Transmission dynamics of HIV infection

Robert M. May and Roy M. Anderson

Simple mathematical models of the transmission dynamics of human immunodeficiency virus help to clarify some of the essential relations between epidemiological factors, such as distributed incubation periods and heterogeneity in sexual activity, and the overall pattern of the AIDS epidemic. They also help to identify what kinds of epidemiological data are needed to make predictions of future trends.

DESPITE remarkable advances in understanding the basic biology of human immunodeficiency virus (HIV — the aetiological agent of AIDS, acquired immune deficiency syndrome)[1-5] public health planning continues to be hampered by uncertainties about epidemiological parameters[4-6]. Accurate information about the typical duration and intensity of infectiousness, or about the fraction of those infected who will go on to develop AIDS (and after how long), will emerge only from carefully designed studies on these same timescales, which is to say many years. In the absence of such information, mathematical models of the transmission dynamics of HIV cannot be used at present to make accurate predictions of future trends in the incidence of AIDS, but they can facilitate the indirect assessment of certain epidemiological parameters, clarify what data is required to predict future trends, make predictions under various specified assumptions about the course of infection in individuals and patterns of sexual activity within defined populations (or changes therein) and, more generally, provide a template to guide the interpretation of observed trends[7,8].

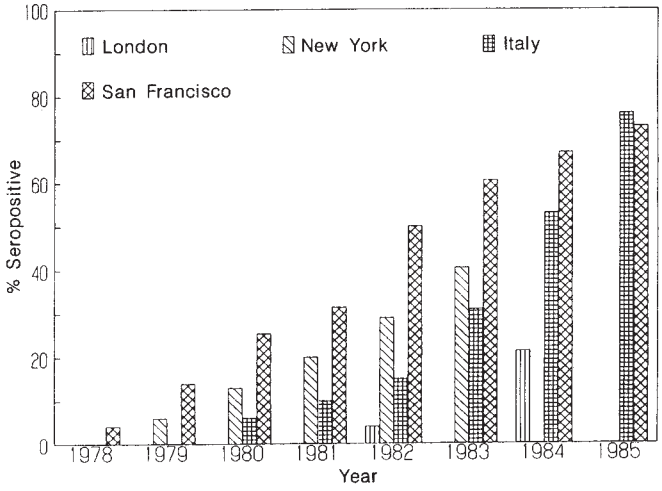

Fig. 1 The rise in seropositivity to HIV antigens in cohorts of patients over the period 1978–1985. The studies in San Francisco[12], London[23] and New York[35] were of homosexual/bisexual males. The study in Italy[24] is of drug addicts.

Whether an infection can establish itself and spread within a population is determined by the key parameter R_0, the basic reproductive rate of the infection[7]. R_0 is the average number of secondary infections produced by one infected individual in the early stages of an epidemic (when essentially all contacts are susceptible); clearly the infection can maintain itself within the population only if R_0 exceeds unity[9,10]. For a sexually transmitted disease (STD), R_0 depends on c, which is essentially the average rate at which new sexual partners are acquired, on β, the average probability that infection is transmitted from an infected individual to a susceptible partner (per partner contact) and on D, the average duration of infectiousness[7,11]. In what follows, we mainly restrict attention to the spread of HIV among homosexual males, now responsible for the bulk of AIDS cases (about 70–80% in the United States, and a similar proportion in European countries[6,12,13]).

Initial stages of the epidemic

The characteristics of most STDs cause their epidemiology to differ from that of common childhood viral infections[11,14,15]. Unlike infections caused by airborne transmission, the rate at which new infections are produced is not dependent on the density of the host. Second, the carrier phenomenon in which certain individuals harbour asymptomatic infection is often important (as in the spread of herpesvirus). Third, many STDs induce little or no acquired immunity on recovery (for example, gonorrhoea) and fourth, net transmission depends on the degree of heterogeneity in sexual activity prevailing in the population.

As Hethcote and Yorke[11] have shown in their studies of gonorrhoea, mathematical models for the dynamics of STDs need to take account of the substantial variations of sexual activity within the population at risk. A particular risk group (such as male homosexuals in San Francisco[12]) of total size N may be roughly partitioned into subgroups of size N_i, each of whom on the average acquire i new sexual partners per unit time (when $N = \sum_i N_i$). The probability that susceptible individuals in this ith group will become infected, per unit time, is thus $i\lambda$, where λ is the probability that infection is acquired from any one new partner. In turn, λ is equal to the product of the transmission probability β defined above and the probability that any one randomly-chosen partner is infected (with such partners being more likely to come from the sub-groups of individuals with high degrees of sexual activity).

Exponential growth

When these assumptions are incorporated into a model for the transmission dynamics of HIV infection, the infected fraction of the population at risk (who are seropositive in tests for HIV) rises exponentially, as $\exp(\Lambda t)$, in the early stages of the epidemic. The exponential growth rate, Λ, is related to the basic epidemiological quantities defined above by:

$$\Lambda = \beta c - 1/D \qquad (1)$$

The effective average over the distribution by degrees of sexual activity, c, is given explicitly as

$$c = \Sigma i^2 N / \Sigma_i N_i = m + \sigma^2/m \qquad (2)$$

where m is the mean and σ^2 the variance of the distribution of the number of new sexual partners per unit of time[8]. Thus, c is not simply the mean but the mean plus the ratio of variance to mean, which reflects the disproportionate role played by highly active individuals (in the tail of the probability distribution of sexual activity), who are both more likely to acquire infection and more likely to transmit it. The basic reproductive rate for HIV infection, R_0, is related to the parameters β, c and D, and hence to Λ by the formula

$$R_0 = \beta c D \qquad (3)$$

In contrast with standard epidemiological models in homogeneous populations (where the exponential phase of rising incidence lasts until something like half the pool of susceptibles have been infected), the early exponential phase is of

relatively short duration in our HIV models, giving way to a more nearly linear rise in the fraction infected (see Fig. 2).

This is because most susceptibles in the sexually highly active categories are infected in the early stages of the epidemic, producing saturation effects in these categories which decrease the exponential rise in incidence within them; although the incidence of infection continues to rise among individuals in less sexually active categories, the overall rate of increase is now slower than exponential.

Much less information is available about the rise in the number of individuals infected with HIV, as a function of time, than about the rise in the subsequent incidence of AIDS[46,12,13], largely because information about infection requires serological examination for antibodies to the HIV virus. Although the initial infection may produce symptoms[3,16,17] and, in some cases acute encephalopathy[18] and meningitis[16], it is not clear that such symptoms are always evoked: in any event, the symptoms are usually sufficiently mild to preclude systematic reporting. By contrast, the opportunistic infections cancers[4,5,12,19], and subsequent mortality characteristic of the destruction of the immune system in AIDS, leads to fairly reliable reporting[20]. There is, however, one study of hepatitis B virus (HBV) in a cohort of 6,875 homosexual and bisexual males in San Francisco, which resulted in serum samples being taken and preserved as early as 1978[12,2122]; stored sera of a representative sample of 785 of these individuals gives the rise in the fraction seropositive for HIV, from 1978 to 1985, shown in Fig. 1.

Table 1 Doubling time of the HIV epidemic (in the early stages)

Area	Period	Doubling time t_d (in months)
Serological data		
(a) Male homosexuals		
San Francisco, USA	1978 – 80	10 – 11
New York City, USA	1979 – 80	10 – 11
London, UK	1982 – 84	9 – 10
(b) Intravenous drug users		
Italy	1980 – 83	15 – 16
London, UK	1983 – 85	11 – 12
Switzerland	1983 – 84	8 – 9
Case notifications		
(a) All risk groups		
Australia	1982 – 85	4 – 5
Austria	1983 – 85	15 – 16
Belgium	1982 – 84	11 – 12
Canada	1981 – 85	9 – 10
Denmark	1982 – 84	13 – 14
Europe (EC)	1982 – 84	8 – 9
France	1982 – 84	8 – 9
Italy	1982 – 84	7 – 8
Netherlands	1982 – 84	7 – 8
Spain	1982 – 84	6 – 7
Sweden	1983 – 85	8 – 9
Switzerland	1983 – 85	9 – 10
United Kingdom	1982 – 84	10 – 11
United States	1982 – 83	5 – 6
West Germany	1982 – 84	6 – 7
(b) Heterosexuals		
United States	1982 – 84	9 – 10
Average		9 – 10

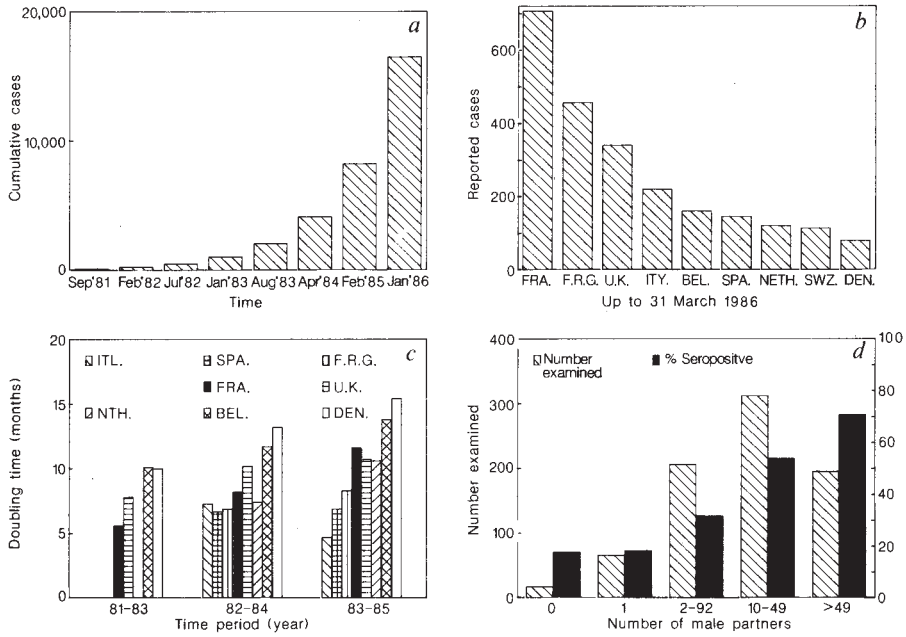

Fig. 2 *a*, The rise in the cumulative number of reported cases of AIDS in the USA over the interval September 1981 – January 1986[13]. *b*, Reported cases of AIDS in 9 countries of the European Community up to 31 March 1986[40]. *c*, Doubling times in the cumulative incidence of AIDS (t_d) recorded in months for various European countries[40] over various time intervals (1981–83, 1982–84, 1983–85; DEN, Denmark; BEL, Belgium; NTH, Netherlands; FRA, France; E.C., European Community in total; F.R.G.. Federal Republic of Germany; SPA, Spain; ITL, Italy; UK, United Kingdom). *d*, The relationship between sexual activity amongst a sample of homosexual/bisexual males (from San Francisco, USA) as measured by the number of male partners over a two-year period, and the percentage of each group (based on sex partners) who were seropositive for HIV antibodies (data from ref. 26).

The pattern of roughly linear rise shown in Fig. 1 is uncharacteristic of standard epidemics (in homogeneously mixed populations), but is suggested by our HIV models. In Britain and other countries in Europe, the virus seems first to have appeared several years later than in the United States (Fig. 2*a*, *b* and *c*), and the spread of infection is still in its early stages. As a result there are serological studies focused on HIV roughly from its initial appearance in Europe[23,24] (Fig. 1). The initially exponential rise in HIV infection may be characterized by a doubling time, t_d, related to the growth rate, Λ, of (1) by $t_d = (\ln 2)/\Lambda$.

Table 1 summarizes information about doubling times deduced from serological and case notification studies, which lead to a surprisingly consistent estimate of t_d ~8–10 months in the early stages of the epidemic (Fig. 2*d*) giving an estimate of Λ of about 1.0 yr^{-1}. The characteristic duration of infection (and infectiousness), D, is probably not significantly less than the characteristic time from HIV infection to manifestation of AIDS.

But D may be significantly longer if a substantial proportion of infected individuals remain asymptomatic carriers (with the epidemiology similar to hepatitis B virus[25]). On the other hand, recent studies observing that measurable HIV antigen (HIV-Ag, the presence of which indicates the presence of the virus) appears early and transiently in primary HIV infection, that antibody production follows (1–3 months after infection) and that HIV-Ag may then disappear could imply lower estimates of D, as would the apparent correlation of this persistence or reappearance of antigen with clinical, immunological and neurological deterioration[3].

In the absence of conclusive data on infectiousness during the incubation period, we shall assume that D is equal to the incubation period. Studies of cases of AIDS associated with transfusion suggest that the average incubation period is 4–5 years[31], but as such studies are extended, this estimate will rise (Fig. 3). The true average may be 8–10 years or more. Our estimate of Λ in conjunction with equation 1 then leads to the rough estimate

$$\beta c \simeq 1\ \mathrm{yr}^{-1} \qquad (4)$$

Note that $\Lambda \simeq \beta c$ provided D is large (4–5 years plus). Thus data on changes in seropositivity over time have allowed us to infer the approximate magnitude of the combination of epidemiological parameters β and c, neither of which can easily be estimated directly.

Is this estimate consistent with what is known about β and c separately? Unfortunately, nearly all the information about degrees of sexual activity among male homosexuals has focused on average numbers of sexual partners, as distinct

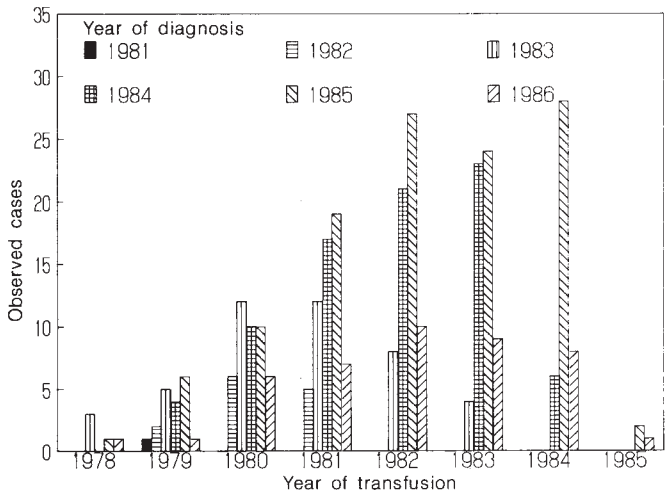

Fig. 3 Data on the distribution of the incubation period of AIDS derived from longitudinal studies of transfusion recipients (data from ref. 31). Observed cases of AIDS are recorded as a function of the year of transfusion (the assumed point of acquisition of infection) and the year of diagnosis. A Weibull distribution provides a good empirical description of this data with a mean incubation period of ~ 4–5 years.

from average number of new partners per unit time[26,27] (Fig. 4). For less active individuals (say, 1–3 partners per 6-month interval), the rate of acquisition of new partners will be seriously overestimated by the average number of partners. On the other hand, the quantity c is disproportionately influenced by highly active individuals, most of whose partners are likely to be new, so that studies based simply on numbers of partners may give a rough guide to the magnitude of c (Fig. 4). Quantitative information on average values of β, whether for homosexuals or heterosexuals, is very limited at present. Estimates vary widely (from 0.05 to 0.5) although it appears that the average probability of transmission per partner contact is higher among male homosexuals than among heterosexuals, perhaps as a result of more frequent sexual activity that results in epithelial damage (for example anal intercourse)[26,29]. Our estimates of $\beta c \sim 1 yr^{-1}$ together with the high estimates of c for homosexuals suggest that β may be small (~ 0.05). But estimates of c based on the reported number of partners per unit of time may significantly overestimate the number of new partners per unit of time, which, or that equating D to the incubation period, may overestimate the average duration of infectiousness. It may also be that the high values of c arise from sampling biased towards the high activity groups of homosexual communities.

As public awareness about AIDS has increased, there have been changes in patterns of sexual activity among male homosexuals in the United States (reflected, for example, in marked decreases in the incidence of rectal gonorrhoea[29,30]) which have presumably resulted in changes both in β and in c[27]. Our discussion, therefore, pertains mainly to the relatively early stages — 1978 to the early 1980s — of the epidemic.

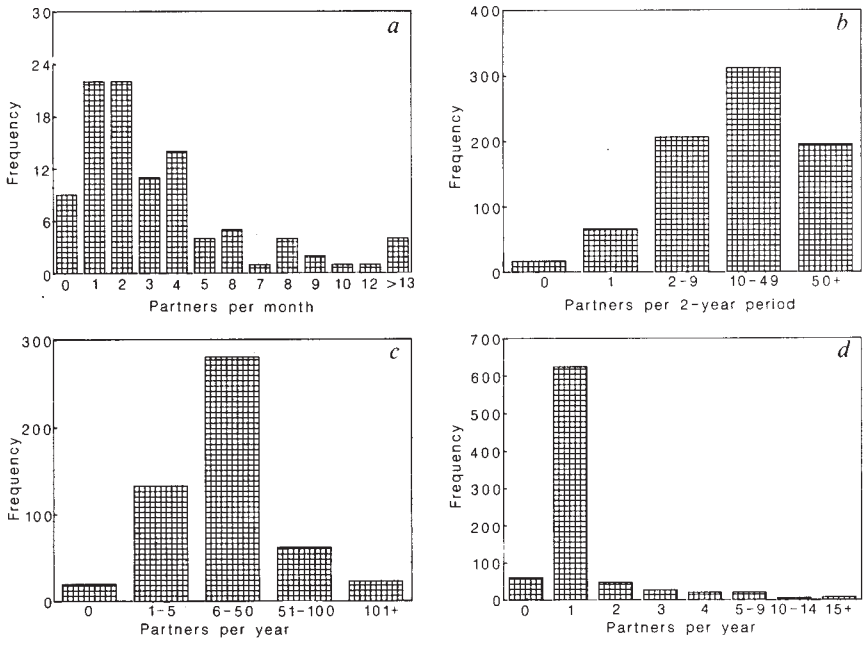

Fig. 4 Studies of sexual activity amongst a–c, male homosexual and d heterosexual communities. The graphs record frequency distributions of the number of sex partners per defined time period in samples of homosexual/bisexual males and heterosexuals. a, Homosexual/bisexual males resident in London surveyed in 1986 (unpublished data from C.A. Carne and I.V. Weller) (m = 4.7 per month, σ^2 = 56.7). Data denote male partners per month. b, Homosexual/bisexual males resident in San Francisco surveyed in 1984–85 (data from ref 26). Data denote male partners per 2-year period. c, Homosexual/bisexual males resident in London surveyed in 1984 (unpublished data from T. McManus). Data denote male partners per year. d, Heterosexuals between the ages of 18–44 years in England surveyed in November 1986 (unpublished data, Harris Research Organisation; R.M.A. and G.F. Medley). Data denote partners of the opposite sex per 1 year period (sample size = 823, m = 1.41, σ^2 = 4.36). A further survey of homosexual men (ref. 27) in San Francisco reveals a decline in the mean partners per month over the period November 1982 to November 1984 from 5.9 to 2.5.

Incubation period

Although much more information is available about the incidence of AIDS than of HIV infection (Table 1), it is harder to tease estimates of epidemiological parameters out of these. Incidence of AIDS depends not only on the transmission factors β and c, but also on the incubation period and on the fraction, f, of those infected who will eventually develop AIDS.

Significantly, estimates of both the incubation period and f have tended systematically to increase since the epidemic was first recognized[4,5,28,31,32]. Estimates of f range grom 10% to 75% or more[19,28,33,37], with an incubation period of 4–5 years or more[31]. The progressive sequence of steps which eventually impair the ability of the immume system to respond to opportunistic infection seem not to be reversible. But whether all those infected with HIV are moving toward AIDS at different rates, or whether some will develop AIDS while others never will, remains unclear. Variability in the incubation period, and whether or not an infected person develops AIDS, could be accounted for by genetic heterogeneity within the host population (HLA-linked[38,39]), or could be associated with specific strains of the antigenically variable HIV virus[1,6].

Studies of the incubation period for those who develop AIDS suggest that the 'hazard function', the probability of the disease manifesting itself as a function of the time since infection, increases with time (Fig. 3). Lui and co-workers[31] have assumed a Weibull distribution (a flexible two parameter probability distribution) for the incubation period with probability density function

$$h(t) = \gamma\nu^{-\gamma}t\exp[-(t/\gamma)^{\nu}] \qquad (5)$$

If indeed the probability per unit time, to develop AIDS (for that fraction f who do indeed develop it) increases linearly with time from infection as αt, the result is a Weibull distribution with $\gamma = 2$ and $\nu = \alpha$ for the hazard function[8]. This

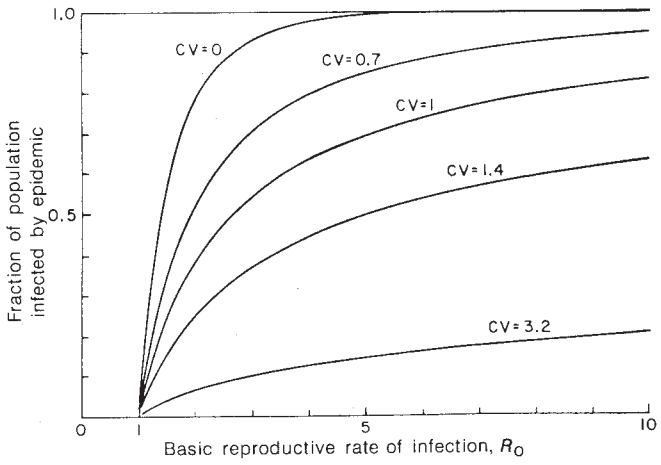

Fig. 5 The relationship between the eventual fraction infected (seropositive) in an epidemic of HIV and the basic reproductive rate R_0 (see text). The predictions are based on a model which assumes that sexual activity (defined as the number of new partners per unit of time) obeys a gamma distribution with varying coefficients of variation (CV) (mean m and variance σ^2; see text).

assumption differs from conventional epidemiological models, where infected individuals move through the incubation interval either at a fixed rate, or in a fixed time. But none of this resolves the question of what proportion of those infected will develop AIDS on what timescale: That issue will be resolved only by very long term (many decades) studies.

Fraction eventually infected

In a closed and homogeneously mixed population, the total fraction eventually infected depends only on the basic reproductive rate of the infection, R_0, defined above as shown[7] by the uppermost curve in Fig. 5. For sexually-transmitted infections such as HIV, the result can be extended to include the complications associated with a wide diversity in degrees of sexual activity.

In a closed population, the eventual fraction seropositive will depend both on R_0 and on the actual distribution of rates of acquisition of sexual partners. Assuming a gamma distribution[8], we may characterize it by c and by its coefficient of variation (CV = σ/m). The resulting overall fraction infected is shown as a function of R_0, for a range of values of CV, in Fig. 5; for fixed R_0, the eventual seropositive fraction can be much lower than for CV = 0, if the variability in degrees of sexual activity (measured by CV) is high. This makes intuitive sense: the highly active individuals acquire infection, and eventually are removed, relatively early in the epidemic; transmission among the remaining, less active, individuals may be relatively weak.

Figure 5 may be used, in combination with two factual observations, to make a rough assessment of R_0 for HIV among male homosexuals. First, the studies indicate great variability in degrees of sexual activity among male homosexuals (with CV significantly in excess of unity[8,26,27]), thus confining attention to the lower curves. The second observation is that levels of seropositivity to HIV among male homosexuals in San Francisco in 1985 are variously reported as 70% or more[12] (in the HBV study which is probably biased towards more active individuals) and as around 50% (in a study carefully constructed to avoid bias[26]), providing a lower bound of 50–70% on the proportion ever seropositive. For CV noticeably in excess of unity, this can be achieved only if R_0 is in excess of 5.

Thus our early estimate of $\beta c \sim 1\ \mathrm{yr}^{-1}$, in conjunction with the assessment that R_0 exceeds 5, leads to an indirect estimate that D exceeds 5 years. Although R_0, like β and c changes with changing social and sexual habits, the data leading to our earlier estimate for βc come from the early stages of the rise in HIV infection, before such changes were significant. The estimate of R_0 depends importantly on observed levels of seropositivity, but these were also high before social changes became pronounced. Consequently, our estimate of D which depends only on the basic biology of HIV, is reasonably consistent. This independent estimate of $D \sim 5$ years accords with current estimates that the incubation period is 4–5 years or more.

An estimate of the value of R_0 in the early stages of the epidemic is also valuable in indicating the magnitude of the social changes needed to bring R_0 below unity. If R_0 is around 5–10 or more, then reductions by a factor of 5–10 or more in βc are needed. Because c depends disproportionately on those in the highly sexually active category, programmes aimed at getting them to change their habits — both to fewer partners and to "safe sex" — are most efficient. But if such individuals are less likely to respond to public health education, it will be harder to bring R_0 below unity.

Mortality

The frequent assumption that the severity of the epidemic, in terms of cumulative mortality, will be greatest if all those infected eventually develop AIDS and subsequently die is not necessarily true. Mortality depends critically on the duration of infectiousness of both those infected who develop AIDS and those infected who do not. If the latter have a similar life expectancy to those not infected, but remain infectious for life, they may contribute more to the net transmission of the virus, R_0, than those who die of AIDS. Much may be understood by recognizing that the overall net reproductive rate of the virus, R_0, is made up of two components, the reproductive rate of those who develop AIDS (R_{01}) and the equivalent rate of those who do not (R_{02}). If a fraction f develop AIDS

$$R_0 = fR_{01} + (1-f)R_{02} \qquad (6)$$

where the two reproductive rates are defined by equation (3) with different parameters for the separate groups. Even if the asymptomatic carriers are less infectious than those who develop AIDS, if they remain infectious over, say, a 30-year span of sexual activity, R_{02} may be much larger than R_{01}, and, depending on f, the contribution of the asymptomatic carriers to R_0 may be dominant.

At present, it is not possible to tell whether the severity of the epidemic will be increased or decreased if a larger fraction of those infected develop AIDS, for the relative infectiousness of the two categories is unknown. For public health planning it is clearly important to attempt to acquire such data.

Dynamics of the epidemic

The dynamics of an HIV epidemic within a homosexual community are represented by the results of our calculations given in Fig. 6, which shows the proportion seropositive and the incidence of cases of AIDS as a function of time since the start of the epidemic. It is assumed that 30% of those infected eventually manifest AIDS, with the incubation intervals obeying a Wiebull distribution such that the average incubation period is 5 years[31]. Individuals who are incubating AIDS are assumed infectious throughout the incubation interval, and the 70% who remain asymptomatic are assumed to remain infectious for similar periods.

Many of the features presented in Fig. 6 show qualitative agreement with observation. The rise in incidence of infection (seropositivity) is initially exponential, but soon shows a more linear rise. And, the rise in incidence of AIDS lags that in the proportion infected, as seen.

It is easy to build epidemiological models of arbitrary complexity, which may appear beguilingly realistic, but we think there is little point in constructing them until more is known about the relevant epidemiological parameters. We distrust predictions made by using statistical procedures to fit polynomial or exponential curves to existing data on the incidence of AIDS, and then extrapolating[40,44]. The HIV epidemic is a dynamic process; to predict future trends, models

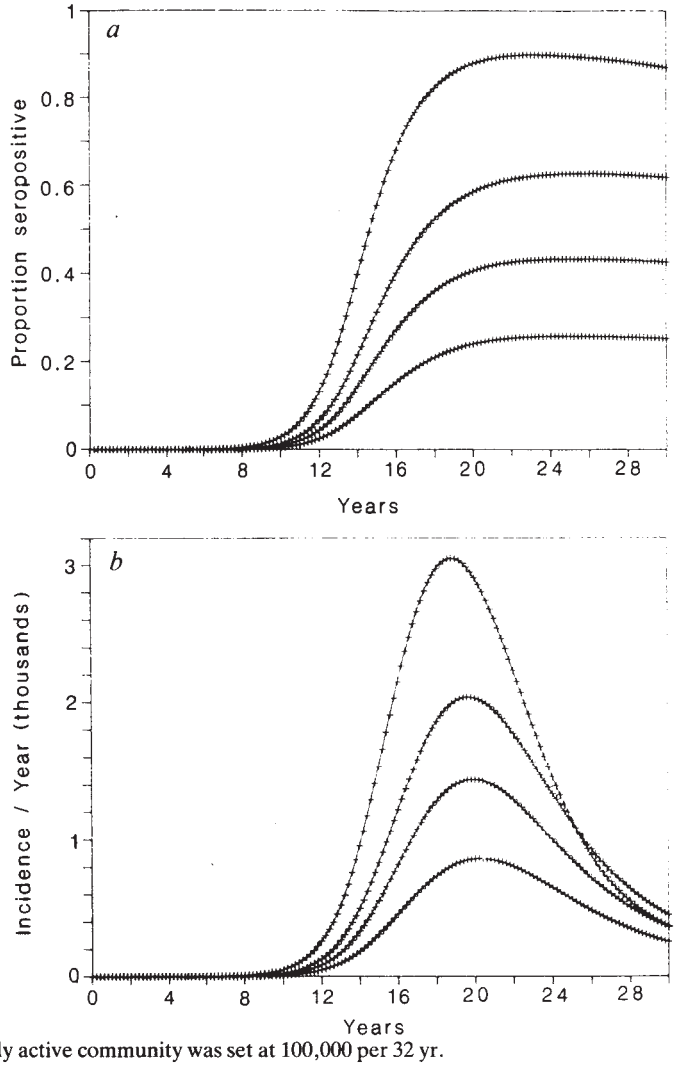

Fig. 6 The predictions of a model (see ref. 8) incorporating variable incubation periods, heterogeneity in sexual activity and recruitment of susceptibles. The two graphs record changes in seropositivity through time from the point of introduction of HIV into a community of 100,000 homosexual/bisexual males (graph *a*) and the incidence of AIDS yr^{-1} (graph b). Heterogeneity in sexual activity is described by a gamma distribution with a mean fixed at 5 partners yr^{-1} and variances 5, 25, 50 and 100 representing the predictions recorded by the four lines depicted in each graph. In *a* and *b* the smallest epidemic arises when the variance is largest and vice versa. Parameter values, $R_0 = 5$, $D = 5$ yr, $f = 0.3$ with the life expectancy of AIDS patient set at 1 yr from diagnosis and for the susceptible sexually active community at 32 yr from the point of joining the sexually active class. The 70% of infecteds who do not develop AIDS are assumed to be infectious for a period equal to D. The immigration of new susceptibles into the sexually active community was set at 100,000 per 32 yr.

must be based on the underlying epidemiological phenomena.

Heterosexual transmission

In developed countries, the extent to which HIV infection can be transmitted by heterosexual contacts is uncertain[43,48]. HIV infections in females come from contact with bisexual males (the dominant sexually-transmitted route at present), transfusion recipients, haemophiliacs and intravenous drug users[45]. If such females are not themselves a significant source of infection back into the homosexual/bisexual community (through contacts with uninfected bisexuals), we would expect the incidence of HIV infections among the female partners of bisexuals initially to rise roughly in proportion to the incidence among homosexual males.

Specifically, we would expect the ratio of HIV infection among female partners of bisexuals to that among bisexual males to be $\sim \beta' c'/\beta c$, where β and c are as previously defined, β' is the transmission probability for male-to-female contact, and c' is the mean number of new female partners acquired by a bisexual male, per unit time. We expect this ratio to be significantly less than unity, because β' is less than β, and c' significantly less than c. The data (Table 1) suggest the doubling time for heterosexually-transmitted HIV infection is roughly equal to that among homosexual males, which is in accord with our simple expectation.

As the epidemic progresses, a high proportion of homosexual males become infected (Fig. 2*e*). On the other hand, the pool of female partners for bisexual males may be large, and we would expect incidence of HIV infection still to be rising roughly linearly among women at and beyond the point at which saturation effects limit the epidemic among homosexual/bisexual males. How long this rise continues, and how many females are eventually likely to be infected, depends on the average duration of infectiousness in the transmitting group of males (which could be long if a substantial fraction remain asymptomatic carriers).

The transmission of HIV infection to females by bisexual males is a process whose initial dynamics is essentially determined by R_0 for transmission among homosexual males, thereafter the question of its transmission and maintenance by purely heterosexual contact arises. The basic reproductive rate for such heterosexual transmission of HIV, R_0', is given by

$$R_0' = (\beta_1\beta_2 c_1 c_2)^{1/2} D \qquad (7)$$

Here β_1 and β_2 are the transmission parameters for contacts between infected females and susceptible males and between infected males and susceptible females, respectively; c_1 and c_2 are as before given by (2), for the distribution in rates of acquiring new partners of the other sex by females and males, respectively.

Data are very limited on the transmission and sexual activity parameters, but the data in Fig. 4 suggest that c_1 and c_2 are significantly smaller than c among homosexual males. Further, it seems likely that $\beta_1 < \beta_2$ and that both are less than the MM for homosexual males. Thus overall, the factor $(\beta_1\beta_2 c_1 c_2)^{1/2}$ seems likely to be much smaller than c for homosexual males, which suggests that in developed countries, R_0 for purely heterosexual transmission is probably significantly smaller than R_0 for purely male homosexual transmission. Whether R_0 is greater than unity, such that HIV infection can maintain itself and spread by purely heterosexual transmission, is at present unclear. There is an urgent need for studies to measure c_1 and c_2 in different communities (stratified by age and social status) and to assess how these parameters change as a consequence of educational programmes and publicity campaigns on AIDS. The use of professional opinion poll organizations to gather quantitative data on rates of partner change over a series of specified time intervals by interview and questionnaire (Fig. 4*d*) could help to fill this gap in our knowledge, but estimates of β_1 and β_2 will come only from long term studies of the heterosexual partners of infected patients.

If R_0' does exceed unity, the incidence of HIV infection in the heterosexual community will initially grow exponentially, at a rate given by the analogue of equation (1):

$$\Lambda = (\beta_1\beta_2 c_1 c_2)^{1/2} - 1/D = (R_0' - 1)D \qquad (8)$$

The estimates above indicate that initial doubling times will be significantly longer than the 9 months or so for HIV among homosexual males; the slow initial growth will be difficult to discern against a background of homosexual transmission among males and bisexual transmission to females.

These observations are not necessarily inconsistent with the epidemiological situation for HIV in sub-Sahara Africa[6,49,56,57]. In contrast to the United States and the United Kingdom, where male/female ratios of AIDS cases have been of the order of 14:1 to 20:1, in certain parts of central Africa including areas in Zaire, Rwanda and Uganda, sex ratios approaching unity have been reported[45,50-53,56,57]. Very high prevalences of HIV antibodies have been found in males and

females from surveys in urban and rural areas[52,56,57]. These points suggest that heterosexual transmission has been frequent in both directions and horizontal studies have shown that infection is associated with the age-related degree of sexual activity amongst heterosexuals[27,47,48,56,57]. We note, however, that in the early and approximately exponential phase of the epidemic, the ratio of the number of seropositive males to seropositive females is not unity, but is roughly $(\beta_1 c_1/\beta_2 c_2)^{1/2}$.

It is generally thought that β_1 is less than β_2 for HIV, although the facts are uncertain (for gonorrhea, for instance, male-to-female transmission, β_2 is roughly twice β_1). Obviously the average number of heterosexual partners of females and males, m_1 and m_2, are equal, but c_1 could significantly exceed c_2 if the variance of the distribution of rate of acquiring new sexual partners by females (associated with the concentrated activities of female prostitutes) is greater than that for males. This effect could partly offset β_1 being smaller than β_2. Although there is no *a priori* reason to expect the ratio $\beta_1 c_1/\beta_2 c_2$ to be exactly unity, its square root could easily be close to unity, which would explain the roughly equal proportions of seropositive males and females. Alternatively, the roughly equal proportions could be explained if homosexual transmission among males had coincidentally raised the seropositive proportion among males to around the level among females, or by transmission by contaminated needles in public and private medical services[6]. In any event, the rough equality of the seropositive proportions among males and females is a puzzle to be explained, and is not by itself evidence for purely heterosexual transmission

Discussion

The ideas presented above are based on relatively simple mathematical models, with the aim of making clear some of the essential relations between epidemiological parameters and the overall course of HIV infection within various populations. Such models help to clarify what kinds of epidemiological data are needed to make predictions. As such data become available, the models can be made more detailed and realistic.

For public health planning, the dominant unknown is f, the fraction infected who will eventually develop AIDS. Estimates of this parameter have been increasing in recent years, but on present evidence the possibility cannot be ruled out that it is as low as 20% or as high as virtually 100%. Thus any current predictions about the number of homosexuals likely to acquire AIDS are uncertain by at least a factor 5 or so. Better understanding of the mechanisms of interactions between virus and host may help to determine f, but it is possible that only epidemiological data gathered on a decade-long timescale, as cases accumulate, will resolve this question.

The duration of infectiousness, and the way this duration is distributed among different infectives, is also relevant to estimates of R_0 and thence of the eventual number infected; more studies directed towards eliciting this information, including looking for virus in the blood, excretions and secretions of infected individuals over time, together with longitudinal studies of the partners of infected patients, are needed[3].

More generally, there is need for more studies that combine information about the epidemiological history of individuals with information about their sexual habits, such as the important study by Winkelstein *et al.*[26] of an unbiased sample of homosexuals in San Francisco, which demonstrated the association between the number of sexual partners and probability of acquiring infection. We emphasize that what is epidemiologically important is the average rate of acquiring new sexual partners, not necessarily the same as the average number of partners per unit time. Some authors have recognized that sexually highly active individuals play a disproportionate role in the transmission dynamics; equation (2) quantifies this observation, making it clear that the epidemiologically relevant quantity is not the mean number of new partners but, rather, the mean-square divided by the mean.

In developed countries, at present and into the near future, it is probable that sexually-transmitted HIV infections among females are likely to come mainly from bisexual males. Whether subsequent spread of infection from such females to heterosexual male partners is likely to reach significant levels, and more importantly whether purely heterosexual transmission of HIV infection may be selfsustaining ($R_0' > 1$), depends on estimates of the transmission parameters β_1, β_2, c_1 and c_2.

We have shown how c for transmission among homosexual males can be estimated indirectly from data on initial doubling times, but corresponding estimates of $\beta_1 c_1$ and $\beta_2 c_2$ are much harder, partly because the corresponding doubling rates are likely to be longer and partly because these infections are likely to be masked by homosexual/bisexual transmission among males, and by bisexual-to-female transmission among females (both of which processes depend simply on c). Attempts to estimate these quantities directly, and thence to estimate R_0, are urgently needed.

From present knowledge, it is not possible to assess whether R_0' is greater or less than unity in developed countries, and it is thus not possible to say whether HIV infections could spread epidemically by purely heterosexual transmission. The evidence from Africa, however, clearly argues that the sexually active population as a whole should be regarded as at risk[6,47,56,57].

We have greatly benefited from discussions with Anne Johnson, Mike Adler, John Pickering, Graham Medley, Stephen Blythe and Jenny Crombie. Financial support from the MRC and the NSF is gratefully acknowledged. We thank C.A. Carne, I.V. Weller and T. McManus for permission to quote unpublished data (Fig. 4). □

Robert M. May is in the Biology Department, Princeton University, Princeton, New Jersey 08544, USA. Roy M. Anderson is in the Parasite Epidemiology Research Group, Department of Pure and Applied Biology, Imperial College, University of London, London SW7 2BB, UK.

1. Hahn B.H. *et al. Science* **232,** 1548-1552 (1986).
2. Goedert, J.J. *et al. Lancet* **ii** 711-715 (1984).
3. Goudsmit, J. *et al. Lancet* **ii** 177-180 (1986).
4. Peterman, T.A., Drotman, D.P. & Curran, J.W. *Epidem. Rev.* **7,** 1-21 (1985).
5. Curran, J.W. *Ann. Intern. Med.* **103,** 657-662 (1985).
6. Acheson, E.D. *Lancet* **i** 662-676 (1986).
7. Anderson, R.M. & May, R.M. *Nature* **318,** 323-329 (1985).
8. Anderson, R.M., May, R.M., Medley, G.F. & Johnson, A. *IMA J. Math. Med. Biol.* (in the press).
9. Dietz, K. *Lect. Notes. Biomaths.* **11,** 1-15. (1976).
10. Anderson, R.M. in *Theoretical Ecology* (ed. May, R.M.) 318-355 (Blackwell, Oxford, 1981).
11. Hethcote, H.W. & Yorke, J.A. *Lect. Notes. Biomaths.* **56,** 1-105 (1984).
12. Centers for Disease Control MMWR **34,** 573-589 (1985).
13. Centers for Disease Control MMWR **35,** 17-20 (1986).
14. Anderson, R.M. & May, R.M. *Nature* **280,** 361-367 (1979).
15. May, R.M. & Anderson, R.M. *Nature* **280,** 455-461 (1979).
16. Ho, D.D. *et al. New Engl. J. Med.* **313,** 1606 (1985).
17. Ho, D.D. *et al. Ann. Intern. Med.* **103,** 880–883 (1985).
18. Carne, C.A. *et al. Lancet* **ii** 1206–1208 (1985).
19. Wong-Staal, F. & Gallo, R.C. *Nature* **317,** 395-403 (1985).
20. Centers for Disease Control MMWR **34,** 373-375 (1985).
21. Schreeder, M.T. *et al., J. Infect. Dis.* **146,** 7-15 (1982).
22. Jaffe, H.W. *et al., Ann. Intern. Med.* **103,** 210-214 (1985).
23. Carne, C.A. *et al., Lancet* **i**, 1261-1262 (1985).
24. Angarano, G. *et al., Lancet* **ii** 1302 (1985).
25. Francis, D.P. *Rev. Infect. Dis.* **5,** 322-329 (1983).
26. Winkelstein, W. *et al., J. Am. Med. Assoc.* **257,** 321–325 (1987).
27. McKusick, L. *et al., Pub. Hlth. Repts.* **100,** 622-628 (1985).
28. Curran, J.W., Morgan, W.M. & Hardy, A.M. *Science* **229,** 1352-1357 (1985).
29. Weller, I.V.D., Hindley, D.J. & Adler, M.W. *Brit. med. J.* **289,** 1041 (1984).
30. Centers for Disease Control MMWR **34,** 613-615 (1985).
31. Lui, K.J. *et al., Proc. natn. Acad. Sci. U.S.A.* **83,** 3051-3055 (1986).
32. Peterman, T.A. *et al., J. Am. med. Ass.* **254,** 2913-2917 (1985).
33. Moss, A.R. *et al., J. infect. Dis.* **152,** 152-161 (1985).
34. Weber, J.N. *et al., Lancet* **i** 1179-1182 (1986).
35. Stevens, C.E. *et al., J. Am. med. Assoc.* **255,** 2167-2171 (1985).
36. Goedert, J.J. *et al., Science* **231,** 992-995 (1986).
37. Brodt, H.R. *et al. Dtsch. med. Wschr.* **111,** 1175-1180 (1986).
38. Weiss, R.A. in *Virus Resistance* (eds Mahy, B.W.J., Mison, A.C. & Dorby, G.K.) 267–288 (Cambridge Univ., London, 1982).
39. Scorza Smeraldi, R. *et al. Lancet* **ii** 1187-1189 (1986).
40. McEvoy, M. & Tillett, H. *Lancet* **ii,** 541-542 (1985).
41. Artalego, F.R. *et al., Lancet* **i** 378 (1986).
42. Downs, A.M., Ancelle, R. & Brunet, J.B. (in the press).
43. Barnes, D.M. *Science* **232,** 1589-1590 (1986).
44. Gonzalez, J.J. & Keoch, M.G. *W.H.O. Report on Euro meeting on AIDS containment* (WHO, Geneva, 1986).
45. Vogt, M.W. *et al., Lancet* **i,** 525-527 (1986).
46. Wofsy, C. *et al., Lancet* **i,** 527-529 (1986).
47. Centers for Disease Control MMWR, **34,** 561-563 (1985).
48. de Perre, P.V. *et al., Lancet* **ii** 524.
49. Harris, C. *et al., New Engl. J. Med.* **308,** 1181-1184 (1985).
50. Centers for Disease Control MMWR **31,** 697-698 (1984).
51. Newmark, P. *Nature* **322,** 6 (1986).
52. Biggar, R.J. *Lancet* **i** 79-82 (1986).
53. Kreiss, J.K. *et al., New Engl. J. Med.* **314,** 414-417 (1985).
54. Papervangelou, G., Roumeliotou-Karayannis, A., Kallinikos, G. & Papoutsakis, G. *Lancet* **ii** 1018 (1985).
55. Tirelli, U. *et al., Lancet* **ii** 1424 (1985).
56. Quinn, T.C., Mann, J.M., Curran, J.W. & Piot, P. *Science* **234,** 955-963 (1986).
57. Melbye, M. *et al. Lancet* **ii** 1113-1117 (1986).