Statistica Applicata Vol. 8, n. 1, 1996

PROJECTING THE AIDS EPIDEMIC IN ENGLAND AND WALES: A BAYESIAN APPROACH

Walter Richard Gilks, Daniela De Angelis

Medical Research Council Biostatistics Unit, Cambridge, UK.

SUMMARY

Data available for projecting the AIDS epidemic in England and Wales include individual–level registry data on individuals diagnosed with AIDS, and periodic aggregate–level seroprevalence survey data on individuals infected with the HIV virus. Analyses of these data is complicated through the mixture of aggregate–level and individual–level data; by the retrospective nature of the registry data; by reporting delays; and by lack of information on the date of HIV infection in those reported with AIDS. We describe a Bayesian approach to projecting the AIDS epidemic, in which we develop a single, coherent model accounting for all these complications. We use the model to project the epidemic in homosexual/bisexual men up to the end of 1999, using Markov chain Monte Carlo estimation techniques.

Keywords: *AIDS, HIV, Markov Chain Monte Carlo, Bayesian model, projection, back–projection, reporting delay.*

1. INTRODUCTION

The Human Immunodeficiency Virus (HIV) epidemic began in England and Wales in the mid–1970's. An HIV–infected individual may appear healthy for several years after infection, but during this *incubation period* the virus progressively disrupts the immune system of its host. Eventually, clinical symptoms emerge and a diagnosis of acquired immune deficiency syndrome (AIDS) may be made.

Since recognition of the viral origin of the disease in 1981, there has been an urgent need to project the future course of the disease, to inform policy makers and to plan medical provision. A register of AIDS diagnoses in England and Wales was set up in 1982 at the Cummunicable Disease Surveillance Centre (CDSC) in London. The first round of projections for England and Wales (Cox, 1988) were made at a time when there was little information on the prevalence of HIV and its incubation distribution. Predictions were made under a range of possible scenarios,

but with hindsight were all found to overestimate AIDS incidence. Two more rounds of projections followed (Day, 1990, 1993), each benefitting from accumulating information on the biology of the disease and its dynamics in the population. Statistical methodology for projecting the epidemic has also steadily accrued since the report of Cox (1988).

Projecting the AIDS epidemic is problematical for several reasons.

- AIDS reports are subject to reporting delay. Only about 60% of reports are received at CDSC within six months of diagnoses. Thus AIDS reports severely underestimate recent trends.
- Future short–term AIDS predictions depend mainly on the numbers already infected with the HIV virus, and on the times of those infections. However, AIDS reports do not record the date of HIV infection, since this cannot be reliably determined for the majority of AIDS cases.
- Infected individuals who have not yet developed AIDS are not included in the AIDS register. Another register maintained by CDSC contains reports on HIV–positive tests both for AIDS and non–AIDS cases. Since HIV infection must precede a positive test, this register provides only a lower bound on the numbers infected.
- To obtain some indication of the extent of the epidemic, HIV seroprevalence surveys of various kinds have been conducted, involving anonymised testing of blood samples and the completion of questionnaires. The data are assimilated by CDSC and result in aggregate–level interval estimates of the numbers currently infected (Day, 1993). The mixture of aggregate–level and individual–level data represents an additional challenge to the statistical analysis.

A fourth round of projections is now imminent, and here we report briefly on the statistical methodology we are developing for it. A central aim of our methodology is to take account of all sources of uncertainty in the projections, which has not hitherto been possible with available methodology. A full account of the methodology will be reported separately (Gilks and De Angelis, 1996). We present results and projections based on a preliminary analysis of the data relating to HIV infections and AIDS diagnoses in homosexual/bisexual men.

2. DATA

We focus the discussion in the remainder of this paper entirely on the epidemic in homosexual/bisexual men in England and Wales. By end of 1994, a total of 7,295 diagnoses of AIDS in this population had been reported to CDSC. Each report

records, amongst other things, the date of the diagnosis and the date that the report was received at CDSC. In the present analysis we do not make use of the HIV–test reports collected by CDSC.

Using unpublished seroprevalence data kindly supplied to us by CDSC, we calculated that up to the end of 1991 there were around 16,000 individuals in our study population who had been infected by HIV. This is an estimate of the cumulative incidence of infection, and includes all those infected, with or without AIDS, at that time, and those who had already died with the infection. Similarly, we also calculated cumulative HIV incidence up to the end of 1993.

3. MODELLING

We assume that HIV infections occur according to a Poisson process with intensity $h(t^H)$, where t^H denotes the time of infection in months since the beginning of 1970. We assume that the time of AIDS diagnosis, t^A, together with the time that the diagnosis is reported to CDSC, t^R, occur independently amongst individuals, given the time of infection. We denote the conditional density of the time of AIDS diagnosis, given the time of infection, by $a(t^A \mid t^H)$. We denote the conditional density of the time of the report, given the time of diagnosis, by $r(t^R \mid t^A)$. Thus we assume that t^H does not affect t^R, controlling for t^A.

Let $n = 7295$ denote the number of AIDS reports received up to time $t^* = 300$ months (the end of 1994). Then the AIDS–report data may be denoted

$$\left\{t_k^A, t_k^R\right\}_{k=1}^{n} \tag{1}$$

Let $\hat{s}_1, \hat{s}_2$ denote seroprevalence–survey–based interval estimates of cumulative HIV incidence up to times $t_1^S = 264$ and $t_2^S = 288$ (see Section 2). Let s_1, s_2 denote the corresponding true cumulative incidences of HIV infection.

Let $P\left(\hat{s}_1, \hat{s}_2 \middle| s_1, s_2\right)$ denote the probability distribution of the cumulative–incidence estimates conditional on the true cumulative incidence at times t_1^S, t_2^S. Finally, let m denote the number of individuals infected before time t^*, but for whom an AIDS diagnosis was not made or not reported by time t^*.

We will use subscript k to denote individuals diagnosed with AIDS and reported to CDSC by time t^*; subscript j to denote individuals infected before time t^* but for whom an AIDS diagnosis was not made or not reported by time t^*; and subscript l to denote individuals infected after time t^*.

With the assumptions described above, it can be shown that the joint distribution:

$$P\left[s_1, s_2, \hat{s}_1, \hat{s}_2, m, n, \left\{t_j^H, t_j^A, t_j^R\right\}_{j=1}^m, \left\{t_k^H, t_k^A, t_k^R\right\}_{k=1}^n\right] =$$
$$P\left(\hat{s}_1, \hat{s}_2 | s_1, s_2\right) \times \mathbf{exp}\left\{-\int_0^{t^*} h\left(t^H\right) dt^H\right\}$$
$$\times \prod_{j=1}^m h\left(t_j^H\right) a\left(t_j^A | t_j^H\right) r\left(t_j^R | t_j^A\right) \qquad (2)$$
$$\times \prod_{k=1}^n h\left(t_k^H\right) a\left(t_k^A | t_k^H\right) r\left(t_k^R | t_k^A\right).$$

For the analysis presented here, we assumed a piecewise–constant function for the HIV infection intensity $h(t^H)$. Changepoints were set at the end of each year from 1980 to 1991, after which the intensity was assumed constant. Thus $h(t^H)$ was specified by 13 model parameters, which we denote by η.

We assumed that the distribution on time of AIDS, $a(t^A \mid t^H)$, is given by a gamma(α_1, α_2) distribution on the incubation time $t^A - t^H$. This assumption could be refined by taking account of Zidovudine therapy, which was introduced in the mid–1980's, and which has been shown to delay the diagnosis of AIDS.

For the reporting–time distribution $r(t^R \mid t^A)$, we assumed a piece–wise constant function on the delay $t^R - t^A$. Changepoints were set at quarterly intervals up to 3 years, then yearly up to 7 years, and we assumed that all reports would be received within 10 years of diagnosis. Thus the reporting delay distribution was parameterised by 16 free parameters, which we denote by ρ. The reporting delay distribution could be refined by allowing for the effects of initiatives to bring the register up to date.

For the seroprevalence estimates, we assumed that $\hat{s}_1$ and $\hat{s}_2$ correspond to symmetric 95% intervals from independent normal distributions with means s_1 and s_2, respectively.

We assumed vague priors for the model parameters η and ρ, and moderately informative priors for α_1 and α_2, based on an incubation distribution estimated from another cohort (Hendriks *et al.*, 1993).

4. ESTIMATION

In principle, inference and projections might be based on the likelihood of the parameters η, α and ρ given the AIDS–report data (1) and seroprevalence data $(\hat{s}_1, \hat{s}_2)$. Evaluation of this likelihood would involve integrating (2) over the latent data

$$\left\{t_j^H, t_j^A, t_j^R\right\}_{j=1}^m; \quad \left\{t_k^H\right\}_{k=1}^n. \qquad (3)$$

However, even with the simple assumptions described above, this integration is not possible analytically. Instead we employ a Markov chain Monte Carlo approach (see, for example, Gilks *et al.,* 1996), and work with (2) directly. This involves running a Markov chain over the space of the latent data (3) and model parameters η, α, ρ. At each iteration new values for the latent data and the model parameters are generated, conditional on the current values, in such a way that the stationary distribution of the chain is the distribution proportional to (2), which is the posterior distribution of the unknown quantities in the model, given the data.

An interesting feature of this model is that (2) contains $31 + n + 3 \times m$ unknown quantities, which is itself unknown since the number of unreported individuals m is unobserved. Thus the size of the model varies dynamically as the Markov chain proceeds, at each iteration a new value for m being generated: see Green (1995) for theoretical justification. We omit further computational details here, but note that the generation of the chain can be organised to involve only elementary manipulations.

Having run the chain long enough (see, for example, Raftery and Lewis, 1992; or Gelman and Rubin, 1992), we can estimate any posterior functional of interest simply by averaging. Specifically, let $x^{(i)}$ denote the values of (3) and η, α, ρ at iteration i, and suppose we are interested in some function $f(x)$. Then the posterior mean of f can be estimated by

$$\bar{f} = \frac{1}{N - M} \sum_{i=M+1}^{N} f\left(x^{(i)}\right), \tag{4}$$

where N denotes the total number of iterations of the chain, and M denotes the number of iterations in the 'burn–in', during which the samples are judged to be materially influenced by the starting values of the chain. For example, the posterior mean of α_1 can be calculated from (4) with $f(x) = \alpha_1$. Similarly the posterior variance and centiles of α_1 can be easily calculated by appropriate choices of $f(.)$ in (4).

The estimate $\bar{f}$ can be made as accurate as desired by increasing the run length, N. Note that the accuracy of the estimate $\bar{f}$ is not the same thing as the posterior variance of f, which is related to the amount of data in (1).

We performed a single run of the Markov chain with $M = 320$ and $N = 40000$ iterations. This took 5 days on a Sun SPARCstation 20, using a specially written Fortran program. We report the results in Section 6.

5. PROJECTION

We can also use simulation to project the HIV—AIDS epidemic up to some projection horizon t^{**}. This involves, at each iteration of the Markov chain, sampling a number m^* of individuals who will be infected in $(t^*, t^{**}]$, along with future data:

$$\left\{t_\ell^H, t_\ell^A, t_\ell^R\right\}_{\ell=1}^{m^*}, \tag{5}$$

conditional on the current values of the model parameters. As before, the simulation can be organised to involve only elementary manipulations.

We can now project any functional of interest by letting $x^{(i)}$ denote the values of (3), (5) and η, α, ρ at iteration i, and using (4) as before. For example, to project the number of individuals infected in any time interval $(t_1,t_2]$, we set

$$f(x) = \sum_{j=1}^{m} I\left(t_1 < t_j^H \le t_2\right) + \sum_{k=1}^{n} I\left(t_1 < t_k^H \le t_2\right) + \sum_{\ell=1}^{m^*} I\left(t_1 < t_\ell^H \le t_2\right), \tag{6}$$

where $I(.)$ denotes the indicator function, being unity when its argument is true, and zero otherwise.

Equation (4) will then provide an estimate of the posterior predictive mean of the number infected in $(t_1,t_2]$. Replacing t^H with t^A in (6) gives an estimate of the posterior predictive mean of the numbers diagnosed with AIDS in $(t_1,t_2]$. Posterior predictive standard deviations, centiles, etc. can be similarly calculated. Note that if $t_2 \le t^*$ then (6) is 'back–projection'.

6. RESULTS

In this section we report the results from our Bayesian model.

Figure 1 projects AIDS incidence for each quarter up to the end of 1999. This suggests that the number of AIDS cases in homosexual/bisexual men has been stable for the last two years, and may begin to fall slightly over the next few years. Figure 1 also plots the raw data: the number of AIDS cases reported up to the end of 1994. This clearly shows the influence of reporting delays: naively interpreted, the raw data would suggest that AIDS incidence began to fall in 1993. Estimated AIDS incidence for the last quarter of 1994 is three times reported AIDS incidence for the same period. This clearly shows the importance of adjusting for reporting delays, as we have done in our Bayesian model.

Figure 2 gives 95% interval estimates for AIDS incidence for each quarter up to the end of 1999. This shows uncertainty bands progressively widening. For the period up to the end of 1994, uncertainty is due only to reporting delays. For the

1980's there is virtually no uncertainty, as by now reporting of these early diagnoses is essentially complete. For the period after 1994 the credible bands are much wider, since they are based only on forwards projections from HIV infections which have essentially been projected backwards from reported AIDS cases by the Markov chain.

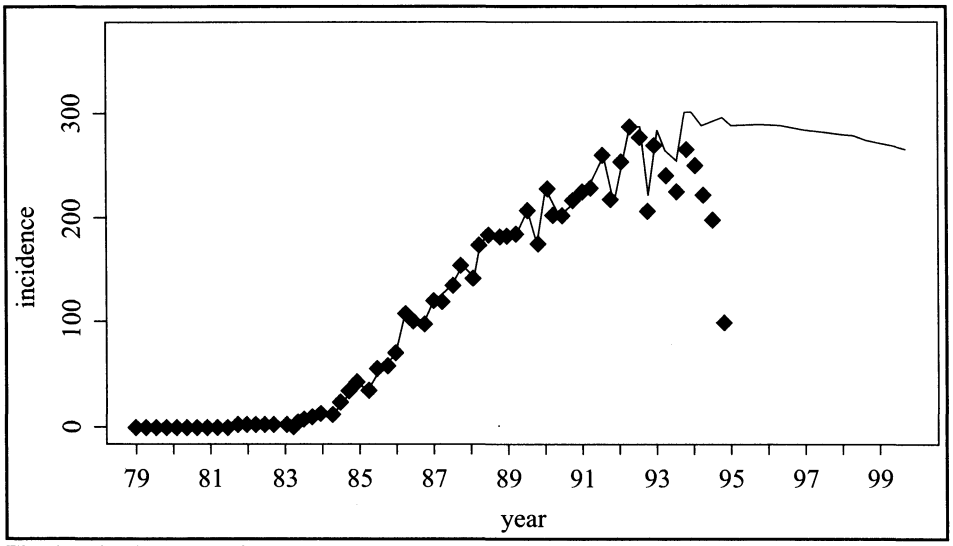

Fig. 1: Aids incidence for each quarter up to the end of 1999. Solid line: posterior predictive mean incidence estimated from the Bayesian model; dotted line: AIDS reports.

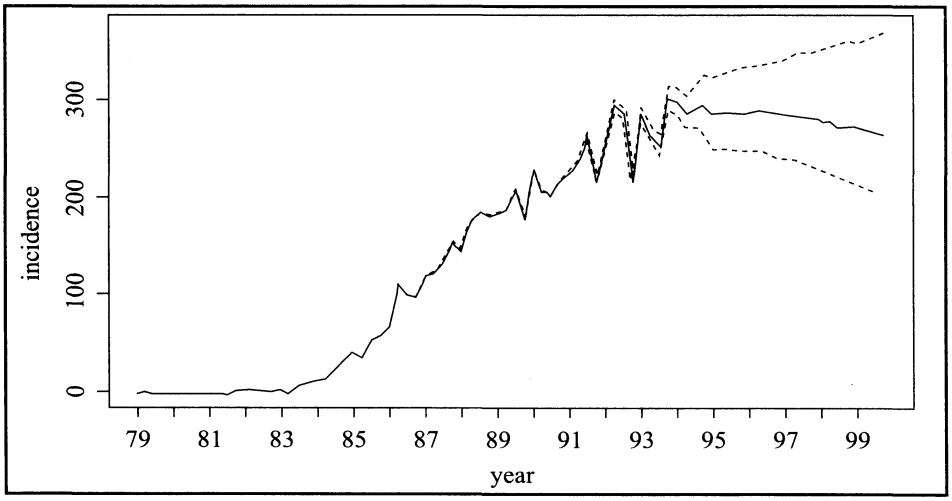

Fig. 2: Aids incidence for each quarter up to the end of 1999. Solid line: posterior predictive mean estimated from the Bayesian model; broken lines: 95% Bayesian credible intervals.

Figure 3 gives point and interval estimates for HIV incidence for each quarter up to the end of 1999. The piecewise–constant appearance is a direct result of the form assumed for $h(t^H)$. The forwards and backwards projections show a remarkable concentration of HIV infections during 1983, and another minor peak in 1989—90. Both of these are in accordance with expert views on the epidemic, although the 1983 peak may be somewhat over–concentrated.

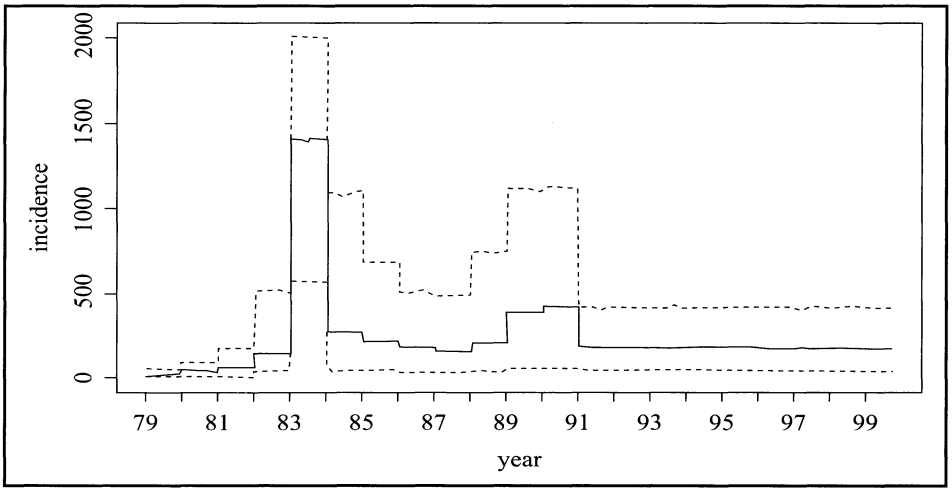

Fig. 3: HIV incidence for each quarter up to the end of 1999. Solid line: posterior predictive mean estimated from the Bayesian model; broken lines: 95% Bayesian credible intervals.

7. DISCUSSION

We have developed a framework for analysis which takes account of different but related sources of data, and which acknowledges many sources of uncertainty. For example, we do not calculate point estimates of the parameters of the incubation distribution, and then use these for projections as if they were the truth. The Bayesian approach allows uncertainty in these parameters, after observing the data, to be propogated into the predictions.

In our future work, we will apply the model of Section 3 to data on other risk groups. We also intend to develop a more flexible form for the infection intensity $h(t^H)$, to experiment with alternative incubation distributions $a(t^A \mid t^H)$, and to incorporate a treatment effect. We will also expand the reporting delay model $r(t^R \mid t^A)$, along the lines suggested in Section 3.

Our analytical framework will withstand considerable generalisation. In particular it is possible to formally include in the model the HIV–test registration data, referred to in Section 1. We also aim to extend the model to encompass all risk–groups simultaneously.

REFERENCES

COX, D. R. (Chairman of working group) (1988), Short–term prediction of HIV infection and AIDS in England and Wales. London: HMSO.

DAY, N. E. (Chairman of working group) (1990), Acquired Immune Deficiency Syndrome in England and Wales to end 1993: projections using data to end September 1989. *Communicable Disease Report (Suppl.).*

DAY, N. E. (Chairman of working group) (1993), The incidence and prevalence of AIDS and other severe HIV disease in England and Wales for 1992–1997: projections using data to the end of June 1992. *Communicable Disease Report (Suppl.).*

GELMAN, A. and RUBIN, D. B. (1992), Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457—472.

GILKS, W. R. and De ANGELIS, D. (1996), Projecting the AIDS epidemic in England and Wales: a Bayesian approach. *In preparation.*

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1996), Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall (in press).

GREEN, P. J. (1995), Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Technical Report,* University of Bristol.

HENDRIKS, J. C. M., MEDLEY, G. F., van GRIENSVEN, G. J. P. *et al.* (1993), The treatment–free incubation period of AIDS in a cohort of homosexual men. AIDS **7**, 231—239.

RAFTERY, A. E. and LEWIS, S. M. (1992), How many iterations of the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 641—649. Oxford: Oxford University Press.

PREVISIONI SULL'EPIDEMIA DI AIDS IN INGHILTERRA E GALLES: UN APPROCCIO BAYESIANO

RIASSUNTO

I dati disponibili per fare previsioni sull'epidemia di AIDS in Inghilterra e Galles provengono sia da registro individuale riferito ai soggetti cui è stata fatta la diagnosi di AIDS sia da indagini periodiche di sieroprevalenza sugli individui affetti dal virus HIV. L'analisi di questi dati risulta complicata a causa:

- *della mescolanza fra dati a livello individuale ed altri a livello aggregato*
- *della natura retrospettiva dei dati del registro*
- *dei ritardi nella segnalazione sulla data dell'infezione HIV nei soggetti con diagnosi di AIDS.*

Viene qui descritto un approccio bayesiano per la proiezione dell'epidemia di AIDS mediante lo sviluppo di un unico, coerente modello che tenga conto di tutte queste complicazioni.

Il modello proposto per proiettare l'epidemia negli uomini sia omo che bisessuali fino alla fine del 1999 utilizza le tecniche di stima Monte Carlo per catene di Markov.

Parole chiave: *AIDS, HIV, Markov Chain Monte Carlo, modello bayesiano, previsioni, back–projection, ritardo di notifica.*