) and median incubation periods (years) as indicated. The horizontal extensions indicated by ● – ● – ● are the infection intensities assumed for the predictions in Table III

Table II. Australian observed (y) and expected ($\hat{\mu}$) AIDS counts with standardized residuals $[R = (y - \hat{\mu})/\sqrt{\hat{\mu}}]$

		Median (years)							
		8		10		12		10	
				with smoothing				no smoothing	
Year*	y	$\hat{\mu}$	R	$\hat{\mu}$	R	$\hat{\mu}$	R	$\hat{\mu}$	R
1982–4	34	51·7	– 2·46	54·4	– 2·77	55·8	– 2·92	38·7	– 0·76
1985	114	104·7	0·91	104·7	0·91	104·4	0·94	106·4	0·74
1986	203	202·7	0·02	199·0	0·28	196·8	0·44	207·3	– 0·30
1987	338	321·8	0·90	315·2	1·28	311·7	1·49	323·3	0·82
1988	469	444·1	1·18	440·7	1·35	439·3	1·42	443·9	1·19
1989	520	554·0	– 1·44	565·0	– 1·89	571·0	– 2·13	558·6	– 1·63
$\sum R^2$			11·19		15·61		18·34		5·95

* Here the year is taken from November of the previous year to October of the current year

years rises sharply at the end, which also seems implausible. In fact, estimates for the last few months are known to be very imprecise. This is intuitively clear from the fact that progression to clinical AIDS is very rare in the first two to three years following infection, so that AIDS counts provide very little information about recent infections, an observation which is supported by the results from simulation studies. Fortunately, the values of the $\hat{\lambda}_t$ for more recent times have relatively little effect on short term predictions of the number of AIDS cases, this being a consequence of the long incubation period.

Estimates of the total number of HIV infected individuals are of interest. These are given by $\sum_{t=1}^{T} \hat{\lambda}_t$, which is the area under each of the curves in Figure 1(a). However, since the estimates of

λ_t lack precision when t is near T, we consider it more appropriate to compare estimates of the total number of HIV infected individuals to month $T - 24$, say. Estimates of 4250, 6500, 10,000, 14,600 and 20,650 are obtained for the number infected to the end of October 1987 when using median incubation periods of 6, 8, 10, 12 and 14 years, respectively. The estimated total is 9100 for a median incubation period of 10 years when no smoothing ($k = 0$) is used.

Table II compares the observed annual AIDS counts with the fitted values obtained when using median incubation periods of 8, 10 and 12 years and smoothing with a window width of $k = 8$. The fit of the expected AIDS counts to the observed is adequate. The largest standardized residuals arise near the ends of the time interval. They seem related to the degree of smoothing used, as estimation without smoothing does not display this feature in general; see the last column of Table II.

Predictions of annual AIDS counts are given in Table III under the assumption that the infection intensity has been constant for some time and will remain at this level. The four choices for the constant infection intensity at T and beyond are zero, the largest value of the estimated infection intensity and two intermediate values. To preserve the continuity of the infection intensity curve we have assumed that the current constant level of infection has been operating since the time when the estimated infection curve first declined to this level. This approach is somewhat arbitrary, but seems as meaningful as any other in view of the imprecise nature of estimates near T. The forms assumed for the infection intensity are shown as dotted lines in Figure 1(b) for the incubation distribution with a median of 10 years. Table III shows predictions to October 1992 for median incubation periods of 8, 10 and 12 years.

AIDS incidence in U.S.A.

The quarterly data for AIDS incidence in the U.S.A. shown in Table IV are taken from Table 2 of Rosenberg and Gail.[8] They have been adjusted for reporting delays. It seems preferable to work with monthly AIDS data when applying the method of smoothed non-parametric back-projection, but in this application we use quarterly data to enable more meaningful comparison with the results given by Rosenberg and Gail. In our application of the smoothed EMS algorithm to the U.S.A. data we disaggregated the AIDS count given for the four year period 1977:1–1981:4. Our results are not affected appreciably when other reasonable choices of the quarterly configuration for this period are used.

A Weibull distribution with $\gamma = 2{\cdot}516$ and a median of 10 years is assumed for the incubation period. With the aggregation of data into quarters less smoothing is required. Accordingly we choose a window width of $k = 2$, but we also show AIDS counts estimated without smoothing ($k = 0$). The fit is good when no smoothing is used, as expected when using so many parameters without restrictions. The fit obtained from the smoothed estimates is also satisfactory, although showing some lack of fit for the period 1977–1981. This could be improved by using a weight function that is more peaked than the weight function (4), thereby reducing the degree of smoothing. Inspection of the residuals reveals that the observed AIDS count in each 4th quarter tends to be lower than expected. This seems to be an artefact of the data and not due to the method of back-projection.

Graphs of two estimates for λ_t $t = 1, 2, \ldots$ are shown in Figure 2. The irregular curve is the result of non-parametric back-projection without smoothing, while the other curve is based on smoothing with window width $k = 2$. The smoothed graph indicates that the infection intensity has peaked and is starting to decrease, which contrasts with the estimates displayed in Figures 1–3 of Rosenberg and Gail.[8] This is largely the result of different model assumptions, but the fact that estimates $\hat{\lambda}_t$ are imprecise for t near T is also relevant.

Table III. Predictions of AIDS incidence for Australia

	Median (years)											
	8				10				12			
	Pattern of recent infection				Pattern of recent infection				Pattern of recent infection			
Year*	zero	low	medium	high	zero	low	medium	high	zero	low	medium	high
1990	630	700	820	1000	680	750	880	1060	700	780	910	1090
1991	680	790	980	1230	760	890	1080	1340	820	940	1140	1410
1992	680	860	1110	1440	820	1010	1290	1640	910	1110	1390	1750
Total HIV to oct 89	6,550	8,750	11,700	15,300	10,000	13,800	18,850	24,800	14,600	20,500	28,200	37,200

* Here the year is taken from November of the previous year to October of the current year.

Table IV. USA observed (y) and expected ($\hat{\mu}$) AIDS counts with standardized residuals $[R = (y - \hat{\mu})/\sqrt{\hat{\mu}}]$

Quarterly calender period	y	$k = 2$		$k = 0$	
		$\hat{\mu}$	R	$\hat{\mu}$	R
1977:1–81:4	374	537·7	−7·06	374·6	0·00
1982:1	185	195·0	−0·72	164·6	1·59
2	200	252·6	−3·31	225·1	−1·68
3	293	323·1	−1·67	303·5	−0·60
4	374	408·3	−1·70	400·6	−1·33
1983:1	554	510·4	1·93	514·6	1·74
2	713	631·3	3·25	643·1	2·76
3	763	773·3	−0·37	785·6	−0·80
4	857	938·4	−2·66	943·9	−2·83
1984:1	1147	1129·0	0·54	1122·5	0·73
2	1369	1347·2	0·59	1324·0	1·24
3	1563	1595·0	−0·80	1552·1	0·28
4	1726	1874·4	−3·43	1815·5	−2·10
1985:1	2142	2187·0	−0·96	2123·5	0·40
2	2525	2534·3	−0·19	2475·7	0·99
3	2951	2917·3	0·62	2868·5	1·54
4	3160	3336·4	−3·05	3303·1	−2·49
1986:1	3819	3792·0	0·44	3783.3	0·58
2	4321	4283·4	0·57	4304·1	0·26
3	4863	4809·7	0·77	4857·3	0·08
4	5192	5369·4	−2·42	5436·3	−3·31
1987:1	6155	5960·6	2·52	6035·9	1·53
2	6816	6580·6	2·90	6651·8	2·01
3	7491	7226·9	3·11	7279·8	2·48
4	7726	7896·4	−1·92	7916·1	−2·14
1988:1	8483	8586·2	−1·11	8556·9	−0·80
$\sum R^2$			148·50		72·16

Predictions to January 1 1993 are shown in Table V under three different assumptions: optimistic; intermediate; pessimistic. For each of these predictions we have used the estimated infection intensity curve, modified to be constant over the later period. The choice of constants is shown in Figure 2 and explained in the first column of Table V. Our predictions are lower than those provided by Rosenberg and Gail[8] under corresponding assumptions; see their Table 4.

Haemophilia-associated AIDS incidence in Japan

Our method is now applied to the haemophilia-associated AIDS data given by Tango.[2] There are a total of only 45 cases which is not really large enough to warrant estimation of the HIV infection curve. However, it is of interest to see how well smoothed non-parametric back-projection works when applied to a small dataset. There is another feature of interest in this application. Blood products given to haemophiliacs have been heat treated in Japan since December 1985, which makes it reasonable to assume that the infection intensity is zero thereafter.

Following Tango we let $t = 1$ correspond to January 1979. We can set $\lambda_t = 0$ for $t = 85, 86, \ldots$. We allow for the possibility that heat treatment for blood products was introduced

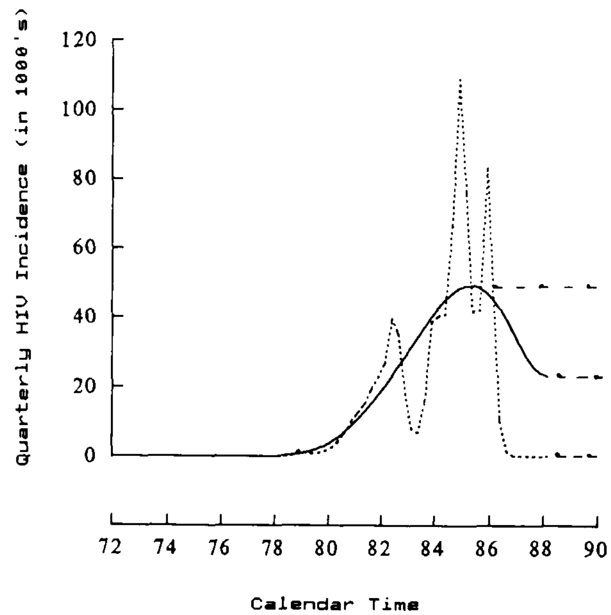

Figure 2. Estimated HIV incidence curves for U.S.A. using $T' = T - 8$ and $\varepsilon = 10^{-6}$ in convergence criterion (5). The solid curve is the smoothed estimate ($k = 2$) and the dotted curve is the unsmoothed estimate. The horizontal extensions indicated by ● - ● - ● are the infection intensities assumed for the predictions in Table V

Table V. Cumulative AIDS counts and HIV incidence for U.S.A.

Assumption about future HIV incidence	AIDS counts to 31 Dec 1992	HIV incidence to 31 Mar 1988
No new HIV infection from 1 Apr 1988	363,000	1,006,000
HIV constant per quarter at 24,000 from 1 Jan 1988	377,000	1,006,000
HIV constant per quarter at 49,000 from 1 Apr 1985	416,000	1,156,000

instantaneously and therefore do not smooth the estimates down to zero near $t = 84$. Monthly AIDS counts are considered reliable to December 1987, so that $T = 108$.

To allow comparison with the results given by Tango we choose a Weibull distribution with $\gamma = 2{\cdot}286$ and a median of 8 years for the incubation period distribution. Tango indirectly postulates an infection intensity of the form $\lambda_t = ct^\theta$, where θ needs to be estimated. Non-parametric back-projection, because it is less restrictive, is likely to provide expected counts which are closer to the observed counts than those given by any parametric model, but the goodness of fit will tend to decline as the degree of smoothing is increased.

Figure 3 shows three HIV infection curves that were estimated by assuming a median incubation period of 8 years. The dashed curve is $\hat{\lambda}_t = 0{\cdot}0055\,t^{1{\cdot}7}$, which is the parametric curve fitted by Tango. The irregular curve is the non-parametric estimate without smoothing. The remaining curve is the smoothed non-parametric estimate of the HIV infection intensity curve, using a window width of $k = 8$. The last twenty-four λ_t's were constrained to be zero for each of these estimated curves, to reflect the heat treatment of blood products after December 1985. Note

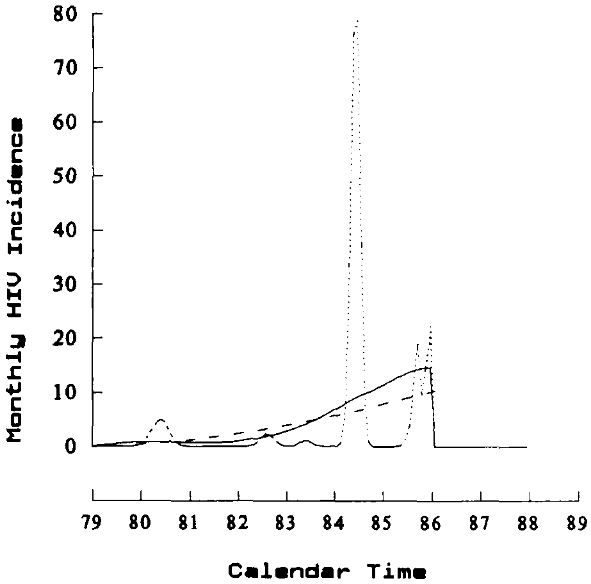

Figure 3. Estimated HIV incidence curves for Japanese haemophiliacs using a median incubation distribution period of 8 years and $\varepsilon = 10^{-6}$ in convergence criterion (5)

——— smoothed ($k = 8$)

...... unsmoothed

— — — parametric curve of Tango[2]

Table VI. Japanese observed (y) and expected ($\hat{\mu}$) haemophilia associated AIDS counts, calculated on monthly data

Calendar time in years	y	Smoothing window width ($k = 8$) $\hat{\mu}$	($k = 0$) $\hat{\mu}$
1979–82	1	1·0	1·4
1983	2	1·6	1·5
1984	3	3·3	2·9
1985	6	6·8	7·0
1986	14	12·7	12·8
1987	19	19·4	19·5

that the smoothed non-parametric curve is initially flatter than the parametric curve and then displays a sharper rise. The parametric form chosen by Tango is clearly unable to fit closely the shape suggested by the smoothed non-parametric curve. Note that concerns about the imprecision of estimates $\hat{\lambda}_t$ near T are less relevant in this type application, because the parameter values near T are known to be zero.

Table VI shows observed and expected AIDS counts. Here each of the versions of the back-projection method is seen to give a close fit to the data. Annual predictions to 1995, not shown, increase gradually to 36 in 1991 and thereafter decline, according to our smoothed non-parametric estimates. The total number infected is estimated to be 380 which is larger than the 320 predicted by Tango.

DISCUSSION

This article has described a method of smoothed non-parametric back-projection that is simple to implement and gives apparently reasonable estimates of past incidence of HIV infection. In particular, the method guarantees that the estimated incidence, $\hat{\lambda}_t$, is non-negative, and that it has a smooth form. On the other hand, the shape of the incidence curve is not constrained to fit a particular parametric form. The use of splines as proposed by Rosenberg and Gail[8] provides another way of fitting a flexible smooth curve to the HIV infection curve. The degree of flexibility is determined, in their approach, by making choices about number and placement of knots, as well as degrees of polynomials to be used. In the approach proposed here, only the degree of smoothing needs to be chosen.

Our method uses a smoothed EM algorithm, labelled EMS by Silverman *et al.*,[12] who proposed it in the context of other indirect estimation problems, in particular the reconstruction of images in classical stereology and in positron emission tomography. These authors and others[17] have discussed the inherently unsatisfactory nature of maximum likelihood (ML) in non-parametric curve estimation, where the high dimensionality of the parameter space leads to very irregular, implausible ML estimates. Silverman *et al.* show that the EMS technique may be related to a maximum penalized likelihood approach, which in turn can be motivated as introducing a 'roughness penalty' to the likelihood function or (from a Bayesian point of view) as introducing a prior density for the λ_t curve. Unfortunately, it is difficult to make these connections explicit, so to a large extent the theoretical properties of the method remain to be investigated.

It is clear in our context that the incidence curve should be a relatively smooth function, since the data are generated by infectious individuals causing further infections. Ideally we would like to capture this dependence among the λ_t's by modelling the infection mechanism that generates the data. This is not practical because of the resulting mathematical complexity and because there have been significant behavioural changes during the course of the epidemic. Hence we are left with the strategy of fitting the incidence curve empirically. The EMS method at least ensures that a smooth estimate is obtained, while allowing the data to influence the shape of the curve. The EMS method does make the assumption that the (unobserved) numbers of HIV infections, $N_1, N_2, \ldots, N_T$ are independent Poisson variates (given the λ_t's). The same assumption is made by Medley *et al.*[20] and Kalbfleisch and Lawless[21] in their analysis of transfusion-associated AIDS data. In the context of back-projection (and using parameterized forms for λ_t), Rosenberg and Gail[8] show that this Poisson assumption gives the same estimates as those based on a quasi-likelihood approach. They also show by example that it gives essentially the same estimates as those given by multinomial maximum likelihood.

Several aspects of the application of the EMS algorithm to estimating HIV incidence require further research. In particular, it is important to develop a method for quantifying the precision of predictions. We require a method for determining a range of HIV incidence curves that could reasonably have led to the observed AIDS incidence data. Simulation studies can help in this regard. To conduct simulations we first obtained the estimates $\hat{\mu}_t$ corresponding to the estimates $\hat{\lambda}_t$. Then data were generated from the Poisson ($\hat{\mu}_t$) distributions, giving simulated sets of AIDS counts. To each simulated set of AIDS counts we then applied the EMS back-projection to arrive at a set of simulated HIV incidence curves. These curves were close together for small t, diverged as t increased and showed extreme fluctuations near T. It is clear from these simulations that estimates for the λ_t become less precise for times approaching the present, and so the 'range' of plausible HIV incidence curves over recent times is large. In our applications we have introduced some arbitrary measures to reflect this fact. The estimation program enables specification of a fixed value of λ_t for any number of recent time points. We illustrated a particularly appropriate application of this feature with the data for haemophiliacs in Japan.

It would also be valuable to study further the choices of both the smoothing kernel and the width of the smoothing window. Work in this direction may help illuminate a degree of lack of fit that is observed when comparing observed and expected numbers of AIDS cases at the extremes of the time sequence of AIDS incidence. Preliminary work in this direction suggests that the goodness of fit to the observed data tends to decline as more smoothing is used. Note however that some smoothing is necessary to arrive at a plausible HIV incidence curve. By increasing the width of the smoothing window one obtains a later peak in the estimated HIV incidence curve and the estimate of the total number infected with HIV increases, although the magnitude of these effects appears small. In a similar vein, only minor changes to the major outcomes, such as goodness of fit and the estimated peak of HIV incidence, appear to result from replacing the binomial kernel with double exponential, quadratic and uniform functions (details not reported). These matters deserve further, more systematic, investigation, especially in light of the suggestion by Day *et al.*[7] that the location of the peak in HIV incidence is important for short-term AIDS forecasts.

Another important detail is the treatment of endpoints in the smoothing algorithm. Equation (3*a*) defines λ_t^{new} for $t = 1, 2, \ldots, T$. The subscript of $\phi_{t+i-k/2}^{\text{new}}$ in equation (3*b*) falls outside this range when t is near 1 or near T. This problem always arises near the end points of the interval when smoothing estimates. The choice of time origin, in the present context, should be such that the infection intensity λ_t is zero prior to $t = 1$. Hence we can define all the weighted averages (3*b*) when t is a near 1 by setting $\phi_{t+i-k/2}^{\text{new}} = 0$ whenever the subscript is less than 1. It is not quite so clear what should be done for t near T. We suggest that $\phi_{t+i-k/2}^{\text{new}}$ be set to ϕ_T^{new} when the subscript is greater than T. This should work well in most applications, although the question of whether another way of smoothing near T can improve the fit needs to be investigated. More specifically, if the ϕ_t^{new} show a clear trend just prior to T, then it may be better to compute the weighted average (3*b*) by extrapolating this trend in some way.

The method of smoothed non-parametric back-projection can readily accommodate changes in the incubation period distributions occurring over time. Simply replace formula (3*a*) by

$$\phi_t^{\text{new}} = \frac{\lambda_t^{\text{old}}}{F_{t,T-t}} \sum_{d=0}^{T-t} \frac{Y_{t+d} f_{t,d}}{\sum_{i=1}^{t+d} \lambda_i^{\text{old}} f_{i,t+d-i}},$$

where $f_{t,d}\, d = 0, 1, 2, \ldots$ is the incubation period distribution for an individual infected in month t and $F_{t,d} = \sum_{i=1}^{d} f_{t,i}$. This provides scope for modelling changes in this distribution over time, a feature that may be required to incorporate a change in the definition of AIDS or changes in the natural history of HIV infection, caused for example by the introduction of new treatments. Solomon and Wilson[22] have suggested that the introduction of the drug zidovudine (AZT) for the treatment of certain HIV-infected individuals in Australia may have led to an important change in the incubation time distribution since June 1987.

As mentioned, the fit could be improved by reducing the amount of smoothing. However, even with the level of smoothing used in the above three applications the fit obtained by the EMS method compares well with those obtained by other methods, apart from a tendency to produce overestimates of AIDS numbers at the extremes. On the other hand, the EMS-based estimates of HIV incidence are somewhat different from previous estimates. In particular, for the U.S. data, we find some evidence for a downturn in the rate of infection, following a peak in the mid-1980's by contrast to the estimates of Rosenberg and Gail,[8] which show a continuing increase. For the Australian data, the indication of a decline in the infection intensity is much stronger. Further work is required to investigate the degree to which these results are an artefact of the method and to what degree they reflect genuine behavioural trends. There is much anecdotal evidence and some epidemiological support for behaviour changes having occurred: for instance, in San Francisco it seems clear that the incidence of new HIV infections among homosexual men declined rapidly in the period 1983–1984.[14] On the other hand, in U.S.A. as a whole, infection rates in other categories (especially intravenous drug users) may have continued to increase for some time, even until the present. In Australia, the epidemic has remained much more confined to the homosexual community (89 per cent of AIDS cases to February 1990), and so it is more reasonable to expect behavioural trends in that group, such as a presumed decline in the rate of infection, to be reflected in the national incidence of HIV infection.

It must be remembered that, although we have experimented with a range of median values, we have assumed throughout that the incubation period distribution has the Weibull shape and does not depend on the time of infection. Further sensitivity analyses should be performed with other parametric or non-parametric forms for this distribution, including specifications that allow for changes due to treatment effects. In view of our preliminary experience suggesting a lack of sensitivity of major conclusions to tuning of the smoothing method (choice of weight function, width of smoother), we expect that the major uncertainties may lie in this area.

ACKNOWLEDGEMENTS

The authors thank Mitchell Gail and Doug Nychka for providing preprints of their papers; References 8 and 12, respectively. This work was supported by Commonwealth AIDS Research Grant number 18 from the Department of Community Services and Health, Australia.

REFERENCES

1. Report of a Working Group: 'Short-term prediction of HIV infection and AIDS in England and Wales', Her Majesty's Stationery Office, London, 1988.
2. Tango, T. 'Estimation of haemophilia-associated AIDS incidence in Japan using individual dates of diagnosis', *Statistics in Medicine*, **8**, 1509–1514 (1989).
3. Curran, J. W., Morgan, M. W., Hardy, A. M., Jaffe, H. W., Darrow, W. W. and Dowdle, W. R. 'The epidemiology of AIDS: current status and future prospects', *Science*, **229**, 1352–1357 (1985).
4. McEvoy, M. and Tillett, H. E. 'Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the U.K.', *Lancet*, **ii**, 541–542 (1985).
5. Isham, V. 'Mathematical modelling of the transmission dynamics of HIV infection and AIDS: a review', *Journal of the Royal Statistical Society, Series A*, **151**, 5–30 (1988).
6. De Gruttola, V. and Lagakos, S. W. 'The value of AIDS incidence data in assessing the spread of HIV infection', *Statistics in Medicine*, **8**, 35–43 (1989).
7. Day, N. E., Gore, S. M., McGee, M. A. and South, M. 'Predictions of the AIDS epidemic in the U.K.: the use of the back projection method', *Philosophical Transactions of the Royal Society of London, B*, **325**, 123–134 (1989).
8. Rosenberg, P. S. and Gail, M. H. 'Backcalculation of flexible linear models of the human immunodeficiency virus infection curve', *Applied Statistics*, **40**, 269–282 (1991).
9. Taylor, J. M. 'Models for the HIV infection and AIDS epidemic in the United States', *Statistics in Medicine*, **8**, 45–58 (1989).
10. Brookmeyer, R. and Gail, M. H. 'A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic' *Journal of the American Statistical Association*, **83**, 301–308 (1988).
11. Isham, V. 'Estimation of the incidence of HIV infection', *Philosophical Transactions of the Royal Society of London, B*, **325**, 113–121 (1989).
12. Silverman, B. W., Jones, M. C., Wilson, J. D. and Nychka, D. W. 'A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography', *Journal of the Royal Statistical Society, Series B*, **52**, 271–324 (1990).
13. Kalbfleisch, J. D. and Lawless, J. F. 'Estimating the incubation time distribution and expected number of cases for transfusion-associated acquired immune deficiency syndrome', *Transfusion*, **29**, 672–676 (1989).

14. Bacchetti, P. and Moss, A. R. 'Incubation period of AIDS in San Francisco', *Nature*, **338**, 251–253 (1989).

15. Payne, C. D. *The GLIM System Release 3.77 Manual*, Numerical Algorithms Group, Oxford, 1985.

16. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977).

17. Vardi, Y., Shepp, A. and Kaufman, L. 'A statistical model for positron emission tomography (with discussion)', *Journal of the American Statistical Association*, **80**, 8–34 (1985).

18. *Australian HIV Surveillance Report*, 23 March 1990, National Centre in HIV Epidemiology and Clinical Research.

19. Brookmeyer, R. and Goedert, J. J. 'Censoring in an epidemic with application of hemophiliac-associated AIDS', *Biometrics*, **45**, 325–335 (1989).

20. Medley, G. F., Billard, L., Cox, D. R. and Anderson, R. M. 'The distribution of the incubation period for the acquired immunodeficiency syndrome', *Proceedings of the Royal Society London, Series B*, **233**, 267–277 (1988).

21. Kalbfleisch, J. D. and Lawless, J. F. 'Inference based on retrospective ascertainment. An analysis of the data on transfusion related AIDS', *Journal of the American Statistical Association*, **84**, 360–372 (1989).

22. Solomon, P. J. and Wilson S. R. 'The back projection method for estimation of the incidence of HIV infection', *Biometrics*, **46**, 1165–1170 (1990).