JIA 117 (1990) 319-405

ACTUARIAL REVIEW OF MODELS FOR DESCRIBING AND PREDICTING THE SPREAD OF HIV INFECTION AND AIDS

By S. Haberman, M.A., Ph.D., F.I.A., A.S.A., F.S.S., A.F.I.M.A.

ABSTRACT

The paper reviews the mathematical models of transmission of infection that have been put forward for representing the spread of HIV infection and AIDS. It describes and compares the main models that have been proposed and thereby provides some guidance on how such models might be constructed and utilised. There is also discussion of the importance of constructing such mathematical models of transmission of infection which further our understanding of the transmission dynamics of the epidemic and help to identify important epidemiological parameters and their likely influence on the epidemic's course.

KEYWORDS

Mathematical models; AIDS; Prediction

1. INTRODUCTION

The aim of this paper is to provide a review of the mathematical models of transmission of infection that have been proposed for representing the spread of the AIDS epidemic. The literature on this topic is very large and is continuing to expand and so, necessarily, this review is not intended to be exhaustive. The aim is to describe and compare the principal models that have been proposed; however, it is likely that new approaches may be put forward while this paper is being written and printed. Further, the review is not intended to be a comprehensive guide to the mathematical techniques that have been used. The excellent review paper by Isham (1988) would fit these terms of reference more adequately. It is hoped that the paper will be of value to actuaries who would like some guidance on how models might be constructed and, indeed, how the models proposed by the Institute of Actuaries AIDS Working Party fit into the wider umbrella of suggested approaches (Daykin *et al.*, 1988). At this point, it should be mentioned that the author is a member of the Working Party and has benefited considerably from the input of the other members of the Working Party; however, the views expressed here are the sole responsibility of the author (as are the remaining errors).

What is the point of attempting to model the spread of the AIDS epidemic? Firstly, it must be admitted that this mathematical modelling is being carried out even though much of the underlying numerical information needed is not available, and, therefore, the models cannot yet be used to give reliable predictions of the future incidence of AIDS. Nevertheless, such modelling is of considerable epidemiological importance. The most important purpose of

mathematical modelling is to provide a means of gaining understanding about the transmission dynamics of the infection and, in particular, of learning which features are likely to have substantial influence on the course of the epidemic and what sort of effects are to be expected. It is thus possible to investigate how changes in the various assumptions and parameter values of the models would affect the course of the epidemic. This understanding, in turn, helps to clarify what behavioural changes are needed and what intervention strategies should be pursued in order to reduce the spread of the infection, and to enable sensible decisions to be made on the sort of data that should be collected in order that better predictions can be obtained and the consequences of the epidemic can be better managed. Thus, the whole process of modelling and data collection is iterative, but the act of modelling helps to structure thoughts about the spread of infection and provides a framework within which to consider questions of the form: what would happen if A were changed to B? The modelling process then provides a guide to the sorts of data that should be collected to make better information available to society about the epidemic. From an actuarial point of view, such modelling is of vital importance in investigating the implications of the AIDS epidemic for insurance and pensions.

2. GENERAL POINTS ON EPIDEMIOLOGY

The earlier paper by Daykin (1990) provides a full review of the epidemiology of HIV infection and AIDS.

It is worth noting the following definitions relating to the course of AIDS in a particular susceptible individual (Figure 1 provides an illustration of the corresponding stages):

> There is a *latent period* extending from infection until the individual becomes infectious. This is followed by an *infectious period* during which the individual is termed an infective and can pass on the disease to susceptibles.
> The individual infected with HIV is described as being *seropositive* when antibodies are normally detectable in the blood. This tends to occur a few weeks after infection.
> The period from infection until overt symptoms appear is the *incubation period.* For many diseases, the individual will be isolated and unable to pass on the infection once symptoms appear, so that, essentially, infectivity ceases at this time.
> The individual may in due course recover or die (from the infection or from other causes) and, if he recovers, he may be immune to the disease or may return once more to the susceptible state.

As Daykin (1990) notes (Section 6), the incubation period may be very long and it is still too early to say definitively whether all those infected with HIV will ultimately progress to AIDS or whether only a fraction will. Given the likely

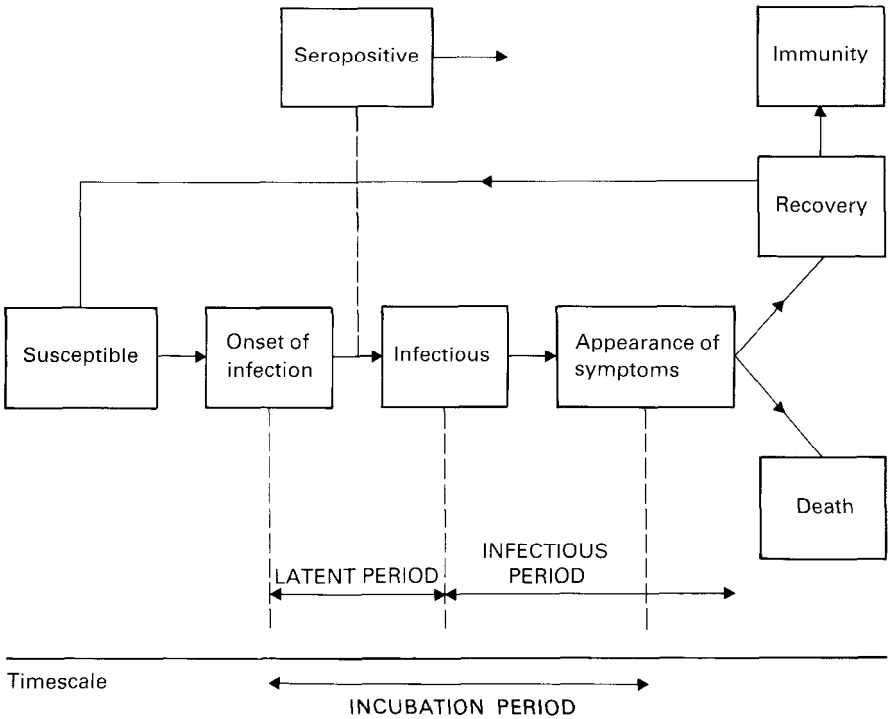

Figure 1. Stages for an infectious disease.

length of the incubation period, it is convenient in modelling the epidemic to assume that all those infected with the virus are seropositive.

For modelling purposes, it is also convenient to make the simplifying assumption that, after AIDS is diagnosed, the individual concerned is effectively isolated and unable to infect further susceptibles. This means that the infectious period is contained within the incubation period. It is thought (Anderson *et al.*, 1986) that the latent period is a matter of days or a few weeks, which again is negligible when compared with the incubation period, but it is not known how long the infectious period lasts and whether or not infectivity is a constant throughout. For any seropositive individuals who do not develop AIDS, there is the further question of whether or not they are infectious continuously after the short latent period (Hyman & Stanley, 1988; Daykin *et al.*, 1989).

At present, recovery from AIDS does not appear possible and so models have tended to avoid including a class of 'recovered-immune' individuals.

The models described in detail in later sections have been progressed to providing numerical results by a variety of means including analytical

approaches, numerical approximation methods and simulation. Further information on mathematical models of the spread of infectious diseases (both in general and for specific cases) can be found in one of the standard textbooks, for example Bailey (1975), Hethcote & Yorke (1984) and Dietz & Schenzle (1985), each of which contains a substantial bibliography.

At this point, it is worth noting the type of data that are being collected in respect of AIDS cases in the United Kingdom. Thus, a surveillance scheme was introduced in 1982 at the Public Health Laboratory's Communicable Disease Surveillance Centre (CDSC) and at the Communicable Diseases (Scotland) Unit (CDSU). An AIDS case reported to the U.K. surveillance scheme will have *six* key dates, five of which are recorded in the scheme where possible:

(1) Date of infection: this will be known for a few patients, for example where infection is due to a single episode of blood transfusion.
(2) Date of seroconversion, production of antibody to HIV. Seroconversion takes place some weeks or months after infection. It is sometimes accompanied by a mild illness, but this illness is seldom, if ever, recorded on a surveillance report form.
(3) Date of onset: the first symptom likely to be HIV related is noted and the date coded. When the onset is insidious or the early symptoms are unusual, then this date may not be accurate, and information on early symptoms is not always recorded.
(4) Date of diagnosis: this is coded as the month in which the patient was recognised to fulfil the current official case definition. It is possible that a patient may have been at that stage for some months before the fact is recognised, but, with an infection which has been present in that patient possibly for many years, some imprecision has to be accepted. Cases by date of diagnosis are the data used internationally to describe the epidemic curve of AIDS.
(5) Date of report: the date at which the surveillance form is received at CDSC and accepted as being an AIDS case may be a few days or months after diagnosis. This is the only date which is known for each patient.
(6) Date of death: if the patient has already died at the time of reporting then this fact is coded together with the date, where known. Doctors are asked to inform CDSC or CDSU if a patient dies after the initial report is made. This is not always practical, especially if the patient has moved. Copies of death entries which mention AIDS, Kaposi's sarcoma or HIV infection as a cause of death are sent to CDSC by the Office of Population Censuses and Surveys (OPCS). This identifies some additional deaths in surveillance patients.

3. DIFFERENT APPROACHES TO MODELLING AND PREDICTION

Three distinct approaches have been put forward in the literature for

modelling and predicting the spread of HIV infection and AIDS. These are as follows:

(1) extrapolation over time by regression or associated methods of the number of reported or diagnosed cases of AIDS so far;
(2) multi-state mathematical models of the transmission of infection and progression of disease; and
(3) back projection methods.

We review each of the approaches in the following sections. Methods (2) and (3) enable properties of the epidemic to be understood as well as assisting with prediction of the future course of the epidemic. Method (1) would be useful only for short-term extrapolation.

The model-based approaches rely upon estimates of the principal parameters being made, which will depend on the case data available. Extrapolation forecasts (method (1)) are not necessarily model-free, as is described in the next section.

This paper does *not* set out to provide a review of the adequacy of the predictions made so far by the various models. Such a review is provided by the General Accounting Office's (1989) report (in respect of the United States of America). However, it should be noted that the likely accuracy of a forecast depends on both the model that is used *and* the quality of data supporting the model. It is possible to have a mathematical transmission model that is comprehensive, in the sense that it takes into account all the relevant parameters for estimating the future course of the HIV/AIDS epidemic, but which relies on poor data, makes unreasonable assumptions and/or fails to allow for known inadequacies in the data. The uncertainty attaching to predictions (or forecasts) stems from statistical fluctuations, but also from 'structural' errors, including uncertainty about the choice of a model, the correct value for a key parameter, or the assumptions regarding the future spread of the epidemic.

A useful summary of some of the models that have been proposed under each of the three headings is provided by Gail & Brookmeyer (1988).

PART I

4. REGRESSION AND ASSOCIATED METHODS FOR SHORT-TERM EXTRAPOLATION

4.1 *Description of the Methods Used*

The approach here is to use the pattern of recent trends in AIDS cases and, with a variety of statistical methods of curve-fitting, project forward the future course of the epidemic.

Short-term projections, based on an extrapolation of the data for reported cases of AIDS in the U.K. have been published by a number of authors, including

McEvoy & Tillett (1985), Tillett & McEvoy (1986) and Healy & Tillett (1988). Similar projections have been made for other countries—for example for the U.S.A. by Curran *et al.* (1985, 1988), Hyman & Stanley (1988), Hellinger (1988) and Fuhrer (1988).

The principal limitation of these empirical 'curve-fitting' approaches is that many different curves are consistent with the historical data and will lead to widely differing predictions over even a modest time range. However, the approaches can be useful for predictions over a short time span, for example the next few years.

McEvoy & Tillett (1985) fit a linear regression equation to $\log N(t)$, where $N(t)$ is the annual number of new patients with AIDS first presenting for medical advice in a calendar year. The model is:

$$\log N(t) = a + bt + e_t$$

where the errors, e_t, are assumed to be independent, identically distributed (IID) normal random variables with zero mean and constant variance (as in a conventional regression analysis). Using observed values of $N(t)$ for 1979–84, estimates of a and b are made and then predictions are made from $\log \hat{N}(t) = \hat{a} + \hat{b}t$ for 1985–88.

Tillett & McEvoy (1986) repeat the analysis, using updated observed values of $N(t)$ for 1979–84. Updating is necessary because of the variability in delays between the date of presentation for medical advice and the date of reporting of cases to CDSC. Thus, the observation for 1984 is revised from 58 to 111 cases. Predictions, $\hat{N}(t)$, for 1987 and 1988 are similarly revised upwards.

The model advocated in these two papers is a poor one in that the errors are unlikely to be symmetrically distributed—a Poisson assumption might be more satisfactory and also the correlation between $N(t)$ and $N(t+1)$ (i.e. between e_t and e_{t+1}) is unlikely to be zero (as the conventional regression approach would assume) and is likely to be an increasing function of t. It should also be noted that both papers use data collected by date of first presentation for medical advice. This date was sometimes closer to the onset of the first HIV-related symptom than to the diagnosis of full AIDS. This date is no longer recorded.

Iverson & Engen (1986) similarly use a least squares approach to the fitting of a linear function to $\log H_t$, where H_t is the number of persons infected with HIV by blood transfusions during a time interval indexed by t. Data from Peterman *et al.* (1985) for the period 1978 to 1983 are used for the estimation of the parameters.

Mortimer (1985) predicts the number of reported U.K. AIDS cases in a simpler way, viz. by dividing the U.S.A. cases by 4 (to allow for the difference in population size) and assuming a time lag of 3 years between the U.S.A. and U.K. epidemics. This appears to be equivalent to assuming exponential growth with a doubling time of about 2 years. The results are in agreement with the earlier results of McEvoy & Tillett (1985)—as noted above, these were subsequently rejected by Tillett & McEvoy (1986).

Curran *et al.* (1985) report on the epidemiology of AIDS and mention some

estimates of the future number of AIDS cases for the U.S.A. The numbers of cases reported per month to CDC, the public AIDS registry at the Centre for Disease Control in Atlanta, are adjusted, on the assumption that the distribution of delays between the actual diagnosis of AIDS and the report of these cases to CDC remains constant over time, to give estimates of the cases actually diagnosed per month.

A polynomial model is then fitted to these adjusted case counts, as transformed by the Box–Cox method, in order that the errors become homoscedastic (Box & Cox, 1964):

$$\frac{N(t)^{\beta}-1}{\beta} = a + bt + ct^2 + e_t$$

where $N(t)$ is the number of cases diagnosed in month t. The data for fitting relate to the period June 1981 to April 1985, and predictions are made up to the end of June 1986.

Morgan & Curran (1986) extend this work. The data up to April 1986 are used for re-fitting the above model after the adjustment for reporting delays and then predictions are made up to the end of 1991. The method of fitting is described as being by weighted least squares.

Both papers use *ad hoc* methods for obtaining confidence intervals: standard results from the theory of non-linear regression are not used. In this sense, the estimates quoted are not 'optimal'.

The predicted values of $N(t)$ for the total U.S.A. population are used then to estimate the predicted values for various sub-populations, defined by risk group, region, sex, race and age group. For each month from January 1983 to April 1986, the logit of the proportion of AIDS cases in each sub-population are calculated, $p(t)$ say, and fitted by the method of weighted linear regression to the following model:

$$\log_e\left(\frac{p(t)}{1-p(t)}\right) = c + dt + e_t.$$

Predicted $\hat{p}(t)$ from this model are then applied to the predicted totals $\hat{N}(t)$ to obtain predictions for each sub-population. These projections are likely to be conservative in that they are based only on cases reported to CDC. Also, patients are identified according to the surveillance criteria then prevailing. Thus, Morgan & Curran (1986) view their projections as being underestimates by about 20%.

WHO Collaborating Centre (1988) provides an outline report of projections of AIDS cases made for a series of western European countries. The methodology of Downs *et al.* (1987) is followed. For each country with at least 50 cases in the region as at the end of 1987, reporting delays are assessed and the reported cases are adjusted to give an estimate of the number of cases diagnosed. Unweighted

linear regression analyses of these adjusted data (after taking logarithms) are carried out over successive (overlapping) 3-year time periods for each country in order to monitor the evolution of the doubling time, t_d, defined by

$$\frac{N(t + t_d)}{N(t)} = 2.$$

Short-term predictions (up to the end of 1989) are then made on the basis of estimated current doubling times for each country and for principal risk groups within each country (where the data permit). The results for the E.C. as a whole show doubling times steadily increasing. Trends within individual countries are less evident, although doubling times in most countries have clearly lengthened since the start of the epidemic. It is not clear from the analyses presented to what extent this trend is continuing.

Whyte *et al.* (1987) assume that the number of cases of AIDS diagnosed among homosexual and bisexual men in Australia at time t has a Poisson distribution with mean

$$\mu(t) = \alpha + \beta \log t.$$

They fit this model using the GLIM package and 14 quarterly data points, from the first quarter in 1983 to the second quarter in 1986 inclusive. The fit is not particularly good! For other risk groups, they use a linear model, also with unsatisfactory results.

Fuhrer (1988) uses a logistic formula fitted by the method of least squares to the observed numbers of AIDS cases in the U.S.A. diagnosed per calendar quarter, t, viz:

$$N(t) = \frac{a}{1 + e^{(b+ct)}} + e_t$$

with t measured from 1982. The parameter estimates are not given in Fuhrer's paper. Fuhrer also provides bounds for the projection: the lower bound is based on a quadratic formula for $\log N(t)$, whereas the upper bound is based on a linear equation with a Box–Cox transformation.

Hellinger (1988) uses reported cases in the U.S.A. (adjusted by a flat inflation factor to allow for reporting delays, under reporting and for illnesses not defined as AIDS until September 1987) to fit polynomial trends. With t representing calendar months measured from 1984, Hellinger provides two bounds for the trend, viz:

LOWER $$N(t) = a + bt + ct^2 + e_t$$

UPPER $$\sqrt{N(t)} = a + bt + e_t$$

where the parameters are again estimated by the method of least squares. The average of these two forecasts is offered as a 'best' estimate of the future trend.

Hyman & Stanley (1988) use the number of AIDS cases diagnosed in the

U.S.A. to fit a cubic polynomial to the cumulative number of cases, for $t \geqslant 1982{\cdot}5$:

$$A(t) = a(t - 1981{\cdot}2)^3 + b + e_t$$

where t is considered as a continuous variable and the rate of new AIDS cases p.a. would be represented by $A'(t)$ or a quadratic polynomial. In deriving this formula, a Box–Cox transformation is used initially, with a non-linear regression approach; however, these results are found to approximate those given by the above simpler approach based on a cubic formula. The authors also find that this polynomial growth pattern is evident in nearly every CDC-defined category, including risk behaviour, age, region and ethnic group.

Healy & Tillett (1988) provide short-term predictions for the epidemic in the U.K., using date of diagnosis as the basis for the forecasts. They investigate the reporting delays (i.e. the difference in months between date of report and date of diagnosis) over the period 1984–86. The distributions over these 3 years are pooled and smoothed and then random samples are used to impute dates of diagnosis to the cases for which the date is partially or wholly unknown. Thus, a series of cases reported in 1982–86 by date of diagnosis is created. Further adjustments are made to allow for the fact that a number of cases diagnosed during 1986 would not have been reported by the end of December 1986. The *ad hoc* procedure used leads to large upwards adjustments being made to the most recent counts.

The procedure is then to fit a simple exponential curve to the monthly numbers of reports up to the end of 1986. This is then extrapolated forwards to give the expected numbers of reports in each month in 1987 and 1988. For each month, the distribution of delays is then used to work backwards to give the expected numbers of diagnosed during 1985 and 1986.

Two regression models are then fitted:

$$\log(N(t) + 1) = a + bt + e_t$$

and

$$\log(N(t)) = a + bt + e_t$$

where, in the first model normal errors are assumed and in the second model Poisson errors are assumed. GLIM is used to carry out the fit. Predictions are made up to the end of 1988.

The authors investigate the inclusion of a quadratic term, ct^2, in both models. The differences in the fitted values are small over the data period; however, the differences become large when the model is projected forward over 2 years. In both models the estimate $\hat{c}$ is negative, which confirms the results of Downs *et al.* (1987) and the WHO Collaborating Centre (1988) who found parabolic fits which correspond to increasing doubling times.

Healy & Tillett also experiment with adjustments so that the early data, even with the weighting implied by the two models, do not have a high degree of influence in forecasting the future. Accordingly, they repeat the above linear fits,

but impose weights that are geometrically decreasing into the past (with common ratio 0·8) so as to give greatest importance to the data from the recent past. (This approach is in the style of an exponentially weighted moving average model.)

This paper is accompanied by a discussion. Here it is mentioned that the set of geometrically decreasing weights, mentioned above, places low weight even on the recent past and so the results can be questioned. Transformations of the Box–Cox type (as used by Curran *et al.* (1985) and by Morgan & Curran (1986)) are advocated. Further, Downs compares and contrasts his methods for adjusting for reporting delays with those of Healy & Tillett (1988). Both methods use period of diagnosis as an independent variable and adjust the data for reporting delays. Downs' method of data adjustment, however, is conceptually different. Rather than using the observed distribution of reporting delay times among cases already reported (requiring predictions of the number of cases to be reported in forthcoming months to correct for reporting delays), he attempts to estimate the distribution of delay times among the (unknown) number of cases actually diagnosed. In this approach, the adjusted case numbers are estimated along with the reporting probabilities.

In more recent work, published as an Appendix to the Cox report, Healy (1988) extends these analyses. The same methodology is applied, except that the series of numbers of diagnosed cases by month are derived in a more sophisticated manner. The monthly numbers of diagnosed cases from 1984–87 are fitted by a log-linear model with Poisson errors using GLIM:

$$\log_e(N(t)) = a + bt + c\sin\frac{2\pi t}{12} + d\cos\frac{2\pi t}{12} + f\sin\frac{2\pi t}{6} + g\cos\frac{2\pi t}{6} + e_t$$

which incorporates seasonal terms. The series is projected forwards for 30 months and the monthly predictions are then combined with the estimated delay distribution (derived as before) to give the numbers of diagnoses for 1986–87 expected to be reported during 1988–90.

Then, log-linear models with Poisson errors of the form:

$$\log_e(N(t)) = a + bt + e_t$$

are fitted to successive 24-month slices of the data starting at January 1984. After an initial decrease, the slopes, $\hat{b}$, are roughly constant. Forecasts are made based on this model using the average of the estimates of $\hat{a}$ and $\hat{b}$. The improvement in fit arising from the inclusion of a quadratic term is investigated—the statistical significance of the quadratic coefficient is not high ($p = 0{\cdot}09$).

Zeger *et al.* (1989) provide an interesting set of statistical techniques for monitoring the AIDS epidemic and for providing short-term predictions. As with earlier works reviewed in this section, they propose using log-linear models to estimate the number of AIDS cases and trends in incidence while correcting for reporting delays. The cases for a given sub-population are arranged in a two-way table by date of diagnosis and length of reporting delay—as for a claims run-off triangle in general insurance. The marginal distribution by date of diagnosis is of

interest and parametric models for estimating this distribution are considered. The data set used is the publicly accessible AIDS registry obtained from CDC in Atlanta as at December 1987.

Let N_{tu} be the number of cases diagnosed in quarter t and reported u quarters later. N_{tu} is assumed to be a Poisson random variable with:

$$E(N_{tu}) = \mu_{tu}$$

where

$$\mu_{tu} = s(t;\beta) + d(u;\theta). \tag{4.1}$$

Here $s(t;\beta)$ is a function of t with unknown parameters, β, which characterises the trend in AIDS incidence. The authors use a cubic spline function with 2 knots for $s(t;\beta)$. $d(u;\theta)$ is a delay function with unknown parameters, θ. The authors use a non-parametric step model for the delay function:

$$d(u;\theta) = \theta_u \qquad \text{for } u = 0, 1, \ldots, 12.$$

The additive form of the link function (4.1) assumes that the reporting delay distribution has not changed over the period under discussion (1982–87): this assumption can be checked by an analysis of residuals.

The model is fitted using GLIM and the predicted values for the missing lower triangle are obtained for each subpopulation.

Trends in incidence can be monitored by, for example, considering:

$$\left(\frac{\hat{\mu}_{t_0+t}}{\hat{\mu}_{t_0}}\right)$$

where

$$\hat{\mu}_t = \sum_{u=0}^{12} \hat{\mu}_{tu}.$$

The variability in such an index derives from the missing data and from the assumed randomness in the observed Ns. The index is likely to be smoother than one based on $\hat{N}$s rather than $\hat{\mu}$s because the log-linear model, represented by equation (4.1), has been used both to fill in the missing triangle and to estimate a parametric and smooth time trend which is assumed to underlie the observations.

The authors consider trends in incidence for the latest year for different subpopulations. In particular they consider:

$$\hat{c}_{ij} = \log\left(\frac{\hat{\mu}_{12}}{\hat{\mu}_8}\right)$$

for risk group i in region j. A separate two-way model is examined for these 'growth rates':

$$\hat{c}_{ij} = \mu + g_i + r_j + e_{ij}.$$

In subgroups with few cases, such trend estimates will be highly imprecise. The

authors propose an Empirical Bayes approach—this idea underpins modern Credibility Theory and has been suggested for the actuarial analogue of claim run-off triangles by Verrall (1988).

The aim is to borrow strength, from the data available for other subgroups, to improve the precision of the trend estimate for the particular subgroup under focus. Of course, the AIDS epidemic has different features among different risk groups and regions, owing to different modes of transmission, levels of public awareness and so on. It would be desirable to borrow strength only from similar subgroups; however, by introducing limited bias through relying on groups with possibly different growth rates, it may be possible to achieve a dramatic improvement in the precision of estimates of the rates for risk groups and/or regions with very sparse data. The technique is thus employed to produce an Empirical Bayes estimate of the trend for risk group i in region j: $\tilde{c}_{ij}$.

Stroinski (1990) also considers the two-way table by date of diagnosis and length of reporting delay, but in a more restrictive way. He proposes a series of ANOVA type models, similar to Kremer's (1982) analysis of the claims run-off triangle:

$$\log(N_{tu}) = \alpha_t + \beta_u + e_{tu}$$

which are used to obtain predicted values for the missing lower triangle for some U.S. populations. Such an approach (like the Chain Ladder method) relies on assumptions about the constancy of the delay distribution, represented by β_u, across dates of diagnosis.

A number of the models proposed for short-term extrapolation of the incidence of AIDS cases fall into the category of log-linear models. These log-linear models for the expected AIDS incidence over calendar time lead to exponential growth. This is unrealistic, mainly because the underlying epidemic of aggregated infections is probably not growing exponentially (see Part II), but also because the presence of extended incubation period may lead to subexponential growth of AIDS incidence, even if the underlying infections are growing exponentially (Gonzalez & Koch, 1987): we shall return to this point in a later section.

Cox & Medley (1989) describe a rather formal method of prediction of numbers of diagnoses on the basis of data with reporting delays. (A summary of this paper appeared earlier as Appendix 7 to the Cox Report—Department of Health (1988).) It is assumed that diagnoses occur in a Poisson process of rate $\lambda(t;\rho)$, where ρ are unknown parameters, and that reporting delays are independent random variables with constant probability density $f_x(x;\theta)$, where θ are unknown parameters. Then, if in a time period ending at t_0, diagnosis-delay pairs (t_i,x_i) for $i = 1, 2, \ldots n$, are observed, the log likelihood is:

$$\Sigma \log \lambda(t_i;\rho) + \Sigma \log f_x(x_i;\theta) - \int_{-\infty}^{t_0} \lambda(\omega;\rho)\, F_x(t_0-\omega;\theta)d\omega \qquad (4.2)$$

where F is the cumulative distribution function associated with f. In practice,

there are some complications associated with the discreteness of the data, for example being recorded in calendar month intervals.

For the delay distribution, f_x, a mixture of two gamma distributions of index 1 is found convenient. (It is necessary to restrict the upper tail of the distribution to avoid indeterminancy in λ.) For the incidence curve, λ, three forms are used:

(a) $\lambda(t;\rho) = \rho_0 \exp(\rho_1 t)$ representing the simplest case, namely of exponential growth;

(b) $\lambda(t;\rho) = \rho_0 \exp(\rho_1 t - \rho_2 t^2)$, very convenient for representing small perturbations from exponential growth; and

(c) $\lambda(t;\rho) = (\rho_0 + \rho_1 t)(1 + \rho_2 \exp(-\rho_3 t))^{-1}$, representing an initial exponential phase converting ultimately into linear growth.

Full numerical results are not given here. The authors place most emphasis on (c), with the above suggested form for f_x, viz.:

$$\theta_0 \theta_1 (\theta_1 x) e^{-\theta_1 x} + (1-\theta_0)\theta_2 (\theta_2 x) e^{-\theta_2 x}.$$

7 parameters are being fitted and the maximum likelihood estimates are:

$\theta_0 = 0{\cdot}5654$, $\theta_1 = 11{\cdot}35$, $\theta_2 = 1{\cdot}496$
$\rho_0 = 71{\cdot}43$, $\rho_1 = 162{\cdot}0$, $\rho_2 = 13{\cdot}49$ $\rho_3 = 0{\cdot}8233$.

Approximate values for the confidence limits for the prediction for, say, 1992, underestimating the variability involved, are obtained by fixing $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$ and then computing a profile likelihood function for the 1992 value essentially by computing the maximised likelihood for a large grid of values in the ρ-space, and filling in the maximised curve by eye. The θ-values are fixed to cut down on the substantial computer time involved.

The differences between the forecasts based on (a), (b) and (c) are significant even after only one year. The logistic curve ((c) with $\rho_1 = 0$) moves relatively quickly to an asymptote after departing from simple exponential growth (a); this is probably unrealistic in view of the heterogeneity in progression from infection to disease in individual patients. The quadratic exponential (b) is valuable for representing small perturbations from exponential growth, but it results in an early peak and then numbers that decline symmetrically—these features are probably unrealistic in the current context in the U.K. The linear-logistic model ((c) with $\rho_1 \neq 0$) seems the most plausible choice at present. Cox & Medley conclude that the choice of incidence function represents a major source of variability and that present predictions are, at best, reasonably accurate for homosexual men and for perhaps 3 years ahead; current information on other risk groups is considered too scanty for separate prediction to be sensible.

Cox & Davison (1989) present a method of prediction for small subgroups of the population. Prediction limits are calculated for the number of events likely to occur in a specified period in an exponentially growing epidemic. The basis for the prediction is the total number of events observed in the past and a binomial probability model for the number of events occurring in the past and the future.

4.2 *General Comments*

In considering these forecasts based on extrapolations, we need to note three general sources of uncertainty that are present:

(1) The forecasts are of expected values; there will be variability about these means, which will be roughly Poisson in nature.
(2) The parameters of the forecasting curves are subject to sampling error.
(3) The forecasts make assumptions about the underlying epidemic curve and about the pattern of variability about it; both of these are subject to uncertainty.

For projecting some time into the future, with fairly large numbers, the third type of error is likely to dominate. This would be the case, for example, in forecasting the number of AIDS cases in the U.K. for 3–4 years ahead. This arises because, essentially, many different shapes of curve are consistent with the data on which the forecast is based. Thus, Healy & Tillett's (1988) attempt to introduce a quadratic term into the exponential growth model for $N(t)$ indicates how quite minor changes in the model assumptions can lead to very large differences in forecast, even over short periods of time. For shorter-term forecasts, the third error becomes less important and, ultimately, when forecasting rather small numbers a short time ahead, Poisson-type errors will be the major source of uncertainty. This is particularly relevant if one wishes to forecast events within a small geographical area for a fairly short time ahead.

Apart from these methodological uncertainties, there are others connected with the data themselves. A major concern is the completeness of the surveillance data that are used for the fitting of models.

Uncertainty also comes from the use of date of diagnosis as the time origin for forecasting. The adjustments, for example used by Healy & Tillett (1988) and Healy (1988), can be large and must be influential in assessing the recent course of the epidemic and the forecasts made.

A comparison of different forecasting models would, however, be useful in assessing how sensitive the short-term forecasts are to changes in the basic model. Also, it would be useful to incorporate collateral information for other countries or regions in making forecasts—in particular, for small countries or for subgroups of countries like the U.K. Such an approach would be hindered by difficulties stemming from differences of case-finding methods and recording conventions, but, nevertheless, may be worthwhile.

Finally, it should be noted that these forecasting methods based on short-term extrapolation implicitly assume that the epidemiological features of the epidemic remain constant and take no account of any changes in the epidemic's spread. For example, they assume that the past trends of reported cases will continue over the short-term future in a similar pattern—however (within a risk group), the ratio of infected to susceptible subjects will clearly change as the epidemic spreads and control measures may affect the rate of spread of the infection between and within risk groups. Further, these forecasting methods do not allow

for differences between the pattern of growth in different at-risk groups, changes in the surveillance definition of AIDS, temporal changes in the distribution of the reporting delay, changes in the level of under-reporting of cases of the disease. Only models that are founded on the transmission mechanisms of HIV can show how the early infection of high-risk groups, behavioural changes, and future medical advances such as treatments and vaccines will affect the future course of this epidemic. The effects will be highly nonlinear functions of the parameter values and, at times, may even lead to changes that are counter to both intuition and simple extrapolated predictions. Forecasts of these counterintuitive mechanisms, using a mathematical model, may greatly improve our understanding of the observations.

A by-product of the development of mathematical models would be the creation (as noted in Section 1) of a logical structure that organises existing information on AIDS into a coherent framework and suggests new information that should be collected about a wide variety of topics, such as drug use, sexual activity, and the interactions between HIV and the immune system. Models can provide qualitative insights, even when data are lacking, and can help to focus priorities in terms of the data to be collected.

Short-term forecasting methods are likely to be of practical use in many developing countries, but, increasingly in developed countries, it will be necessary to stratify past trends carefully by risk group for the purposes of extrapolation. For developed countries, the reliability of short-term projections is likely to decrease as the epidemic slows in particular risk groups. It is then of importance to use models based on the transmission dynamics of the disease, to allow for the increasingly non-linear patterns that are emerging. The real observed epidemic is formed from a complex network of separate, identifiable, but interlinked subepidemics within the different at-risk groups and classes within a specific group, defined by activity/behaviour/geography. Only transmission models can represent this network of relationships and the spread of the infection from high-risk to lower-risk groups. This brings us to Part II of this paper.

PART II

5. MATHEMATICAL MODELS OF EPIDEMIC TRANSMISSION: INTRODUCTION

5.1 *General Comments*

In this section we consider the transmission models that have been proposed for representing the spread of HIV infection and AIDS.

As has been described by Daykin (1990) in his epidemiological review, there are a number of modes of transmission of the infection, each of which would be differentiated in a transmission model, viz:

(a) by sexual contact between male homosexuals and bisexuals, one infected and one not;
(b) by injecting drug use (IDU) involving the sharing of needles or syringes contaminated with the blood of an infected person;
(c) by receiving blood or blood products contaminated with HIV, this having been a particular hazard to haemophiliacs. This route of transmission has now been virtually eliminated in the U.K. as a result of measures taken to treat and/or screen blood and blood products;
(d) by sexual contact between infected and susceptible heterosexuals; and
(e) by vertical transmission from seropositive pregnant women to their babies during pregnancy or around the time of birth. In the U.K. the great majority of women transmitting the infection in this way have been IDUs.

The models proposed tend to focus on homosexual males, considered as a risk group on their own, and aim to represent the progression of the disease mathematically by systems of equations in which the rates of transition between the various 'states' are specified quantitatively. This broad idea has a long history, but the application to AIDS brings in many new features. The rate of occurrence of new infected individuals is determined by the numbers of susceptible and infected individuals, the pattern of interaction between these two groups and the magnitude and distribution in time of infectivity of an already infected individual. Any such mathematical representation is, of course, idealised: the more realistic the models, the more quantities, e.g. transition rates, have to be specified numerically. As has been stressed, there is a paucity of information on many of these points. Note, for example, that if relevant aspects of sexual behaviour are changing over time, this should be introduced into the model.

The approach of this section of the review paper is to introduce the simple deterministic epidemic model, consider the choice between a deterministic and a stochastic approach to modelling and then consider in a broad fashion the models that have been proposed by various workers in the field. It is not intended to provide complete mathematical descriptions of each model. However, comparisons of approach will be made.

5.2 *A Simple Deterministic Epidemic Model*

A basic deterministic model is as follows. Suppose that at time t, a fixed population, of size n, can be separated into a group of $x(t)$ susceptibles and $y(t)$ infectives, where $x(t) + y(t) = n$, and where both $x(t)$ and $y(t)$ are sufficiently large that they can be regarded as continuous variables. Assume the population mixes homogeneously, so that in any small time interval $(t,t + h)$ the number of contacts between a susceptible and an infective is proportional to both $x(t)$ and $y(t)$ (and h) and that a fixed proportion of these contacts results in the susceptible becoming infected. Then, the number of new cases of infection in the time interval is $\alpha(t)\, x(t)\, y(t)\, h$ for some constant of proportionality $\alpha(t)$, so that $x(t)$ satisfies the differential equation:

$$\frac{dx(t)}{dt} = -\frac{dy(t)}{dt} = -\alpha(t)\, x(t)\, y(t) = -\alpha(t)\, x(t)\, (n-x(t)). \tag{5.1}$$

Note that the rate of increase of $y(t)$ is equal to the rate of decrease of $x(t)$. We may regard $n\alpha(t)$ as a rate of contact between the two groups. Then, an intuitive explanation of the above equation is that each of the $x(t)$ members of the not infected population are in contact with $n\alpha(t)$ other people in unit time, chosen at random, and there is a probability $y/(x+y)$ that each of them is infected. Alternatively, each of the $y(t)$ members of the population is in contact with $n\alpha(t)$ other members chosen at random, and there is a probability that $x/(x+y)$ of these are not infected, but become infected by the contact.

If $\alpha(t)$ has a constant value, α, then the solution of (5.1) is:

$$x(t) = n\, x(0)[x(0) + (n-x(0)) \exp(n\alpha t)]^{-1}. \tag{5.2}$$

It then follows that, if we assume that initially the number of infectives is relatively small so that $x(0) \simeq n$, then for small t:

$$y(t) \simeq y(0)e^{n\alpha t}. \tag{5.3}$$

Thus, the prevalence of the infection increases exponentially in the early stages of the epidemic. As before, we define *the doubling time*, t_d, of the epidemic to be the time taken for the number of infectives to double. We then find that in the early stages of the infection:

$$e^{n\alpha t_d} \simeq 2,$$

or

$$t_d \simeq (n\alpha)^{-1} \log_e 2. \tag{5.4}$$

Since individuals will not usually remain infectious indefinitely, the model can be made more realistic by assuming that individuals leave the infective class at a rate υ to play no further part in the epidemic. It is irrelevant, at this stage, whether they are recovered but immune, have been isolated or withdrawn from the population, or are dead. Then equation (5.1) is replaced by:

$$dx(t)/dt = -\alpha x(t) y(t),\ dy(t)/dt = \alpha x(t) y(t) - \upsilon y(t). \tag{5.5}$$

The solution of (5.5) is much more difficult than that of (5.1). An approximate solution was obtained by Kermack & McKendrick (1927) and a full discussion is given in Bailey (1975). However, suppose, as before, that we assume that $x(0) \simeq n$. Then, for small t, it follows from (5.5) that:

$$y(t) \simeq y(0)e^{(n\alpha - \upsilon)t} \tag{5.6}$$

where again the prevalence increases exponentially, as long as $n\alpha > \upsilon$. In this case the doubling time in the initial stages of the epidemic will satisfy:

$$t_d \simeq (n\alpha - \upsilon)^{-1} \log_e 2. \tag{5.7}$$

A few authors have used the above simple deterministic model to represent the

AIDS epidemic. Thus, Bailey (1988) applies equation (5.1) to data on the prevalence of HIV in the San Francisco City Clinic Cohort of homosexual and bisexual men (enrolled between 1978 and 1980 for a series of studies connected with hepatitis B).

The equation is modified to refer to the prevalence of HIV antibody $p(t) = y(t)/n$ and to allow for a fraction, k, of the population having a relatively safe lifestyle:

$$\frac{dp(t)}{dt} = \alpha p\left(1 - \frac{p}{1-k}\right)$$

where α is taken to be a constant. As before, this can be solved to give a logistic type formula, which Bailey has fitted to the empirical data.

Chin & Mann (1989) also provide numerical projections based on simple mathematical models such as the above.

Some general comments about the interpretation of the doubling time statistic would be appropriate here, since this is a widely quoted figure used to indicate the rate of spread of the epidemic.

As de Gruttola & Lagakos (1989) report, the doubling time of the AIDS epidemic in the U.S.A. increased from 5 to 13 months between 1982 and 1987 (CDC, 1986). What can be inferred from this increase? Does it mean that the epidemic may have begun to 'run its course', or that behavioural changes have had a major impact in reducing the incidence rate? More generally, how is the unobserved epidemic of HIV infection reflected in the observed epidemic of AIDS?

De Gruttola & Lagakos (1989) consider the value of AIDS incidence data in estimating and interpreting the extent of HIV infection. Apart from modifications of behaviour, changes in the cumulative incidence of AIDS are influenced by three phenomena:

(1) as the prevalence of HIV infection among individuals at highest risk increases, the rate of growth in incidence of infection in that population decreases;
(2) the populations at risk for AIDS are highly heterogeneous: some, such as homosexually-active men practising high-risk behaviour with many partners can be almost entirely infected, while others have lower prevalences and rates of spread of HIV infection; and
(3) the incubation time between infection with HIV and the onset of AIDS can last many years.

The first two of these phenomena are characteristic of many epidemics resulting from the introduction of a new infectious agent into a population, and tend to cause the doubling time for the cumulative infection rate to increase over chronologic time. The third phenomenon, which is not characteristic of most infectious diseases, also can increase the doubling time. Thus, the observed

increase in doubling time for AIDS is influenced by several factors that are unrelated to behavioural changes. De Gruttola & Lagakos show, from examples, that, currently, it is not possible to determine the degree to which behavioural changes may have contributed to the increase in doubling time, and that changes in doubling time are not particularly informative about the future spread of the epidemic to lower-risk populations such as heterosexually-active individuals.

5.3 *Stochastic Models: Background Comments*

In using a deterministic model rather than a stochastic one, it is assumed that the numbers of persons at risk and infected are sufficiently large that they can be approximated by continuous variables and that the spread of any infection, starting from specified initial values, will always take exactly the same course. We obtain only an approximation to the average behaviour of the underlying stochastic model and, in some situations, the variation between realisations of the epidemic could be such that knowledge of the behaviour of the average is not particularly helpful. This would be true for small subgroups of the population or at the very beginning of the epidemic.

Thus, HIV infections can persist (apparently dormant) in a few isolated individuals with low sexual activity for extended periods—unlike many other infectious diseases. This particular feature can cause sporadic local epidemics, whenever the infected individual passes the virus to a highly sexually active person. In such situations, the virus could spread rapidly and widely without warning, infecting many people. These sporadic outbursts should be represented by a stochastic rather than a deterministic model, which would smooth over the sporadic effects of such local random events.

The justification for the deterministic approach (and it is worth noting that a deterministic approach has been used by most investigators in this field) is threefold:

(1) solutions are more difficult to find for stochastic than for deterministic models;
(2) given the large populations involved in the AIDS epidemic and, once the epidemic is established, the large numbers infected, the deterministic models should give results that are approximately valid, especially when modelling is at its present embryonic stage; and
(3) the considerable uncertainty attaching to estimates of the important parameters means that further sophistication may not be warranted.

Comment (2) has to be tempered with a qualification. Populations under consideration can be large, but what is important is the number of people with whom an individual interacts: this number is probably small as far as HIV and AIDS are concerned. This points to an advantage of the stochastic approach over the deterministic one. Further, for models involving a stratified population (see later sections), it would be necessary for *all* the subpopulations to be large. This may be difficult to achieve in practice, especially for models that involve a

continuous stratification. Ball points out in his contribution to the discussion of Isham's (1988) review paper, that a deterministic approach (to the equations of the next section) will underestimate the doubling time and will *overestimate* the spread of the infection, relative to a corresponding stochastic approach.

A further problem in relation to (2) is knowing exactly *which* stochastic process the deterministic approach is approximating. As Mollison points out in the discussion of Isham's (1988) paper, it is not at all clear what assumptions have to be made in connection with a particular deterministic model. Isham comments that, for any completely specified stochastic model with state $(X(t), Y(t), \ldots)$ at time t, the set of conditional expected increments of the form $E\{dX(t)|X(t) = x(t), Y(t) = y(t), \ldots\}$ can be written down. The deterministic model is then obtained by equating each such expectation to the corresponding increment $dx(t)$. While each stochastic model determines a unique deterministic model, of course the converse does not apply and there will be a whole family of stochastic models corresponding to the same deterministic model. Thus, there can be problems, for example, in interpreting the parameters of a deterministic model. In particular, these would not have to be interpreted as corresponding to the specific stochastic model originally described, to motivate the deterministic model! In view of these difficulties, Isham comments that it would be important to study the applicability of solutions to deterministic models in stochastic solutions.

A few stochastic studies have been reported in the literature. Barrett (1988) has developed a stochastic simulation model of the heterosexual spread of HIV, focusing on the beginning of the epidemic in a small population. The model does not incorporate mortality or the removal of infected persons (with AIDS). The results indicate, primarily, the wide range and variability in the number of persons infected directly or indirectly by one infector. The simulations indicate a small degree of selection for risk of infection among infected women. The implication is that 'lifestyle underwriting' may not be able to eliminate a high percentage of heterosexual persons in the 'at risk' category, since the characteristics of infected persons considerably overlap, as found here, those of uninfected ones. Further, discrimination between characteristics of seronegative and seropositive persons may not be possible in studies of less than a few dozen seropositive persons.

This may explain the absence of a significant positive correlation between seropositive status and frequency or number of sexual contacts with infected spouses, as noted by Daykin (1990) in his review paper.

Barrett's model represents variations between people in risk of infection per partner month for both new and old partners. It incorporates variability in the length of partnership and in the rate of partner change. The results show whether infection tends to spread first to those with the highest risks of infection (per infected partner) or to those with the most partners. Correlations between the distribution of risks of infection and of numbers of partners or of partner changes have not been included, because of the lack of empirical data to provide any

guidance. It is possible that the monthly risk of infection could be negatively correlated with the number of partners or with the rate of partner change (because promiscuous people may have fewer acts of intercourse per partner month) or be positively correlated (because of the presence of venereal disease, for example).

Mode *et al.* (1988) formulate a stochastic model of an AIDS epidemic in a population of male homosexuals. Computer-intensive (rather than analytical) methods are used to investigate some properties of the model. Three factors of importance in the evolution of the epidemic are studied in a numerical factorial experiment, viz: distribution of the incubation period; probability of infection with HIV per sexual contact with an infected individual; and the distribution of the number of contacts per sexual partner per month. The numerical results suggest that the distribution of the incubation period will have a decisive impact on the evolution of the AIDS epidemic, but that this impact depends critically on the levels of the other two factors. A Monte Carlo experiment suggests that, if forecasts of an epidemic were made solely on the basis of the deterministic non-linear difference (or differential) equations embedded in the stochastic process, then predictions of the number of individuals infected with HIV and AIDS cases may be too pessimistic, reinforcing the earlier point attributed to Ball (discussion of Isham, 1988).

Tan & Hsu (1989) go further in the development of stochastic models. They propose a model for the spread of AIDS in a homosexual population. The probability generating function of the numbers of susceptible persons (seropositive latent, but not infective), infective persons and AIDS cases is derived. It is then shown that the expected numbers, variances and covariances of these persons satisfy a set of ordinary differential equations, which are then solved numerically to assess the effects of various factors on the spread of AIDS (e.g. the initial number of infective persons, the rate of sexual contact between susceptible and infective persons, the mortality rate from AIDS). They show that, if the number of susceptible persons is large, then, as expected, the deterministic approach is equivalent to working with the expected numbers in the stochastic model. The stochastic model is able to indicate how various factors affect the variances and covariances of infective persons and AIDS cases—results which are obviously not attainable using a deterministic approach. In particular, the relative size of the variance terms in certain of the simulations suggests that a deterministic approach would *not* be adequate.

5.4 *Deterministic Models with Aggregate Representation of Infections*

Certain modellers have taken a half-way house in their approach to the mathematical representation of the spread of AIDS, half-way, that is, between the extrapolation approach of Section 4 and the detailed transmission models to be presented in Section 6.

These models simulate epidemiological processes at the society-wide or 'macro' level—without addressing the individual-level events that combine to

determine the epidemic trend. Two steps are typically involved. First, distributions are used to model or project the spread of HIV infection during past, present, and future years. Second, having projected HIV infections (including future infection), the models then proceed to estimate the numbers of AIDS cases that will result from these infections, as of a specific future date. This approach will be seen to resemble closely the back calculation method (to be presented in Part III of this review). Like the back calculation method, the second step of these models consists of combining the estimated number of HIV infections with information on the time-to-AIDS distribution in order to predict AIDS cases for selected future years.

Unlike back-calculation models, however, these macro-level models typically predict *future* HIV infections, and they include in their forecasts AIDS cases that are expected to result from these new infections. Also, unlike back-calculation models, macro-level models can use distributions that have built-in alternative scenarios for the future course of the epidemic. For example, in Artzrouni & Wykoff's (1988) model, HIV-infected persons decrease or entirely cease risky behaviours when they contract AIDS.

While risk-transmission groups have been examined separately in some macro-level models, other demographic groups (such as geographically defined subpopulations) have not been examined.

The principal 'macro' models that have appeared have actuarial origins. The first such model was proposed by Cowell & Hoskins (1987). Cowell & Hoskins focus on modelling the progression from HIV infection to onset of AIDS and ultimate death, using a discrete-term Markov chain based approach. Individuals progress through successive states of morbidity (according to the Walter Reed Staging Method):

1A seronegative and at risk,
1B seropositive and asymptomatic,
2A with HIV infection and lymphadenopathy syndrome,
2B with HIV infection and AIDS-related complex, and
3 with AIDS.

There is no skipping of stages allowed and no reverse transitions.

Transition probabilities are assumed to depend only on the stage and the current duration in that stage (hence, they are assumed to be independent of age, sex, duration since seroconversion, calendar year) and are estimated by Cowell & Hoskins in an approximate (and non-optimal) way from published data from the University of Frankfurt Centre for Internal Medicine study of homosexual males at risk of HIV infection (Brodt *et al.*, 1986). Mathematically, Cowell & Hoskins consider $I(t)$, the number of infected individuals without AIDS at time t, split into three sections for states 1B, 2A and 2B (say, $I_j(t)$, $j = 1, 2, 3$).

If $S(t)$ is the number of susceptible individuals at time t and $A(t)$ the number of AIDS cases at time t, then Cowell & Hoskins effectively consider the discrete time versions of the following set of differential equations:

$$\left.\begin{aligned}
\frac{d}{dt}S(t) &= -\lambda S(t) \\
\frac{d}{dt}I_1(t) &= \lambda S(t) - \gamma_1 I_1(t) \\
\frac{d}{dt}I_2(t) &= \gamma_1 I_1(t) - \gamma_2 I_2(t) \\
\frac{d}{dt}I_3(t) &= \gamma_2 I_2(t) - \gamma_3 I_3(t) \\
\frac{d}{dt}A(t) &= \gamma_3 I_3(t) - \delta A(t).
\end{aligned}\right\} \tag{5.9}$$

The model has been represented in this form to facilitate comparison with the transmission models described in Section 6. Here, the γ_i are transition intensities (analogous to the force of mortality) and δ is the force of mortality from AIDS. The flow of new cases into state 1B is represented by $\lambda S(t)$ and here Cowell & Hoskins take λ to be a parameter to be specified, whereas, as we will see later, the transmission epidemic models would allow λ to be a function of time t *and* to depend on the relationship between the numbers of persons infected at time t and the total population at time t. This latter point is a particularly significant deficiency which, as will be seen, is allowed for in the models of Section 6.

The Canadian Institute of Actuaries set up a Task Force on AIDS. Three subcommittees were formed in 1988. The Subcommittee on Modelling published two reports in November 1988, dealing with Canada and the U.S.A., respectively. Their approach to modelling the HIV epidemic in Canada was to focus on the male homosexual population without identifying particular 'at risk' groups. As with Cowell & Hoskins (1987), an aggregate approach to representing the spread of infections is used. The cumulative number of persons infected with HIV up to time t is assumed to follow a stochastic process with mean:

$$\Lambda(t) = k(t-t_0)^{\beta} \qquad \text{for} \qquad t > t_0 \tag{5.10}$$

where t_0 represents the origin. k, t_0 and β are estimated by maximum likelihood methods from the published population case data up to 1987. For Canada, $\hat{\beta} = 2{\cdot}5$ (to be compared with the corresponding work of Hyman & Stanley (1988), who found $\hat{\beta} = 3{\cdot}0$ for the U.S.A., when modelling the cumulative number of AIDS cases up to time t).

The incubation period, from infection to development of full AIDS, is represented by a Weibull model with probability density and cumulative distribution functions:

$$\left.\begin{aligned}
g(t) &= ab(bt)^{a-1}\exp[-(bt)^a] \\
G(t) &= 1-\exp[-(bt)^a]
\end{aligned}\right\} \tag{5.11}$$

respectively. From a review of the literature, a was taken to be 2·5 and b was chosen so that the median of the distribution was 10 years, i.e. $b = 0{\cdot}0864$.

The time from development of full AIDS to death is represented by an exponential distribution with parameter μ, i.e. with probability density and cumulative distribution functions:

$$h(t) = \mu e^{-\mu t}$$

and

$$H(t) = 1 - e^{-\mu t}$$

respectively. From mortality data for Canada, μ was taken to be 1/1·40.

Then it can be shown that the AIDS cases form a stochastic process with mean $\Gamma(t)$ where:

$$\Gamma(t) = \int_{t_0}^{t} \lambda(s)\, G(t-s)ds$$

and

$$\lambda(s) = \Lambda'(s)$$

and the AIDS deaths form a stochastic process with mean $D(t)$ where:

$$D(t) = \int_{t_0}^{t} \gamma(s)\, H(t-s)ds$$

and

$$\gamma(s) = \Gamma'(s)$$

for $t > t_0$.

Alternatively, these formula may be written:

$$\gamma(t) = \frac{d}{dt}\Gamma(t) = \int_{t_0}^{t} \lambda(s)\, g(t-s)ds$$

$$d(t) = \frac{d}{dt}D(t) = \int_{t_0}^{t} \gamma(s)\, h(t-s)ds.$$

Up to this point, the model has been used for all ages combined. In order to obtain age specific results, the deaths in each year are distributed in accordance with the currently (at 1988) experienced age distribution of deaths in the Canadian populations (which are graduated using a lognormal distribution). In studying the future course of mortality, the Sub-Committee on Modelling follows an approximate approach (for reasons of expediency) which has serious shortcomings.

The key assumption is that new infections are assumed to stabilise at the level estimated to have been effective in the first half of 1984. In other words, behavioural change is assumed to have been so marked from the beginning of

1984 as to have resulted in a flattening off of new infections at that point. This level is then sustained indefinitely, initially by new infections from the group originally at risk, but subsequently by infections from new entrants to the at-risk group.

There are a number of problems with this approach. First and foremost, there is no rationale for assuming a levelling off of new infections; a more likely pattern would be a rise to a peak, followed by a long decline, even with assumptions involving an early change of sexual behaviour.

Furthermore, although there is some evidence of behavioural change having begun as early as 1984, it would have had to have been very dramatic at about that time in order to achieve a sudden levelling out of the number of new infections. If it were as dramatic as that, then the number of new infections could be expected to fall in future years, rather than to remain at that level. Thus, implicitly, continuing new entrants to the risk group are assumed (at 1·7% of the male population reaching age 20), all of whom are assumed to develop HIV and die of AIDS.

The approach of Artzrouni & Wykoff (1988) is similar to the above. They assume that the new-infection rate is a decreasing function of the cumulative number of HIV infections; using the national surveillance data, together with a model based on this assumption, results in an estimated epidemic in which new infections peak in *1983*. The results of this model are linked to two key ideas: (1) virtual 'containment' of the epidemic within high-risk groups (male homosexuals and IV-drug users)—that is, little or no expansion of the epidemic in the non-drug-using heterosexual population—and (2) relatively early 'saturation' of the high-risk groups (homosexuals and IV-drug users). Within these parameters, Artzrouni & Wykoff's best estimate is based on a time-to-AIDS distribution with a median of 8 years.

Dahlman *et al.* (1987) also use an aggregate approach to the modelling of infections.

6. MATHEMATICAL MODELS OF EPIDEMIC TRANSMISSION

6.1 *General Comments*

The models described in this section involve the representation of individual risk behaviours, HIV transmission from infected to previously uninfected persons, and the development of the disease (AIDS) among those infected with HIV. In this way, the spread of HIV infection is modelled for past, present and future years, and future AIDS cases are projected. These models require estimates of the size of the various risk groups (such as homosexuals, IV-drug users, and non-drug-using heterosexuals) and the frequency of corresponding risk behaviours, as well as of the 'transmission efficiencies', or probabilities that HIV transmission will occur when an uninfected person becomes exposed to the virus.

Unlike the aggregate models described in Section 5.4, these models of HIV transmission also involve specifying the 'mixing' of individuals within or across the risk groups (for example, the number of homosexual partners that a male has or the percentage of non-drug-using heterosexuals who have sexual contact with IV-drug users). Mixing behaviour can also be defined in terms of racial, geographical or other kinds of population subgroups. Finally, in addition to modelling HIV transmission, the models consider disease development (progression or conversion to AIDS) among HIV-infected persons. Disease development is linked back to HIV spread in terms of the decreasing tendency of an infected person to continue the risk behaviour as the disease develops.

Because of their detailed depiction of the individual-level processes involved in the HIV epidemic, these models are useful for producing considerable policy and research-related information, such as examining and/or comparing the likely or potential effects of the different kinds of possible interventions on different components of the epidemic.

A complete model of the spread of the AIDS virus in the sexually active and IV-drug-using community must account for the complicated interactions between people. However, one must begin by understanding the behaviour of simple models before going on to explore more complex ones. Two different approaches to this modelling have been developed.

One approach considers the behaviour of individuals as they form and break partnerships. Here, paired individuals become infected through multiple contacts when one partner is infected, but remain protected for the duration of the partnership if both are uninfected and also individuals cannot become infected between partnerships—see, for example Dietz (1988). These models are difficult to stratify, because of the wide variations in risk behaviour within the community.

The other approach considers the risk to the individual. The population is stratified according to the amount of risk that individuals incur, but this approach does not represent well the risk (or protection) of longer-term relationships.

In this and subsequent sections we concentrate on the second approach, being primarily concerned with the models proposed for the spread of HIV in high-risk populations. Account is partially made for partnership duration by allowing a variable number of contacts in each partnership (see later).

6.2 *Deterministic Epidemic Transmission Models: Basic Approach*

The description here begins with simple models for the spread of HIV infection within a closed group of male homosexuals.

The model is then made more realistic and complex by allowing for immunity from infection, allowing for an open population with migration and deaths being incorporated, allowing for variations due to the progression of the infection and for variations in the population according to risk behaviour (i.e. heterogeneity in

the population). Failure to segment the population according to risk behaviour or other important characteristics, like geographic region, would have the effect of overstating the number of HIV infections, because the model would fail to take into account the fact that, to some degree, the epidemic may be limited or contained within a subgroup.

The models may be adapted to deal with heterosexual spread in a two-sex population and with needle-sharing associated with IV-drug abuse (to be discussed in more detail in Section 6.7).

In this and later sections, models will be described in terms of stochastic behaviour, but the equations considered will usually be those of a deterministic approximation. This approach, while standard in the literature on AIDS modelling, does mean that potential ambiguities hover in the background. If a specified stochastic model is approximated by a deterministic process, then the interpretation of the latter is fairly clear. However, as noted earlier, since a particular deterministic process may reasonably approximate a variety of stochastic models, a unique set of stochastic assumptions cannot be deduced from a set of deterministic equations, and this may lead to problems, for example, in interpreting the parameters of those equations.

We start with a simple model for the spread of HIV infection within a *closed male homosexual* community, and assume that the total population has a fixed size, n.

We use the following notation:

t = time,
$S(t)$ = number of susceptible individuals at time t,
$I(t)$ = number of infected individuals without AIDS at time t,
$A(t)$ = number of AIDS cases at time t,
$AC(t)$ = accumulated number of AIDS cases up to time t,
γ = rate of developing AIDS for infected individuals,
β = probability of infection from a sexual contact with an infected individual,
c = average number of contacts between sexual partners, and
r = average number of new sexual partners per year.

Suppose that susceptibles become infected through sexual contacts with partners, whom they choose randomly at a fixed rate from the community. If $N(t)$ is the number, at time t, at risk of being chosen in this way, then two extreme values for $N(t)$ would be $N(t) = n$, the whole population, or $N(t) = I(t) + S(t)$ if each individual who develops AIDS is withdrawn from the class of infectives. We shall take the latter case as representing a reasonable approximation to reality.

Then a deterministic approximation to the underlying stochastic process governing the behaviour of $S(t)$, $I(t)$ and $A(t)$ is provided by the following set of ordinary differential equations:

$$\left.\begin{aligned} \frac{d}{dt}S(t) &= -\lambda(t)\,S(t) \\ \frac{d}{dt}I(t) &= \lambda(t)\,S(t) - \gamma I(t) \\ \frac{d}{dt}A(t) &= \gamma I(t) \end{aligned}\right\} \tag{6.1}$$

where

$$\lambda(t) = \beta c r \frac{I(t)}{N(t)}$$

and

$$N(t) = S(t) + I(t).$$

γ can be interpreted as the parameter for an exponentially distributed random variable representing the length of time that infected individuals remain infective.

The behaviour of the epidemic in the early stages, when $S(t) \simeq N(t)$ is given by:

$$I(t) \simeq I(0)\exp((\beta c r - \gamma)t)$$

and

$$t_d \simeq \frac{\log_e 2}{(\beta c r - \gamma)}.$$

Thus an infective in an otherwise wholly susceptible population will pass on the infection to an average of:

$$R = \frac{\beta c r}{\gamma} \text{ susceptibles.} \tag{6.2}$$

R is the reproductive rate of the infection and must satisfy $R > 1$ if an epidemic is to develop. Some empirical estimates of R are provided by Anderson & May (1987) (R may be compared with the gross and net reproduction rates used in demography).

Lemp *et al.* (1988) provide some numerical projections of the epidemic in San Francisco among male homosexuals up to 1993, using the above simple mathematical model.

A more general model separates the infectives into two classes, according to whether or not they ultimately develop AIDS. This allows for the possibility that the mean incubation period for AIDS (γ_1^{-1}) is different from the mean infectious period among those seropositives who do not develop AIDS (γ_2^{-1})—in the extreme case, $\gamma_2 = 0$, so that the seropositives remain infectious indefinitely, as investigated by Bailey & Estreicher (1987). We now modify the notation, so that I_1, I_2, A_1, A_2 denote, respectively: the numbers of infected individuals who will ultimately develop AIDS; the number of infectives who do *not* develop AIDS; the number of AIDS cases; and the number of non-infectious seropositives.

Then the ordinary differential equations become:

$$\left.\begin{aligned}
\frac{d}{dt}S(t) &= -\lambda(t)\,S(t)\\
\frac{d}{dt}I_1(t) &= \rho\,\lambda(t)\,S(t)-\gamma_1 I_1(t)\\
\frac{d}{dt}I_2(t) &= (1-\rho)\,\lambda(t)\,S(t)-\gamma_2 I_2(t)\\
\frac{d}{dt}A_1(t) &= \gamma_1 I_1(t)\\
\frac{d}{dt}A_2(t) &= \gamma_2 I_2(t)
\end{aligned}\right\} \qquad (6.3)$$

where $I(t) = I_1(t) + I_2(t)$ is the total number of infectives at time t, ρ is the probability that an individual enters the class of potential AIDS patients on withdrawal from the class of infectives and a suitable choice for $N(t)$ in the definition of $\lambda(t)$ might be as before, or with:

$$N(t) = S(t) + I_1(t) + I_2(t) + A_2(t) = n - A_1(t).$$

If $\gamma_1 = \gamma_2$, then equations (6.3) reduce to (6.1).

Then, for this model, the overall reproductive rate of HIV infection is given by:

$$R = \beta c r\left(\frac{\rho}{\gamma_1} + \frac{1-\rho}{\gamma_2}\right).$$

The model represented by equations (6.3) has been studied numerically by Anderson & May (1986), Anderson *et al.* (1986) and by Blythe & Anderson (1988).

So far, the model has applied to a *closed* population. In order to apply the model to time periods beyond the initial stages of the spread of infection, it is necessary to allow immigration to the class of susceptibles and deaths from all classes.

Suppose:

$m(t)$ = rate of immigration to the class of susceptibles at time t
μ = death rate of individuals without AIDS (in the form of the force of mortality)
δ = extra death rate of individuals with AIDS (with $\delta \gg \mu$).

Then the differential equations (6.3) are modified as follows:

$$\left.\begin{aligned}
\frac{d}{dt}S(t) &= m(t)-\mu S(t)-\lambda(t)\,S(t)\\
\frac{d}{dt}I_1(t) &= \rho\,\lambda(t)S(t)-(\mu+\gamma_1)I_1(t)\\
\frac{d}{dt}I_2(t) &= (1-\rho)\,\lambda(t)\,S(t)-(\mu+\gamma_2)\,I_2(t)\\
\frac{d}{dt}A_1(t) &= \gamma_1 I_1(t)-(\mu+\delta)\,A_1(t)\\
\frac{d}{dt}A_2(t) &= \gamma_2 I_2(t)-\mu A_2(t)\\
\frac{d}{dt}AC_1(t) &= \gamma_1 I_1(t)\\
\frac{d}{dt}AC_2(t) &= \gamma_2 I_2(t)
\end{aligned}\right\} \qquad (6.4)$$

where $N(t) = S(t) + I_1(t) + I_2(t) + A_2(t)$ in the definition of $\lambda(t)$.

This model has been presented and analysed by many authors for the spread of various sexually transmitted diseases: the interested reader is referred to Hethcote & Yorke (1984) for a full review.

This model has been studied numerically by Anderson & May (1986) and by Anderson *et al.* (1986) using the simpler form $\gamma_1 = \gamma_2$ and $N(t) = S(t) + I_1(t) + I_2(t) + A_1(t) + A_2(t)$ and by Hyman & Stanley (1988) using the simpler form $m = \mu S_0$ where S_0 is the population size before the AIDS virus was introduced (so that there is a balance between flows into and out of the population) and $\rho = 1$ (so that $I_2(t) \equiv 0$ and $A_2(t) \equiv 0$). Similarly, Thompson (1987) explores numerically $\rho = 1$ and $m(t) = m$, but does not test the model for goodness-of-fit against observed data.

An analytic solution has been found by Birkhead (1987) under certain further restrictive assumptions. Firstly, it is assumed that $N(t) = S(t) + I_1(t)$, so that, not only the AIDS patients but also the non-infectious seropositives, are excluded. Secondly, it is assumed that the immigration of susceptibles is at a rate proportional to $N(t)$, i.e. $m = m_0 N(t)$, rather than being constant. Neither modification will be significant in the initial stages of the epidemic, and as the epidemic progresses it is plausible that changing behaviour could result in a reduced level of immigration into the homogeneously mixing male homosexual community being modelled. Then, explicit analytic formulae for $I(t)$ and $N(t)$ can be derived. Birkhead uses a different interpretation of ρ: a proportion of the seropositives are assumed to develop full-blown AIDS and then cease sexual

mixing, with the rest remaining infective and sexually active. Thus, the equation for $I_1(t)$ becomes:

$$\left.\begin{aligned} \frac{d}{dt} I_1(t) &= \lambda_1(t)S(t) - (\rho\gamma_1 + \mu)I_1(t) \\ \text{where} \quad \lambda_1(t) &= \frac{\beta c r\, I_1(t)}{(S(t) + I_1(t))}. \end{aligned}\right\} \tag{6.5}$$

Equations (6.4) can be modified to model purely heterosexual spreading, by splitting the population according to sex and including 'partnership balance' relationships. These balances are necessary to take account of situations where there are not enough women, so that men cannot have as many partners as they might like, and vice versa.

Other modifications have been introduced to the above basic model by van Druten *et al.* (1987), to allow for the fact that, at the start of the spread of infection in the population, few of its members are at risk and that the size of the population at risk should, therefore, be a dynamic variable. The effect of modelling the risk population in this way is to slow down the initial exponential rate of increase in the number of infectives (relative to the models described earlier), with most effect when the average number of sexual partners per infective is 'small' (this parameter is proportional both to r and the mean duration of a partnership and, for a fixed value of r, would be small when partnerships are short lived).

6.3 *Deterministic Epidemic Models: Allowing for Time Since Infection*

If we include the time since infection or AIDS, then variable infectivity and the distributions of times from infection to AIDS and of times from AIDS to death may be explicitly modelled. Following Anderson *et al.* (1986), we break down the infected population $I(t)$ according to the, time, τ, since infection, $I(t) = \int I(t,\tau)d\tau$. $I(t,0)$ is now the rate at which people become infected, and $I(t,\tau)$ has the units of people/year. Similarly, we distribute AIDS patients according to time τ since AIDS, $A(t) = \int A(t,\tau)d\tau$. Defining:

$I(t,\tau)$ = distribution of infecteds according to time τ since infection. $I(t,\tau)$ is the number of people infected per year τ years before time t,
$A(t,\tau)$ = distribution of AIDS cases according to time since they developed AIDS,
$\beta(\tau)$ = probability of infection from a contact with a person infected τ years ago,
$\gamma(\tau)$ = rate of developing AIDS at a time τ after infection, and
$\delta(\tau)$ = death rate at time τ after developing AIDS,

the system of equations (6.4) becomes:

$$\left.\begin{aligned}
\frac{dS(t)}{dt} &= m(t)-\mu S(t)-\lambda(t)S(t)\\
I(t,0) &= \lambda(t)S(t)\\
\frac{\partial I(t,\tau)}{\partial t}+\frac{\partial I(t,\tau)}{\partial \tau} &= -[\gamma(\tau)+\mu]I(t,\tau)\\
A(t,0) &= \int_0^\infty \gamma(\tau)I(t,\tau)\,d\tau\\
\frac{\partial A(t,\tau)}{\partial t}+\frac{\partial A(t,\tau)}{\partial \tau} &= -\delta(\tau)A(t,\tau)\\
\lambda(t) &= \frac{cr}{N(t)}\int_0^\infty I(t,\tau)\,\beta(\tau)\,d\tau\\
N(t) &= S(t)+\int_0^\infty I(t,\tau)\,d\tau\\
\frac{dAC(t)}{dt} &= A(t,0)
\end{aligned}\right\} \qquad (6.6)$$

where, for convenience, we have followed Hyman & Stanley (1988) and put $\rho = 1$. A corresponding version of equations (6.4) can be set up with $\rho \neq 1$. The infectivity $\beta(\tau)$ is an average over all individuals infected at time τ. For constant γ, β and δ, $I(t)$ and $A(t)$ would satisfy the corresponding version of (6.4). It would be possible to vary c and r also with time since infection and thus to take account of behavioural changes induced by infection. Similarly, for transmission in a heterosexual population, the model can be generalised to incorporate numbers of infected individuals distributed according to time since infection and the AIDS cases distributed by time since diagnosis.

An early example of the use of this model with $\rho \neq 1$ is van Druten *et al.* (1986). They fit the model by simulation to the data from the San Francisco hepatitis B study cohort of homosexual and bisexual men, which numbered 6875 in 1978. The infectious period was assumed random and exponentially distributed with a mean of 3, 5 or 10 years and the incubation period was assumed to follow a similar distribution. The value of ρ was 0·10, which now seems a gross underestimate. Van Druten *et al.* conclude that, if the numbers of sexual contacts were halved, the predicted numbers of AIDS cases (and infected cases) for 1988 would fall by only 13% (and 14%).

Anderson *et al.* (1986) have studied the version of equation (6.6) with $\rho \neq 1$. Their numerical investigations compare the model with a variable incubation period (represented by a Weibull distribution) with the simpler model, in which

the incubation period is assumed to be constant, γ^{-1}, and independent of the duration of infection. They show that, *ceteris paribus*, the main effect of the variable incubation period is to alter the shape of the epidemic over time, in such a way that the rise in cases of AIDS follows the patterns set by the rise in seropositivity (i.e. incidence of HIV), but with a pronounced delay: the simpler model displays a less marked lag between the rise in seropositivity and rise in AIDS cases. However, the time to peak incidence is little changed. Anderson *et al.*'s Weibull model is:

$$\gamma(t) = ct.$$

Such a choice follows the intuitive belief that the incubation period should have an increasing hazard function, representing progressive deterioration of the infective's immune system, and that a linear approximation to the function might be adequate. The hazard function for a Weibull distribution with cumulative distribution function $[1-\exp(-(bt)^a)]$, is $ab^a t^{a-1}$, so Anderson *et al.*'s (1986) choice corresponds to $a = 2$.

Blythe & Anderson (1988) take this work further and compare four different parametric representations of the incubation period distribution:

$$\text{exponential } \gamma(t) = \gamma$$

$$\text{Weibull } \gamma(t) = ab^a t^{a-1}$$

$$\text{Erlang } \gamma(t) = \frac{k^{n-1}t^n}{n!}\left(\sum_{j=0}^{n}\frac{k^j t^j}{n_j!}\right)^{-1}$$

$$\text{rectangular } \gamma(t) = \frac{1}{T+\frac{1}{2}h-t} \qquad \text{for} \qquad T-\tfrac{1}{2}h < t < T+\tfrac{1}{2}h$$

and their effects on the models of the transmission dynamics of HIV.

A numerical solution of the full non-linear models suggests that the four distributions listed above yield similar results with respect to the steady-state behaviour of the respective models and their local stability properties. The full nonlinear transient (i.e. t finite) behaviour of the four models is also similar, given the same mean incubation period for the distributions. In this study, the authors prefer the flexibility of a low-order Erlang distribution—it should be noted that in this case, as $t \to \infty$, $\gamma(t)$ increases, but is bounded above. The exponential distribution is found to provide a crude, although not unreasonable, first approximation, despite the fact that a constant $\gamma(t)$ is at odds with observations.

However, it should be noted that uncertainty concerning the precise choice of parameter values, created by data limitations and inadequate parameter estimation procedures, can cause variation at least as large as that generated by the four different incubation period distributions described above. This is discussed further in a later section. Also, it is possible that in the presence of heterogeneity of sexual activity, the differences between various parametric representations of the incubation period distribution may be more significant.

The Weibull distribution has the advantage of a hazard function which is analytically simple. On the other hand, the gamma distribution, at least for integer values of the index, can be modelled using the method of stages which exploits the fact that a gamma variable can be regarded as a sum of exponential variables. This model is described by Bailey & Estreicher (1987), who also describe a preliminary investigation in which the incubation period is constant, corresponding to the limit as the number of stages tends to infinity, although this may be biologically implausible. In Bailey (1988) (in an extension to the simple model referred to in Section 5.2), the incubation period is split into two parts, the first of which is regarded as the period when immune defences are breaking down and is modelled by a gamma variable via a series of exponentially distributed stages. During the second part, the individual is at major risk of opportunistic infection and, after an exponentially distributed interval, develops AIDS. In general, this representation will not lead to a gamma distribution for the incubation period, since the parameter for the latter exponential distribution is not assumed to be the same as the (common) parameter of the earlier stages. However, in this model, which is illustrated by fitting data from San Francisco, Bailey assumes that all those who are infected will ultimately develop AIDS, so obtaining a much longer-tailed distribution for the incubation period. (The optimum fit involves 7 stages for the first part of the incubation period.)

Similarly, Panjer (1987, 1988) uses a multi-stage approach to modelling the incubation period (and the time from development of AIDS to subsequent death). In work related to the report of the Canadian Institute of Actuaries Subcommittee on Modelling (1988) and described in Section 5.4, Panjer adopts the Walter Reed Staging Method for the classification of AIDS—this leads to the identification of 5 states labelled, as before, 1A, 1B, 2A, 2B and 3.

Panjer represents the transitions from stage to stage by a continuous time Markov process. An exponential distribution is used—specifically if T_j denotes the time in stage j:

$$\Pr(T_j > t) = \exp(-t\mu_j)$$

so the transition intensity (or force of transition) depends only on stage and *not* on the other factors such as age, sex or length of time in the stage. Hence, the incubation period would be represented by a sum of three exponential random variables ($T_{1b} + T_{2a} + T_{2b}$)—i.e. a generalised Erlang distribution. Using the grouped data from the Frankfurt study (Brodt *et al.*, 1986), Panjer uses maximum likelihood methods to estimate the parameters. Comparisons are made with the more *ad hoc* methods of Cowell & Hoskins (1987).

A similar approach is used by Salzberg *et al.* (1989).

Longini *et al.* (1989) fit a five-stage, time-homogeneous Markov model to heavily censored data from a cohort of 513 homosexual and bisexual men from the San Francisco area found to be HIV seropositive, 73 individuals known to have received transfusions of HIV-infected blood, and 17 haemophiliacs who received HIV-infected factor VIII. The model partitions the infected period into

four progressive stages. The first stage is HIV infection, but with antibody-negative status. Stage 2 is antibody-positive status, but asymptomatic. The third stage occurs when an individual develops an abnormal haematologic indicator and/or prodromal illnesses (pre-AIDS symptoms), such as persistent generalised lymphadenopathy, Stage 4 is clinical AIDS, and a fifth stage is included in the model—death due to AIDS.

The homosexual and bisexual men come from a random sample from the larger cohort of men who were enrolled at the San Francisco City Clinic between 1978 and 1980 for studies of hepatitis B. It is assumed that infected individuals progress irreversibly through the stages of infection. According to this model, the waiting time from when an individual enters stage 1 until reaching stage 4 is the AIDS incubation period. The exact transition times are not available in these data i.e. there is interval censoring. In addition, data may be right censored (that is, at the last observation an individual may still be in one of the infected stages) or left censored (that is, at the time of the first observation an individual may have already been in that stage for an indeterminate amount of time).

If the transition intensities are λ_i (independent of time) and the probability that an individual who is in stage i at time t will be in stage $k(\geqslant i)$ at time $(t_0 + t)$ is $p_{ik}(t)$, then the probability density function for the AIDS incubation period is:

$$f(t) = \lambda_3 p_{13}(t).$$

(This density is of the form of an Erlang distribution.)

Using numerical methods, maximum likelihood estimates of the parameters λ_i are obtained.

Longini *et al.*'s estimated mean incubation period is 9·8 yrs (for primarily sexually-infected homosexual and bisexual men and transfusion-infected individuals) and is longer than the earlier estimates of Medley *et al.* (1988, 1988a) and Lui *et al.* (1986).

6.3.1 *Statistical Problems in Modelling the Incubation Period*

Daykin (1990), in his Section 6, provides a full review of estimates of the incubation period. Our purpose here is to focus on the statistical and modelling issues raised by attempts to represent the incubation period distribution.

Historically, the first information on the incubation period distribution was derived from studies of transfusion-associated cases of AIDS, so that models fitted to these data are not necessarily applicable to other categories of AIDS patients. For example, the suggestion is often made that the mean incubation period for transfusion-associated cases of AIDS will be less than that for other methods of transmission, since the transfusion recipient is likely to receive a much larger number of infected cells. The other obvious problem with these data is that a length-biased sampling is in operation, with the longer incubation periods less likely or impossible to be observed while the epidemic is still in its early stages, since most infections will have occurred not long before, or during, the observation period. Most attempts to fit these data focus on the Weibull and

gamma distributions, which are flexible in shape and are both reasonably tractable and plausible. However, Rees (1987, 1987a) fits a normal distribution using trial and error methods. His fitted distribution has a mean of 15 and a standard deviation of 5, so that the chance of negative incubation periods is very small; however, Lawson & Dagpunar, in the discussion on Isham's (1988) paper, point out that these parameters are sensitive to the use of an unbounded distribution like the normal—they obtain maximum likelihood estimates of 6 years and 2 years for the mean and standard deviation using a *truncated* normal. Further, they point out the problems caused by a flat likelihood function and the imposition of a symmetric distribution.

Rees' approach has been extensively criticised also by Barton (1987), Lui *et al.* (1987), Costagliola & Downs (1987) and Beal (1987).

Another (perhaps more subtle) problem with studies of transfusion-associated AIDS cases, is that inclusion in these studies is *conditional* on developing AIDS. Hence, the data cannot provide information on the probability that an infected individual develops AIDS unless further assumptions are made—the technical issues underlying this point have been discussed by Kalbfleisch & Lawless (1988) and by Lagakos *et al.* (1988) *inter alia.*

These technical problems must be borne in mind when interpreting the results of the various studies.

An early attempt to fit a Weibull distribution to the transfusion-related incubation period data is reported by Lui *et al.* (1986), who obtain a mean of 4·5 years. However, although their method takes account of the fixed period of observation in which the diagnoses of AIDS occur, it does not allow adequately for the increasing rate at which transfused blood was infected.

In order to emphasise the importance of considering the increasing rate of infection, we shall digress and consider the incubation period distribution from first principles.

Suppose infections occur in a Poisson process of rate $\alpha(t)$. To each infection is attached an incubation time, x, having a distribution with density $f(x)$ and cumulative distribution function $F(x)$. Observations are made over a period $(-\infty, t_0)$ and suppose the incubation times are sorted by time of onset of AIDS. Then the distribution of the incubation time for individuals observed to have AIDS at time t has the truncated distribution $f(x)/F(t_0-t)$ for $0 \leq x \leq t_0-t$ (if we ignore mortality of AIDS cases). But if individuals are sorted by time of report, then biases arise. If S is a random variable representing time of report, then the conditional distribution of X, given $S = s$, is:

$$\frac{f(x)\,\alpha(s-x)}{\int_0^{\infty} f(u)\,\alpha(s-u)du} = f(x)$$

for all x and s if and only if $\alpha(t)$ is constant. In particular if $\alpha(t) = \alpha e^{\beta t}$, the conditional density is, for all s:

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

So, if $\beta > 0$, the bias is in favour of the shorter incubation times, the conditional distribution showing no trend in time, i.e. it is independent of s. If infections grow sub-exponentially so that, in effect, β is slowly decreasing with s, then the conditional distribution of x, given time to report s, shows increasing incubation times as s increases. This is a natural consequence of the 'biased' sampling. It can be shown that, more generally, this conditional density is independent of s, if and only if $\alpha(t)$ is exponential.

Using data from the U.S.A. (CDC—Atlanta) on patients who have no other risk of acquiring infection other than having received a transfusion of infected blood or blood products, Medley *et al.* (1987, 1988) fit various distributions, making an explicit allowance for the effect of the increasing rate of infected transfusions (as discussed above). The rate at which transfusions of infected blood occur is modelled by a linearly or exponentially increasing function (appropriate until such transfusions cease because of mandatory screening of donated blood) and the incubation period is represented by a Weibull or gamma distribution. The probability of diagnosis is allowed to be time-dependent (via a probit function, i.e:

$$\Phi\left(\frac{t-\theta_0}{\theta_1}\right)$$

where $\Phi(\,)$ is the standard normal integral) and the model is fitted by numerical maximisation of the log likelihood. The data are divided by sex and age (under 5 years, 5–59 years, 60 and over) and, as is to be expected from the theory described in Section 6.2, *inter alia*, Medley *et al.* find that an exponential increase in the rate of transfusion-related infections fits the data better than a linear growth function.

Medley *et al.* conclude that, for the data available at present, the Weibull and gamma distributions for the incubation period both give equally good fits and it is interesting to note that the index of the fitted Weibull distribution (for the whole data set) is 2·4 as compared with $a = 2$ used by Anderson *et al.* (1986, 1987). These results indicate that incubation periods for men may be shorter than for women. Using an exponentially increasing rate function and Weibull incubation period, Medley *et al.*'s fitted means are 8·9 years for females and 5·6 years for males. There is no obvious explanation for this difference, since the method of infection (blood transfusion) is the same in each case. One possibility is that there is an immunological difference between men and women, but, alternatively, perhaps there are biases in diagnosis which result in men being diagnosed earlier in the course of the disease. Further investigation of existing and subsequent data is needed to resolve this point. When a single Weibull

distribution is fitted to all the data (297 observations thought to be reliable, 112 females and 185 males, corresponding to diagnoses made up to mid-1986), the mean is 6·3 years. As expected on theoretical grounds, this mean is larger than that obtained by Lui *et al.* (1986) (4·5 years) although the bigger data set may be partly responsible. The incubation periods for the youngest age group (fitted mean 2·3 years) and the oldest group (5·6 years) are found to be shorter than those for the central (5–59 years) age group, for whom the fitted mean is 8·1 years.

In interpreting these fitted mean values, it is important to bear in mind that the data being fitted lie in the lower tails of the fitted distributions, while the mean is estimated on the assumption that the fitted distribution would work equally well in the upper tail, on which it is highly dependent. Estimation of the mean from data in the lower tail alone is likely to be unreliable. As more data become available, it should be possible to obtain better information on the shape of the upper tail of the distribution, enabling a distinction to be made between the Weibull and gamma alternatives and better predictions to be made of the eventual magnitude of the epidemic.

Boldsen *et al.* (1988) explore the parametric fitting of the incubation period distribution to data from Peterman *et al.* (1985) (referring to transfusion-associated cases) and from Curran *et al.* (1985) (referring to longitudinal data from the San Francisco City Clinic Cohort of homosexual and bisexual men). Using a Weibull distribution with hazard function $\gamma(t) = ab^a t^{a-1}$, with the former data set, an exploration of the form of the likelihood region suggests that the data mainly tend to provide information on a (rather than b), i.e. the data determine the shape of the left-hand tail of the distribution, but not the total probability in that tail. Experiments with the latter data set indicate that a Weibull curve is more satisfactory than either a gamma or a log-normal. They comment that Brookmeyer & Gail (1986) choose a value of the parameter b for their back projection calculations (see Section 7) that is unreasonably high (being based on the same data). Lagakos *et al.* (1988) fit Weibull models to data on 258 adults with transfusion-related AIDS and they confirm that the likelihood function is very flat over a range of parameter values, representing a wide range of possible incubation distributions.

Costagliola *et al.* (1989) estimate the incubation time distribution from a data-set of transfusion associated cases provided by the French Ministry of Health. Using a Weibull distribution, they estimate the mean to be 5·3 years with a 90% confidence interval of [4·4, 8·9], following the methodology of Medley *et al.* (1987, 1988, 1988a).

Consideration of the infection rate $\alpha(t)$ leads to further statistical problems, which have been the subject of some debate in the literature.

Suppose transfusion-associated infections occur at a Poisson process of rate $\alpha(t)$. Kalbfleisch & Lawless (1988) point out the identifiability problems arising from attempts to estimate simultaneously the infection rate $\alpha(t)$ and the probability density function, for the time from infection to diagnosis of AIDS: these problems lead to unreliable estimates. The data available from the

transfusion studies are only sufficient to identify the shape of the early distribution and cannot *strictly* be used for estimation of the mean or median incubation period. It is thus not possible to discriminate between high infection rates coupled with long incubation times on the one hand, or low infection rates coupled with short incubation times on the other.

In reply, Medley *et al.* (1988a) point out the possibility of using external collateral information to eliminate the identifiability problem. They apply the methods described earlier to an extended data set of transfusion associated cases from CDC, comprising 512 cases aged over 12 years at diagnosis of AIDS. The most satisfactory fit to the incubation period distribution is a Weibull with mean 7·6 years, allowing for an implied annual rate of infective transfusion which is consistent with other data. The fit based on a gamma distribution is associated with a much higher rate of new infections, since the introduction of screening, than seems likely from other information. Until more data become available and more of the incubation period has been observed, these results must remain as tentative. The confidence intervals (not presented) are likely to be very wide, reflecting the fact that only a portion of the incubation period has been observed in most patients.

As noted earlier, the data sets used by Lui *et al.* (1986) and Medley *et al.* (1987, 1988, 1988a) are subject to length-biased sampling, since all individuals are ascertained because they had an AIDS diagnosis. Thus, these cohorts include only those individuals who had relatively short incubation periods. Although both analyses attempt to adjust for bias, the corrective measures are indirect and may have achieved only partial success. In addition, other sources of bias may have occurred due to the increasing hazard function of the Weibull distribution used in both analyses (this is pursued further by Brookmeyer & Gail (1987) and by Brookmeyer *et al.* (1987)). In some studies, there are biases arising from the fact that all the individuals are infected prior to enrolment and the exact dates of infection are unknown—also discussed by Brookmeyer & Gail (1987) and by Brookmeyer *et al.* (1987).

As described in Section 6.3, Panjer (1987, 1988) and Longini *et al.* (1989) use a staged Markov model, with the incubation period split into 3 stages, where each stage is associated with a constant hazard function. This renders the analysis less subject to those forms of bias that affect estimation based on models that do not have constant hazard functions. Ascertainment of individuals used in Longini *et al.*'s analysis occurs only because of their known infection and, thus, length-biased sampling is probably not a problem either. If, however, the assumption of constant hazard functions within stages is violated, then this analysis would be subject to model mis-specification biases.

A number of other investigators have estimated the AIDS incubation period as the waiting time from stage 2 to the entering of stage 4 (according to the Walter Reed Staging Method). De Gruttola & Mayer (1988) fit a Weibull distribution to data for men from the San Francisco cohort observed to seroconvert; the estimated mean waiting time is 9·1 years. (This is relatively close to Longini *et*

al.'s corresponding estimate of 9·6 years.) Lui *et al.* (1988) fit a Weibull distribution to data for 84 men from the San Francisco cohort observed to seroconvert within a one-year period. Their estimate of the mean waiting time from the first seropositive blood test, i.e the early part of stage 2 to AIDS diagnosis, is 7·8 years, with a 90% confidence interval of [4·2, 15·0] years. Harris (1988) combines cohorts of HIV-infected individuals from the San Francisco cohort, transfusion recipients and adults with haemophilia. He estimates the mean AIDS incubation period as 9·8 years, which is identical to Longini *et al.*'s estimate.

Longini *et al.*'s data, as noted in Section 6.3, exhibit considerable censoring. This makes it impossible to specify exact transition times for individuals, requires the use of a time-homogeneous model and makes it very difficult to test the appropriateness of such assumptions. As more data with less censoring become available, it may then be possible to use a time-dependent specification where $\lambda_i(t)$ are functions of time.

Peto (1988), in his contribution to the discussion of Anderson's (1988) paper, also criticises the parametric modelling of the incubation period distribution, because of the inappropriateness of the distribution chosen in the presence of very high mortality rates from other causes of death, and because the assumed exponential increase in the underlying incidence rate is not appropriate—as noted elsewhere, the incidence rate increases very rapidly and then, because of the saturation effects among the most promiscuous, the rate slows down.

Kalbfleisch & Lawless (1988) and Peto mention the importance of non-parametric methods for estimating the incubation period distribution, but, it should be noted, these cannot provide estimates of the progression to AIDS at durations greater than have been observed. Parametric methods do enable such extrapolation to take place, although widely different results can be obtained using different parametric distributions. As noted earlier in this section, it is sometimes possible to use the chosen distribution to make estimates about other quantities which can then be checked for reasonableness.

Lagakos *et al.* (1988) consider the estimation of the incubation period distribution, using information on persons infected with the AIDS virus from a contaminated blood transfusion. Of persons infected in this way, *only* those who develop AIDS by a certain date can be identified. The statistical problem is one of making inferences about a stochastic process of infection and subsequent disease, in which realisations are right-truncated in time. Lagakos *et al.* consider the process in reverse time and transform the problem to one of survival data that are left-truncated in 'internal' time. This approach is used to develop non-parametric methods for estimating and comparing the identifiable aspects of the incubation distribution for various groups.

Generally, in many studies of infectious diseases, the times of infection are not known precisely, although they are known up to an interval. Brookmeyer & Goedert (1989) consider this problem and develop a methodology for analysing data generated by cohort studies during an epidemic in which the exact times of infection cannot be ascertained, although the times may be known up to an

interval. A 2-stage parametric regression model is proposed for jointly estimating the effects of covariates on risk of infection, as well as risk of progression to clinical disease once infected. The two stages refer to *infection* followed by *onset of clinical disease*. The methodology tackles three important features of the data:

(1) the calendar dates of infection are interval censored;
(2) the times from infection to clinical disease are not known precisely because of the interval censoring in (1) and because some infected individuals may not have developed the disease by the last date of follow-up (right-censored); and
(3) there are two types of covariates: X_1 which modify risk infection and X_2 which modify risk of disease once infected. Some covariates may be included in both vectors.

Brookmeyer & Goedert fit the model to data available from the Hershey Hemophilia Cohort, San Francisco City Clinic Cohort and National Cancer Institute Multicentre Hemophilia Cohort Studies.

In a parallel paper, de Gruttola & Lagakos (1989a) consider issues in the non-parametric estimation of the distribution function of the time to onset of clinical disease following infection, allowing for the interval- and right-censoring features of the data, but *not* allowing for the presence of covariables.

A further statistical problem that arises is the difficulty of interpreting events near to the start of an epidemic in the presence of a skewed incubation period distribution. Gonzalez & Koch (1987) and Gonzalez *et al.* (1987) consider this and related problems, by investigating the effect that a non-exponentially distributed incubation period has on the start of the epidemic.

They make the following (reasonable) assumptions about the initial stage of the epidemic, for a sufficiently homogeneous subset of the population: the number of virus carriers is very small compared to the number of susceptibles and the factors which govern the transmission of the disease are roughly constant. In these circumstances, the number of *virus carriers* grows exponentially. They show that the observed slowing in the rate of growth of recorded AIDS cases within the first 3–6 years of the epidemic can be explained as a spurious effect (rather than being attributable to optimistic influences e.g. changes in the behaviour of risk groups, to prophylactic measures and/or the depletion of subsets of the population). Thus, Gonzalez & Koch (1987) show that the initial exponential increase in AIDS incidence can, itself, be temporarily increased so that the doubling time for AIDS cases is less than that for HIV incidence, and they suggest that this is one factor contributing to the observed slowing in the rate of growth of AIDS incidence. Section 5.2 has indicated that another important factor is heterogeneity of sexual activity, which can lead to a decrease in the growth rate of the epidemic once most of the highly active individuals are infected. The doubling time of the epidemic in its early stages is one easily observed property which is used in estimating parameter values, so that, if there is a non-negligible distortion due to non-exponential incubation

periods, then this should be taken into account. In Gonzalez & Koch (1987) and Gonzalez *et al.* (1987), curves are fitted to the total incidence of AIDS for a number of countries, using a gamma distribution for the incubation period (mean 5 years, variance 8 years) and it is shown that their model is consistent with data for the early part of epidemics, during which exponential growth of HIV infections applies.

We consider the convolution equation that forms the basis of the back projection method (Section 7):

$$a(t) = \int_0^\infty h(t-u)\, f(u)\, du$$

where $a(t)$ is the intensity of the point process for new diagnoses of AIDS, $f(u)$ is the pdf of the incubation period distribution and $h(u)$ is the HIV infection rate. Then it can be verified, from the above equation, that for a broad class of incubation distributions, f, $a(t)$ can grow exponentially over much of its range if $h(t)$ is exponential. Exponential $h(t)$ corresponds to a constant doubling time in the cumulative number of HIV infections. Even if $h(t)$ were exponential, however, $a(t)$ may be dominated by non-exponential terms for small values of t. As Gonzalez & Koch (1987) point out, the doubling time in the number of AIDS cases may initially be shorter than the doubling time in the number of people infected with HIV. Although exponential $h(t)$ might ultimately induce an exponential $a(t)$, exponential growth in $h(t)$ will only occur early in the epidemic, when the number of infectives is negligible relative to the number of susceptibles. This is true even in simple epidemics in homogeneous populations (Bailey, 1975). As the epidemic spreads, the reduction in the proportion of the population that remains susceptible causes $h(t)$ to become subexponential; that is, the doubling time in the number of infecteds will increase with chronological time. The proportion infected does not have to be very large for the doubling time in AIDS incidence to increase substantially. For example, with a simple epidemic model for infection and a Weibull induction distribution with a median of 9·6 years, the doubling time in cumulative incidence of AIDS increases by more than 50% by the time the prevalence of infection is 25% (see de Gruttola & Lagakos (1989, 1989a) for further details).

An additional factor that can cause $h(t)$ to be subexponential is that the populations at risk are heterogeneous. Heterogeneity refers not only to different 'risk groups' for AIDS (for example, homosexually-active men, IV-drug users, etc.), but also to variations in behaviour within groups. For example, variations in the rate of new partner acquisition, type and frequency of sexual acts per partner, and duration of relationship all affect $h(t)$. This heterogeneity is believed to be the reason that the observed rise in infection in some cohorts of homosexually-active men is more nearly linear than exponential (Anderson & May, 1987; Peto, 1986).

If h is linear or quadratic in t, it can be shown that, for $a(t)$ to grow exponentially, $f(t)$ must also grow exponentially over its range—but such an

incubation distribution is uncommon in biological processes, and is inconsistent with what is currently known about the incubation period for AIDS. So, the conditions needed for a constant doubling time in AIDS incidence seem unrealistic.

6.4 *Deterministic Epidemic Models: Allowing for Heterogeneity of Sexual Activity*

So far, the models presented do not treat variations in risk behaviour between different people in the group. These models would be sufficient if the variations in risk behaviours were not large and did not play such a significant role in the epidemic. However, surveys of risk behaviour in the homosexual communities demonstrate that the variance in the number of sexual partners per year is large.

In the AIDS epidemic, it is significant that the people with many partners tend to become infected first and then become carriers who infect less active people. This feature can have a marked effect on the course of the epidemic and on which risk group is currently at highest risk of infection.

To model risk behaviour, we suppose that the population can be distributed according to their numbers of new sexual partners per year. Letting:

r = number of new sexual partners per year,
$S(t,r)$ = distribution of susceptibles according to the number of partners per year,
$I(t,r,\tau)$ = distribution of infecteds according to the number of partners per year and the time since infection,
$c(r,r')$ = total number of contacts in a partnership between people with r and r' partners per year, and
$m(t,r)$ = immigration rate at time t of people with r new partners per year,

we have the model:

$$\left.\begin{aligned}
\frac{\partial S(t,r)}{\partial t} &= m(t,r) - \mu S(t,r) - \lambda(t,r)S(t,r) \\
I(t,0,r) &= \lambda(t,r)S(t,r) \\
\frac{\partial I(t,\tau,r)}{\partial t} + \frac{\partial I(t,\tau,r)}{\partial \tau} &= [\gamma(\tau) + \mu]\, I(t,\tau,r) \\
A(t,0) &= \int_0^\infty \int_0^\infty \gamma(\tau) I(t,\tau,r)\, d\tau\, dr \\
\frac{\partial A(t,\tau)}{\partial t} + \frac{\partial A(t,\tau)}{\partial \tau} &= -\delta(\tau)A(t) \\
\frac{dAC(t)}{dt} &= \int_0^\infty \int_0^\infty \gamma(\tau) I(t,\tau,r)\, d\tau\, dr \\
N(t,r) &= S(t,r) + \int_0^\infty I(t,\tau,r)\, d\tau
\end{aligned}\right\} \qquad (6.7)$$

where, for convenience, we have followed Hyman & Stanley (1988) and put $\rho = 1$.

We repeat here the model represented by equations (6.7), but allow for the presence of an immune group, i.e. we are combining equations (6.4) and (6.7), viz.:

$$\left.\begin{aligned}
&\frac{\partial S}{\partial t}(t,r) = m(t,r) - \mu S(t,r) - \lambda(t,r)S(t,r)\\
&I_1(t,0,r) = \rho\,\lambda(t,r)\,S(t,r)\\
&I_2(t,0,r) = (1-\rho)\,\lambda(t,r)\,S(t,r)\\
&\frac{\partial}{\partial t} I_1(t,\tau,r) + \frac{\partial}{\partial \tau} I_1(t,\tau,r) = -(\gamma_1(\tau) + \mu)\, I_1(t,\tau,r)\\
&\frac{\partial}{\partial t} I_2(t,\tau,r) + \frac{\partial}{\partial \tau} I_2(t,\tau,r) = -(\gamma_2(\tau) + \mu)\, I_2(t,\tau,r)\\
&\frac{\partial}{\partial t} A_1(t,r) = \int_0^\infty \gamma_1(\tau) I_1(t,\tau,r)\, d\tau - (\mu + \delta)\, A_1(t,r)\\
&\frac{\partial}{\partial t} A_2(t,r) = \int_0^\infty \gamma_2(\tau)\, I_2(t,\tau,r)\, d\tau - \mu\, A_2(t,r)\\
&N(t,r) = S(t,r) + \int_0^\infty [I_1(t,\tau,r) + I_2(t,\tau,r)]\, d\tau + A_2(t,r)
\end{aligned}\right\} \tag{6.8}$$

where the notation has been extended with I_1, I_2, A_1, A_2, representing respectively the number of infected individuals who will ultimately develop AIDS, the number of infectives who do *not* develop AIDS, the number of AIDS cases and the number of non-infective seropositives.

If we are not interested in I_2 and γ_2 as functions of τ, then the equations simplify as follows:

$$\left.\begin{aligned}
&\frac{\partial}{\partial t} I_2(t,r) = (1-\rho)\lambda(t,r)S(t,r) - (\gamma_2 + \mu)\, I_2(t,r)\\
&\text{and}\\
&\frac{\partial}{\partial t} A_2(t,r) = \gamma_2 I_2(t,r) - \mu A_2(t,r).
\end{aligned}\right\} \tag{6.9}$$

These models have been pursued by many investigators—in particular by Anderson *et al.* (1986), May & Anderson (1987), Blythe & Anderson (1988b) and Hyman & Stanley (1988).

We must still define $\lambda(t,r)$. We discuss below two possible choices: random partner choice and a bias of people towards partners like themselves. Note that now $S(t,r)$ has the units people time/partners, and $I(t,\tau,r)$ has the units people/partners.

6.4.1 *Random Choice of Partners*

If we assume that partners are chosen at random from the entire population, then $\lambda(t,r)$ is given by:

$$\lambda(t,r) = \frac{r\int_0^\infty c(r,r')r' \int_0^\infty \beta(\tau)\, I(t,\tau,r)\, d\tau\, dr'}{\int_0^\infty r\, N(t,r)\, dr}. \tag{6.10}$$

This model, except with no differences in partnership durations and no variability in infectiousness [$c(r,r')$ and $\beta(\tau)$ constant], was first proposed by Anderson *et al.* (1986), i.e.:

$$\lambda(t,r) = \frac{rc\beta \int_0^\infty r'\, I(t,r')\, dr'}{\int_0^\infty r\, N(t,r)\, dr}.$$

The model assumes that the average $(r-r')$ partnership is sufficiently short and infectivity is sufficiently low that the probability that a person has already become infected in the partnership is small, i.e.:

$$\max_\tau \{i(\tau)\}c(r,r') \ll 1.$$

Furthermore, the epidemic cannot grow so fast that the chance that a partner is infected becomes significantly different, during the course of the partnership, from an unmatched person from the same risk group. Anderson *et al.* (1986) show that the initial growth of this model is determined, not by the average number of partners/year, $\bar{r}$, but instead by $(\bar{r} + \sigma^2/\bar{r})$, where σ^2 is the variance about this mean. They then proceed to approximate the model by replacing r with $(\bar{r} + \sigma^2/\bar{r})$.

Thus, for small t, it can be shown that:

$$\left.\begin{aligned} I(t) &\simeq I(0)\exp(\beta c(\bar{r} + \frac{\sigma^2}{\bar{r}} - \gamma)t) \\ \text{and} \qquad t_d &\simeq \frac{\log_e 2}{\beta c(\bar{r} + \sigma^2/\bar{r}) - \gamma}. \end{aligned}\right\} \tag{6.11}$$

Anderson *et al.* (1986) investigate numerically the results from a model based on equations (6.8), that allows for recruitment and heterogeneity of sexual activity and a model that allows for recruitment, variable incubation periods and heterogeneity of sexual activity. A continuous gamma distribution is used to represent the variability of sexual activity within the population: for numerical work this is then approximated by a discrete distribution. The distribution of r is

used to determine the distribution of new susceptible migrants to the various discrete classes. The authors show how the prevalence of HIV in the population and the incidence of AIDS are affected by changes in the value of $\sigma^2/\bar{r}$. In particular, the time at which the maximum incidence of AIDS cases occurs is little affected by heterogeneity of activity. On the other hand, if the initial doubling time of the epidemic, the mean activity rate $\bar{r}$ and the withdrawal rate γ are fixed, then the magnitude of the epidemic decreases as the heterogeneity of sexual activity in the population increases (see May & Anderson, 1987). The intuitive argument for this is that, if σ^2 is large, then the infection spreads quickly through the most highly active individuals, who are relatively rapidly withdrawn into the AIDS or non-infectious seropositive classes. The remaining individuals will then have an average activity that is noticeably lower than the original mean rate $\bar{r}$, so that the epidemic will ultimately affect fewer individuals than in a homogeneous population all of whose members have activity rate $\bar{r}$. Mathematically, of course, if t_d, γ and $\bar{r}$ are all fixed, then increasing the heterogeneity of the population corresponds to decreasing the value of the parameter λ, which represents the probability that a susceptible acquires the infection from an infected partner.

Thus, a slowing up in the rate of increase of the incidence curve is an inevitable consequence of heterogeneity in sexual activity and, when observed, is, therefore, not necessarily a consequence of individuals changing their behaviour (see May & Anderson (1987) for further discussion). This same effect is illustrated by the results of a simulation study described by Peto (1986) using a simple model in which r takes just two values, with a small minority of individuals having a very high activity rate while the majority of the population have a low rate.

In the models described so far, the transmission coefficient, λ, represents the (average) probability that a particular infective transmits the infection to a specific susceptible partner, while the activity rate, r, measures the rate at which the infective acquires new sexual partners. Thus, this probability should depend on the number of acts of intercourse with the susceptible partner; one might also expect it to depend on other quantities whose possible variability has so far been ignored, for example, the infectivity or susceptibility of the individuals concerned. Peto (1986) discusses the effect of changing the model so that an infective has a rate, k, of acts of intercourse, with a fixed probability of ε of transmitting infection to a susceptible partner per act, which is assumed to be the same for all infective-susceptible pairs.

The model may be further detailed by allowing simultaneously for heterogeneity in both r and k and correlations between these factors—e.g. it is possible that, in heterosexual populations, these factors may be negatively correlated; the penalty attaching, however, to models that incorporate greater realism is the increased number of parameters, about whose values and/or distributions there is much uncertainty—thus, here there is little information available on the suitable form of the joint distribution for r and k.

Anderson & May (1988) display a scatter plot of the logarithm of the variance

in reported numbers of different sexual partners, *versus* the logarithm of the mean number of sexual partners from several published and unpublished studies. The variances and means are drawn from studies that utilise a wide range of sampling methods, sample different populations and record numbers of different sexual partners over differing time intervals. Excluding the data from Africa, they find that a best fit to the scatter plot is obtained by a power model of the form:

$$\sigma^2 = a\bar{r}^b$$

where $a = 0{\cdot}555$ and $b = 3{\cdot}231$. This relationship is then built in to a series of projections, which allow for heterogeneity of sexual activity.

Blythe & Anderson (1988b) describe in detail a proportionate mixing one-sex model of sexual transmission of HIV, in which sexual activity (new partners per unit time) is defined as a continuous variable in a set of integro-partial-differential equations, as above. They point out that there are considerable computational problems surrounding the numerical solution of equations like (6.7) or (6.8). An alternative approach proposed by Anderson *et al.* (1986) is to assign individuals to groups depending on whether they had 0, 1, 2, . . ., M partners in a chosen time interval such as one year. The difficulty with this approach is that M is often very large for male homosexual communities (c. 1,000), producing a system of equations which is unmanageably large. Some progress can be made by summation and reformulation to arrive at a small set of equations that denote changes in the moments of the statistical distribution of the rate of partner change (as in Anderson *et al.* (1986)). Blythe & Anderson (1988b) follow a different course and set up a discrete activity-class approximation to the continuous variable model—this is developed by matching equilibrium state and rate variables as closely as possible with the continuous variable model, and consists only of ordinary differential equations. Activity-class boundaries are arbitrary, and each class is characterised by a single level of activity—thus the range of activity, r, is partitioned into N discrete classes with boundaries, r_i, such that:

$$0 = r_1 < r_2 < \ \ldots \ < r_N < r_{N+1} = \infty \, .$$

Given these N classes, the level of sexual activity of $(N-1)$ of them is such that the steady-state susceptible class sub-population is equal to the population in the equivalent section of the continuous model. (Technical reasons prevent the level of activity for all N classes being fixed in this way.) The activity level for the remaining class (or 'balancing' class) is chosen so that the condition for endemicity of the infection (related to the reproductive value—see equation (6.2)) in the approximation, is equal to that for the equivalent continuous-variable model; this minimises errors in the steady state population.

The relationship between the discrete and continuous-variable models is explored via numerical and analytical studies, in order to evaluate the accuracy of the approximation. The numerical studies relate to the U.K. Comparisons are

made using a sequence of increasingly fine partitions of r, beginning with $N = 1$ (homogeneous mixing approximation). For $N > 1$, each class is considered in turn and used as the 'balancing' class. Blythe & Anderson find a consistent and convergent approximation to the continuous variable model for $N = 6$ (with the second highest activity class being used as the 'balancing' class). The particular partition of the continuum of values of r is:

$$r_1 = 0,\ r_2 = 2,\ r_3 = 5,\ r_4 = 10,\ r_5 = 20,\ r_6 = 50,\ r_7 = \infty\ .$$

Among the conclusions arising from their numerical experiments are the following:

(1) The time taken to attain a peak incidence in AIDS varies greatly between the different sexual activity classes.
(2) There are marked differences in the numerical magnitude of the incidence rate in each class, where a high fraction of the higher-activity classes will acquire AIDS while a low fraction of the lower-activity classes will do so.
(3) The absolute incidence of AIDS in the lower-activity classes is higher than in the higher-activity classes, as the system approaches equilibrium, because a greater proportion of the total population resides in the lower-activity classes.
(4) The distribution of sexual activity in the susceptible and infected classes changes as the epidemic progresses, because of the increased likelihood of a more active person becoming infected and dying prematurely of AIDS. Changes in sexual behaviour may thus become apparent in the susceptible subpopulation which do *not* reflect changing habits (as a result of education, for example).

The model first introduced by Anderson *et al.* (1986) and extended later by Blythe & Anderson (1988b) involves what may be called 'proportionate mixing'. It is assumed that individuals choose sexual partners according to their own and their partner's activity class, such that, for any individual in the population, the fraction of partners chosen who have activity r, say, is proportional to r times the proportion of people with activity r in the population. The assumptions lie somewhere between the two extremes of homogeneous mixing (individuals choose partners at random regardless of their activity level) and of partner choice entirely restricted to those with identical activity levels. It has the useful property that the total number of partners chosen *by* people with activity r_1 *from* among people with activity r_2, is equal to the total number of partners chosen *by* people with activity r_2 *from* among people with activity r_1.

6.4.2 *Biased Partner Selection*

The possibility that non-promiscuous individuals may preferentially select non-promiscuous partners (and similarly for the promiscuous) can be allowed for by a suitable choice of the $\lambda(t,r)$ function.

The $\lambda(t,r)$ given by equation (6.10) takes no account of the fact that people do not choose partners at random from all groups, but, instead, prefer partners of a certain type and choose them when available. Ideally, the partner selection in any model should be based on sociological data. As a first step towards addressing this question, Hyman & Stanley (1988) present a model with a strong bias of people toward partners of similar risk behaviour.

They introduce a partnership acceptance function, $f(r,r')$, which represents the frequency that partners of risk r' are accepted by persons of risk r, and take a particular Gaussian form for $f(r,r')$, viz.:

$$f(r,r') = \exp\left[-\frac{(r-r')^2}{8\epsilon r^2}\right]$$

and a particular form for $N(t,r)$, i.e. $N(2\bar{r}+r)^{-4}$, derived from some empirical data quoted by May & Anderson (1987). With these assumptions, an approximate form for $\lambda(t,r)$ is obtained. This is described as a 'diffusion risk model'.

Hyman & Stanley (1988) also present a 'biased mixing model', with contact function:

$$c(r,r') = 1 + (c_0 - 1)\exp(-c_1(r+r'))$$

which only permits individuals to have contact with individuals of similar 'risk behaviour'. c_0 and c_1 are here constants to be estimated from empirical data.

Neither totally random choice nor biased choice solely from neighbouring risk groups captures individuals' true behaviour. In the absence of data, however, it is worthwhile to postulate these two extremes and compare the epidemics that each predicts, but it is also necessary to look at mixtures of the two behaviours.

Hyman & Stanley (1988) compare the two extremes via a series of numerical projections and show that there are considerable differences in the results. Jacquez *et al.* (1988) have used a linear combination of the two extremes to examine the transition from pure random selection to pure self-selection, using a model with four discrete activity levels. They see a large difference in epidemic growth rates, the time to spread across the different activity groups, and the endemic state when the pure self-selection term dominates (over 90%).

A number of other investigators have considered the effects of biased partner selection.

Thus, Colgate *et al.* (1989) employ models similar to those introduced by Hyman & Stanley (1988) in order to understand why the early growth of AIDS in the U.S.A. can be represented by a polynomial function of time, rather than an exponential function (see results quoted in Section 4.1). Hyman & Stanley find that cumulative cases of AIDS grow as a cubic function of time, leading to a decreasing growth rate. Colgate *et al.* (1989) use a biased mixing model for male homosexuals that reproduces this initial growth curve on the assumptions that the level of risk behaviour is distributed as r^{-3}, that either new partner frequency or sexual contact frequency dominates the risk behaviour separately (or when both are positively correlated) and that the probability of conversion to AIDS

per unit time is approximately constant (i.e. a constant hazard rate for the incubation period distribution, $\gamma(\tau)$). Several results of the model agree with observations, e.g. decreasing risk behaviour with time, numbers infected, decreasing growth rate. The authors comment that a decreasing doubling time may not reflect efficiency and efficacy of education and a decreasing participation in risk behaviour, but may be the natural consequence of such a biased mixing model, in which the virus is transmitted from high-risk to low-risk groups over time. The assumption regarding the correlation between new partners and sexual contacts and regarding the incubation period distribution are somewhat crude and are critical as regards the use of the model for extended projections.

Roberts & Dangerfield (1988) describe a simulation model of the spread of AIDS among homosexual males, parametrised with respect to available U.K. data and based on the paradigm of system dynamics. The model is used as a tool for exploring the relative effects of varying scenarios. Despite their reference to system dynamics, the equations used correspond to the difference equations mentioned earlier. Like Longini *et al.* (1989), they effectively use an Erlang distribution (type 3) for the incubation period. The model allows for the incorporation of varying parameter values (e.g. proportion of HIV infecteds contracting AIDS, average time to death of newly diagnosed AIDS cases), disaggregation of the susceptible population in order to reflect heterogeneity in sexual behaviour, changes in number of sexual partners, scenarios concerning anti-viral therapy (e.g. a vaccine for seronegative members of the at-risk population, a drug to reduce the infectivity of seropositive members of the at-risk population). Dangerfield & Roberts (1990) extend the model to allow for a variable infectivity profile and describe a parameter optimisation experiment in relation to the quarterly time series data on cumulative reported cases of AIDS in the U.K.

Stigum *et al.* (1988) investigate the effect of selective choice of partners on the spread of AIDS in a male homosexual population. The model follows a simplified form of equations (6.7). The development of the HIV epidemic is simulated, under various assumptions with regard to partner choice, in a population structured in groups according to promiscuity. Three scenarios are compared: a homogeneous population, a group structured population with random partner choice; and a group structured population with positive selection of partners within the same group.

The model describes changes over time in the prevalence of HIV infection in groups of sexually active individuals. The model contains biological parameters (transmission rates, incubation periods and fractions of the infected who develop AIDS) and population parameters (group sizes, frequency of intercourse, frequency of partner change and preferences in choice of partner).

The following simplifying assumptions are used:

(1) The number of individuals in each group is constant over time except for death due to AIDS.

(2) A fraction of the infected individuals develop AIDS after σ years and are no longer contagious.
(3) The transmission rate is independent of the time since infection.
(4) Sexual behaviour does not change over time.
(5) Within each group, behaviour is described by average parameters for the group.

The rationale behind assumption (1) is that the majority of the infected individuals belong to an age group with low mortality (from other causes than AIDS). Other causes of mortality are not considered, nor are other factors that may change group size. Assumption (2) describes a simplified incubation time distribution (with density function given by the Dirac generalised δ-function: $\delta(t-\sigma)$). Assumption (3) is not necessarily realistic, but represents a straightforward assumption in the absence of precise knowledge. The effects of behaviour changes are not considered (assumption (4)). A population with large variance in sexual behaviour will, in this model, be subdivided into groups so that each group is reasonably described by its average (assumption (5)).

These assumptions simplify the model beyond the point of realistic predictions, but make it a useful tool for investigating the effects of different phenomena (such as long short incubation times, group structure, selective partner choice, age structure). In particular, it should be noted that assumption (2) would not take into account the wide variability in duration of infections preceding AIDS. It is necessary to allow for this factor in order to predict, with some accuracy, the correct distribution of people developing AIDS and to ensure that infected people in the model remain infectious for lengths of times that reflect the actual infectious periods. A person who is healthy, but infected for a long period, has a higher probability of infecting someone else than a person who develops AIDS relatively early.

Stigum *et al.*'s study shows that group structure and selective partner choice need to be considered in a model describing the spread of HIV, and that *without* such information model predictions may be unreliable, even when other parameters are known (or can be estimated reliably).

Data on this type of selective choice of partners are difficult to obtain directly from populations. In some instances, they may be obtained indirectly, if both promiscuity and choice of partners correlate with a third variable that is more easily assessed. An example of this may be age: promiscuity tends to vary with age, and there may be strong positive selection for partners within the same age group.

In a different type of study, Eisenberg (1989) analyses the probability of contracting the AIDS virus in relation to the number of sexual contacts and the number of different partners. Using the basic mathematical theory of probability, it is shown that in the case of a *fixed* number, n, of sexual contacts there is the following ranking from least risk to greatest risk:

(1) monogamous relationship with a non-infected partner,

(2) monogamous relationship with a randomly selected partner,
(3) relationship with more than one randomly selected partner,
(4) n randomly selected partners, and
(5) monogamous relationship with an infected partner.

All of these values are estimated to be much higher 'on the average' for the male homosexual population than for the heterosexual population.

6.5 *Institute of Actuaries Working Party Model*

The Institute of Actuaries Working Party, of which the author is a member, has presented a mathematical model representing the transmission of HIV and the spread of AIDS. The model has been widely published and full details can be found in Daykin *et al.* (1988), Wilkie (1988, 1988a, 1989).

The Institute of Actuaries Working Party model is of a similar type to the models described earlier in Sections 5 and 6. However, because of the emphasis on applying this model to assessing the effect of HIV and AIDS on life insurance underwriting, life insurance premiums and reserving (as well as permanent health insurance and pension provision), the focus of this actuarial model has been different. The model follows on from the terms of reference of the Working Party; these include:

> "To show the potential impact of HIV on mortality and morbidity and the implications for the use of existing actuarial bases and standard tables for premium rating and reserving."

Actuaries require such a model to be age-specific, in order to consider the progress of individuals of a given age and sex through future calendar years, to consider the longer-term trend in transmission and to produce numerical results (although not necessarily by analytical means). Thus, equilibrium models would be of less interest. It is also important for the model to reflect the type of data that would normally be available to an insurance company.

For the above reasons, the Working Party's model is age-specific and the resulting numerical complexity has meant that elements that depend on detailed assumptions about sexual behaviour (as described in Section 6.4) have been avoided.

The model belongs to the family of stochastic processes that have been introduced by others, but it addresses male homosexuals only. Each cohort (of a single age) is dealt with independently of other cohorts. It is assumed that infection occurs from a contact between two individuals within a single age group. This assumption is artificial, but, if infections between those of different ages balance out, it may be considered to be a reasonable representation of reality. The transition intensities between states are allowed to vary with attained age and time. The model allows for immigration of susceptibles and for normal mortality as well as extra mortality from AIDS.

A further simplification is the assumption that all those males described below as being 'at risk' of infection behave in the same manner at any one time, so that the chance of infection depends on the age of the individual at risk and the

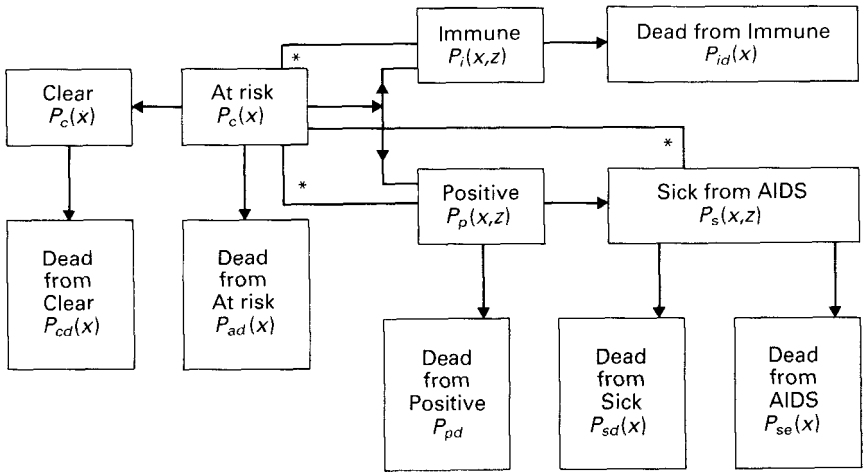

Figure 2. Institute of Actuaries AIDS Working Party model: States and Transitions (*denotes possible infection).

particular calendar year, but not on any sub-division according to frequency of sexual contact or frequency of change of sexual partner.

The members of one cohort at age x may be in any one of the eleven discrete states that are indicated in Figure 2. Five of these are live states: 'clear'; 'at risk'; 'immune'; 'positive'; and 'sick' from AIDS. Six are dead states; these are kept separate simply to show the live state that someone died in. The dead states are: 'dead from clear'; 'dead from at risk'; 'dead from immune'; 'dead from positive'; 'dead from sick' (other than from AIDS); and 'dead from AIDS'. It may not be possible to distinguish the last two categories, but calculated deaths, other than from AIDS, of those who suffer from AIDS are comparatively trivial.

Those in the clear state are those whose assumed sexual activity is such that they run no risk whatever of becoming infected with HIV. They form the 'normal' pre-AIDS population for comparative purposes. Those in the at risk state are treated as exposing themselves to the risk of acquiring HIV infection by reason of sexual contact with infected people. Those in the immune state are assumed to have acquired HIV infection and to be infectious, but to be wholly immune from becoming sick from AIDS or dying from AIDS.

Those in the positive state are HIV seropositive, but not yet sick from AIDS; they are infectious and not immune. It is assumed that it is possible to distinguish between those who are HIV seropositive and those who are sick from AIDS. In reality, there are several stages in the transition from HIV infection to death from AIDS. Those who are suffering from AIDS are thought to be highly infectious, but it is possible that their sexual activity may be considerably reduced. The model makes it possible to choose whether those sick from AIDS are treated as contributing to further infections or not.

It is assumed that the current age is part of the status, and that transition intensities can all vary by current age. In addition, since each age cohort (or year of birth cohort) is treated separately, each transition intensity can also be varied by calendar year, so that each cohort has its own set of transition intensities.

Duration since entry to the states, immune, positive and sick from AIDS are also relevant to the transition intensities. This duration is denoted in each case as z.

Possible transitions are as shown in Figure 2. Those in any of the live states may die, and those who are sick from AIDS may die from AIDS or from causes other than AIDS. Those who are at risk may change their behaviour and become clear, for example, by giving up sexual activity altogether, or by restricting themselves to one equally monogamous partner. There is no representation in the model of transfer from clear to at risk. Those who are at risk may become infected, and at that point are immediately allocated either to the immune state or to the positive state, in proportions that may depend on age (and on calendar year, though it seems unlikely that this would actually exercise any influence).

Those in the positive state may become sick from AIDS, if they do not die first. Infection is possible from the immunes, positives and sick to the at risk.

A series of ordinary and partial differential equations (in the form of equations (6.6)) are then set up and solved by numerical means, given assumptions about the form of the transition intensities. Various sets of assumptions have been explored and the sensitivity of the results tested. For example, in their Bulletin No. 4, Daykin *et al.* (1989) consider the effects of:

(1) a transition intensity from 'at risk' to 'clear' that varies with time,
(2) a transition intensity from 'at risk' to 'positive' that varies with age, time and duration since becoming infected,
(3) a transition intensity from 'positive' to 'sick' that varies with duration in the 'positive' state, and
(4) a transition intensity from 'sick' to 'dead' that varies with age and time.

The force of mortality from the 'clear', 'at risk' and 'positive' states is represented by a current level of population (England and Wales) mortality. The transition intensity from 'positive' to 'sick' is the hazard rate for the incubation period distribution—Gompertz and Weibull forms have been used in experiments:

$$\left.\begin{aligned}\gamma(\tau) &= \text{Min}[\exp(-a + b\tau), c]\\ \gamma(\tau) &= ab^a\tau^{a-1}.\end{aligned}\right\} \tag{6.12}$$

Other critical assumptions needed in this model are the proportion of the starting population assumed to be at risk or positive and the proportion of future new entrants (at age 15) assumed to be at risk or positive.

The model assumes, through the equations describing transmission from the infected to the susceptible, homogeneous mixing as far as sexual activity is concerned. However, it should be noted that the 'clear' group could be said to have zero new sexual partners per unit time ($r = 0$) while the 'at risk' and 'positive' groups have $r > 0$. Allowing the transition intensity from 'at risk' to

'positive' (via infection) to vary downwards over time, allows the model to attempt to mimic the rate of spread of the infection from the very promiscuous to the less promiscuous, as would happen in a heterogeneous (real) population.

The model has proved to be both flexible and practicable and very useful for its intended purposes of assisting actuarial applications and for providing projections for the total population. However, it should be pointed out that the model's failure to allow for epidemiological features of the spread of HIV and AIDS (in particular, the heterogeneity of sexual activity in the population) means that it is unlikely to produce forecasts (in terms of cases or deaths) that will be exactly fulfilled. However, it does serve the purpose of providing a benchmark set of forecasts.

The separate treatment of age cohorts contains the major restriction that infection can only be transmitted within cohorts and not between them. It would be possible to augment the model to permit some assumption about mixing of partners between cohorts—the complications caused to the numerical solution of the resulting equations have meant that this is yet to be pursued.

6.6 *Allowing for Variable Infectiousness*

As Daykin (1990) notes, there are suggestions in the literature of a pattern of two peaks in the antigen concentration in the serum of infected persons and that this may be mirrored in a pattern of variable infectiousness with similar peaks (Anderson & May, 1988).

The inclusion of variable infectiousness in models has been considered by Blythe & Anderson (1988a) and Hyman & Stanley (1988) in detail and by Daykin *et al.* (1989).

Both Blythe & Anderson (1988a) and Hyman & Stanley (1988) base their model of fluctuating infectiousness during the long and variable incubation period on the simple homogeneous mixing model in a male homosexual community as described in Sections 6.2 and 6.3. Hyman & Stanley (1988) represent variable infectivity by a piecewise linear approximation and then use numerical techniques to carry out the evaluation of the necessary convolution integrals. Blythe & Anderson (1988a) divide the infected population into a series of subclasses with different levels of infectivity (fixed for each subclass) and different durations of occupancy (exponentially distributed)—this corresponds to considering a generalised Erlang distribution for the incubation period distribution, as described in Section 6.3.1. Blythe & Anderson also adopt a second approach, which is more mechanistic in character and is based on an attempt to relate changes in viral abundance within an infected person to the duration of the incubation period and hence the likelihood that the AIDS develops. Variable incubation is induced by variation in the rate of change of viral abundance in the infected population.

Numerical projections of changes in the incidence of AIDS through time, generated from both types of model, are compared with projections based on the assumption of constant infectivity throughout the incubation period of AIDS.

Blythe & Anderson's models, in particular those with parameter values reflecting two peaks in infectiousness, show how temporal variation in infectivity can influence the qualitative shape of the AIDS epidemic. In broad terms, the initial phase of infectiousness will tend to drive the early doubling time of the epidemic (i.e. the rate of increase in the incidence of the disease), while both episodes of infectiousness will determine the overall magnitude of the epidemic and the level of the endemic equilibrium state. The models also show how much variability influences the derivation and interpretaton of estimates of the basic reproductive rate, R, of infection. If infected persons are infectious for only a fraction of the incubation period, then past estimates of the transmission probability must be revised upwards, to take account of the shorter duration of infectivity. In other words, it may be that infectiousness is of shorter duration, but of greater intensity, than has previously been envisaged.

The impact of variable infectivity on the possible spread of HIV infection could be considerable (Daykin, 1990). The initial stages of the epidemic would be determined by the early phase of infectiousness, and some levelling out could be expected before increasing numbers of the infections begin to occur, as a result of the second phase of infectiousness. This pattern will tend to be blurred by the variability in the incubation period, but the general effect will be to slow down the rate of new infections after the initial surge. Whether or not a further increase in new HIV infections will be observed, as a result of more people entering the second phase of infectiousness, will depend on whether or not behavioural change is sustained.

Given the limitation in our quantitative understanding of variability in infectiousness, it is likely that Blythe & Anderson's first (and simpler) model, based on a multi-stage classification, would currently be of more value. It permits a certain degree of analytical investigation (e.g. the determination of the probability distribution of the incubation period of the disease, in terms of the number of infected substages, and the parameters controlling the rates of movement of individuals between them), and its structure facilitates numerical study. At present, Blythe & Anderson believe that three subclasses ($N = 3$) are probably sufficient to capture current epidemiological understanding, with high, low, and high infectivity, respectively, as an individual follows the progression from infection to AIDS (but using the six stages of the Walter Reed Staging System would be an obvious elaboration). Whichever model proves itself to be the more valuable in the future, it will be necessary to allow also for heterogeneity of sexual activity as well in the modelling (Section 6.4) in order to produce a more refined analysis and improved understanding of the dynamics of the epidemic.

One further point, that does emerge from these preliminary investigations, is the limited nature of current understanding of the factors that control the length of the incubation period in different patients and the infectiousness of infected persons throughout the incubation period of the disease.

6.7 *Models Allowing for Heterosexual Spread*

The models for the transmission of HIV infection, described in previous

sections, have been devised with a male homosexual population in mind, in which the infection is spread by sexual contact and in which any infective member of the population can pass on the infection to any susceptible member. For transmission between heterosexuals, the model has to be made more complicated, since there are two groups of individuals with the infectives in one group spreading the infection to the susceptibles in the other group. If we denote the numbers of female susceptibles and infectives at time t by $FS(t)$ and $FI(t)$, with $MS(t)$ and $MI(t)$ being the corresponding quantities for males, then in the analogue of the basic homogeneous mixing model (described by equations (6.1)), the following would hold:

$$\left.\begin{aligned}
\frac{d}{dt}MS(t) &= -\lambda_1(t)MS(t) \\
\frac{d}{dt}MI(t) &= \lambda_1(t)MS(t) - \gamma^M MI(t) \\
\frac{d}{dt}FS(t) &= -\lambda_2(t)FS(t) \\
\frac{d}{dt}FI(t) &= \lambda_2(t)FS(t) - \gamma^F FI(t)
\end{aligned}\right\} \tag{6.13}$$

where

$$\lambda_1(t) = \frac{\beta^F C^F r^M FI(t)}{FN(t)}$$

$$\lambda_2(t) = \frac{\beta^M C^M r^F MI(t)}{MN(t)}$$

$$FN(t) = FS(t) + FI(t)$$

and

$$MN(t) = MS(t) + MI(t).$$

In the above, the mean incubation periods $(1/\gamma^M)$ and $(1/\gamma^F)$ and the transmission coefficients $(\beta^i c^i \ i = M,F)$ for the spread of infection from male to female and from female to male are assumed to be different.

May & Anderson (1987) suggest that $\beta^F c^F < \beta^M c^M < \beta c$, the corresponding coefficients for homosexual males. Similarly the rates of partner change r^M and r^F are likely to be different and much smaller than the homosexual rate of partner change.

However, the actual numbers of partner changes for the two groups must be exactly the same (assuming that there are no other groups of potential partners) and so, if there are approximately equal numbers of males and females (that is, if $MN = FN$), then necessarily $r^M = r^F$. The reproductive rate for female → male infections is $\beta^F c^F r^F/\gamma^F$, while that for male → female infections is $\beta^M c^M r^M/\gamma^M$. Whether or not the epidemic is self-sustaining (by purely heterosexual contact) in

the heterosexual community depends on the product $(\beta^F c^F r^F \beta^M c^M r^M / \gamma^M \gamma^F)$ which represents the average number of females infected indirectly, via a male intermediary, by a single female infective (or, equivalently with female and male interchanged) at the start of the epidemic when all contacts are with susceptibles. It appears likely that the reproductive rate $\beta cr/\gamma$ of the infection in the male homosexual community is appreciably greater than 1, but much less is known about heterosexual transmission. If we assume that $\gamma^M = \gamma^F = \gamma$ then it is crucial to know the relative magnitudes of β^i and c^i as compared with β and c and of r^i as compared with r, in order to predict whether or not a purely heterosexual epidemic will occur.

In equations (6.13), only the spread of HIV infection is modelled. To extend these equations to include the incubation period distribution of AIDS, non-infectious seropositivity, together with immigration of susceptibles and natural mortality and heterogeneity of sexual activity, extra variables and parameters must be introduced, as in the more complex models of Sections 6.3 and 6.4: the mathematical details are not pursued here.

It is of interest to investigate the effect on the heterosexual community of the current epidemic in male homosexuals, as bisexuals transfer infection between the two groups, and for this the model must be extended still further. The model described by Knox (1986) has a population made of 12 groups or 'behavioural classes', each having its own rate of partner change. It includes classes of female prostitutes and bisexual males and separates heterosexual men and women and homosexual men into promiscuous and non-promiscuous classes. Inevitably, such a model involves many parameters, about which little is known. Knox discusses the information which is available and suggests sets of suitable parameter values, which are then used to estimate the incidence and prevalence of HIV infection when the system is in equilibrium. In this process, assumptions are made that various parameters are constant over time, which is restrictive—this particularly applies to the mortality rate and the rate of change of sexual partners.

Perhaps of more immediate concern while the AIDS epidemic, particularly among heterosexuals, is in an early stage is the transient process. A simulation study, concentrating on the German (Federal Republic) population, is reported by Kiessling *et al.* (1986) and Stannat *et al.* (1987). In this study, there are six groups: bisexuals; female prostitutes; homosexual males; and heterosexuals (no distinction is made between the sexes), where the last two classes are each divided in two by the rate of sexual activity. Again, suitable parameter values have to be assumed. For example, it is assumed that only the bisexuals and higher activity rate heterosexuals have contacts amongst the prostitutes. Those infectives who do not develop AIDS are assumed to remain infectious indefinitely. One difference from the models previously described is that although, as before, the newly infected are divided into those who will develop AIDS and those who will remain seropositive, there is also a constant drift from the seropositive class across into the AIDS class. The aim of this is to make some allowance for the non-exponential nature of the incubation period distribution.

Another exploratory simulation study is described by Gonzalez *et al.* (1987) (referred to in Section 6.3.1), in which individuals are divided by age and sexual preference and a class of intravenous drug users is included, and where a random (non-exponential) incubation period is incorporated. The study investigates the effects of changes in parameter values on the growth of the epidemic.

The above formulation of the model, leading to equation (6.13), follows that for the homosexual spread of the epidemic. The presentation focuses on the risk of transmission per partner rather than per sexual contact. The implications of switching the model to examine the latter would be dramatic for the heterosexual spread of the epidemic. Thus, it should be more likely to pass on the infection with one partner over n years than with a single contact. The pattern of heterosexual sexual activity among young adults could be that someone lives with a person for a year and then changes partners and lives with someone else for a year and so on—the risk of infecting each partner may then be high relative to a promiscuous person who may have 10 different partners occasionally and hence have fewer sexual acts. However, the model may be extended to allow for this finer level of description.

Farmer & Emami (1987) describe a simulation model for the heterosexual transmission of HIV and the evolution of the AIDS epidemic. The population is considered to be made up of four groups: homosexual males; bisexual males; heterosexual males; heterosexual females. The model allows for biased selection of partners. Given somewhat arbitrary estimates of various parameters, including: the probability of transmission of HIV from group i to group j; and the average number of sexual encounters of a person in group i per unit of time, Farmer & Emami produce graphs of the trend over time in the level of seroprevalence, the numbers of new cases of HIV arising and so on. The graphs depict clearly the likely pattern of a wave of cases among homosexual and bisexual men, followed by a wave of heterosexual cases.

This model is relatively simplistic. It does not allow for inflow of susceptibles; mortality of infecteds; movement of individuals into and out of the sexually active pool; presence of prostitutes or IV-drug users in the population and for transmission of HIV via needle-sharing; heterogeneity of sexual activity and behaviour.

Lorper (1987, 1989) sets up a model for the spread of AIDS in the Federal Republic of Germany, to assist with the actuarial discussions within his employer, a reinsurance company. The model projects new infections, new AIDS cases, deaths and so on for the sexually active population. The model is *not* age specific. It allows for eight risk groups within the general population, viz.:

Heterosexual females (not drug addicts or prostitutes),
Prostitutes, not drug addicts,
Intravenous drug-addicted prostitutes,
Other intravenous drug-addicts (inclusive of homosexual drug addicts),
Heterosexual males with no prostitute contacts,

Heterosexual males with prostitute contacts,
Bisexual males, and
Homosexual males.

It allows for heterosexual spread of the disease and random selection of partners, but it does not allow for biased selection. A series of difference equations are used, with the time variable measured in days, and simulations are run once the various transition probabilities have been estimated.

Kanouse *et al.* (1988) set up a simulation-based model which incorporates heterosexual and IV-drug sharing transmission of HIV. The model classifies the resident population of the U.S.A. using two sets of categories. The first set describes key demographic, behavioural, and other characteristics than can be used to stratify the population according to the risk of acquiring or transmitting HIV infection. People are classified according to their gender and sexual behaviour (five categories), IV needle sharing status (two categories), region (ten categories), and age group (nine categories). Thus, the model contains 900 risk categories that are based on combinations of factors considered to be important in the epidemiology of HIV infection.

By focusing on transmission through homosexual and heterosexual intercourse and the sharing of intravenous needles, the model addresses the dominant transmission modes, which account for at least 95% of the adult cases reported to date in the U.S.A.

The second set contains epidemiological and clinical categories that describe serological status and, for those infected, stage of disease. The model distinguishes 13 stages, which are treated as mutually exclusive and exhaustive. Once infected, individuals move through these stages probabilistically and unidirectionally, at rates that are estimated from cohort studies. In the initial stage of infection (Stage 2), individuals have not yet developed antibodies, though they are capable of infecting others. In Stages 3–10, they have developed antibodies but have no clinical symptoms. The successive asymptomatic stages can be thought of as representing a progressive deterioration in the underlying state of the immune system, much like the deterioration represented by the categories in the Walter Reed Staging System. When HIV-infected persons first develop symptoms, their symptoms may be ones that immediately classify them as having AIDS by the CDC definition (Stage 12), or they may develop other HIV-related symptoms, including those often referred to as AIDS-related Complex or ARC (Stage 11). The final stage, death from HIV-related causes, may occur as a result of AIDS, ARC, or other conditions that are induced or complicated by HIV infection.

The spread of HIV infection over time is tracked by computing, for each month, the number of people in each risk category who shift their epidemiological status from uninfected to infected. These computations require estimates of the values of key epidemiological and behavioural parameters, describing how many people engage in behaviours that could expose them to

HIV and how likely it is that these behaviours will actually result in transmission. The estimated values differ across specific populations, reflecting differences in both behaviour and exposure.

The simulations presented illustrate how the virus spreads from the risk groups and regions that serve as initial 'epicentres' to other risk groups and regions with which they are linked. In the process, the epidemic growth gradually slows—a feature that may be of particular relevance to recent developments in a number of western countries.

The base case version of the model assumes, as a plausible hypothesis, that infectivity through sexual transmission increases exponentially over the course of infection. Results demonstrate that this hypothesis is consistent with the known history of the epidemic. If true, the hypothesis implies much slower future decline in the rates of growth in sexually—and especially heterosexually—transmitted cases than would otherwise occur. But this is not the only hypothesis or associated future scenario that fits available facts. If infectivity follows some other time pattern, such as a sharp initial 'spike' shortly after the initial infection, followed by a long period of low infectivity, and then a substantial increase in the late stages of infection, the implications for future spread and control are quite different.

Other key parameters, whose values have an especially important bearing on the future course of the epidemic, include (1) the extent of population heterogeneity in infectivity and susceptibility; (2) the degree of asymmetry in the efficiency of male to female *v.* female to male transmission; and (3) possible sex differences in natural history. If women transmit the virus to men less efficiently than men do to women, that is 'good news' from an epidemiological standpoint, because any weakness in the chain of transmission slows the overall rate of spread. If, on the other hand, women are slower than men to progress through the stages of infection, this is not entirely 'good news', because it implies that more women are now infected than might otherwise be inferred from the number of reported cases, and, because it means that there is a longer delay in the transmission chain, warning of rapid growth in the future.

Plumley (1989) presents a discrete-time model of the progression of the AIDS epidemic in the U.S.A. It is developed by projecting forward the spread of the epidemic within various major categories of the population, based on the sexual and intravenous drug activities that cause its transmission from one person to another. Hence, it is more comprehensive than the models of Bongaarts (1989) and Gail *et al.* (1989) described below.

The model begins with a population in 1981 and an assumed number of persons infected with HIV at that time. Using assumed frequencies of the various types of sexual and IV-drug behaviour, combined with probabilities of risk of infection from these acts and rates of progression from infection to AIDS, the number of AIDS cases and other relevant data can be calculated year by year. Next, the number of modelled AIDS cases for each year through 1988 are compared with known data, in order to validate the assumptions. It then is

possible to project the epidemic based on any desired assumptions as to future sexual and IV-drug behaviour.

Plumley's model would not accurately depict the spread of the epidemic for various subgroups of the U.S.A. population—this restriction is imposed by the desire to avoid extra complications and also by the availability of data for estimating (or guessing) parameter values.

Among the simplifying assumptions that Plumley adopts are the following:

(1) Age-specific effects can be ignored. The ageing of the population and inflow of new cohorts of young adults are not allowed for, although the model does incorporate a global population growth factor.
(2) Within each racial group, it is assumed that the various types of sexual and IV-drug activity are not segregated by geography, social class or other strata such as degree of promiscuity.
(3) It is assumed that no interracial sexual activity or sharing of IV-drug needles takes place.
(4) No specific allowance is made for prostitution, in particular among the IV-drug using community.
(5) No explicit allowance is made for heterogeneity of sexual/drug activity within each population subgroup considered.
(6) It is assumed that homosexual men engage in sexual activity only with other homosexual and bisexual men.
(7) Bisexual men are assumed to engage in sexual activity with other bisexual men, with homosexual men, and with heterosexual women; however, it is assumed that their risk of infection comes only from other men, thus ignoring the much smaller risk of becoming infected from engaging in vaginal sex with an infected woman.
(8) IV-drug users are assumed to risk infection only from engaging in needle-sharing with other IV-drug users, thus ignoring the much smaller risk of becoming infected from vaginal sex.
(9) Those homosexual or bisexual men who are also IV-drug users are split between those two categories, so that the IV-drug users have a risk only from IV drugs, and the homosexual and bisexual men have a risk only from receptive anal intercourse. This ignores the possible increased susceptibility to the AIDS virus because of the combination of anal intercourse and IV-drug use.
(10) For years prior to 1987, it is assumed that persons with AIDS are just as likely to engage in sexual or needle-sharing activities as are those who were HIV positive but who had not actually contracted AIDS. However, beginning in 1987, it is assumed that public education has progressed to the point that persons who actually have AIDS would no longer be having any sexual or needle sharing activity unless their sexual or IV-drug partner were fully protected from acquiring the disease.
(11) For years prior to 1987, it is assumed that heterosexual men engage in

sexual activity at random with heterosexual women, both IV-drug users and others. However, for years beginning with 1987, a 'discrimination factor' is added to the formula, so that the model could measure the effect of various degrees of avoidance of sexual activity with female IV-drug users.

(12) For years prior to 1987, it is assumed that heterosexual women engage in sexual activity at random with heterosexual and bisexual men, including IV-drug users. However, for years beginning in 1987, two 'discrimination factors' are added to the formula. One of these permits the model to measure the effect of various degrees of avoidance of sexual activity with male IV-drug users, and the other allows the model to measure the effect of various degrees of avoidance of sexual activity with bisexual men.

(13) For years prior to 1987, it is assumed that persons generally do not have knowledge of their HIV status, and therefore that, in choosing a sexual or IV-drug partner, one would not be able to exclude anyone because he or she were infected. However, for years beginning in 1987, a 'knowledge factor' is added to the formulae for all categories, to permit the model to reflect the availability of information on HIV status.

Plumley's approach is based on the discrete-time difference equation version of the continuous-time partial differential equations described above. For each racial group (white, black, other), difference equations are set up for estimating:

(i) the number of new infections from homosexual activity for the racial group in year n;
(ii) the number of new infections from bisexual activity for the racial group in year n;
(iii) the number of new infections for males from IV-drug activity for the racial group in year n;
(iv) the number of new infections for females from IV-drug activity for the racial group in year n;
(v) the number of new infections for male drug-free heterosexuals from heterosexual activity for the racial group in year n; and
(vi) the number of new infections for female drug-free heterosexuals from heterosexual activity for the racial group in year n.

The mode of transmission in (i) and (ii) is via unprotected receptive anal intercourse; in (iii) and (iv) is via unprotected IV-drug activity (i.e. needle sharing); in (v) and (vi) is via unprotected penile-vaginal sexual activity.

Plumley's paper provides a very useful section on sensitivity analysis, in which the sensitivity of the results to changes in parameters and assumptions is discussed—this includes variation in the rate of progression from HIV seropositive to AIDS, undercounting of AIDS cases, variation in the size of the high-risk groups, variation in the rate of population growth, consideration of the use of 'average' number of sexual and IV-drug acts per year, measurement of the risk of sexual activity, variation in HIV seroprevalence by race. Finally, he notes that,

because his model fails to segment the population by degree of promiscuity, geographic reason or other characteristic, it is likely to *overstate* the number of HIV infections, since it does not take into account, to some degree, the potential saturation effect of the epidemic becoming contained within a particular subgroup.

Although the approach of Kanouse *et al.* (1988) and of Plumley (1989) is useful for the assessment of different 'what if?' scenarios, it suffers from the clear disadvantage of having a very large number of parameters which need to be specified.

A model with fewer parameters would clearly be preferable: a model that is more parsimonious as regards numbers of parameters is that due to Bongaarts (1989).

Bongaarts (1989) introduces a model that describes the transmission of HIV infection by different modes, including homosexual and heterosexual sexual contact and transmission from infected mother to newborn baby. The model is based on a subdivision of the population into compartments and leads to a set of linear differential equations—as in equations (6.8) and (6.13). The model proceeds by computer-based simulations and its objective is to project, for periods up to one or more decades, the annual incidence and prevalence of HIV infection and AIDS in a population with given epidemiological, behavioural and demographic characteristics. In addition, the epidemic's impact on a range of demographic variables is calculated. The demographic framework in which the model operates is based on the standard, cohort dependent, component method of population projection. The population is stratified by age, sex, sexual behaviour, marital status and infection/disease status. An important point here is the inclusion of the age variable, which enables the model to allow for age-specific dependence in a number of the critical assumptions.

To take account of heterogeneity in sexual activity, each cohort is divided into the following strata:

Males

(1) homosexuals (including bisexuals);
(2) heterosexuals with high sexual mobility (for example, clients of prostitutes);
(3) partners in monogamous unions; and
(4) sexually inactive males.

Females

(1) monogamous partners of bisexual males;
(2) women with high sexual mobility (for example, prostitutes);
(3) monogamous partners of sexually mobile males;
(4) partners in monogamous unions; and
(5) sexually inactive females.

As a short cut to modelling the incubation period distribution, the number

of infected individuals at risk of developing AIDS at time t is subdivided into four states with constant transition intensities from (a) → (b) → (c) → (d), and zero transition intensities in the opposite direction (as in the work of Panjer (1987, 1988), Longini *et al.* (1989), De Gruttola & Mayer (1988), discussed in Section 6.3.1). The states used by Bongaarts are:

(a) asymptomatic, normal immune function;
(b) asymptomatic, impaired immune function;
(c) persistent generalised lymphadenopathy (PGL); and
(d) AIDS-related complex (ARC).

The distribution of the incubation time is given by the convolution of four exponential distributions (which yields a gamma distribution if the transition intensities are taken to be equal—which is the path chosen by Bongaarts).

The model allows for the primary mode of transmission being sexual and different, but closely related equations are used to estimate infection rates for groups with multiple partners and frequent partner change, and for groups with single partners and infrequent partner change.

The model can allow also for other transmission routes:

(1) infected women who become pregnant can infect their newborns *in utero* or at birth or via breastfeeding;
(2) sharing of contaminated needles among IV-drug abusers;
(3) blood transfusions with infected blood; and
(4) medical injections with contaminated needles (e.g. for vaccination).

Bongaarts provides an illustrative application of the model to a Central African population with *no* homosexuality and *no* IV-drug abuse. In this hypothetical simulation covering the period from 1975 to 2000, HIV prevalence in the adult population rises from 0% to 21%. By the end of the projection period, mortality is about double the level that would have prevailed in the absence of the epidemic, but, owing to the very high birth rates that prevail in most of Africa, the growth rate of the population remains substantially positive, although reduced from 3% p.a. to 1·9% p.a. The total number of parameters required for this simulation is 32.

Gail *et al.* (1989) use discrete-time epidemic models (with an allowance for migration) of the form discussed in Sections 5 and 6, to evaluate the potential benefits of voluntary confidential testing (VCT) for HIV. The discrete-time models rely on the use of transition matrices and are closely related to the continuous-time models that have been presented here (however, it should be noted that the long-run behaviour of discrete time and continuous time models may differ: Gani (1978)).

Gail *et al.* consider the number of tests required to prevent one case (called the economic ratio, ER) and calculate the number of cases prevented by the screening programme. The methods differ from the epidemic models described earlier, because susceptibles and infecteds are subdivided according to testing

status, in order to permit modelling of the efficacy of knowledge of HIV status in retarding epidemic spread. The methods also allow for subdivision into several homogeneous subpopulations, whose members may be attracted in varying degrees to members of other subpopulations. Thus, the common assumption of free mixing across subpopulations is relaxed.

The models make a number of assumptions. They assume that the forces of mortality from susceptibles and infecteds are constant (a common assumption); they make no allowance for age-specific effects; and they assume that the probability of transmission is independent of how long an infected subject has been infected. There is evidence (from the experience of cohorts of homosexual men and haemophiliacs) that the hazard rate for AIDS (and hence for mortality) increases with the duration of infection and (from the experience of cohorts of haemophiliacs) that the risk of transmission may also increase with duration of infection—see Daykin's (1990) review paper for further comments on the epidemiology of AIDS. The models of Gail *et al.* could be generalised to accommodate these features, by subdividing the infected state into additional categories, defined by duration of infection.

The model could also be extended to allow for distinct modes of transmission like sharing of IV-drug equipment. Thus, subgroups of the population might possess each mode of risky activity to varying degrees and separate matrices could be specified for each mode of transmission. Extensions of this nature would be useful for describing the role of IV-drug users, whose internal risks derive both from drug paraphernalia and sexual contact, but whose main threat to the non-drug using population would be through sexual transmission.

Gail *et al.*, on the basis of a number of simulations, reach the following conclusions which refer to a population of 100,000 persons over a period of 5 to 15 years, and are robust to wide variations in assumed parameter values and other aspects of modelling:

(1) VCT prevents hundreds or thousands of infections in isolated high-risk populations and ER values are typically less than 100, making VCT very attractive economically.
(2) VCT prevents only a few infections in isolated low-risk populations with initial prevalence 0·1% or less, and the ER values are well above 2,000. However, in 'low-risk' populations with 1% initial prevalence, tens or hundreds of infections may be prevented, and ER values fall below 2,000 for plausible spread rates, indicating that VCT may be economically feasible in such settings.
(3) In a mixed population, a VCT programme that aims primarily at the homosexual/bisexual subpopulations prevents more disease in the homosexual/bisexual populations, prevents more disease in the heterosexual populations, and requires fewer tests per case prevented than a VCT programme that tests all subpopulations equally.

The model of Gail *et al.* can incorporate changes in behaviour as the epidemic

spreads. Indeed, Gail & Brookmeyer (1988) describe the results when the spread rate is allowed to decrease over time—leading to subexponential growth early in the epidemic and delayed onset of saturation.

In an unpublished paper, Fuxman (1989) produces a series of models for populations that are:

(i) unisexual and homogeneous;
(ii) unisexual and heterogeneous;
(iii) heterosexual and homogeneous; and
(iv) heterosexual and heterogeneous,

using the same mathematical formulation as presented in Sections 5 and 6 of this paper. Fuxman's approach is conventional, but there is no reference to the work of others in this field. A significant departure from the work of others in this field is Fuxman's use of somewhat unrealistic incubation time distributions: either $\gamma(t) = \gamma$, where $\gamma = 1/7{\cdot}5$, corresponding to an exponential distribution; or $\gamma(t) = \delta\ (t - 7{\cdot}5)$, the Dirac generalised δ function.

As noted earlier, failure to allow for an incubation period distribution with sufficient tail, can lead to misleading results and these deficiencies are noted by Fuxman in his conclusions, where he suggests a gamma distribution as being more appropriate and that the "choice of the distribution has a significant influence on the bottom and middle parts of the dynamic curve".

The models discussed so far in this section are straightforward, although detailed, extensions of those considered in Section 5 and the earlier part of Section 6, which were devised mainly with a male homosexual population in mind, where individuals change partners frequently and where all members of the population are continuously at risk of infection. As noted earlier in Section 6, the assumption, that the transmission coefficient βc is a constant for each partnership between individuals of specific types, is most appropriate in a highly promiscuous community. Also, if two susceptible partners form a pair, they are at no risk of infection as long as they have no sexual contacts with other partners. In particular, this influences the early stages of the epidemic, when most existing pairs will be of two susceptible partners, who will play no part in the transmission of infection until a new partnership is formed with an infective. Thus, models are needed for populations in which longlasting partnerships are common, be they homosexual or heterosexual.

Daykin (1990) and Wiley *et al.* (1989) note recent studies which indicate variation in the infectivity of HIV among heterosexual couples (β). Wiley *et al.* represent this heterogeneity by modelling β as a random variable. Using data on the number of contacts and seroconversion of couples from two partner studies, the model is fitted by maximum likelihood estimation with a beta distribution and a discrete distribution for β. The models which fit best are those indicating extreme forms of heterogeneity, i.e. transmission after a few contacts or not at all. Wiley & Herschkorn (1988) show that this type of heterogeneity implies that the risk of acquiring HIV infection depends mainly on the number of sexual

partnerships formed and not on the frequency of intercourse within the partnership. This result supports the models proposed, for example by May & Anderson (1987). However, it should be noted that the studies used here for the fitting of the models are of modest sample size only and the results will need to be checked on other, comparable data sets.

Readers interested in a full description of micro models of partnerships (including their formation and separation) are referred to the extensive work of Dietz (1987, 1988) and Dietz & Hadeler (1988).

Many of the models described above, of the heterosexual transmission of the epidemic, allow only for sexual contacts within partnerships. However, casual sexual contact may also take place. This is considered in some of the models of Dietz (1988) and Dietz & Hadeler (1988), who divide up the heterosexual population into eight compartments according to whether an individual is male or female, infected or not infected, paired in a sexual relationship with a member of the opposite sex or not in a sexual relationship.

This approach highlights the importance of sexual pair formation, but, even in this restricted domain of heterosexual transmission, too much remains presently unknown for reliable projections of the spread of HIV infections to be feasible. Further, the models involve the introduction of a large number of parameters. However, these models can be of value in the qualitative insight which they provide.

A further feature that needs to be incorporated into models of the heterosexual spread of the epidemic is the mixing of groups of individuals with different levels of sexual activity (as in the homosexual population—see Section 6.4). Also, sexual behaviour changes over time, and people with many partners one year may have only a few the next, or vice versa. Social groups within which mixing is strong, and between which it is weak, may cause low-activity people in one group to be infected before high-activity people in another group.

The social/non-social mixing behaviours modelled by Sattenspiel (1987) and Sattenspiel & Simon (1988) may also play an important role in the spread of this disease. Models with a variety of mixing assumptions need to be developed and compared, both with each other and with behavioural and serological studies, to ascertain what complexities are really necessary for modelling HIV spread and which are not.

6.8 *Models for the Spread of AIDS in Developing Countries*

A further step in modelling the spread of the HIV epidemic in heterosexual populations would be to consider the population *in toto* and focus on its (reduced) ability to reproduce itself in demographic terms. This would have particular relevance to the developing part of the world. In such a population, it would be important to augment the model to allow for vertical transmission of HIV from mother to baby—the probability of which has been estimated by some workers to be in the range 30%–70% (Anderson *et al.* (1988)).

The deaths in the epidemic and much of the morbidity will be in the segment of

the population which has not yet reproduced. Moreover, the rate of successful reproduction in seropositive mothers will be greatly reduced for several reasons. Firstly, they may be unwilling to have children. Secondly, the risk of vertical transmission from an infected mother to her baby is very high. Thirdly, the seropositive progeny may have a low probability of reaching sexual maturity, and those who do will have a reduction in their own probability of reproduction. Over several generations this may have an enormous impact on the fertility of the HIV-infected segment of a population.

The negative effects of HIV on reproduction may be observed in males as well as in females, as a consequence of the significant (currently unknown) probability of infection of the latter by the former.

At some point, the effect of HIV infection on the fertility of a population, coupled with the direct effects of the epidemic, may impact seriously on the chances of the population to survive, resulting in a possible net decline in the population, as well as a shift in mean age.

Thus, the effects and consequences of HIV positivity on the long-term growth and stability of a closed human population could be significant over several generations, especially in the absence of the ability to screen and detect infected parents. The effective birth rate could fall and may not be correctable because of the heavy toll already taken in the young adult population.

This is explored by Bongaarts (1989), and in detail by Anderson *et al.* (1988), who conclude that AIDS is capable of changing population growth rates in the developing world from positive to negative values over timescales of a few decades. They estimate that the disease would have little effect on the dependency ratio of a population, defined as the number of children (say, aged under 15) and elderly (say, aged over 65), divided by the number of adults (say, aged 15–64).

6.9 *Concluding Comments*

Simple mathematical models of the type described in Section 6 can yield important qualitative insights about the spread of HIV infection. In particular, such models show:

(a) that the rate of spread of the epidemic depends strongly on initial HIV seroprevalence rates and on transition rates;
(b) that the epidemic will die out if the rate of new infections is less than the rate of death of infected individuals;
(c) that the underlying prevalence of infection will grow subexponentially in many scenarios;

[Subexponential growth of infections can arise because:

(i) the subpopulation becomes saturated with infected individuals, or
(ii) individuals modify their behaviour to reduce the risk of HIV transmission, or
(iii) the epidemic spreads from high-risk subpopulations to low-risk subpopulations with lower rates of infection, or

(iv) the total incidence of infection arises from the aggregation of infections in temporally separated subepidemics.]

(d) the effects of heterogeneous risk behaviours—and, for example, the slower growth of the epidemic in a high-risk subpopulation than when it is isolated, because some of its contacts are diluted by low-risk contact with the low-risk population; and
(e) the effects of temporal lags between the epidemics in subgroups of the population—for example the aggregation of two subepidemics with a temporal lag, can lead to the saturation of one group while the second group enters its exponential phase (Gail & Brookmeyer (1988)).

It is thus worth noting that the aggregate annual incidence rate in a population may take many forms, including sustained exponential growth, subexponential growth, and even undulation, depending on the nature and modes of cross-infection of contributing subpopulations. Moreover, these patterns may be occurring in underlying infection rates well before they are detectable by their effects on AIDS incidence.

The mathematical models we have described can be considered as being very useful for organising facts related to epidemic spread, for identifying gaps in epidemiological knowledge, and for determining the kinds of behaviour, subexponential growth in particular, that might be anticipated in various circumstances. However, it is unlikely that such models will be reliable for quantitative predictions of annual incidence of HIV infection in the immediate future, though researchers are attempting to devise epidemic models for this purpose. One potential application of compartmental models is the comparison of alternative strategies for disease prevention, because such comparisons may be less sensitive to changes in model assumptions than are quantitative predictions of annual HIV incidence (Gail *et al.*, 1989).

Sections 5 and 6 have indicated how a basic deterministic model of epidemic spread may be extended to incorporate more critical and realistic features of the transmission of HIV, through sexual and other routes.

In drawing this section to a close, it may be worth highlighting certain themes:

(i) Extending the model to a widely varying heterosexual population results in an increase in the number of parameters because of asymmetries between the sexes (Section 6.7). It becomes desirable to take into account partnerships (this is also the case for the non-promiscuous homosexual group), since, during a long partnership between two susceptibles, neither is at risk of infection if there are no outside contacts. It is then natural to model the probability of transmission of infection per sexual contact, rather than per partnership.
(ii) As the model is generalised to apply to increasingly broad communities, it is straightforward mathematically to include more and more sources of variation between individuals, in an attempt to mimic to an ever greater

extent 'actual' behaviour. From the perspective of the use of such models, however, it is important to identify those sources of variation which are critical in their effect on the spread of infection and those which can effectively be assumed constant, in order to obtain a broad-brush picture, which is the most that is needed for many practical purposes. For example, it would be necessary to separate out the highly promiscuous individuals from the rest, dividing individuals into 'high' and 'low' activity, but whether further division is advantageous is less obvious.

(iii) Another important aspect of modelling still to be tackled is the incorporation of spatial features. The models have tended to envisage a fairly small closed community, but it is also of great relevance to see how epidemics in different spatial locations are linked together. The idea of an epidemic in one community feeding infectives into another community, in which a self-sustaining epidemic may or may not be generated, has already been mentioned in the context of homosexual and heterosexual populations, but the idea applies equally to communities distinguished by spatial location. The geographical spread of infection has been studied in particular with regard to influenza in the U.S.S.R. (see, for example, Bailey (1975)), but no systematic investigation has been attempted for HIV infection.

(iv) Numerical studies of the models described in this paper have been made to judge the effects of changes in assumptions and parameter values and to gain a feeling for which of these have a critical influence on the course of the epidemic and to which it is reasonably robust. Such studies have an important part to play in enabling sensible decisions to be reached on intervention strategies to be used, with the aim of reducing the impact of the epidemic. There have also been many numerical studies, using various models, in conjunction with estimates of parameter values based on observational or survey data, to predict the course of the infection in particular populations. In these studies there are often large numbers of parameters to be estimated, on the basis of rather small amounts of data, and it is here that questions of insensitivity to specific parameter values need to be addressed.

(v) The data that are available, relating both to parameter values and to the incidences of HIV infection and AIDS, need careful use and interpretation for many reasons. The following points merit attention:

(1) As the epidemic proceeds, with its associated publicity, the underlying parameter values relating to sexual behaviour may be changing.
(2) Much of the data relate to groups that are self-selected to some extent—for example, individuals attending STD clinics.
(3) Some parameter values will vary substantially between subpopulations and regions.
(4) There are problems caused by under-reporting and non-detection of

cases, time lags between diagnosis and reporting and the differential operation of these factors over time and between subpopulations.

(vi) The use of mathematical transmission models to provide reliable estimates of the future course of the epidemic depends on the collection and analysis of more data than are currently available. Many authors have written persuasively on the importance of this topic (e.g. Bailey, 1988; May & Anderson, 1987). Collection of data is not, however, merely a statistical exercise and needs to be considered in the light of other, broader issues that cannot be addressed here.

PART III

7. THE BACK PROJECTION METHOD

7.1 *Introduction*

The basis of this method is that, if the distribution of the incubation period is assumed to be known, then the distribution over time of HIV infection and of the appearance of AIDS are directly linked. Knowledge of one distribution allows the other to be estimated. Starting with the observed number of AIDS cases by time of diagnosis, one can then estimate the number of HIV infections by time of infection. The number of new HIV infections in future successive time intervals can then be estimated by extrapolation and the annual future numbers of AIDS cases can, in turn, be estimated via the distribution of the incubation period.

To use a hypothetical example, suppose that at the very start of the epidemic in England 100 cases are diagnosed in one year—and that it is known that only 5% of all HIV-infected convert to AIDS within a year of infection; calculating backwards in time, one would estimate that 20 times 100 persons (or 2,000) had become infected during the prior year. Then, to continue the hypothetical example, suppose that a forecaster wishes to make a 5-year prediction for cumulative AIDS cases and that 35% of HIV-infected persons contract AIDS within 6 years; if, as of a year ago, 2,000 were infected and, of these, 100 (5%) have already been diagnosed with AIDS, then in 5 more years, an additional 600 (30% of 2,000) are expected to contract the disease—for a predicted cumulative total of 700 cases in 5 years from the forecast date.

Knowledge of the prevalence of HIV infection in a particular population, and the rate at which new infections are occurring, is clearly of great importance. However, for most populations of interest, this information is not available. The serological data that have been collected usually relate to extremely specific groups of mostly high-risk individuals and these data cannot easily be combined to predict seroprevalence in a wider, more heterogeneous, population.

It is reasonable to suppose that new cases of HIV infection occur in a point process. For the moment we consider only those cases for which AIDS will eventually be diagnosed and denote the intensity of the corresponding point

process by $h(t)$. Let the lengths of the incubation periods (between infection and diagnosis) for these individuals be independent, identically distributed variables with probability density function f. Then new diagnoses of AIDS will occur in a point process with intensity $a(t)$, where $a(t)$ is given by:

$$a(t) = \int_0^{\infty} h(t-u)f(u)du. \tag{7.1}$$

Thus, if the density of the incubation periods were known, together with the HIV infection rate $h(u)$, for $u \leq t$, we could calculate the distribution of the number of new diagnoses of AIDS in any time period up to time t. In many ways, this would be the most natural method of predicting AIDS incidence.

Conversely, we can use the above equation to deduce h if the functions a and f are known. This forms the basis of the method known as *back projection*, in which knowledge about AIDS incidence and the distribution of incubation periods is used to make inferences about the incidence of HIV infection. It must be stressed, however, that, because the proportion of those infected who will ultimately have AIDS diagnosed is not known, this method only provides information about the process of HIV infections that do subsequently lead to diagnosis of AIDS.

Equation (7.1) shows how growth in the incidence of AIDS lags behind the growth in infections. The convolution filters the changes in the underlying infection rate, $h(t)$, so that even abrupt changes in $h(t)$ appear only several years later as gradual changes in AIDS incidence, $a(t)$.

7.2 *Applications of the Back Projection Method*

Iversen & Eugen (1986) use this approach to estimate the proportion of those infected with HIV who develop AIDS within a period of 8–10 years, using data from Curran *et al.* (1985) and Peterman *et al.* (1985). They assume a normal distribution for $f(t)$, with parameters estimated from the same data set. The truncation at 8–10 years occurs because, at their time of writing, there were no observed cases with incubation times of more than 8 years.

Rees (1987, 1987a) similarly uses a normal distribution for $f(t)$ to estimate the numbers of new and total infections in the U.K. and the U.S.A. by the method of back projection. As has already been noted, Rees' methodology for fitting $f(t)$ and estimating the parameters (based on data from Peterman *et al.* (1985)) has been subject to widespread criticism. A normal distribution seems inappropriate for $f(t)$ (see Section 6.3.1) and, as will be noted later, the back projection method is particularly sensitive to the choice of distribution for the incubation period.

Boldsen *et al.* (1988) consider the total numbers infected in the U.S.A. and in Denmark using a back projection method. They assume a Weibull distribution for the incubation time distribution, $f(t)$, and a logistic growth curve for $h(t)$.

Brookmeyer & Gail (1986) use a back projection method in discrete time for estimating the number of those already infected with HIV, which is then used to project the short-term future number of AIDS cases in the U.S.A. This work

yields a *lower* bound on the size of the epidemic, in the sense that *no* allowance is made for future infections. Specifically, they postulate Weibull or log-logistic forms for $f(t)$, and use equation (7.1) to determine $h(t)$, which is assumed to be a piecewise uniform step function, viz. taking values:

1977–1980	h_1
1981–30 June 1981	h_2
1 July 1982–30 June 1984	h_3
1 July 1984–31 December 1985	h_4

where h_i are to be determined.

Given an assumed known parametric form for the incubation period distribution, $f(t)$, Brookmeyer & Gail use a maximum likelihood procedure to estimate the levels of the step function representation, h_i of $h(t)$. The procedure relies on the observation that the number of AIDS cases, diagnosed by the jth interval, follows a multinomial distribution, with cell probabilities given by convolution time integrals of f and h. The projections use counts of AIDS cases diagnosed before 1 January 1986, which are obtained from incidence data reported to the CDC surveillance system in the U.S.A., after adjustment for reporting delays.

The sensitivity of the results to changes in the cut-off points for the step function, $h(t)$, and to assumptions about the incubation period distribution are tested. The results are found to depend strongly on the latter—on both the mean of the distribution and its shape. The log-logistic distribution, with its longer tail, leads to higher predictions for future numbers of AIDS cases diagnosed.

This approach is criticised by Anderson *et al.* (1987) on technical grounds and because the approach ignores "the number and dynamics of the seropositive individuals in the population". Brookmeyer & Gail (1987) reply that the simplicity of their approach is a strength rather than a weakness and that their integral equations are correct and not an approximation. The former seems to be a valid point—the 'back projection' method is of some value, providing its limitations are recognised.

In general, solving the convolution equation (7.1) is mathematically ill-posed, in the sense that small changes in $a(t)$ or $f(t)$ may cause large changes in $h(t)$.

In Hyman & Stanley's (1988) use of the back projection method on U.S.A. data from CDC, $h(t)$ is approximated by piecewise cubic splines (in fact, Hermite polynomials). Good results are obtained when *10–30* piecewise polynomials are used. When fewer than 10 polynomials are used, the approximation is too coarse and above 30, the ill-posed nature of the problem creates high frequency oscillations in the solution. The estimates of the cumulative number of infected individuals (from $h(t)$) are most sensitive to the extrapolated estimates of $a(t)$, the fraction of the infected population that is eventually reported to CDC as AIDS cases and the most likely conversion time to AIDS, i.e. t_a, such that $f'(t_a) = 0$. The estimates are relatively insensitive to the width of the $f(t)$ distribution about t_a.

Isham (1989) uses the method of back projection specifically in the context of the AIDS epidemic within the population of the U.K., partly to see what can be inferred about the number of individuals who have been infected with HIV over the past few years and the number of those who may become infected in the near future, but also to investigate the implications of specific assumptions about the functions a and f. The theoretical development is as follows.

Suppose that $\tilde{a}(s)$ denotes the Fourier transform of $a(t)$:

$$\tilde{a}(s) = \int_{-\infty}^{\infty} e^{ist} a(t)dt$$

and similarly for $\tilde{h}(s)$ and $\tilde{f}(s)$. Then it follows immediately from (7.1) that:

$$\tilde{a}(s) = \tilde{h}(s)\tilde{f}(s) \tag{7.2}$$

and therefore that $h(t)$ can be obtained by taking the inverse transform of $\tilde{a}(s)/\tilde{f}(s)$.

Some basic parametric forms are assumed for a and f. In particular, for $a(t)$ it is assumed either that:

$$a(t) = a_0 \exp(a_1 t - a_2 t^2) \tag{7.3}$$

or that:

$$a(t) = (b_0 + b_1 t)\,[1 + \exp(b_2 - b_3 t)]^{-1} \tag{7.4}$$

(see Cox & Medley (1989)—discussed in Section 4). As already noted, there are good theoretical grounds for expecting that the curve of $a(t)$ should be close to an exponential curve in the early part of the epidemic and that this rapid growth should gradually slow down as the infection spreads. The quadratic exponential function defined by equation (7.3) is a mathematically convenient way of representing a curve that, for low values of t, increases exponentially, but has a slower growth rate as t increases. The linear logistic function given by equation (7.4) also increases exponentially for low values of t, but becomes more nearly linear for higher values of t. This curve corresponds approximately, for moderate values of t. to the solution of a fairly simple epidemic model (as in equation (5.2)). It is not assumed that either of the functions represented by equations (7.3) and (7.4) is appropriate for arbitrarily high values of t, *but only* over the range of values for which the form of h is to be deduced.

Two flexible parametric families of distributions are often used to model the incubation period; the gamma distribution, denoted by $\Gamma(\alpha,\lambda)$, has density:

$$f(t) = \lambda(\lambda t)^{a-1}\exp(-\lambda t)/\Gamma(a) \qquad (t \geqslant 0),$$

whereas the Weibull distribution, denoted Wei(β,ρ), has density:

$$f(t) = \beta\rho(\rho t)^{\beta-1}\exp[-(\rho t)^{\beta}] \qquad (t \geqslant 0).$$

Modelling the incubation period distribution has been discussed in more detail in Section 6.3.1.

If $a(t)$ follows a quadratic exponential form, as in equation (7.3) and a gamma distribution is used for $f(t)$, then explicit expressions for $h(t)$ can be obtained.

With a Weibull choice, analytic determination of $h(t)$ is not feasible and it is necessary to proceed numerically.

Thus, Isham (1989) uses the method of back projection to estimate the expected numbers of infections, H_i, and diagnoses, A_i, in year i, from the integrals:

$$H_i = \int_i^{i+1} h(u)du, \qquad A_i = \int_i^{i+1} a(u)du$$

using the alternatives of a quadratic exponential or linear logistic for $a(t)$ (with parameters estimated by Cox & Medley (1989)) and a gamma or Weibull distribution for $f(t)$ with specified parameters.

There are a number of points to note. Firstly, negative values of H_i are obtained in later years, when the quadratic exponential function is used for $a(t)$, which means that this function is not compatible with the various assumed incubation period distributions over the whole 1980–93 time period. Essentially, in such cases, if the early values of $h(t)$ are determined to give the quadratic exponential function $a(t)$ for low values of t, then too many AIDS diagnoses will occur later. To compensate for these extra cases, negative numbers of infections are then needed if the function $a(t)$ is to have the chosen increasing doubling time. When $a(t)$ has the linear logistic form, no negative values of H_i are obtained over the 1980–93 time period, although there is still an implausible oscillation in the later values. However, the function $a(t)$ has been fitted using data only up to June 1988, and only the lower tail of the incubation period distribution curve can be fitted. This lack of compatibility is not, therefore, surprising. On the other hand, the figures for expected HIV incidence up to 1986 might be hoped to be reasonably reliable. Further, the similarity between the fitted values of the two forms of $a(t)$ up to 1987, results in a corresponding similarity in the values of H_i up to 1986 (using a particular distribution for the incubation periods).

Secondly, with the assumption of the gamma distribution, $\Gamma(2,0{\cdot}14)$ or $\Gamma(3,0{\cdot}21)$, only 41% or 35% respectively of incubation periods will be of length 10 years or less, as compared with 76% for the Weibull distribution. Thus, the annual incidence of HIV infection with either of the gamma distributions must be much higher than that using the Weibull distribution during the early stages of an epidemic, to produce the same function $a(t)$. Thus, the total numbers of HIV infections occurring during 1980–87 show that the totals for the gamma distributions are almost double, or more, those for the Weibull distribution.

Thirdly, it is of interest to note that the use of an exponential curve for $a(t)$ with no quadratic term, together with the Weibull distribution for the incubation period, results in a total of some 22,000 HIV infections over the years 1980–87. This number is very similar to those obtained using the quadratic exponential or linear logistic curves, although the numbers in individual years follow a different pattern.

The particular incubation period distributions fitted by Isham (1989) are obtained from the earlier work of Anderson & Medley (1988), from data relating

to recipients of blood transfusions. It is possible that this distribution may vary with the mode of transmission, and so these parametric forms may not be applicable to a numerical projection for the whole population of the U.K. However, there is some evidence that the mode of transmission of infection does *not* have a strong effect on the distribution of the incubation period, and so the assumption of a common distribution for the aggregate population may be reasonable.

Day *et al.* (1989) approach back projection using a discrete time framework:

$$a_i = \sum_{j=1}^{i} f_{i-j} h_j \tag{7.5}$$

where

$$f_{i-j} = \int_{t_{i-j}}^{t_{i-j+1}} f(u)\, du$$

and a_i and h_i are the number of new AIDS cases and HIV infections in the interval (t_{i-1}, t_i) for $i = 1, \ldots, n$. This discrete equation is an approximate version of the convolution equation (7.1) and can also be written in matrix form in order to facilitate the inversion.

Day *et al.* apply the methodology to AIDS among U.K. residents. They assume a certain form for $f(t)$ and then, on the basis of observed values of $a(t)$ up to the present, a range of values of $h(t)$ consistent with the observed $a(t)$ is calculated. This range is then examined, in the light of available knowledge on HIV infection in the population, to indicate which ranges are implausible. For the projection of future AIDS cases, assumptions are needed regarding future HIV infections—the importance of these assumptions for future AIDS cases is examined.

Data on the number of new diagnoses of AIDS by year over the period 1982–88 are obtained. Healy's (1988) estimates, based on reports up to December 1987 (and discussed in Section 4), are untenable in the light of cases reported during 1988. New estimates, based on a linear extrapolation of new cases reported by quarter and an assumed constancy of the delay distribution, are calculated by Day *et al.* as replacements.

Three different incubation period distributions are used: based on an observed empirical distribution from CDC data, and based on fitting Weibull and gamma distribution to observed data (as for Isham (1989)).

The CDC empirical distribution leads (cumulatively) to 64·4% of cases progressing to AIDS after 12 years. A distribution based on half the underlying rates lead to a 32·2% progression rate after 12 years. These two distributions have been smoothed by a gamma distribution. The Weibull incubation period distribution used has an 84·6% cumulative progression rate after 12 years.

Day *et al.* assume that $h(t) = a \exp(bt + ct^2)$, with a time origin at January 1981 and, for each choice of incubation period distribution, obtain values of a, b and c, which are consistent with the number of AIDS cases in 1982–87 (using a χ^2 goodness of fit measure).

The results of this analysis illustrate three points. Firstly, for predictions of AIDS cases four to five years into the future, the back projection method is largely insensitive to the assumption one makes about the incubation period distribution. The two extreme distributions considered represent the fast and slow extremes of incubation period distribution usually proposed; distributions that lie between these two give predictions within the range of predictions that the two generate. The estimated number of new HIV infections, however, is highly sensitive to the assumed incubation period distribution; prediction of AIDS cases in the long term will be similarly sensitive.

Secondly, each prediction of AIDS cases within the range of consistent predictions corresponds to a particular form for the HIV epidemic. This correspondence enables one to use a variety of data, of varying degrees of reliability and direct relevance, to assess the plausibility of each prediction. This use of additional data is a major attraction of the back projection method, which would otherwise be just another form of extrapolation. Assuming a form for the HIV epidemic and for the incubation period distribution, is equivalent to assuming a certain form for the yearly number of AIDS cases. Unless further information is used, the back projection method would be equivalent to straightforward extrapolation of the AIDS cases. In this straightforward extrapolation, however, there would be no way of incorporating such further information—for example, accumulating data on the incubation period, changing levels of transmission or rates of HIV infection. Dissection of the extrapolation process in the back projection method focuses attention on where this extra information can be used.

Thirdly, the AIDS reports in 1988 make a major impact on short- and medium-term projections. The range of predictions is much narrower, and considerably lower, than when based on data available at the start of 1988. Preliminary estimates for 1988 diagnoses, based on 1988 reports, would generate a range of only two-fold for 1992 predictions, compared to the 15-fold range based on data to the end of 1987.

The range of uncertainty in the predictions for 1992 would, of course, be reduced if the 1989 data were available *and* followed current trends.

Brookmeyer & Damiano (1989) extend the earlier work of Brookmeyer & Gail (1986, 1987) and Gail & Brookmeyer (1988). The AIDS cases reported to the CDC surveillance system in the U.S.A. by 1 January 1988 are adjusted, using an estimated reporting delay distribution to give estimates of the number of AIDS cases diagnosed over time. Different reporting delay distributions are used for each major geographical region. The incubation period distribution used for $f(t)$ is a Weibull form, as used to fit the data from the Hershey hameophilia cohort study, with:

$$F(t) = 1 - \exp(-0{\cdot}004t^{2{\cdot}438})$$

using a method developed by Brookmeyer & Goedert (1989), which allows for the censored nature of these data (discussed in Section 6.3.1). Two different

parametrisations for the incidence rate, $h(t)$, are considered. As for the earlier work of Brookmeyer & Gail, a step function is used with $h(t)$ constant over the following intervals:

1 January 1977–31 December 1980	h_1
1 January 1981–30 June 1982	h_2
1 July 1982–30 June 1984	h_3
1 July 1984–30 June 1987	h_4

Also, a log logistic model (similar to one of the parametrisations used by Isham (1989)) is employed:

$$h(t;\theta) = \frac{\theta_1\theta_2(\theta_1 t)^{\theta_2 - 1}}{K(1 + (\theta_1 t)^{\theta_2})^2}$$

where K is a normalising constant, to ensure that the function integrates to 1 over the $10\frac{1}{2}$-year period under consideration. If $\theta_2 > 1$ this density increases monotonically to a single mode; if $\theta_2 \leqslant 1$, the density decreases monotonically.

The same methodology, based on maximum likelihood estimation, as for the earlier work of Brookmeyer & Gail, is used to estimate the parameters (h_i or θ). For short-term projections, new infections can be incorporated by extrapolating the infection rate obtained from the estimate of the last step of the step function version, (h_4), of $h(t)$ described above.

Taylor (1989) also builds on the earlier work of Brookmeyer & Gail, and again in a U.S.A. context. For his back projection calculations, Taylor uses AIDS incidence data reported to CDC in six-monthly intervals from July 1978 to June 1987, with upward adjustments to allow for the lag in reporting times (based on the analysis of Harris (1987)). For the incubation period distribution, a set of 21 different distributions is used, representing the likely range of the 'true' distribution and representing the fits that have been carried out on the various cohort studies in the literature: all 21 distributions are truncated at 9 years. For the incidence of HIV infections five models are considered:

(1) $h(t) = he^{-e^{a+bt}}$ double exponential incidence

(2) $h(t) = he^{bt^{1/4}}$ root exponential incidence

(3) $h(t) = \dfrac{he^{a+bt}}{1 + e^{a+bt}}$ logistic incidence

(4) $h(t) = \dfrac{hbe^{a+bt}}{(1 + e^{a+bt})^2}$ logistic prevalence (as for Brookmeyer & Damiano (1989))

(5) $h(t) = ht^2$ quadratic incidence.

Strictly speaking, $h(t)$, which is the number of new infections at time point t, should be integer valued: however, models (1)–(5) do not restrict $h(t)$ 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, there are advantages in using 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 1984–87, in cohorts which are under intense study. 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 (ht) and exponential incidence (he^{bt}) are used by Taylor, but give very poor fits to the data and the results are not presented.

For each of the 21 incubation period distributions, Taylor uses maximum likelihood procedures to estimate the parameters θ for each of the five chosen incidence models $h(t;\theta)$. The range of estimates from the 105 fits (together with χ^2 values with degrees of freedom = 18 − number of parameters) are then examined—tabulations are provided of the predicted number of AIDS cases in the second half of 1987 (on the basis of the numbers currently infected and the assumption that AIDS cannot develop 9 years after infection), of the number of currently infected people, of the minimum predicted total number of AIDS cases by 1991 (on the basis of the currently infected and the assumption that AIDS cannot develop 9 years after infection). Taylor also investigates the sensitivity of the results to underreporting of AIDS cases and an inappropriate estimate of the number of cases diagnosed in the latest time interval (first half of 1987).

Taylor also proposes a Bayesian approach to incorporating the uncertainty in the knowledge of which HIV incidence model to choose.

The most satisfactory fit results from the use of model (1) for $h(t)$: the double exponential model.

7.3 *General Comments*

There are several important caveats associated with the method of back projection.

Projections of this type depend, of course, entirely on the assumptions being made. If the functions $a(t)$ and $f(t)$ are known exactly, then the HIV infection rate $h(t)$ could be *exactly* determined. Error in the *assumed* forms of $a(t)$ or $f(t)$ will be reflected in errors in the estimated $h(t)$ and the estimated annual incidence of HIV

infection, H_i. In some cases, the assumed forms will be clearly incompatible, at least over part of their ranges, e.g. if negative values or unreasonable oscillations result (Isham, 1989).

The estimates provided by back projection for the numbers of individuals infected early in the epidemic are more reliable than the estimates for the numbers infected in approximately the last year.

It should be noted that empirical information is only available about the shape of the lower part of the incubation period distribution curve. So the predicted values of the rate of HIV infection, $h(t)$ (and the associated H_i), obtained by using the values of $a(t)$ or $f(t)$ for t lying outside the restricted ranges for which data are currently available, must be treated with considerable caution. Thus, Isham (1989) notes that even up to 1987 (i.e. for the most recent few years), the projected values of H_i vary considerably with the particular incubation period distribution chosen to model $f(t)$.

It is important, therefore, to investigate the sensitivity of the projections to the assumed form for $f(t)$.

Further, an underlying assumption of the method implicit in the convolution given in equation (7.1) is that the calendar date of infection is independent of the incubation period: the incubation period distribution for individuals infected early in the epidemic is assumed to be the same as that for those infected later. The assumption could be violated, for example, if cofactors affecting the incubation period distribution are more likely or less likely to be present among individuals infected in the 1970s compared with those infected in the 1980s.

Secular change in the incubation period distribution is a definite possibility, given that the clinical definition of AIDS is an endpoint of different nature affecting diverse risk groups—thus, as Daykin (1990) reports, there are major temporal, regional and demographic differences in the epidemiology of AIDS.

Also, cofactors may be identified which affect the incubation period distribution e.g. age at onset, in which case the analysis should be stratified according to such important cofactors.

Short-term projections of AIDS incidence are not very sensitive to the choice of model for $h(t)$, because such projections depend mainly on the numbers of infected individuals several years in the past. However, longer-term projections require data on the numbers infected in approximately the last year, and are, therefore, much more sensitive to the choice of model for $h(t)$.

For the same reasons, estimates of current HIV seroprevalence derived from back calculation are uncertain and highly dependent on the model chosen for $h(t)$. For example, estimates of current U.S. HIV seroprevalence at the end of 1987 based on logistic, log-logistic, and damped exponential models for $h(t)$ are 420,000, 853,000 and 1,649,000 individuals, respectively (CDC, 1987). The damped exponential model allows for very rapid growth of infections in the most recent time periods, whereas the logistic model forces $h(t)$ to plateau in this region. Gail & Brookmeyer (1988) favour flexible models, such as the piecewise constant model for $h(t)$, which do not force estimates of $h(t)$ in the final intervals

to be strongly determined by data affecting earlier intervals. Nevertheless, estimates for the most recent intervals are highly uncertain. The longer-term projections of AIDS cases would be greatly strengthened if HIV seroprevalence data could be used to estimate the numbers infected in recent time intervals, to complement estimates based on back calculation of the numbers infected in earlier years.

De Gruttola & Lagakos (1989) discuss the problems that arise from the uncertainty about the appropriate form for $h(t)$ and they indicate the size of the ranges in estimates that can rise from the back projection method through the use of different $h(t)$: experiments are considered with linear, quadratic, linear-cubic and epidemic transmission models. They conclude that without additional information, AIDS incidence data are of *limited* value in estimating and interpreting the extent of the HIV infection growth (in the back projection method, for example). Even if the incubation time distribution were known exactly, AIDS incidence data cannot accurately determine the number of persons recently or currently infected with HIV by working backwards.

Further, they comment that the interpretation of $h(t)$ (even if known precisely, rather than estimated) is severely limited by a lack of knowledge of the natural history of HIV and of behavioural practices.

The shape of $h(t)$ is influenced by a variety of factors—none well characterised—including heterogeneity of the population at risk, efficiencies of transmission by various routes, and variability over time in the infectiousness of an infected person. Population heterogeneity includes variability in behaviour among individuals and behavioural changes of individuals over time. Thus, $h(t)$ can be thought of as being composed of contributions of many inter-relating subepidemics, defined by type of individual, region, route of infection, etc. Even if individual behaviour were constant over time, too much is still unknown about the sizes or behaviours of these subpopulations to identify their individual effects from $h(t)$. When the possibility of time changes is also considered, the problem is even more complex. Thus, a decrease in the rate of growth of $h(t)$ might be due to changes in behaviour (for example, greater use of condoms), but also might be a consequence of near saturation of the pools most at risk, or to a decrease in the number of new sexual partners by promiscuous individuals. Similarly, $h(t)$ gives no information about the degree of spread to lower risk populations (homosexual men practising 'safe sex', promiscuous heterosexuals and so on) even though such groups may account for an increasingly large proportion of cases in the future.

REFERENCES

ANDERSON, R. M. (1988). The epidemiology of HIV infection; variable incubation plus infectious periods and heterogeneity in sexual activity. *J.R.S.S.* Series A, **151**, 66–93.

ANDERSON, R. M. & MAY, R. M. (1986). The invasion, persistence and spread of infectious diseases within animal and plant communities. *Phil. Trans. Roy. Soc.* B, **314**, 533–570.

ANDERSON, R. M. & MAY, R. M. (1987). Plotting the spread of AIDS. *New Scientist*, 26 March 1987, 54–59.

ANDERSON, R. M. & MAY, R. M. (1988). Epidemiological parameters of HIV transmission. *Nature*, **333,** 514–519.

ANDERSON, R. M., MAY, R. M. & MCLEAN, A. R. (1988). Possible demographic consequences of AIDS in developing countries. *Nature*, **332,** 228–234.

ANDERSON, R. M., MAY, R. M., MEDLEY, G. F. & JOHNSON, A. E. (1986). A preliminary study of the transmission dynamics of the Human Immunodeficiency Virus (HIV), the causative agent of AIDS. *I.M.A. J. Math. Appl. Med. & Biol.* **3,** 229–263.

ANDERSON, R. M., MEDLEY, G. F., BLYTHE, S. P. & JOHNSON, A. E. (1987). Is it possible to predict the minimum size of the AIDS epidemic in the U.K.? *Lancet*, **1,** 1073–1075.

ANDERSON, R. M. & MEDLEY, G. F. (1988). Epidemiology, HIV infection and AIDS: the incubation and infectious periods, survival and vertical transmission. *AIDS*, **2,** S57–S63.

ARTZROUNI, M. & WYKOFF, R. (1988). 'A two state infective age-structured model for the spread of AIDS in the U.S.A.' Poster presentation at the IVth International Conference on AIDS, Stockholm, June 1988. Abstract 4695.

BAILEY, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases*. Griffin, London.

BAILEY, N. T. J. (1988). Simplified modelling of the population dynamics of HIV/AIDS. *J.R.S.S.* Series A **151,** 31–43.

BAILEY, N. T. J. & ESTREICHER, J. (1987). 'Epidemic prediction and public health control, with special reference to influenza and AIDS.' Proc. 1st World Congress of Bernoulli Society (Tashkent, September 1986).

BARRETT, J. C. (1988). Monte Carlo simulation of the heterosexual spread of the human immunodeficiency virus. *Journal of Medical Virology*, **26,** 99–109.

BARTON, D. E. (1987). Striking the balance on AIDS. *Nature*, **326,** 734.

BEALE, S. (1987). On the sombre view of AIDS. *Nature*, **328,** 673.

BIRKHEAD, B. G. (1987). 'A mathematical model of the transmission of the HIV under diminishing recruitment—an exact solution.' Department of Statistical Science, University College, London. Technical Note.

BLYTHE, S. P. & ANDERSON, R. M. (1988). Distributed incubation and infectious periods in models of the transmission dynamics of the human immunodeficiency virus (HIV). *I.M.A. J. Math. Appl. Med. & Biol.* **5,** 1–19.

BLYTHE, S. P. & ANDERSON, R. M. (1988a). Variable infectiousness in HIV transmission models. *I.M.A. J. Math. Appl. Med. & Biol.* **5,** 181–200.

BLYTHE, S. P. & ANDERSON, R. M. (1988b). Heterogeneous sexual activity models of HIV transmission in male homosexual populations. *I.M.A. J. Math. Appl. Med. & Biol.* **5,** 237–260.

BOLDSEN, J. L., JENSEN, J. L., SOGAARD, J. & SORENSEN, M. (1988). On the incubation time distribution and the Danish AIDS data. *J.R.S.S.* Series A, **151,** 42–43.

BONGAARTS, J. (1989). A model of the spread of HIV infection and the demographic impact of AIDS. *Statistics in Medicine*, **8,** 103–120.

BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations. *J.R.S.S.* Series B, **26,** 211–252.

BRODT, H. R., HELM, E. B., WERNER, A. *et al.* (1986). Spontanverlauf der LAV/HTLV-III Infektion. *Deutsche Medizinische Wochenschrift*, **111,** 1175–1180.

BROOKMEYER, R. & DAMIANO, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine*, **8,** 23–34.

BROOKMEYER, R. & GAIL, M. H. (1986). Minimum size of the AIDS epidemic in the United States. *Lancet*, **2,** 1320–1322.

BROOKMEYER, R. & GAIL, M. H. (1987). Methods for projecting the AIDS epidemic. *Lancet*, **2,** 99.

BROOKMEYER, R. & GAIL, M. H. (1987a). Biases in prevalent cohorts. *Biometrics*, **43,** 739–749.

BROOKMEYER, R., GAIL, M. H. & POLK, B. F. (1987). The prevalent cohort study and the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, **126,** 14–24.

BROOKMEYER, R. & GOEDERT, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*, **45,** 325–335.

CANADIAN INSTITUTE OF ACTUARIES TASK FORCE ON AIDS (1988). First report of the Subcommittee on Modelling. November 1988.

CANADIAN INSTITUTE OF ACTUARIES TASK FORCE ON AIDS (1988a). Second report of the Subcommittee on Modelling. An analysis of U.S.A. data. November 1988.

CENTRES FOR DISEASE CONTROL (1986). Update: acquired immunodeficiency syndrome (AIDS)—United States. *Morbidity and Mortality Weekly Reports*, **32,** 17–21.

CENTRES FOR DISEASE CONTROL (1987). Human immunodeficiency virus infection in the United States: a review of current knowledge. *Mortality and Morbidity Weekly Reports*, **36,** 1–48.

CHIN, J. & MANN, J. (1989). Global surveillance and forecasting of AIDS. *Bulletin of WHO*, **67,** 1–7.

COLGATE, S. A., STANLEY, E. A., HYMAN, J. M. *et al.* (1989). A behaviour based model of the cubic growth of AIDS in the United States. *Proc. Nat. Acad. Sci. U.S.A.* **86,** 4793–4797.

COSTAGLIOLA, D. & DOWNS, A. M. (1987). Incubation time for AIDS. *Nature*, **328,** 582.

COSTAGLIOLA, D., MARY, J.-Y., BROUARD, N. *et al.* (1989). Incubation Time for AIDS from French transfusion-associated cases. *Nature*, **338,** 768–769.

COWELL, M. J. & HOSKINS, W. H. (1987). 'AIDS, HIV mortality and life insurance.' Society of Actuaries special report, August 1987 (also in report of the Society of Actuaries Task Force on AIDS).

COX, D. R. & MEDLEY, G. F. (1989). A process of events with notification delay and the forecasting of AIDS. *Phil. Trans. Roy. Soc.* B, **325,** 135–145.

COX, D. R. & DAVISON, A. C. (1989). Prediction for small subgroups. *Phil. Trans. Roy. Soc.* B, **325,** 185–187.

CURRAN, J. W., MORGAN, W. M., HARDY, A. M. *et al.* (1985). The epidemiology of AIDS: current status and future prospects. *Science*, **229,** 1352–1357.

DAHLMAN, G. E., BERGSTROM, R. L. & MATHES, R. W. (1987). 'Projecting extra AIDS mortality for individual ordinary life insurance in force as of December 31 1986.' Milliman & Robertson Research Report (revised version in Report of Society of Actuaries Task Force on AIDS).

DANGERFIELD, B. & ROBERTS, C. (1990). A role for system dynamics in modelling the spread of AIDS. *Trans. of Institute of Measurement and Control.* (To appear.)

DAY, N. E., GORE, S. M., MCGEE, M. A. & SOUTH, M. (1989). Predictions of the AIDS epidemic in the U.K.: The use of the back projection method. *Phil. Trans. Roy. Soc.* B, **325,** 123–134.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1987). AIDS Bulletin No. 1. Institute of Actuaries AIDS Working Party.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1987a). AIDS Bulletin No. 2. Institute of Actuaries.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1987b). The implications of AIDS for life insurance companies (Supplement to AIDS Bulletin No. 2). Proceedings of a seminar on 1 February 1988. Institute of Actuaries.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1988). AIDS Bulletin No 3. Institute of Actuaries AIDS Working Party.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1988a). The Impact of HIV Infection and AIDS on Insurance in the United Kingdom. *J.I.A.* **115,** 727–837.

DAYKIN, C. D., CLARK, P. N. S., EVES, M. J., HABERMAN, S., LE GRYS, D. J., LOCKYER, J., MICHAELSON, R. W. & WILKIE, A. D. (1989). AIDS Bulletin No. 4. Institute of Actuaries.

DAYKIN, C. D. (1990). Epidemiology of HIV Infection and AIDS. *J.I.A.* **117,** 51–94.

DE GRUTTOLA, V. & MAYER, K. H. (1988). Assessing and modelling heterosexual spread of the human immunodeficiency virus in the United States. *Review of Infectious Diseases*, **10,** 138–150.

DE GRUTTOLA, V. & LAGAKOS, S. W. (1989). The value of AIDS incidence data in assessing the spread of HIV infection. *Statistics in Medicine*, **8,** 35–43.

DE GRUTTOLA, V. & LAGAKOS, S. W. (1989a). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45,** 1–11.

DEPARTMENT OF HEALTH/WELSH OFFICE (1988). 'Short-term prediction of HIV infection and AIDS in England and Wales.' Report of a Working Group. HMSO. London.

DIETZ, K. (1987). 'Epidemiological models for sexually transmitted infections.' Proc. 1st World Congress of Bernoulli Society (Tashkent, 1986).

DIETZ, K. (1988). On the transmission dynamics of HIV. *Mathematical Biosciences*, **90,** 397–414.

DIETZ, K. & HADELER, K. P. (1988). Epidemiological models for sexually transmitted diseases. *Journal of Math. Biology*, **26,** 1–25.

DIETZ, K. & SCHENZLE, D. (1985). Mathematical models for infectious disease statistics. In: *A Celebration of Statistics*, eds. A. C. Atkinson & S. E. Fienberg, pp. 167–204. Springer, New York.

DOWNS, A. M., ANCELLE, R. & 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.

EISENBERG, B. (1989). The number of partners and the probability of HIV infection. *Statistics in Medicine*, **8,** 83–92.

FARMER, R. D. T. & EMAMI, J. (1987). 'The transmission of HIV and the evolution of the AIDS epidemic—sexual transmission model.' (Unpublished.)

FUHRER, C. (1988). 'Projecting the number of AIDS Cases.' Presented to the Society of Actuaries Symposium "Insurance and the AIDS Epidemic". Chicago, Illinois. May, 1988.

FUXMAN, Y. L. (1989). 'Generating relations in the mathematical modelling of the AIDS epidemic.' (Unpublished manuscript.)

GAIL, M. H. & BROOKMEYER, R. (1988). Methods for projecting course of acquired immunodeficiency syndrome epidemic. *J. Nat. Cancer Inst.* **80,** 900–911.

GAIL, M. H., PRESTON, D. & PIANTADOSI, S. (1989). Disease prevention models of voluntary confidential screening for human immunodeficiency syndrome. *Statistics in Medicine*, **8,** 59–81.

GANI, J. (1978). Some problems of epidemic theory. *J.R.S.S.* Series A, **140,** 323–347.

GENERAL ACCOUNTING OFFICE (1989). 'AIDS forecasting: undercount of cases and lack of key data weaken existing estimates.' Report to Congress. GAO PEMD 89-13. Washington DC, U.S.A.

GONZALEZ, J. J., KOCH, M. G., DORNER, D., L'AGE-STEHR, J., MYRTVEIT, M. & VAVIK, L. (1987). 'The prognostic analysis of the AIDS epidemic: mathematical modelling and computer simulation.' Proc. E.C. Workshop on Statistical Analysis and Mathematical Modelling of AIDS (Bilthoven, December 1986). Oxford University Press.

GONZALEZ, J. J. & KOCH, M. G. (1987). On the role of transients for the prognostic analysis of the AIDS epidemic. *American Journal of Epidemiology*, **126,** 985–1005.

HARRIS, J. E. (1987). 'Delay in reporting acquired immune deficiency syndrome.' M.I.T. Technical report No. 452, M.I.T., Mass., U.S.A.

HARRIS, J. E. (1988). 'The incubation period for human immunodeficiency virus (HIV)', in Kulstad, R. (ed.) AIDS 1988: AAAS Symposia Papers, AAAS, Washington DC.

HEALY, M. J. R. (1988). 'Extrapolation forecasting. Appendix 6. Short term prediction of HIV infection and AIDS in England and Wales.' Report of a Working Group. HMSO, London.

HEALY, M. J. R. & TILLETT, H. E. (1988). Short-term extrapolation of the AIDS epidemic. *J.R.S.S.* Series A, **151,** 50–65.

HELLINGER, F. J. (1988). Forecasting the personal medical care costs of AIDS from 1988 through 1991. *Public Health Reports*, **103,** 309–319.

HETHCOTE, H. W. & YORKE, J. A. (1984). Gonorrhoca: transmission dynamics and control. Lecture Notes in *Biomathematics*, **56,** 1–105. Springer-Verlag, Berlin.

HYMAN, J. M. & STANLEY, E. A. (1988). Using mathematical models to understand the AIDS epidemic. *Mathematical Biosciences*, **90,** 415–473.

ISHAM, V. (1988). Mathematical modelling of the transmission dynamics of HIV infection and AIDS: A Review. *J.R.S.S.* Series A, **151,** 5–30.

ISHAM, V. (1989). Estimation of the incidence of HIV infection. *Phil. Trans. Roy. Soc.* B, **325,** 113–121.

IVERSEN, O.-J. & ENGEN, S. (1986). Epidemiology of AIDS—statistical analyses. *J. Epidemiol. and Comm. Health*, **41,** 55–58.

JACQUEZ, J. A., SIMON, C. P., KOOPMU, J., SATTENSPIEL, L. & PERRY, T. (1988). Modelling and analysing transmission: the effect of contact patterns. *Mathematical Biosciences*, **92,** 119–199.

KANOUSE, D. E., CARDELL, N. S., GORMAN, E. M. *et al.* (1988). 'Modelling the spread of HIV infection in the United States.' (Unpublished working draft.) Presented to the XVth General Assembly of the Geneva Association, The Hague, June 1988. The Rand Corporation.

KALBFLEISCH, J. D. & LAWLESS, J. F. (1988). Estimating the incubation period for AIDS patients. *Nature*, **333,** 504–505.

KERMACK, W. O. & MCKENDRICK, A. G. (1927). Contribution to the mathematical theory of epidemics. *Proc. Roy. Soc.* A, **115,** 700–721.

KIESSLING, D., STANNAT, S., SCHEDEL, I. & DEICHER, H. (1986). Überlegungen und Hochrechungen zur Epidemiologie des 'Acquired Immunodeficiency Syndrome' in der Bundesrepublik Deutschland. *Infection*, **14,** 217–222.

KNOX, E. G. (1986). A transmission model for AIDS. *European Journal of Epidemiology*, **2,** 165–177.

KOLBYE, J. (1987). 'AIDS mortality and life insurance.' Baltica-Nordisk Re.

KREMER, E. (1982). IBNR claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal*, 47–55.

LAGAKOS, S. W., BERRAJ, L. M. & DE GRUTTOLA, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS. *Biometrika*, **75,** 515–523.

LEMP, G. F., PAYNE, S. F., RUTHERFORD, G. W. *et al.* (1988). 'Projections of AIDS morbidity and mortality in San Francisco using epidemic models.' Poster presentation at the IVth International Conference on AIDS, Stockholm, June 1988. Abstract 4682.

LONGINI, I. M., SCOTT CLARK, W., BYERS, R. H. *et al.* (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, **8,** 831–843.

LORPER, J. (1988). 'Actuarial studies of the AIDS problems.' Publications of the Cologne Re. 14.

LORPER, J. (1989). Projecting the spread of AIDS into the general population—application to life assurance. *J.I.A.*, **116,** 625–638.

LUI, K. J., DARROW, W. W. & RUTHERFORD, III, G. W. (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science*, **240,** 1333–1335.

LUI, K. J., LAWRENCE, D. N., MORGAN, W. M., PETERMAN, T. A., HAVERKOS, H. W. & BREGMAN, D. J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proc. Nat. Acad. Sci. U.S.A.* **83,** 3051–3055.

LUI, K. J., PETERMAN, T. A. & LAWRENCE, D. N. (1987). Comments on the sombre view of AIDS. *Nature*, **329,** 207.

MAY, R. M. & ANDERSON, R. M. (1987). Transmission dynamics of HIV Infection. *Nature*, **326,** 137–142.

MCEVOY, M. & TILLETT, H. E. (1985). Some problems in the prediction of the future numbers of cases of the acquired immunodeficiency syndrome in the U.K. *Lancet*, **2,** 541–542.

MEDLEY, G. F., ANDERSON, R. M., COX, D. R. & BILLARD, L. (1987). Incubation period of AIDS in patients infected via blood transfusion. *Nature*, **328,** 718–721.

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. Roy. Soc.* B, **233,** 367–377.

MEDLEY, G. F., ANDERSON, R. M., COX, D. R. & BILLARD, L. (1988a). Estimating the incubation period for AIDS patients. *Nature*, **333,** 505.

MODE, C. J., GOLLWITZER, H. E. & HERRMANN, N. (1988). A methodological study of a stochastic model of an AIDS epidemic. *Mathematical Biosciences*, **92,** 201–229.

MORGAN, W. M. & CURRAN, J. W. (1986). Acquired immunodeficiency syndrome: current and future trends. *Public Health Reports*, **101,** 459–465.

MORTIMER, P. P. (1985). Estimating AIDS, U.K. *Lancet*, **2,** 1065.

PANJER, H. H. (1987). 'Survival analysis of persons testing HIV positive.' Working Paper Series in Actuarial Science ACTSC 87–14, Faculty of Mathematics, University of Waterloo, Canada.

PANJER, H. H. (1988). 'AIDS: some aspects of modelling the insurance risk.' Research Report 88–10, Institute of Insurance and Pension Research, University of Waterloo, Canada.

PETERMAN, T. A., JAFE, H. W., FEORINO, P. M., GETCHELL, J. P., WARFIELD, D. T., HAVERKES, H. W., STONEBURNER, R. L. *et al.* (1985). Transfusion-associated acquired immunodeficiency syndrome. *J. Amer. Med. Assoc.* **254,** 2913–2917.

PETO, J. (1986). AIDS promiscuity. *Lancet*, **2,** 979.

PLUMLEY, P. W. (1989). Modelling the AIDS Epidemic by Analysis of Sexual and Intravenous Drug Behaviour. *Trans. Soc. Act.* **41** (to appear).

REES, M. (1987). The sombre view of AIDS. *Nature*, **326,** 343–345.

REES, M. (1987a). Describing the AIDS epidemic. *Lancet*, **2,** 98–99.

ROBERTS, C. A. & DANGERFIELD, B. C. (1988). 'Simulation models of the epidemiological consequences of HIV infection and AIDS.' Working Paper No 8901, Dept. of Business and Management Studies, University of Salford.

SALZBERG, A. M., DOLINS, S. L. & SALZBERG, C. (1989). HIV incubation times. *Lancet*, **2,** 166.

SATTENSPIEL, L. (1987). Population structure and the spread of disease. *Human Biol.* **59,** 411–438.

SATTENSPIEL, L. & SIMON, C. (1988). The spread and persistence of infectious diseases in structured populations. *Mathematical Biosciences*, **90,** 341–366.

STANNAT, S., KIESSLING, D., SCHEDEL, I. & DEICHER, H. (1987). 'Computer simulations of the AIDS-Epidemic in the Federal Republic of Germany.'

STIGUM, H., GROEENESBY, J. K., MAGNUS, P. *et al.* (1988). 'The effect of selective partner choice on the spread of HIV.' Poster presentation at The Global Impact of AIDS Conference, London, March 1988.

STRONISKI, K. (1990). 'Delays in reporting of Canadian AIDS cases.' ARCH (to appear).

TAYLOR, J. M. G. (1989). Models for the HIV infection and AIDS epidemic in the United States. *Statistics in Medicine*, **8,** 45–58.

TAN, W. Y. & HSU, H. (1989). Some stochastic models of AIDS spread. *Statistics in Medicine*, **8,** 121–136.

THOMPSON, J. R. (1987). 'AIDS: old disease, new society.' Technical Report 87–1, Dept. of Statistics, Rice University, Texas.

TILLETT, H. E. & MCEVOY, M. (1986). Reassessment of predicted numbers of AIDS cases in the U.K. *Lancet*, **2,** 1104.

VAN DRUTEN, J. A. M., DE BOO, TH., JAGER, J. C. *et al.* (1986). AIDS prediction and intervention. *Lancet*, **1,** 852–853.

VAN DRUTEN, J. A. M., DE BOO, TH., REINTJES, A. G. M., JAGER, J. C., HEISTERKAMP, S. H., COUTINHO, R. A., BOS, J. M. & RUITENBERG, E. J. (1987). Reconstruction and prediction of spread of HIV infection in populations of homosexual men. *Proc. E.C. Workshop on Statistical Analysis and Mathematical Modelling of AIDS* (Bilthoven, December 1986). Oxford University Press.

VERRALL, R. J. (1988). 'Bayesian linear models and the claims run-off triangle.' Actuarial Research Paper No. ?, City University, London.

WHYTE, B. M., GOLD, J., DOBSON, A. J. & COOPER, D. A. (1987). Epidemiology of acquired immunodeficiency syndrome in Australia. *Medical Journal of Australia*, **146,** 65–69.

WILEY, J. A. & HERSCHKOM, S. J. (1988). The perils of promiscuity. *Journal of Infectious Diseases*, **158,** 500–501.

WILEY, J. A., HERSCHKOM, S. J. & PADIAN, N. S. (1989). Heterogeneity in the probability of HIV transmission per sexual contact: The case of male-to-female transmission in penile-vaginal intercourse. *Statistics in Medicine*, **8,** 93–102.

WILKIE, A. D. (1988). An actuarial model for AIDS. *J.R.S.S.* Series A, **151,** 35–39.

WILKIE, A. D. (1988a). An actuarial model for AIDS. *J.I.A.* **115,** 839–853.

WILKIE, A. D. (1989). Population projections for AIDS using an actuarial model. *Phil. Trans. Royal Soc.* B, **325,** 99–112.

WHO COLLABORATING CENTRE (1988). 'Results from the latest half-yearly analysis of European AIDS surveillance data: assessment of temporal evolution and predictions to December 1989.' Paris.

ZEGER, S. L., SEE, L.-C. & DIGGLE, P. J. (1989). Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine*, **8,** 3–21.