PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

A Process of Events with Notification Delay and the Forecasting of AIDS

D. R. Cox and G. F. Medley

Phil. Trans. R. Soc. Lond. B 1989 **325**, 135-145
doi: 10.1098/rstb.1989.0078

Phil. Trans. R. Soc. Lond. B **325**, 135–145 (1989) [135]
Printed in Great Britain

A PROCESS OF EVENTS WITH NOTIFICATION DELAY AND THE FORECASTING OF AIDS

By D. R. COX, F.R.S.[1], and G. F. MEDLEY[2]

[1] *Nuffield College, Oxford OX1 1NF, U.K.*

[2] *Department of Pure and Applied Biology, Imperial College of Science and Technology, London SW7 2BZ, U.K.*

Analysis and prediction of a point process are studied when there is delay in notifying the occurrence of events. A primarily parametric approach is taken to studying the rate of the underlying process and to predicting future properties of the system. Data on AIDS are used in illustration.

1. Introduction

This investigation was motivated by the need for short-term prediction of the AIDS epidemic (Healy & Tillett 1988). Brookmeyer & Damanio (1988) and Downs *et al.* (1987) have also considered this problem. The account here is, however, in a rather more general setting. Further details of the application are given in Cox, Working Group Report (1988).

Consider a point process of events occurring in continuous time in a Poisson process of rate $\lambda(t)$. When we wish to emphasize that $\lambda(t)$ depends on unknown parameters we write $\lambda(t;\rho)$. With a point at time t_i is associated a notification lag x_i; that is, the point is not entered into the data until time t_i+x_i, when both t_i and x_i are recorded. We treat the x_i as independent and identically distributed copies of a random variable X, having probability density function $f_X(x;\theta)$, depending on unknown parameters θ. In particular X_i is independent of T_i.

We consider two problems associated with this system. Suppose that the system is observed, either for $(-\infty, t_0)$ or for $(0, t_0)$. In the former case, all events occurring and notified before t_0 are available for analysis. In the second case only events occurring after the time origin are recorded; note the alternative possibility that events notified after the time origin are recorded regardless of the time of occurrence of the originating event.

In the first problem it is required to study the form of the rate function $\lambda(t)$. In the second problem there may be further properties (durations or costs, for instance) associated with the point events and it is required to predict some aspect of these at a future time, $v > t_0$.

In the application to AIDS that we have in mind, namely the short-term forecasting of the AIDS epidemic in the U.K., the point events are defined by patients newly diagnosed as having AIDS. The corresponding rate $\lambda(t)$ is currently increasing less than exponentially with t. The notification lag here is the delay between the diagnosis of a case and its notification to the Communicable Disease Surveillance Centre (CDSC), Public Health Laboratories. At present this lag is between a few weeks and, in extreme cases, 2 or more years. In the forecasting aspects of the problem we might be interested, for example, in the number of patients with AIDS alive at time v or in the number of patients with AIDS in hospital at time v. We shall see in §3 how these and similar features can be studied in the present framework. A quite different interpretation with a different focus of interest is obtained by identifying originating points with instants of infection and the delay with incubation period (Medley *et al.* 1988).

The main idealizations in the present formulation, again as regards AIDS, are that the process is formulated in continuous time, whereas data become available monthly; that there may be changes with time in the distribution of the delay X, and that there are some dependencies between the delays for different patients arising when collection centres submit their notifications in batches. The analysis could be extended to deal with these complications, but we have not done so.

2. Analysis

Under the above assumptions if the process is observed for a time interval $A(t_0)$ ending at time t_0 and diagnosis–lag pairs $(t_1, x_1), \ldots, (t_n, x_n)$ are observed, the log likelihood is

$$\sum \log \lambda(t_i;\rho) + \sum \log f_X(x_i;\theta) - \int_{A(t_0)} \lambda(w;\rho)\, F_X(t_0 - w;\theta)\, \mathrm{d}w. \tag{1}$$

Note that, as explained in §1, when $A(t_0)$ is finite, the lower limit of the integral is determined in our formulation by the instance of occurrence of the originating event.

The integral will in general have to be evaluated repeatedly and so there is some gain in having parametric forms for which the integral can be found in closed form. In some of the discussion below we shall take the form of the rate function to be determined by semi-theoretical considerations, but for cautious empirical investigation a natural starting point for detecting departures from exponential form is

$$\lambda(t;\rho) = \exp(\rho_0 + \rho_1 t - \rho_2 t^2) = \exp[\gamma_0 + \gamma_1(t_0 - t) - \gamma_2(t_0 - t)^2]. \tag{2}$$

We have taken the negative sign for the quadratic term so that slowing down of an initial exponential growth is represented by $\rho_2 = \gamma_2 > 0$. The reparameterization in the second form has some advantages for computation, but suffers from the obvious disadvantage that the parameters change as t_0 changes.

Note that subject to the convergence of the integral involved, the likelihood is such that (1) associated with (2) would form a full exponential family were the lag distribution known.

Important special cases have (i) $\rho_1 = \rho_2 = 0$ and (ii) $\rho_2 = 0$. Null hypotheses corresponding to (i) and (ii) can be tested via the additional parameters in the usual way.

Although (2) is the natural starting point for testing exponentiality of growth, if departures are found it is important for prediction to consider rate functions broadly consistent with theoretical knowledge (Anderson *et al.* 1986; Isham 1988). We have fitted to the AIDS data not only (1) and (2) but a logistic function and, more importantly, a linear logistic function that is initially exponential switching gradually to a linear phase. It corresponds approximately to one of the more realistic models considered by Isham (1988). Of course, the linear phase will itself be of limited duration.

Now the distribution of X will often be of small intrinsic interest, it being a 'nuisance preventing direct observation of the instants of diagnosis and it will then be sensible to choose a parametric form on the basis of mathematical convenience, or to proceed non-parametrically. Simple forms that link well with (2) are (*a*) an exponential distribution, $\theta \mathrm{e}^{-\theta x}$; (*b*) a linear combination of exponentials, or more generally, a density with a rational Laplace transform, thus including gamma distributions of integer index and mixtures thereof, for example a mixture of two gamma distributions of index one,

$$\theta_0 \theta_1^2 x \mathrm{e}^{-\theta_1 x} + (1 - \theta_0)\, \theta_2^2 x \mathrm{e}^{-\theta_2 x}; \tag{3}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

(*c*) the cumulative distribution function

$$1-\exp(-\theta_1 x-\theta_2 x^2). \tag{4}$$

Here we require both component parameters to be non-negative and at least one to be strictly positive.

In all these cases the log likelihood (1) is available in closed form. We shall not give every possibility in detail but the following special cases are worth noting.

First if $A(t_0)=(-\infty,t_0)$ and $\lambda(t;\rho)=\exp(\rho_0+\rho_1 t)$, with $\rho_1>0$, then the integral in (1) is

$$\rho_1^{-1}\exp(\rho_0+\rho_1 t_0)\,E(\mathrm{e}^{-\rho_1 X};\theta), \tag{5}$$

i.e. it is expressible in terms of the moment generating function of X and so, in particular, is available in simple form for distributions with rational Laplace transforms. If, however, $A(t_0)=(0,t_0)$, the form typically necessary whenever there have been many points before the start of recording, then the integral takes the more complicated form

$$\exp(\rho_0+\rho_1 t_0)\int_0^{t_0}\mathrm{e}^{-\rho_1 z}F_X(z)\,\mathrm{d}z;$$

this can also be evaluated explicitly in important cases.

In the special case in which $A(t_0)=(-\infty,t_0)$, $\rho_1>0$ and $F_X(x;\theta)=1-\mathrm{e}^{-\theta x}$ the log likelihood (1) can be written in the canonical exponential family form

$$\psi_0 n+\psi_1\Sigma t_i+\psi_2(-\Sigma x_i)-\exp(\psi_0+\psi_1 t_0)\,[\psi_1(\psi_1+\psi_2)]^{-1}, \tag{6}$$

where $$\psi_0=\rho_0+\log\theta,\ \psi_1=\rho_1,\ \psi_2=\theta.$$

It follows in particular from this that the simple estimate of the mean reporting lag, $\Sigma x_1/n$, has expectation approximately $(\theta+\rho_1)^{-1}$ rather than θ^{-1} and so is seriously biased unless ρ_1 is small compared with θ. Such biases are likely to arise much more generally if aspects of the distribution of X are estimated directly from the reported lags, unless either the lags are nearly all small compared with ρ_1^{-1} or values near the end of the data are discarded. Of course, the full likelihood analysis automatically corrects for such biases. Some further aspects of the study of reporting delays are discussed in §5.

The form of rate function of considerable interest in connection with AIDS, especially for testing for exponential form, is probably (2) with $\rho_1>0$ and $\rho_2\geqslant 0$. If we take the interval of observation to be $(-\infty,t_0)$ and the distribution of X to be either (4) or a combination of exponentials, the integral in the log likelihood can be evaluated explicitly. For the interval $(0,t_0)$ the results are slightly more complicated, but for $\rho_2>0$, the answer is expressed in terms of Dawson's function

$$D(x)=\exp(-\tfrac{1}{2}x^2)\int_0^x\exp(\tfrac{1}{2}t^2)\,\mathrm{d}t. \tag{7}$$

We give the detailed result only for a single exponential and $\rho_2>0$, when the integral becomes

$$\theta\pi^{\frac{1}{2}}\rho_2^{-\frac{1}{2}}\exp[\rho_0+\tfrac{1}{4}(\rho_1+\theta)^2\rho_2^{-1}]\,\Phi[(2\rho_2)^{\frac{1}{2}}t_0-(\rho_1+\theta)\,(2\rho_2^{-\frac{1}{2}})] \tag{8}$$

where $\Phi(x)$ is the standardized normal integral.

From whichever of these expressions is appropriate maximum likelihood estimates, their asymptotic covariance matrix, likelihood ratio statistics and profile log likelihood functions for

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

individual component parameters can be found. It is possible also to develop simple score tests for null hypotheses such as $\rho_2 = 0$ or $\rho_1 = 0$, assuming that $\rho_2 = 0$, or $\rho_1 = \rho_2 = 0$, but we have preferred to study parameters via their profile log likelihoods, despite the extra computation involved. Adequacy of a model can be tested either by comparing observed and fitted numbers of events in suitable time intervals, or by fitting expanded models in the usual way.

3. Prediction

We now turn to methods for predicting future properties of the system analysed above. Errors of forecasting will be of three types, namely errors arising from the random system as formulated, errors in estimating the parameters in the model and errors arising from mis-specification of the process. In many applications, and certainly in the application to AIDS, the last of these three sources is the most important, even in quite short-term forecasting.

Suppose that attached to each point event is a random function of time, independent and identically distributed from event to event and independent of the point process. For an event at t, i.e. a case diagnosed at t, denote the function by $I(t;v)$ and consider the total at time v of these functions associated with all points occurring up to v, i.e. let

$$Y(v) = \int_{-\infty}^{v} I(t;v)\,\mathrm{d}N(t), \tag{9}$$

where $N(t)$ is the counting function of the originating point process, i.e. the number of points occurring at or before time t. Note that we are estimating actual events and not just notified events. If, for instance, attached to each point is a random survival time, then $I(t;v)$ is one during the survival time of the point and zero otherwise and $Y(v)$ is the number of individual points alive at time v. There is, however, no need to restrict I to be a binary indicator function.

The limits of integration in (9) are appropriate when the function of interest is zero before the originating point, and this will be the case in our applications, but there is no difficulty in dealing with 'two-sided' functions of interest, in fact by taking the integral in (9) to be over $(-\infty, \infty)$.

The main properties of $Y(v)$ follow from Campbell's theorem (Cox & Miller 1965, chapter 9); note that a version for nonstationary point processes is required and that the assumption that the originating point process is a Poisson process could easily be relaxed. The key results are as follows.

For binary I, $Y(v)$ has a Poisson distribution and under the independence assumptions set out above

$$E[Y(c)] = \int_{-\infty}^{v} E[I(t;v)]\,\lambda(t;\rho)\,\mathrm{d}t. \tag{10}$$

Generally if the number of points contributing to $Y(v)$ is large, it follows from the central limit theorem that the process $[Y(v)]$ is approximately a nonstationary Gaussian process with the above mean and with

$$\mathrm{cov}\,[Y(v_1), Y(v_2)] = \int_{-\infty}^{\min(v_1, v_2)} E[I(t;v_1)\,I(t;v_2)]\,\lambda(t;\rho)\,\mathrm{d}t. \tag{11}$$

Thus if $I(t;v)$ indicates whether a point originating at t is still alive at time v, then

$$E[I(t;v)] = \mathscr{F}_S(v-t;\phi),$$

where $\mathscr{F}_S(s;\phi)$ is the survivor function attached to each point, assumed to depend on an unknown parameter ϕ.

Now to predict $Y(v)$ on the basis of observations up to time t_0, in a non-Bayesian treatment we first find the conditional distribution of $Y(v)$ given the data and supposing the unknown parameters governing the process to be known. Then we insert efficient estimates of the unknown parameters. Finally, in principle at least, we adjust the resulting prediction intervals for the estimation errors in the unknown parameters. In practice the last step is often omitted as a minor refinement and sometimes the first step is replaced by an approximation using the 'best' linear predictor and the associated standard deviation. In the Bayesian treatment the posterior distribution of $Y(v)$ is computed directly in a standard way.

In the present situation $Y(v)$ is the sum of three independent terms corresponding to

(a) those points already observed, i.e. occurring and notified before t_0;

(b) those points occurring before t_0 but not notified;

(c) those points occurring in the interval (t_0, v).

We treat in more detail the case corresponding to survival times although the argument generalizes fairly easily. Then contributions (b) and (c) have Poisson distributions with means respectively

$$\left.\begin{aligned}&\int_{-\infty}^{t_0}\lambda(t;\rho)\,\mathscr{F}_S(v-t,\phi)\,\mathscr{F}_X(t_0-t;\theta)\,\mathrm{d}t,\\&\int_{t_0}^{v}\lambda(t;\rho)\,\mathscr{F}_S(v-t;\phi)\,\mathrm{d}t,\end{aligned}\right\}\tag{12}$$

whereas (a) is a sum of independent binary contributions from those individual points occurring and notified before t_0 and known to be still 'alive' at t_0. Provided that notification of death is virtually immediate, the contribution of one such point occurring at t_i, say, and still 'alive' at t_0 has expectation

$$\mathscr{F}_S(v-t_i)/\mathscr{F}_S(t_0-t_i)\tag{13}$$

so that the total contribution has expectation the sum of (13) over all recorded points still 'alive' at t_0. It will often be a reasonable conservative approximation to treat the corresponding distribution to be of Poisson form and hence to conclude that $Y(v)$ has a Poisson distribution with mean, $\mu(v, t_0;\rho,\theta,\phi)$, the sum of the three contributions just specified. If there is a time delay between death and notification, then (13) needs modification.

In one of the simpler special cases in which

$$\lambda(t;\rho)=\exp(\rho_0+\rho_1 t),\ \mathscr{F}_X(x;\theta)=\mathrm{e}^{-\theta x},\ \mathscr{F}_S(s;\phi)=\mathrm{e}^{-s\phi}\tag{14}$$

the resulting Poisson mean is

$$\begin{aligned}r(t_0)\exp[-\phi(v-t_0)]+\exp(\rho_0+\rho_1 v)/(\rho_1+\phi)\\-\theta\exp[\rho_0+\rho_1 t_0-\phi(v-t_0)]/[(\rho_1+\phi)(\rho_1+\theta+\phi)],\end{aligned}\tag{15}$$

where $r(t_0)$ is the number of points notified before t_0 and still 'alive' at t_0. Note that in this particular case, the contribution (a) has a binomial distribution of index $r(t_0)$.

We thus obtain Poisson prediction limits on $Y(v)$ on replacing the parameters by maximum likelihood estimates.

Note finally that if Y is a variable to be predicted having a Poisson distribution with large mean μ and if μ is estimated by m having standard error s, then approximate prediction limits for Y are

$$m \pm k_\alpha (m + s^2)^{\frac{1}{2}}, \tag{16}$$

where k_α is the appropriate normal multiplier. Of course, as noted above, in many situations errors in specifying the model, in particular the form of the rate function $\lambda(t; \rho)$, are likely to predominate.

4. Application to aids

In this section we apply the above to short-term prediction of the total numbers of AIDS cases diagnosed in the United Kingdom. The number of AIDS cases diagnosed and still living is also of considerable interest and can be tackled by the arguments of section 3, but we defer discussion of this.

The AIDS notification data were supplied by the CDSC for all cases up to 1 July 1988. A fairly detailed description of the data and some important *caveats* can be found elsewhere (Cox, Working Group Report 1988; Healy & Tillett 1988; Tillett *et al.* 1988). For each case the calendar months of diagnosis and of the arrival of the report to CDSC were given. We excluded those cases that were known to be visitors to the U.K. from abroad, and those cases with an incomplete date of diagnosis (the date of report is always complete). Thus we used 1470 out of the 1598 available cases. Our final predictions are divided by 1470/1532 to take account of those cases that are not included in the full analysis owing to incomplete data, but who were not visitors. Where necessary the originating event was chosen to be 1 January 1979. The diagnoses were assumed to have been made in the middle of each month, thus half a month was subtracted from the time of diagnosis and from t_0 (1 July 1988).

We have investigated the fit of the exponential model by taking (2) as an empirical incidence function, and testing the hypothesis that $\rho_2 = 0$. We used many different distributions for the reporting lag, and give the results from one here, namely the weighted sum of two first order gamma distributions. Years are the time units used throughout. The improvement in log likelihood by including ρ_2 is 18.54. The profile log likelihood function for ρ_2 is roughly quadratic with a slight skewness towards higher values, thus indicating that the epidemic is not described adequately by simple exponential growth. Healy & Tillett (1988) reached the same conclusion about ρ_2 using least squares methods.

To obtain a reasonable approximation to the lag distribution, it was necessary to take account of the discreteness of the recorded data at lags of 0, 1 and 2 months. The dates are recorded as calendar months, thus a lag of 0 months includes 0 to 30 days, 1 month includes 1 to 61 days and so on. The lag distribution is altered so that for month $n (= 0, 1, 2)$

$$p(n) = \int [\theta_0 \theta_1^2 x \, e^{-\theta_1 x} + (1 - \theta_0) \theta_2^2 x \, e^{-\theta_2 x}] \, dx,$$

where the integral is over the appropriate month.

A very long tail in the delay distribution would be unobservable, and yet would inflate the estimated number of diagnoses. We have arbitrarily supposed that the chance of a delay of more than 3 years is negligible, and deleted lags greater than this from the likelihood (six observations). Figure 1 compares the observed lag distribution for all 133 cases diagnosed before 1985 with that calculated from the distribution in table 1.

Epidemiological theory predicts the initial exponential increase in the number of cases, and

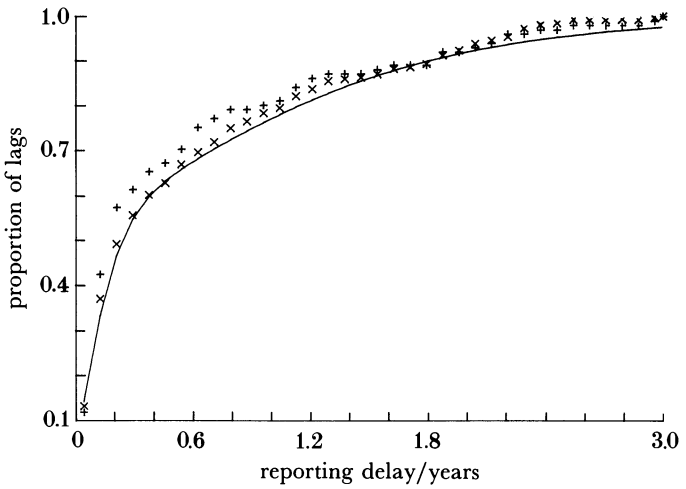

FIGURE 1. Comparison of observed frequency distribution by month of reporting lags with fitted distribution. The plus signs (+) are observed data reported up to June 1988 and diagnosed prior to 1985 (excluding visitors and lags greater than 3 years, $n = 101$). The crosses (×) are observed data reported up to June 1988 and diagnosed before 1986 ($n = 264$). The line is a mixture of two first order gamma distributions fitted with the linear–logistic incidence function (see table 1). The early cases were chosen to minimize bias towards shorter lags in observed reporting distribution, but may not be an accurate representation of the true, prevailing distribution.

TABLE 1. LOG LIKELIHOOD ESTIMATES AND VALUES, ι, FOR DIFFERENT INCIDENCE FUNCTIONS AND REPORTING LAG OF A MIXTURE OF TWO FIRST ORDER GAMMA DISTRIBUTIONS (3)

ρ_1	ρ_2	ρ_3	ρ_4	θ_1	θ_2	θ_3	
quadratic exponential (equation 2)							
−2.57	1.55	0.05232	—	0.57	10.65	1.52	5353.58
logistic (equation 17)							
8.22	0.97	1519	—	0.57	11.36	1.49	5348.09
linear–logistic (equation 18)							
70.99	162.06	13.57	0.82	0.57	11.35	1.52	5350.80

also forecasts that as the epidemic progresses, 'exponential' ceases to be an adequate description of the data and the 'next approximation' is required (Anderson *et al.* 1986). To this end we considered the logistic and the linear–logistic equations:

$$\lambda(t;\rho) = \rho_3/[1+\exp(\rho_1-\rho_2 t)], \tag{17}$$

$$\lambda(t;\rho) = (\rho_1+\rho_2 t)/[1+\rho_3\rho_1\exp(-\rho_4 t)]. \tag{18}$$

Over the period of the data, there is no great difference between the incidence functions, and the maximum log likelihoods are not greatly different. Table 1 gives the estimated parameter values and maximum log likelihoods, and table 2 shows the predictions calculated from them.

It is possible to estimate confidence intervals for the predictions derived from the logistic model (17) by setting

$$\rho_2 = P\Big/\int_B \lambda(t;\rho)\,\mathscr{F}_S(s;\theta)\,\mathrm{d}s\,\mathrm{d}t$$

where B is the time region over which the prediction is being made, and P is the prediction. Essentially, the particular form of the logistic equation allows us to substitute parameter P for

TABLE 2. COMPARISON BETWEEN PREDICTIONS OF THREE INCIDENCE FUNCTIONS[a]

year quarter	observed notifications	notifications quad. exp.	notifications logistic	notifications linear logistic	diagnoses quad. exp.	diagnoses logistic	diagnoses linear logistic
1987 1	134	121	123	123	162	167	164
2	132	141	143	142	187	190	188
3	175	162	164	163	213	214	213
4	147	186	185	186	241	238	239
1988 1	186	211	207	209	271	260	264
2	153	237	228	232	302	281	290
3	196	265	248	256	333	300	315
4	188	294	266	280	365	316	339
1989 1	—	324	284	303	397	331	362
2	—	354	300	326	428	343	384
3	—	383	314	348	458	353	405
4	—	412	326	370	485	361	425
1990 1	—	439	337	390	510	368	443
2	—	465	347	410	532	374	461
3	—	488	355	429	551	378	477
4	—	509	362	447	565	382	493
1991 1	—	526	368	464	575	385	508
2	—	540	373	480	581	387	522
3	—	550	377	496	582	389	536
4	—	555	380	511	578	390	549
1992 1	—	557	383	525	569	392	562
2	—	555	386	539	556	392	574
3	—	548	388	552	539	393	586
4	—	537	389	565	518	394	598

[a] Note that figures beyond the second quarter of 1988 are predictions. The number of notifications received by CDSC are given for comparison. These figures differ slightly from those published previously (Working Group 1988) because they include Scotland, i.e. the predictions are for the U.K.

ρ_2, and obtain a confidence interval for P in the usual manner. Generally the confidence interval is asymmetric, with the upper bound being further from the point estimate than the lower bound.

Approximate confidence intervals for other incidence functions can be obtained by computing the log likelihood for a grid of values in parameter space. When plotted as in figure 2, the confidence interval can be fitted by eye. We chose to fix the lag distribution at the point estimate value, and to only vary the incidence parameters, ρ.

This was partly to reduce the computing resources required and partly because the lag distribution compared well with observation (figure 1). The effects of ρ on both the likelihood and predictions are non-linear and highly correlated. It should be possible to transform the parameters and to choose combinations of them that make their effect on the log likelihood linear and orthogonal, and thus reduce the computation involved greatly.

Table 2 shows predictions of numbers of diagnoses and notifications for the three functions for quarter years to 1992. The differences between the forecasts are major, even after a year or so. The logistic curve moves relatively quickly to an asymptote after departing from simple exponential growth; this is unrealistic in view of the heterogeneity in progression from infection to disease in individual patients. The quadratic exponential is valuable to represent small perturbations from simple exponential growth, but the fact that it peaks relatively rapidly and

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

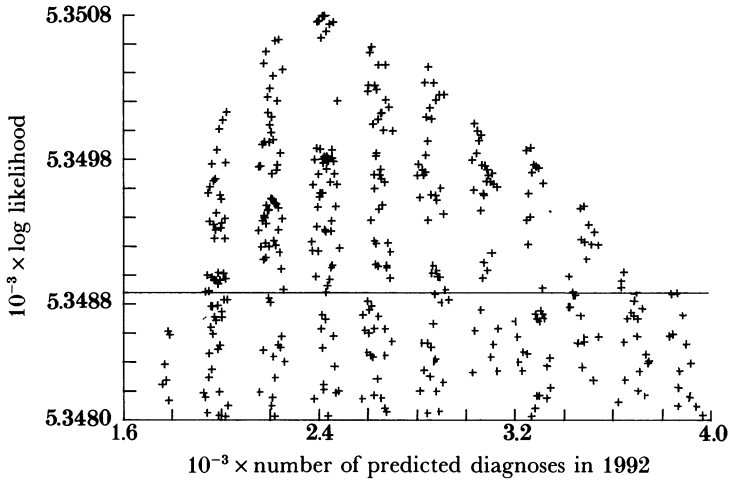

FIGURE 2. The log likelihood values for different combinations of ρ_1, ρ_2, ρ_3 and ρ_4 assuming a linear–logistic incidence (18) and a fixed lag distribution (see table 1). The number of diagnoses expected throughout 1992 is calculated from ρ. Because the effects of ρ on both the likelihood and predictions are nonlinear and highly correlated, many more evaluations were calculated than shown on the figure.

then declines symmetrically is also unrealistic in the present context. The linear–logistic model is the most plausible choice due to theoretical justification.

The major source of variability is the choice of incidence function. The present predictions are at best reasonably accurate for homosexual men for perhaps three years ahead; information on other risk groups (intra-venous drug users, for example) is too scanty for separate prediction to be sensible.

5. Distribution of reporting delays

We now consider briefly some further aspects of the study of reporting delays. In the previous part of the paper these have been studied indirectly via maximum likelihood fitting of a composite model. Suppose, however, we wish to study the delays directly, preferably by some simpler method, and in particular to examine stability of the distribution in time.

Suppose as before that events corresponding here to diagnoses, occur in a Poisson process of rate $\lambda(t)$. To each event is attached a reporting delay, X, having a distribution with density $f_X(x)$ and cumulative distribution function $F_X(x)$. Observations are made over the period $(-\infty, t_0)$. The distribution of delay for individuals observed to be diagnosed (near) t has the truncated distribution $f_X(x)/F_X(t_0-t)$ for $0 \leqslant x \leqslant t_0-t$. Therefore, to examine the constancy of the distribution in time a simple procedure is to truncate all delay times at some suitable x', omitting all diagnoses after t_0-x'. The resulting delay times sorted by time of diagnosis should then have a fixed distribution. Clearly this gives us no information about the delays longer than x'.

If however, individuals are sorted by time s of report and the corresponding distributions of delay examined, biases arise. If T, S, X are random variables representing time of diagnosis, report and delay, then the conditional distribution of X given $S = s$ is

$$f_X(x)\,\lambda(s-x) \Big/ \int_0^\infty f_X(u)\,\lambda(s-u)\,\mathrm{d}u.$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Now this is equal to $f_X(x)$ for all x, s if and only if the rate $\lambda(.)$ is constant. In particular if $\lambda(t) = \alpha e^{\beta t}$, the conditional density is, for all s,

$$f_X(x)\, e^{-\beta x} \int_0^\infty f_X(u)\, e^{-\beta u}\, du.$$

Thus if $\beta > 0$, the bias is in favour of the shorter times, the conditional distribution being independent of s, i.e. showing no trend in time. Note, however, that if the epidemic grows subexponentially, so that in effect β is slowly decreasing with s, the conditional distribution of X given time of report, s, shows increasing delays as s increases. This is a consequence of the 'biased' sampling. More generally, the conditional density is independent of s if and only if the rate is exponential.

To see this last point qualitatively, consider an idealized case in which half the delays are quite small and the other half long. If initially the epidemic grows rapidly, the conditional distribution of delay at a fixed time of report is strongly biased towards the short times. If now there is a switch to a constant rate, the conditional distribution has a transient phase and then tends to the 'true' distribution, i.e. delays will appear to have lengthened.

Plots of the distribution of delay time sorted by date of report show some tendency for delays to lengthen, as might be expected from the above discussion. If, on the other hand, the cases are sorted by quarter of diagnosis and delays over 8 months omitted, the mean delays are those in table 3.

Table 3. Mean and standard deviation of delay distribution, delays truncated at 8 months

	1986				1987			
quarter of diagnosis...	1	2	3	4	1	2	3	4
mean/months	3.02	2.81	2.43	3.03	2.74	2.19	2.14	2.03
s.d./months	2.38	2.36	2.32	2.41	2.31	2.13	2.27	1.97

There is some decrease in the shorter reporting delays in 1987: it is from the data alone impossible to separate changes in 1988 from changes in incidence and it is also effectively impossible to study the frequency and length of very long delays.

Although we shall not explore the matter in detail, it would be possible to fit a model in which the distribution of delays depends simply on the time of diagnosis. If, however, rather arbitrary dependence were allowed there is clearly some practical or total indeterminacy; any recent fall in numbers reported could be explained by a recent increase in reporting lags.

We are grateful to Dr V. Isham for very helpful comments. The data on case notifications was kindly supplied by CDSC.

References

Anderson, R. M., Medley, G. F., May, R. M. & Johnson, A. E. 1986 A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. appl. Med. Biol.* **3**, 229–263.

Brookmeyer, R. & Damiano, A. 1988 Statistical methods for short-term projections of AIDS incidence. *Stat. Med.* **8**, 5–20.

Cox, D. R. (chairman) Department of Health Working Group Report 1988 Short-term prediction of HIV infection and AIDS in England and Wales. London: HMSO.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Cox, D. R. & Isham, V. 1980 *Point processes*. London: Chapman & Hall.
Cox, D. R. & Miller, H. D. 1965 *Theory of stochastic processes*. London: Chapman & Hall.
Downs, A. M., Ancelle, R. M., Jager, J. C. & Brunet, J. B. 1987 AIDS in Europe; current trends and short-term predictions estimated from surveillance data, January 1981–June 1986. *AIDS* **1**, 53–57.
Healy, M. J. R. & Tillett, H. E. 1988 Short-term extrapolation of the AIDS epidemic (with discussion). *Jl R. statist. Soc.* A **151**, 50–61.
Isham, V. 1988 Mathematical modelling of the transmission dynamics of HIV infection and AIDS: a review. *Jl R. statist. Soc.* **151**, 5–30.
Medley, G. F., Billard, L., Cox, D. R. & Anderson, R. M. 1988 The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS). *Proc. R. Soc. Lond.* B **233**, 367–377.
Tillett, H. E., Galbraith, N. S., Overton, S. E. & Porter, K. 1988 Routine surveillance data on AIDS & HIV infections in the U.K.: a description of the data available and their use for short-term planning. *Epidem. Inf.* **100**, 157–169.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES