STATISTICS IN MEDICINE, VOL. 8, 45–58 (1989)

MODELS FOR THE HIV INFECTION AND AIDS EPIDEMIC IN THE UNITED STATES

JEREMY M. G. TAYLOR

Division of Biostatistics, School of Public Health and Jonsson Comprehensive Cancer Center, University of California at Los Angeles, Los Angeles, California 90024, U.S.A.

SUMMARY

Statistical models of the HIV infection epidemic in the U.S. which account for the observed incidence of AIDS cases in the years 1978–1987 are considered. The models assume a known distribution of times from infection to AIDS. The best model estimates that there were approximately 563,000 to 1,110,000 individuals infected in the U.S. in April 1987. These estimates do not take into account underreporting of AIDS cases.

The sensitivity of the conclusions to the model's assumptions is ascertained by investigating a variety of parametric models for the infection epidemic, a variety of likely distributions for the time from infection to AIDS, and some plausible alternatives for the history of AIDS cases in the U.S.. It is concluded that there is too much uncertainty in the data and the models to be able to give highly accurate predictions of the number of people currently infected in the U.S., however, the results from the best fitting models suggest that there are less than the 1 to 1·5 million infected as estimated by the Centers for Disease Control. A Bayesian scheme is suggested for incorporating the uncertainty in the models.

KEY WORDS AIDS modelling HIV infection epidemic Time to AIDS distribution

INTRODUCTION

Various approaches to modelling the AIDS epidemic have been suggested. One approach[1,2] is to fit a smooth curve through the current AIDS incidence data and extrapolate this line into the future. Using this approach it is estimated[1] that there will be a total of 270,000 AIDS cases in the U.S. by the end of 1991. This approach will be useful for short term predictions but may be inaccurate for long term predictions.

A second method is to develop deterministic models[3,4] which take into account the many factors which affect the spread and development of AIDS. Such factors include the infectious period of the virus, the rates of interactions between and within different at-risk populations, the average incubation period and the effect of changes in social habits. This complex approach has the greatest potential for long term accurate predictions, but is subject to too many uncertainties to be of immediate usefulness.

The third approach, which will be pursued in this paper, is to develop a simple stochastic model. This approach attempts to use the approximately known distribution of times from HIV infection to AIDS and estimates a model for the growth of the HIV infection epidemic which will account for the history of AIDS incidence. Other authors adopt this third approach, either formally[5,6] or less formally.[7,8] Brookmeyer and Gail[5,9] give a justification and explain in detail the limited aim of this approach. In particular, they show that the approach can only predict the future number of AIDS cases arising from already infected individuals, and thus is an estimate of the minimum

0277–6715/89/010045–14$07.00 *Received August 1988*

future size of the AIDS epidemic. They estimate this minimum size to be 135,000 AIDS cases by the end of 1991. This paper extends the work of Brookmeyer and Gail,[5] as it estimates not only the future size of the AIDS epidemic, but also attempts to estimate the history of the HIV infection epidemic. The source of data for the distribution of times from infection to AIDS and the specific statistical models for the infection epidemic also differ.

In this paper, using the third approach, we concentrate on three areas:

(a) the estimate of the current number of HIV infected individuals in the United States;
(b) a comparison of parametric curves for the infection epidemic model;
(c) an evaluation of a Bayesian method for incorporating the uncertainty in the choice of model in addition to the usual statistical uncertainty associated with a fixed model.

METHODS

AIDS incidence data

Using the information in the Centers for Disease Control weekly AIDS surveillance reports (2 May 1988) it is assumed that the AIDS incidence in the six-month intervals from July 1978 to June 1987 was 1, 3, 9, 20, 34, 89, 181, 368, 656, 1240, 1616, 2526, 3339, 4717, 6168, 7876, 9456 and 11543. The data from the years 1983 to 1987 have been slightly adjusted upwards, by 0·2, 0·5, 1, 2·5, 4·0, 5·5, 7, 9 and 12 per cent for the nine intervals, to allow for the lag in reporting times. These adjustments for reporting delay were based on the analysis by Harris.[10]

Distribution of times from HIV infection to AIDS

Although this distribution is not precisely known, some estimates are available from cohort studies.[8,11–15] These estimates are in broad agreement with each other: in particular, that approximately 1·5 per cent of infected individuals will develop AIDS within two years of infection, 11 per cent in four years, and 23 per cent in six years. In our analysis it is necessary to assume this is known up to nine years following infection. To ensure that the assumptions concerning this distribution were not critical to the conclusions, 21 different distributions were used in the model. These 21 were chosen to cover the likely range of the possible true distribution. The range of the distributions and hazards are given in Figures 1 and 2. The 21 distributions at selected time points are shown in Table I. These 21 distributions can be thought of as the authors' prior beliefs concerning the distribution. Some of the distributions are higher (or lower) than the median over the whole time scale, while others are higher (or lower) than the median for early times, but lower (or higher) than the median for later times. Thus, the 21 distributions are not an orderly set, but rather they lie in a disorderly manner within the ranges given in Figures 1 and 2.

In Figure 1 the distribution is plotted on a log scale to emphasize short times after infection. Knowledge of this distribution at short times after infection is critical in a model-based approach to estimating the current HIV prevalence. The estimate of the distribution for times less than two years after infection is not exactly known, however, various studies provide valuable information. A study of haemophiliacs[11] ($N=84$) indicates no development of AIDS within 18 months of infection. A cohort of homosexual men[12,13] ($N=155$) also indicates no development of AIDS within 18 months of seroconversion. In both of these studies the time of seroconverion is only known up to an interval which in some cases is as wide as two years, and the midpoint of the interval is assumed to be the date of seroconversion. A prospective cohort study[15] of homosexual men ($N=268$) indicates no AIDS development within one year, 0·9 per cent AIDS development

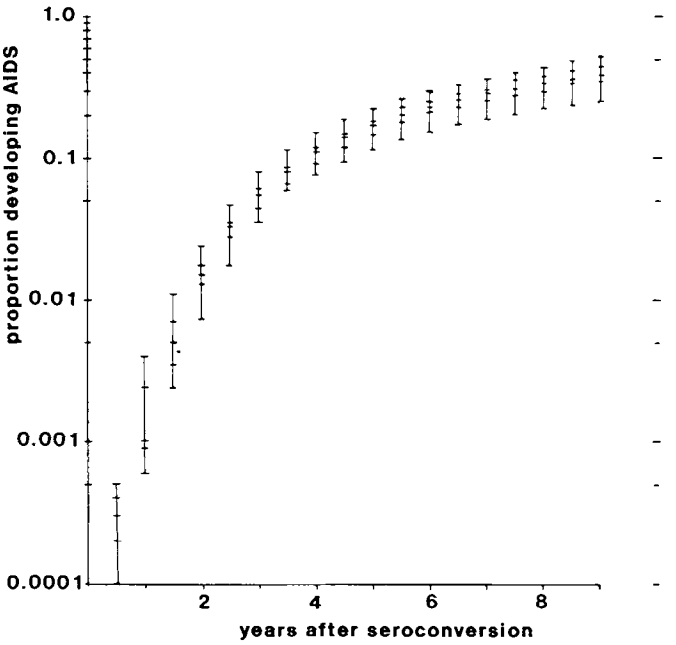

Figure 1. Range of the 21 distributions of the time from HIV infection to AIDS used in the analysis. These represent a range of likely estimates constructed from published estimates of this distribution.[8,11-15] Also shown are the 6th, 11th and 16th values of the distribution, when ordered at each time point

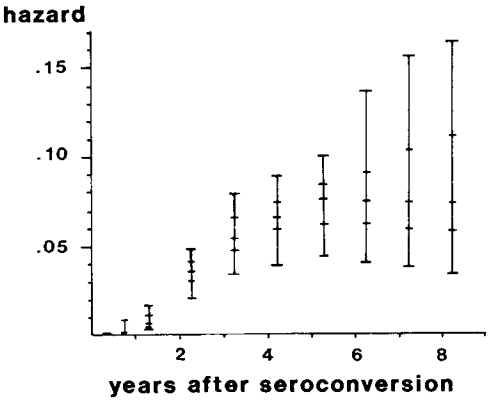

Figure 2. Range of the hazard functions of the 21 distributions of the time from infection to AIDS. Also shown are the 6th, 11th and 16th values of the hazard, when ordered at each time point

within 18 months and 1·4 per cent AIDS development within two years of seroconversion. In this study the date of seroconversion is known up to a six month interval. A study of transfusion-associated AIDS cases in adults[16] indicates that, in the U.S., there have been at least six AIDS cases within six months, at least 18 cases within 12 months and at least 47 cases within 18 months of infection. Together these data suggest that the distribution and hazard functions are very small

Table I. 21 distributions for time from infection to AIDS. 2,4,6 and 8 year values

	2 years	4 years	6 years	8 years
1	0·0074	0·0864	0·2134	0·3624
2	0·0210	0·1410	0·2520	0·3240
3	0·0109	0·1159	0·2269	0·3279
4	0·0190	0·1300	0·2660	0·3870
5	0·0150	0·1370	0·2830	0·4060
6	0·0110	0·1120	0·2280	0·3150
7	0·0117	0·0917	0·1937	0·2667
8	0·0091	0·0851	0·1631	0·2251
9	0·0130	0·0790	0·1750	0·2750
10	0·0145	0·1065	0·2315	0·3735
11	0·0128	0·1098	0·2298	0·3358
12	0·0175	0·1005	0·2045	0·2965
13	0·0220	0·1120	0·2320	0·3340
14	0·0174	0·1094	0·2364	0·3764
15	0·0144	0·1094	0·2514	0·3754
16	0·0170	0·1120	0·2260	0·3560
17	0·0240	0·1510	0·2970	0·4300
18	0·0240	0·1350	0·2740	0·4140
19	0·0150	0·0950	0·1750	0·2410
20	0·0150	0·0760	0·1530	0·2200
21	0·0140	0·0910	0·2460	0·4410

but non-zero in the first two years after infection. They are the basis for the ranges shown in Figures 1 and 2.

Models for the growth of the HIV infection epidemic

It is assumed that infections only occur on 1st January or 1st July of each year. This can be thought of as approximating the actual infections which occur in the six month span surrounding the beginning or midpoint of each year.

It is further assumed that the first date of infection is 1 July 1978, and only infections up to and including those on 1 January 1987 are considered. The assumption is consistent with data from a haemophilia-associated AIDS study[17] and data from a study in Los Angeles of homosexuals in a hepatitis clinic,[18] but it is not consistent with the information from a San Francisco cohort of homosexual men in a STD clinic where there was evidence of some infections prior to July 1978.[19]

Let Y_i denote the incidence of HIV infections at time point i, where $i=1$ corresponds to July 1978, $i=2$ corresponds to January 1979, etc. Five models for Y_i were entertained:

Model 1. Double exponential incidence $Y_i = he^{-e^{a+bi}}$

Model 2. Root exponential incidence $Y_i = he^{bi^{1/4}}$

Model 3. Logistic incidence $Y_i = \dfrac{he^{a+bi}}{1+e^{a+bi}}$

Model 4. Logistic prevalence $Y_i = \dfrac{hbe^{a+bi}}{(1+e^{a+bi})^2}$

Model 5. Quadratic incidence $Y_i = hi^2$.

Strictly speaking Y_i, which is the number of new infections at time point i, should be integer valued, however, models 1–5 do not restrict Y_i to integers.

Model 3 represents an initial exponential growth with the possibility that the incidence is reaching a plateau later on in the epidemic. Model 1 is similar in shape to model 3, but the plateau is reached less abruptly. Model 4 represents an initial exponential growth but with the rate of increase of the incidence reduced in later years with possibly even decreasing incidence. Models 1, 2, 3 and 5 show an increasing rate of incidence, but increasing at less than an exponential rate. None of these models is expected to describe exactly the history of the HIV epidemic, however, in the current state of knowledge it was considered preferable to use simple, plausible, smooth models, rather than more complex models with many parameters.

The limited epidemiologic evidence suggests that incidence of infection is not drastically increasing in the later years of the epidemic, with the possible exception of the drug user risk group. Longitudinal studies of homosexual men have shown that the incidence of HIV infection has decreased over the years 1984–1987 in cohorts which are under intense study.[20] The introduction of screening of the blood supply in March 1985 also had the effect of decreasing the potential spread of the virus. Information concerning the incidence of infection among drug users and their sexual partners is less well known. This is the risk group which is more likely than the others to show a rapid rise in the HIV incidence.

Two further models, linear incidence ($Y_i = hi$) and exponential incidence ($Y_i = he^{bi}$) were used, but gave very poor fits to the data and thus the results are not shown.

Statistical methods

Let F_i, $i = 1, \ldots, 18$ denote the probability of developing AIDS within i six-month time periods of infection, thus F_{18} represents the probability of developing AIDS within nine years of infection. Let $X_j, j = 1, \ldots, 18$ denote the number of AIDS cases which developed in period j, where $j = 1$ is the second half of 1978, denoted 1978(2), and $j = 18$ is the first half of 1987, denoted 1987(1). Let $N(= 49842)$ be the total number of AIDS cases which have developed up to the middle of 1987. With this notation, the distribution of X_j is multinomial $(N:\Pi_1, \ldots, \Pi_{18})$ where Π_j is the probability that one of the AIDS cases, from the set of N cases, occurs in interval j. The probability Π_j is given by

$$\Pi_j = \frac{1}{N} \sum_{i=1}^{j} Y_i (F_{j+1-i} - F_{j-i})$$

where $F_0 = 0$.

This is the same convolution equation used by previous authors.[5,6,20] Inherent in this equation is the assumption that the Y_i are all infected at the beginning of the ith interval, while F_i denotes the probability of developing AIDS any time up to the end of the ith six month period following infection.

This formulation allows one to estimate the parameters of the infection incidence model using maximum likelihood and to obtain appropriate standard errors of quantities of interest. A Fisher scoring algorithm was used to maximize the likelihood.

The fit of the various models is evaluated using a likelihood ratio chi-squared statistic,

$$\chi^2 = -2 \sum X_j \left(\log\left(\frac{X_j}{N}\right) - \log(\hat{\Pi}_j) \right),$$

where $\hat{\Pi}_j$ is the estimate of Π_j from the model. The χ^2 statistic has 15 degrees of freedom for models 1, 3 and 4, 16 degrees of freedom for model 2, and 17 degrees of freedom for model 5. It should be

noted that as the models are non-nested, the difference between two χ^2 statistics has an unknown distribution. However, the individual χ^2 statistics can still be used as a measure of the goodness-of-fit of a model.

The prediction for the number of new AIDS cases A_k developing in the kth six month period following July 1987 is obtained from the equation

$$A_k = \sum_{j=k+1}^{18} Y_j(F_{19+k-j} - F_{18+k-j}),$$

where $k=1$ refers to 1987(2) and $k=9$ refers to 1991(2). The values of A_k are estimates of the minimum future number of AIDS cases. They are underestimates because they do not include any cases which might develop in people who are uninfected in March 1987, and because it is assumed that there is no possibility of developing AIDS greater than nine years after infection. The values of A_1 and A_2 are likely to be only slight underestimates but A_8 and A_9 could be substantial underestimates.

The models were further checked by comparing A_1, the model predicted minimum number of cases in 1987(2), with the range 12600 to 13800 which is the likely number based on extrapolation from the current AIDS incidence surveillance data.

RESULTS

Comparison of the infection epidemic models

Table II shows the comparison of the five models. The first row shows the median and range (over the 21 distributions of F) of the χ^2 goodness-of-fit statistics. The values indicate that model 1 is slightly preferable to model 3 which is slightly preferable to model 4 which is preferable to models 2 and 5. Model 1 gave the smallest χ^2 value for 20 of the 21 distributions. The second row indicates that models 2 and 5 give unreasonably high predictions of the number of AIDS cases in 1987(2), and models 1 and 3 are slightly preferable to model 4, which gives too low a prediction. The third row shows for how many of the 21 distributions the estimated number of AIDS cases in 1987(1) is higher than the observed number (11543). Again Model 1 is the best among the five choices. The fourth row gives the median and range for the estimated total number of HIV infected people in the U.S. up to and including those infected on 1 January 1987. The estimates for models 1, 3 and 4 are lower than the estimate of 1 to 1·5 million given by the U.S. Public Health Service,[20] unless there were a surprisingly large number of new infections since January 1987. The fifth row gives the predicted minimum total number of AIDS cases by the end of 1991. The values for models 1, 3 and 4 are not inconsistent with the estimate of 270,000 cases by the end of 1991.

The conclusions to be drawn from Table II are that models which show a levelling or even a decreasing of the HIV incidence better explain the AIDS incidence data than models in which the HIV incidence is continually rising at an increasing rate.

Comparison of the incubation distributions

There was no obvious pattern to which of the 21 distributions gave the smallest chi-squared value, and in fact some widely different distributions gave very similar chi-squared values and predictions. Thus this method cannot be used with the currently available data to estimate the distribution F.

Table II. Comparison of five models

		Infection growth model				
		Double exponential incidence	Root exponential incidence	Logistic incidence	Logistic prevalence	Quadratic incidence
Chi-squared values	Median*	49	192	65	107	125
	Range*(L)	43	127	46	73	71
	Range*(U)	60	229	80	133	323
Predicted number† of AIDS cases in 1987(2)	Median	13,679	15,043	13,181	12,548	14,324
	Range (L)	13,556	14,741	12,979	12,354	13,596
	Range (U)	13,812	15,218	13,563	12,982	15,796
Fraction of distributions with predicted number of AIDS cases in 1987(1) greater than 11,543		14/21	21/21	0/21	0/21	18/21
Number of currently‡ infected people in United States	Median	729,186	931,317	666,411	527,119	852,191
	Range (L)	562,898	716,679	510,277	413,662	622,826
	Range (U)	1,109,564	1,398,028	1,008,002	834,718	1,205,512
Minimum predicted† total number of AIDS cases by end of 1991	Median	207,232	249,391	193,544	169,940	229,943
	Range (L)	192,474	228,373	175,786	150,907	200,676
	Range (U)	226,015	276,719	214,227	187,221	292,792

* Median and range refer to the values obtained from the 21 different distributions of the time from HIV infection to AIDS
† Predicted number of AIDS cases are minimum estimates as they refer only to those who are currently infected, and assume that AIDS cannot develop more than 9 years after infection
‡ Refers to the number infected up to the end of March 1987

Predictions of future cases and current prevalence from the best fitting model

Fourteen of the 105 choices of model gave χ^2 values less than 50. For these 14 the range of the estimated current prevalence was 549,904 to 887,932 and the range of the total minimum number of AIDS cases by the end of 1991 is 204,064 to 226,015.

The smallest chi-squared value for the double exponential incidence was 43. This χ^2 value, although statistically significant, on 15 degrees of freedom represents a reasonable fit to the data. The pattern of observed AIDS cases shows a seasonal effect, with higher than expected number of cases in the first half of each year from 1983 to 1986. This oscillating effect could not be fit by any smooth model of the form described above, and is the major reason for the high χ^2 values. The 1240 cases in 1983(1) is the particular data point which makes the largest contribution to χ^2 value. Using the same models with discrete units of one year instead of six months eliminates the seasonal effects and gives smaller non-significant χ^2 values for models 1 and 3.

The particular cumulative distribution for the time from infection to AIDS which gave rise to this χ^2 value was $F_1 = 0{\cdot}0004$, $F_2 = 0{\cdot}0030$, $F_3 = 0{\cdot}0080$, $F_4 = 0.015$, $F_5 = 0{\cdot}035$, $F_6 = 0{\cdot}065$, $F_7 = 0{\cdot}100$, $F_8 = 0{\cdot}137$, $F_9 = 0{\cdot}174$, $F_{10} = 0{\cdot}211$, $F_{11} = 0{\cdot}248$, $F_{12} = 0{\cdot}283$, $F_{13} = 0{\cdot}316$, $F_{14} = 0{\cdot}347$, $F_{15} = 0{\cdot}377$, $F_{16} = 0{\cdot}406$, $F_{17} = 0{\cdot}434$, $F_{18} = 0{\cdot}461$. This distribution is described in detail for illustrative purposes; it is chosen because it gave the smallest χ^2 value, not because it is thought to be the most likely estimate. The parameters of the model for the infection epidemic are $h = 83{,}798$,

Table III. Model estimates* of the number of AIDS cases and the incidence of HIV in the years 1978–87

	Number of AIDS cases in each period		
	Model predicted	True number	Model predicted number newly infected during period
1978(2)–1979(2)	7	13	3802
1980(1)	16	20	4520
1980(2)	42	34	7712
1981(1)	97	89	11,933
1981(2)	201	181	17,046
1982(1)	380	368	22,813
1982(2)	666	656	28,945
1983(1)	1092	1240	35,160
1983(2)	1689	1616	41,216
1984(1)	2482	2526	46,931
1984(2)	3486	3339	52,183
1985(1)	4706	4717	56,907
1985(2)	6136	6168	61,083
1986(1)	7763	7876	64,721
1986(2)	9565	9456	67,854
1987(1)	11,517	11,543	70,526

* The estimates are from the double exponential incidence model for HIV infection and the best fitting distribution of times from infection to AIDS (see text)

$a = 1{\cdot}88$ (SE $= 0{\cdot}022$), $b = -0{\cdot}20$ ($SE = 0{\cdot}009$). Table III shows the model predicted number of AIDS cases in the years 1978(2)–1987(1), which are very close to actual values. The table also shows the estimated number of people infected on 1st January and 1st July of each year. The double exponential incidence model predicts that the total number of infected people in the U.S., up to and including those infected on 1 January 1987, is 593,353 (SE = 9672). The model predicted the minimum number of AIDS cases in the six-month time periods 1987(2) to 1991(2) to be 13,552, 15,505, 17,317, 18,965, 19,604, 19,385, 18,626, 17,517, 16,195, respectively, giving a total number of cases by the end of 1991 of 206,509.

Sensitivity analysis

In addition to considering a variety of models for Y the infection incidence, and a range of distributions for F, we also considered the possibility that the reported number of AIDS cases are not a true reflection of the actual number which occurred. Three scenarios were considered:

(i) *Underreporting in the years* 1978–1981 Since many of the reported cases in these years were obtained by retrospective analysis, underreporting is conceivable. To investigate the effect of this, the number of cases in the six-month periods prior to 1982 were changed from 1, 3, 9, 20, 34, 89, 181 to 1, 6, 18, 35, 51, 111, 226, respectively, and the whole analysis repeated. This change slightly increased the chi-squared values for model 1 and left the values approximately the same for models 3 and 4. It also indicated that model 3 is slightly preferable to model 1. The estimates of the minimum predicted number of AIDS cases by the end of 1991 increased by about 2·5 per cent and the estimate of the number of infected people in the U.S. increased by about 3·5 per cent.

(ii) *Underreporting in all years* For a variety of reasons it is likely that a significant proportion of AIDS cases are not reported to local health departments. We hypothesized that the actual number of AIDS cases is 15 per cent higher than the reported number and repeated the analysis. The effect of this is to increase all estimates of the predicted number of cases and the number infected by 15 per cent.

(iii) *Inappropriate estimate of the number of cases in* 1987 The estimated number of cases in the first half of 1987 is assumed to be 11,543 in the calculations; this is calculated as a 12 per cent increase of the 10,306 cases reported in the surveillance report of 2 May 1988. To investigate the effect of uncertainty in the lag times between diagnosis and reporting, the analysis was repeated assuming either 11,100 or 12,000 cases for the first half of 1987. This did not effect the result that the double exponential incidence model was the best of the five considered; however, for the larger number of cases the quadratic incidence model was sometimes the second best. The estimate of the number of infected people in the U.S. was decreased or increased by about 5·0 per cent and the predicted total number of cases by 1991 was decreased or increased by approximately 4·0 per cent.

Incorporating the uncertainty of the selected model

The estimate and standard error of the HIV prevalence in the U.S. from the best fitting model are 593,353 and 9672. This standard error represents only the sampling variability for the fixed choice of F and the HIV infection epidemic model, and does not incorporate the uncertainty in the knowledge of that model, and thus is clearly an underestimate. One way to incorporate the model selection uncertainty is to use the Bayesian approach advocated by Hodges.[21]

Let j denote the model ($j=1, \ldots, 105$), which refers to a fixed combination of F and infection model; let θ_j denote the parameters of model j; let Y denote the observations; and let $Z=h_j(Y, \theta_j)$ be the number of people currently infected. For the distributions, let $g(Y|\theta_j, j)$ denote the density of the observation and $L(\theta_j)$ the log-likelihood, let $\Pi(\theta_j|j)$ denote the prior for θ_j and W_j be the discrete prior for model j. For ease of notation we will assume that both these priors are uniform.

By application of Bayes' rule the predictive distribution of Z is given by

$$P(Z|Y)=\sum_j P(Z|Y,j)\,P(j|Y)$$

where

$$P(Z|Y,j)=\int_{h_j(Y,\theta_j)=Z} P(\theta_j|Y,j)\,\mathrm{d}\theta_j$$

and

$$P(j|Y)=\frac{\int_{\theta_j} g(Y|\theta_j,j)\,\mathrm{d}\theta_j}{\sum_j \int_{\theta_j} g(Y|\theta_j,j)\,\mathrm{d}\theta_j}$$

by application of Bayes' rule.

The predictive mean of Z is

$$\begin{aligned} E(Z|Y) &= \int Z P(Z|Y)\,\mathrm{d}Z \\ &= \sum_j E(Z|Y,j)\,P(j|Y). \end{aligned} \tag{1}$$

It can be shown that the predictive variance of Z is

$$\operatorname{var}(Z|Y)=\sum_j \operatorname{var}(Z|Y,j)\,P(j|Y) + \sum_j [E(Z|Y,j)-E(Z|Y)]^2\,P(j|Y). \quad (2)$$

To evaluate these expressions requires numerical integration, however, as a first attempt, various approximations are used. $E(Z|Y,j)$ is approximated by $h_j(Y,\hat{\theta}_j)$, that is, the MLE of Z for model j. $\operatorname{var}(Z|Y,j)$ is approximated by the usual sampling variability variance for a fixed model based on the information matrix. An approximation to $P(j|Y)$, the posterior probability of model j, or equivalently $\int g(Y|\theta_j,j)\mathrm{d}\theta_j$, is based on the proposal by Schwarz[22] and Tierney and Kadane.[23] Their work suggests using $P(j|Y)$ proportional to $g(Y|\hat{\theta}_j,j)(2\Pi)^{p/2}|\sum|^{1/2}$ where p is the number of parameters in the model and $\sum$ is the inverse of the information matrix. The justification for this choice is to note that

$$\begin{aligned}\int g(Y|\theta_j,j)\mathrm{d}\theta_j &= \int \mathrm{e}^{L(\theta_j)}\mathrm{d}\theta_j \\ &\simeq \mathrm{e}^{L(\hat{\theta}_j)}\int \mathrm{e}^{1/2(\theta_j-\hat{\theta}_j)L''(\hat{\theta}_j)(\theta_j-\hat{\theta}_j)}\mathrm{d}\theta_j \\ &\simeq g(Y|\hat{\theta}_j,j)(2\Pi)^{p/2}|E[L''(\hat{\theta}_j)]|^{-1/2}.\end{aligned}$$

Using these weights in (1) and (2) the predictive mean of Z and its standard error are 602,549 and 72,024. Although this standard error is appreciably larger and more realistic than the standard error from a fixed model, it is not entirely satisfactory for three reasons. First, it does not incorporate the uncertainty in the AIDS surveillance data or the reporting delay assumptions; second, it is dependent on the accuracy of the approximations used in its calculation; and finally, it is dependent on how one chooses the prior set of distributions and infection models. Thus until one feels more confident about the AIDS surveillance data and one knows how to assess one's prior beliefs about F and the models for the infection epidemic, this method, although preferable to the results from a fixed model, should not be considered highly accurate.

DISCUSSION

This paper has investigated a simple model for the growth of the HIV infection and AIDS epidemic in the U.S.. In essence we developed a model for the infection epidemic which best accounts for the history of the reported number of AIDS cases, using the assumed known distribution of times from infection to AIDS.

The conclusions from the analysis are that the number of HIV infected people in the U.S., including only those infected prior to April 1987, is in the range 563,000 to 1,110,000. This range is lower than the 1 to 1·5 million estimated to be currently infected.[20] Even if one takes into account underreporting and a liberal estimate of the number of new infections since March 1987, the estimate is still lower than 1·5 million. This suggests that the lower limit of the 1 to 1·5 million range is the more reasonable value. The second conclusion from the analysis is that the minimum predicted total number of AIDS cases by the end of 1991 is in the range 192,000 to 224,000. This number is not inconsistent with other predictions[1,5] of the future course of the AIDS epidemic. The third conclusion from the analysis is that the incidence of HIV infection is not growing exponentially and could even have reached a constant level, or be decreasing.

An analysis similar to the above was performed by Brookmeyer and Gail.[5] They reached a slightly lower value concerning the predicted minimum number of AIDS cases. There are three significant differences between the analyses. First, they assumed a step function continuous model with four constant levels of incidence for the growth of the infection epidemic. In contrast, we assumed a discrete model based on a variety of smooth curves for the infection epidemic. Secondly, the Brookmeyer and Gail analysis was based on older AIDS incidence data and they assumed no new infections in 1986 and thereafter. This may explain the numerical differences. The third difference is that Brookmeyer and Gail used the distribution of incubation times as estimated using a Weibull model from data on transfusion associated AIDS cases.[24] This distribution is of the times from infection to AIDS of all people who have or will sometime in the future develop AIDS. This is a hard distribution to estimate from retrospective data as indicated by the very large confidence intervals for the mean of the assumed Weibull model.[25] The theoretical problems involved in estimating the distribution of incubation times are discussed in Brookmeyer and Gail.[26] In contrast we use the distribution of times from infection to AIDS, specified non-parametrically. This distribution, which refers to all infected people, is also hard to estimate, but many estimates from different samples[8,11–15] agree quite closely. Using the distribution of times from infection to AIDS instead of the incubation times distribution allows one to estimate the number of people currently infected as well as make predictions about the future size of the AIDS epidemic.

The most interesting and controversial result is the difference between the model-based estimate of the number of infected people in the U.S. (563,000–1,110,000) and the Department of Health and Human Services estimate of 1 to 1·5 million. One explanation of this difference is the discreteness assumption in the model, that is, assigning all infections which occur in a six-month interval to the midpoint of that interval, and assuming that no infectons occurred prior to 1 July 1978. These assumptions were made for mathematical convenience and because of the discreteness of the AIDS surveillance data, and are likely to give only a negligible bias in the results of the model fitting.

A second explanation is that the surveillance data on the number of AIDS cases is inaccurate owing to underreporting. However, any inaccuracy in these data is unlikely to be of such a magnitude so as to fully account for the difference.

A third possibility is that the hypothesized model (double exponential incidence) for the growth of the HIV infection is inappropriate. Although it is certainly not an exact model, it did fit better than four alternative models and it gave good predictions for the number of AIDS cases in the years 1978–1987, thus it is unlikely to be grossly inaccurate. The quadratic and the root exponential infection models, both of which show continually increasing incidence of infections, gave a consistently worse fit to the data than the logistic prevalence model which shows a decreasing incidence of infection since 1984. Also, the quadratic and root exponential models consistently gave too high estimates for the number of AIDS cases in the first half of 1987, and too high predictions for the number of cases in the second half of 1987. This suggests that the recent incidence of HIV infection is not growing as fast as suggested by these two models.

Another explanation is that the distribution of times from infection to AIDS is not known exactly. This is certainly true, however it appeared not to make a substantial difference to the estimated number infected. Twenty-one different likely distributions were tried and the range of the number infected was 563,000 to 1,110,000. Nineteen of the 21 distributions gave estimates less than 1 million, and clearly less than 1·5 million, even if new infections are allowed for.

A fifth possibility is that the distribution of times from infection to AIDS has changed over the course of the epidemic, whereas the model assumed it was constant. This is a distinct possibility

because the clinical definition of AIDS is a heterogeneous endpoint affecting diverse risk groups. There are certainly major temporal, regional and demographic differences in AIDS.[26] For example, the incidence of Kaposi's Sarcoma is higher in the homosexual population with AIDS than in IV drug users with AIDS. Also the proportion of AIDS cases with Kaposi's Sarcoma has been decreasing recently. These and other reasons, such as the effect of behavioural changes on the natural history of infected people, indicate that the distribution of times from infection to AIDS is a dynamically changing distribution influenced by many factors, and thus it may be unreasonable to assume it is constant.

Another possible explanation of the discrepancy is that 1 to 1·5 million is inaccurate; this is certainly conceivable because it is not based on any well designed national survey, but rather on a conglomeration of evidence concerning HIV prevalence rates and estimates of the size of the major risk group populations.

In summary, the most likely explanation of the difference between the two estimates of the number of people currently infected in the U.S. is either that 1 million is an accurate figure and uncertainties in the data and model account for the difference, or that 1 to 1·5 million is an overestimate, or that the distribution of times from infection to AIDS has changed over the course of the epidemic in such a way that the incubation times are now longer than they were in the early years of the epidemic.

A Bayesian analysis which attempts to incorporate the uncertainty in the selection of the model as well as the statistical uncertainty associated with a fixed model was performed. This method gives more weight to those models which give a better fit to the data. Using this technique there are estimated to be 603,000 (SE = 72,000) infected individuals in the U.S.. Although this estimate is preferable to any based on a specific model, it is not entirely satisfactory as there are various mathematical approximations in its calculation, it does not incorporate the uncertainty in the AIDS surveillance data and delayed reporting time assumptions, and it is dependent on the choice of models which are investigated. Despite these reservations it is consistent with the conclusion that there are less than 1 million people infected in the U.S..

The Centers for Disease Control report[20] suggested that modelling of this form could not be used to estimate the HIV prevalence in the U.S.. They argue that since very few people progress to AIDS within 2–3 years of infection, the AIDS surveillance data are only giving information about infections up to 1984. Thus many different models for the infection epidemic which are similar up to 1984 but diverge beyond 1984 are consistent with the AIDS surveillance data. This is certainly true to a degree, but is a pessimistic view of the ability to distinguish between models. The analysis in this paper suggests that the infection epidemic models which fit the data are similar up to about the beginning of 1985. The proportion developing AIDS within two years of infection is known to be close to but not exactly zero[15,16] (see Figures 1 and 2); thus infection epidemic curves which diverge rapidly beyond 1984 may give measurably different contributions to the number of AIDS cases. For example, for the distribution used in Table III, the double exponential incidence model predicts that there were 264,184 infected in July 1985 or later, of which 956 developed AIDS in the first half of 1987, out of a total of 11,517 predicted for that interval. The actual number of cases in the first half of 1987 was 11,543. The root exponential model predicted that there were 423,790 infected in July 1985 or later, of which 1412 developed AIDS in the first half of 1987, out of a total of 12,088 predicted for that interval. This difference in the number of AIDS cases in 1987 from those infected in 1985 or later is at the limit of being measurably different, and is one of the main reasons that the χ^2 values are higher for the root-exponential model than the double exponential model. This example illustrates that using non-zero values for the proportion developing AIDS within two years of infection enables one to distinguish, to a small degree, between infection epidemic curves which diverge rapidly in the later years of the epidemic. However, the ability to

distinguish is reduced the later these curves diverge. In this paper the estimate of the number infected refers to March 1987; faith in the double exponential incidence model is required to bring that estimate from 1985 to 1987, and extrapolation is needed to extend this estimate to times later than March 1987.

This paper describes a simple stochastic model for the AIDS epidemic. As with any model it makes many assumptions and is subject to numerous uncertainties. As further information is gathered on the AIDS epidemic it can be incorporated into this model. Particularly useful would be more precise estimates of the distribution of times from infection to AIDS and better estimates of the number of currently infected individuals based on properly designed surveys. Future applications of this type of modelling will probably give only marginally more accurate estimates of the number currently infected. There will always be uncertainties associated with the AIDS surveillance data, uncertainties about the distribution of times from infection to AIDS at long follow-up times and uncertainties in the estimate of the number of recently infected individuals. Thus this type of modelling approach is a poor substitute for a national survey of HIV prevalence.

The results of the model fitting emphasize the enormity of the AIDS epidemic: first with a prediction of 192,000 to 224,000 AIDS cases at a minimum by the end of 1991, and secondly by estimating that there were approximately 563,000 to 1,110,000 HIV infected individuals in the U.S. in March 1987. This estimate is similar to that based on a non-random sample of military recruits.[27] Although it is not as large as the 1 to 1·5 million estimated by other means, it is still an extremely large number and a cause for public health concern.

The Centers for Disease Control have estimated 1 to 1·5 million for the number infected in December 1987.[20] This paper suggests a range of 563,000 to 1,110,000 for the number infected in March 1987, or 460,000 to 750,000 if the Bayesian analysis is to be trusted. If one adds 15 per cent to the first range to allow for new infections since March 1987 and for underreporting of AIDS cases, and one adds 10 per cent uncertainty to represent the statistical variability and the uncertainty associated with the appropriateness of the double exponential incidence model, then the most conservative range suggested by this paper is 590,000 to 1,390,000 with a median value of 840,000. Thus if the 1 to 1·5 million is accurate, the modelling in this paper would suggest that 1 million is the more likely figure.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the National Cancer Institute, CA-16042.

REFERENCES

1. Morgan, W. M. and Curran, J. W. 'Acquired immunodeficiency syndrome: current and future trends', *Public Health Reports*, **101**, 459–465 (1986).
2. 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).
3. Pickering, J., Wiley, J. A., Padian, N. S., Lieb, L. E., Echenberg, D. F. and Walker, J. 'Modelling the incidence of acquired immunodeficiency syndrome (AIDS) in San Francisco, Los Angeles, and New York', *Mathematical Modelling*, **7**, 661–688 (1986).
4. Anderson, R. M., Medley, G. F., May, R. M., and Johnson, A. M. 'A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS', *IMA Journal of Mathematics Applied in Medicine and Biology*, **3**, 229–263 (1986).
5. Brookmeyer, R., and Gail, M. H. 'Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States', *Lancet* **ii**, 1320–1322 (1986).
6. De Gruttola, V. and Lagakos, S. W. 'The value of doubling time in assessing the course of the AIDS epidemic', Technical Report, Harvard School of Public Health, 1987.
7. Van Druten, J. A. M., de Boa, T. H., Jager, J. C., Heisterkamp, S. H., Coutinho, R. A. and Ruitenberg, E. J. 'AIDS prediction and intervention', *Lancet*, **ii**, 852–853 (1986).

8. Taylor, J. M. G., Schwartz, K. and Detels, R. 'The time from infection with human immunodeficiency virus (HIV) to the onset of AIDS', *Journal of Infections Diseases*, **154**, 694–697 (1986).
9. Brookmeyer, R. and Gail, M. H. 'Methods for projecting the AIDS epidemic', *Lancet*, **ii**, 99 (1987).
10. Harris, J. E. 'Delay in reporting Acquired Immune Deficiency Syndrome (AIDS)', M. I. T. Technical Report No. 452, 1987.
11. Eyster, M. E., Gail, M. H., Ballard, J. O., Al-Mandhing, H. and Goedert, J. J. 'Natural history of human immunodeficiency virus in hemophiliacs: Effect of T-cell subsets, platelet counts, and age', *Annals of Internal Medicine*, **107**, 1–6 (1987).
12. Hessol, N. A., Rutherford, G. W., O'Malley, P. M., Doll, L. S., Darrow, W. W., Jaffe, H. W., Lifson, A. R., Engelman, J. G., Maus, R., Werdegar, D. and Curran, J. W. 'The natural history of human immunodeficiency virus infection in a cohort of homosexual and bisexual men: a seven-year prospective study', Proceedings of the Third International Conference on AIDS, Washington, D.C., 1987.
13. Curran, J. W., Jaffe, H. W., Hardy, A. M., Morgan, W. M., Selik, R. M. and Dondero, T. J. 'Epidemiology of HIV infection and AIDS in the United States', *Science*, **239**, 610–616 (1988).
14. Ward, J..W., Deppe, D., Perkins, H., Kleinman, S., Holland, P.and Allen, J. 'Risk of disease in recipients of blood from donors later found infected with human immunodeficiency virus (HIV)', Proceedings of the Third International Conference on AIDS, Washington, D.C., 1987.
15. Phair, J., Munoz, A., Kingsley, L., Fox, R., Kaslow, R., Visscher, B. and Jacobsen, L. 'Incidence of AIDS in homosexual men developing HIV infection: (Abstract)', Fourth International Conference on AIDS, Stockholm, Sweden, 1988.
16. Peterman, T. A., Lui, K. J., Lawrence, D. N. and Allen, J. R. 'Estimating the risks of transfusion-associated acquired immune deficiency syndrome and human immunodeficiency virus infection', *Transfusion* **27**, 371–374 (1987).
17. Eyster, M. E., Goedert, J. J., Sarngadharan, M. G., Weiss, S. H., Gallo, R. C. and Blattner, W. A. 'Development and early natural history of HTLV-III antibodies in persons with hemophilia', *Journal of the American Medical Association*, **253**, 2219–2223 (1985).
18. De Cock, K. M., Niland, J. C., Lu, H. P., Rahimian, A., Edwards, V., Shriver, K., Govindarajan, S. and Redeker, A. G. 'Experience with human immunodeficiency virus infection in patients with Hepatitis B virus and Hepatitis delta virus infections in Los Angeles, 1977–1985', *American Journal of Epidemiology*, **127**, 1250–1260 (1988).
19. Jaffe, H. W., Darrow, W. W., Echenberg, D. F., O'Malley, P. M., Getchell, J. P., Kalyanaraman, V. S., Byers, R. H., Drennan, D. P., Braff, E. H., Curran, J. W. and Francis, D. P. 'The acquired immunodeficiency syndrome in a cohort of homosexual men. A six-year follow-up study', *Annals of Internal Medicine*, **103**, 210–214 (1985).
20. Department of Health and Human Services. *Human Immunodeficiency Virus Infection in the United States: A Review of Current knowledge and Plans for Expansion of HIV Surveillance Activities*, U.S. Public Health Service, Centers of Disease Control, November 30, 1987.
21. Hodges, J. S. 'Uncertainty, policy analysis and statistics', *Statistical Science*, **2**, 259–275 (1987).
22. Schwarz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
23. Tierney, L. and Kadane, J. B. 'Accurate approximations for posterior moments and marginal densities', *Journal of the American Statistical Association*, **81**, 82–86 (1986).
24. Lui, K. J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W. and Bregman, D. J. 'A model-based approach for estimating the mean incubation period of transfusion associated acquired immunodeficiency syndrome, *Proceedings of the National Academy of Science USA*, **83**, 3051–3055 (1986).
25. 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).
26. Medley, G. F., Anderson, R. M., Cox, D. R. and Billard L. 'Incubation period of AIDS in patients infected via blood transfusion', *Nature*, **328**, 719–721 (1987).
27. Burke, D. S., Brundage, J. F., Herbold, J. R., Berner, W., Gardner, L. I., Gunzenhauser, J. D., Voskovitch, J. and Redfield, R. R. 'Human immunodeficiency virus infections among civilian applicants for United States military service, October 1985 to March 1986. Demographic factors associated with seropositivity', *New England Journal of Medicine*, **317**, 131–136 (1987).