

Infectious disease dynamics: what characterizes a successful invader?

Robert M. May^{1,2*}, Sunetra Gupta² and Angela R. McLean²

¹Office of Science and Technology, Albany House, 94–98 Petty France, London SW1H 9ST, UK

²Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Against the background of human immunodeficiency virus (HIV) and acquired immune deficiency syndrome (AIDS) and other potentially emerging (or re-emerging) infectious diseases, this review will focus on the properties which enable an infectious agent to establish and maintain itself within a specified host population. We shall emphasize that for a pathogen to cross a species barrier is one thing, but for it successfully to maintain itself in the new population is must have a ‘basic reproductive number’, R_0 , which satisfies $R_0 > 1$. We shall further discuss how behavioural factors interweave with the basic biology of the production of transmission stages by the pathogen, all subject to possible secular changes, to determine the magnitude of R_0 . Although primarily focusing on HIV and AIDS, we shall review wider aspects of these questions.

Keywords: HIV; AIDS; emerging diseases; transmission; establishment; basic reproductive number

1. INTRODUCTION

Roughly ten years ago, The Royal Society held a Discussion Meeting on Biological Invasions (Kornberg & Williamson 1987). Like many other such meetings in recent years, it focused on the vexatious ecological problems caused by ‘alien invaders’: species that have been introduced into new places, sometimes accidentally and sometimes deliberately, but almost always as a result of human activities, with adverse consequences for indigenous organisms.

In all such ecological discussions, it helps to distinguish two different questions. First, how did the invader get there? Second, did it succeed in spreading and establishing itself, and how is this success explained? In most ecological situations, the first question is more easily answered (‘in the mud on birds’ feet’, if all other suggestions fail), and often less interesting. Both from the standpoint of basic ecological theory, and for practical purposes of conservation biology or pest control, the difficult problems are usually to understand why some arrivals spread and persist, while others—apparently similar—fail to establish. Many of the relevant questions were first raised in Elton’s (1958) *The ecology of invasions by animals and plants*, but, despite much subsequent progress (e.g. Crawley 1986; Cohen & Carlton 1995; Mooney & Drake 1986), we are still often at a loss to explain why some alien plant and animal species have made such nuisances of themselves while others, despite repeated introductions, have failed to take off.

With a few exceptions, most of this ecological literature has paid little attention to invading micro-organisms, including viral, bacterial and protozoan agents of infectious disease. Studies of ‘emerging diseases’ are relatively recent (e.g. Morse 1993). In contrast with the ecological

work, such studies tend to concentrate on how the pathogen emerged into its new host, rather than on how it subsequently succeeded in establishing itself within this host population. For example, the excellent survey of ‘changing patterns of infectious disease’ by M. L. Cohen (2000) treats environmental and social, as well as immunological and aetiological, aspects of the ‘emergence of infectious diseases’, but gives relatively little consideration to questions of herd immunity (and makes no mention of the ‘basic reproductive number’, as introduced below). We think this is understandable: ecologists tend to look at populations of interacting species, leaving the lives of individuals to the behavioural biologists; in contrast, most work on infectious diseases is focused on individual patients, and their personal combat with pathogens, newly emerging or otherwise.

The present discussion of the origins of human immunodeficiency virus (HIV) and the acquired immune deficiency syndrome (AIDS) epidemic is fairly representative of these trends. Much of the emphasis is on how, exactly, HIV entered human hosts. Many parallel questions can be asked about how other new diseases—‘legionnaire’s disease, Lyme disease, HIV, Nipah virus, hantavirus, *Escherichia coli* O157:H7, “flesh-eating” bacteria, and many others’ (M. L. Cohen 2000)—made their appearance in humans. But arguably more important is the question of how HIV has established itself and developed as an epidemic within human populations, at different speeds in different places.

2. THE BASIC REPRODUCTIVE NUMBER

Central to the discussion of the persistence and spread of any organism is its basic reproductive number, R_0 . This is essentially the average number of successful offspring that the organism—be it human, frog, fish or flea—is capable of producing, in the absence of crowding and

*Author for correspondence (robert.may@zoo.ox.ac.uk).

other 'density-dependent' effects. It is, in effect, Fisher's (1930) 'net reproductive value' for the organism.

For microparasitic infections (*sensu* Anderson & May 1979), which broadly are those where the host population can be partitioned into susceptibles, infectives and recovered-and-immune (or more general compartmental models), R_0 is more precisely defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible. This definition applies to most viral and bacterial infections, including HIV, and also to many protozoan infections (Anderson & May 1991).

It immediately follows that such an infection cannot maintain itself within its host population if $R_0 < 1$. Conversely, if $R_0 > 1$, the infection can take off.

These two cases make very clear the distinction between jumping a species barrier to infect a new host, and the question of whether this infection can subsequently persist and spread within its new host population. Going back to the quote from M. L. Cohen (2000), above (§ 1), we would draw a big distinction between HIV and hantavirus or Nipah virus infections in humans. Clearly HIV has $R_0 > 1$ in many groups. But recent reports of Nipah virus (e.g. Enserink 2000*a*), under the heading of an 'emerging disease', suggest strongly that it has $R_0 < 1$ in human populations; although this paramyxovirus may be endemic ($R_0 > 1$) among bats in Malaysia, and transmitted to humans via pigs, it would seem that human infections are always decaying chains derived from this reservoir. Hantavirus infections in the USA, and the new arenavirus blamed for recent deaths in California (Enserink 2000*b*), also seem to be in this $R_0 < 1$ category, epiphenomena of endemic infections among rodents. We would, indeed, conjecture that the haemorrhagic fever-causing Lassa and Ebola viruses, which sporadically cause nasty deaths, also may currently have $R_0 < 1$ in human populations, being maintained in some other animal reservoir (where they have $R_0 > 1$).

In short, the distinction between R_0 being greater, or less, than unity is central to any discussion of infectious diseases.

We now turn to develop these ideas more fully, in the specific context of HIV and AIDS. We begin with relatively simple models, which lay bare the basic principles. The effects of various social, geographical, immunological, genetic and other heterogeneities are outlined. We discuss the effect upon the likelihood of disease emergence of the number of hosts initially infected, and of possible complications arising from cross-immunity produced by other infections. We conclude by drawing some morals about the difference in control strategies for emerging diseases with $R_0 > 1$ versus those with $R_0 < 1$.

3. EXPLICIT EXPRESSIONS FOR R_0

As reviewed in detail by Anderson & May (1991), R_0 for directly communicated infectious diseases can, in general, be written as

$$R_0 = T(\alpha, N)/(\alpha + \mu + v), \quad (1)$$

where T is the average rate at which infected individuals produce new infections (i.e. the number per unit of time),

not discounted for the possibility that some of these potential infectees are already infected. The denominator represents the average rate at which infected individuals move out of the infectious class, either by recovering (at the rate v) or by dying (from the infection, at the rate α , or from other causes, at the rate μ). That is, $1/(\alpha + \mu + v)$ represents the average duration of infectiousness, D . A variety of realistic complications can modify equation (1) for specific infections, but this simple formula helps to clarify some basic ideas.

We have written the transmission and recovery rates, T and v , as functions of the virulence (α) and of population size (N), $T(\alpha, N)$ and $v(\alpha)$, to emphasize the possible interrelatedness among the components of R_0 . Such interconnections are important in any discussion of the evolution of the virulence of a particular infectious agent. Given that the evolutionary pressures on the pathogen are to increase its basic reproductive number, it is clear from equation (1) that virulence, α , would tend to decrease if T and v are independent of α ; this simplistic argument is the basis for the statements that 'successful parasites and pathogens evolve to be harmless', still commonly found in the medical literature. But, in reality, the damage done to the host, represented by α , is often directly associated with producing transmission stages, and the consequent functional relationships between α , $v(\alpha)$ and $T(\alpha)$ can lead to a multitude of coevolutionary patterns between hosts and pathogens, including virulence increasing over time (for further discussion, see Anderson & May (1991), pp. 648–653).

Direct estimation of the biological components of T for most infections—respiratory infections, for example—is very difficult, if not impossible. But for sexually transmitted diseases (STDs) equation (1) can take a simpler form, with components that are often amenable to direct assessment. The standard formula (see Anderson & May (1991), and references therein) is

$$R_0 = \beta c D. \quad (2)$$

Here, as before, D is the average duration of infectiousness; c is the average rate of acquiring new sexual partners (average number per unit time); and β is the probability that an infected individual will infect a susceptible partner. This oversimplified equation (2) assumes many different kinds of homogeneities—genetic, behavioural, socio-economic, geographical, etc.—within the risk group to which it applies. Understanding the origins of the HIV epidemic, and its development within particular risk groups, of course requires more refined development of the basic equation (2). It nevertheless serves as a useful point of departure (see, for example, Hahn *et al.* 2000).

In particular, it has been clear from the earliest stages of the HIV epidemic that rates of acquisition of new sexual partners vary widely, within both homosexual and heterosexual groups. McKusick *et al.* (1985), for example, found that among homosexual males in San Francisco, the average number of different partners per month was around six, but the standard deviation was of the same order as this average. The earliest models for HIV and AIDS have therefore incorporated a distribution in rates of partner acquisition (May & Anderson 1987), with $p(i)$

being the proportion of the relevant population who on average acquire i new partners per unit time ($\sum_i p(i) = 1$). Adding this refinement to otherwise homogeneous models for HIV, or other STDs, still gives equation (2), but with c now interpreted as

$$c = \langle i^2 \rangle / \langle i \rangle. \quad (3)$$

Here $\langle i^n \rangle = \sum_i i^n p(i)$. Alternatively, equation (3) can be written as

$$c = m + \sigma^2 / m, \quad (4)$$

where m is the mean rate of partner acquisition, and σ^2 the variance.

Treating such populations as homogeneous, which corresponds to interpreting c in equation (2) as simply m , can thus be highly misleading when the variance is high. Thus estimating R_0 by using $c = m$ in equation (2), when in reality equation (4) would give a significantly higher value of c , could seriously underestimate R_0 .

The above discussion of R_0 for HIV and other directly communicated STDs pertains to homosexual transmission, or to heterosexual transmission where both the male-to-female and female-to-male transmission probabilities (β_m and β_f , respectively), and the distributions of rates of acquiring new sexual partners ($p_m(i)$ for female partners of males, and $p_f(i)$ for male partners of females), are identical. More generally, for heterosexual transmission of HIV, equation (2) becomes

$$R_0 = (\beta_m \beta_f c_m c_f)^{1/2} D. \quad (5)$$

Here c_m and c_f are the epidemiologically effective rates of partner acquisition by males and females, respectively, as calculated from equations (3) or (4), with the appropriate distributions. Notice that the average rate of acquiring partners, m , is necessarily the same for males and females. But if the distribution, and hence the variance, of the two behavioural patterns are different, c_m and c_f can differ substantially.

Such results for the spread of HIV in heterosexual populations have some implications that deserve to be more widely recognized. In the early stages of an HIV epidemic in a heterosexual population, growth will be roughly exponential, as given by $\exp[(R_0 - 1)t/D]$, among both females and males. But in these relatively early stages, the ratio of cases among males (C_m) to those among females (C_f) is (see Anderson & May 1991, § 11.3.9):

$$C_m / C_f = (\beta_f c_f / \beta_m c_m)^{1/2}. \quad (6)$$

If we ignore heterogeneities in the sexual-partner distributions among females and males, simply putting $c_m = c_f = m$, we might expect that the incidence of HIV in the epidemic's early stages would be somewhat lower among males, because male-to-female transmission probabilities tend to be somewhat higher than female-to-male (as for other STDs, such as gonorrhoea; see Peterman *et al.* 1988): $C_m / C_f = (\beta_f / \beta_m)^{1/2} < 1$. But, using the correct equation (4) for c_m and c_f , we see that this inequality can easily be reversed if the variance in the partner-acquisition distribution by females is sufficiently greater than that for males, as it arguably often is, by virtue of the role played

by female sex workers. For a more full discussion, see Anderson & May (1991).

4. R_0 AND THE EMERGENCE OF HIV AND AIDS

If HIV has entered the human population through chimpanzees or other primates in the 'bush meat' industry, why did it not do so long ago? One argument, of course, is that such contacts are relatively recent, or at least that earlier contacts were too infrequent for an improbable jump to have occurred. But changes in R_0 for HIV provide an alternative explanation, which we find more convincing.

Of the three components of R_0 in equation (2), one depends largely on the aetiology of HIV infectiousness, one on the sexual behaviour of humans, and one on both biological and behavioural factors.

The duration of infectiousness, D , depends on the way HIV produces transmission stages during its long and variable course in infected individuals. This remains poorly understood. For example, to what extent are individuals infectious during the asymptomatic 'latent period', or does infectiousness have effectively two episodes (around initial infection with HIV, and during the later, AIDS, stages)? Nor can the duration of infectiousness be strictly regarded as a biological given; cultural intervention, in the form of chemotherapy which suppresses viral replication, can clearly and beneficially prolong life, but the corresponding effects on duration and intensity of infectiousness are not entirely clear (Anderson *et al.* 1991).

Although tidily separated in equation (2), the transmission probability, β , ultimately weaves together with considerations of duration of infectiousness. Some aspects of β depend on the aetiology or 'natural history' of HIV itself, while others have behavioural components (use of condoms; circumcision; genital lesions associated with other STDs, whose prevalence depends both on levels of medical treatment for such other STDs and on other factors).

The epidemiologically effective rate of partner acquisition, c , depends on sexual behaviour within the groups under consideration. As seen in equation (4), c can be very sensitive to the distribution of rates of acquiring new partners; those members of the group with exceptionally high such rates are disproportionately important, effectively increasing c (via its σ^2/m term).

In short, in the relatively early stages of the HIV epidemic in Africa, and differently in other places and other groups (for example, homosexual males in developed countries, and needle-sharing drug users), we think that secular changes in D and, to a lesser extent, secular changes in β , are unlikely to be associated with HIV's emergence. But changes in sexual partner-acquisition rates, resulting from increasing movement from rural areas to cities, and other patterns of displacement (especially work-related movements which disrupted earlier patterns of family life), have arguably caused substantial increases in c , compared with the earlier decades of the 20th century. This could easily have produced significant increases in R_0 . Such changes in partner-acquisition distributions, moreover, are such that relatively small differences among distributions, especially

at the high-activity end, $p(i)$ with $i \gg 1$, can result in quite large differences among the consequent R_0 -values.

As an illustrative example, let us assume the transmission probability $\beta = 0.1$ (the same for males and females if transmission is heterosexual), and that $D = 5$ yr. Let us further assume that the population is made up mainly of relatively monogamous individuals who on average acquire, say, 0.5 new partners per year, along with a small fraction or 'core group', f , of very active individuals who acquire on average ten new partners per year (the distribution being the same for males and females). Then, from equations (2) and (3), $R_0 = 0.5(0.25 + 99.75f)/(0.5 + 9.5f)$. If $f = 0.005$, or 0.5%, then $R_0 = 0.75$, and although infectious individuals entering this community may produce subsequent infections, they will be decaying chains, and no epidemic is possible. But if f is 2%, we have $R_0 = 1.6$, and an epidemic can be sustained (provided the initial 'seed' of infection survives initial statistical events; see below). The purpose of this illustration is to emphasize how quite small changes in core groups of people can make very big differences in whether or not an epidemic can take off (and in the fraction of the population affected if it does: see Anderson & May (1991, § 11.9.3)). Notice, incidentally, that if we incorrectly estimated c as the average rate of acquiring partners, even though the situation is heterogeneous, then even if 10% were in the highly active core group, we would nevertheless wrongly estimate $c = 1.45$ and $R_0 = 0.73$, with no epidemic.

More generally, in this two-group model, let individuals in the predominant, relatively monogamous fraction $(1-f)$ acquire on average c_1 new partners per unit time, and the fraction (f) in the highly active core group at the rate c_2 . If we further assume the epidemic is not sustainable within the first group alone (for $f \rightarrow 0$, $R_0 \rightarrow R_1 \equiv \beta D c_1 < 1$), but would be sustainable in a population made up only of the core group (for $f \rightarrow 1$, $R_0 \rightarrow R_2 \equiv \beta D c_2 > 1$), then the core group fraction needed for an epidemic to take off is

$$f > \frac{R_1(1 - R_1)}{(R_2 - R_1)(R_2 + R_1 - 1)}. \quad (7)$$

Assuming R_1 is significantly smaller than one, and R_2 significantly larger than one, the establishment criterion of equation (7) reduces to the useful approximation $f > (1/R_2)(R_1/R_2)$. Had we ignored the nonlinearities in the processes of transmission of infection which amplify the effects of heterogeneity in sexual patterns within this population (i.e. equation (3)), and incorrectly used the simple average rate of partner acquisition in equation (2) for R_0 , the establishment criterion would be $f > (1 - R_1)/(R_2 - R_1)$, or roughly $f > 1/R_2$. This would be larger by a substantial factor, specifically $(R_2 + R_1 - 1)/R_1$ or roughly R_2/R_1 , than the correct criterion of equation (7).

We can rewrite equations (2) and (4) as

$$R_0 = \beta D m [1 + (\text{CV})^2]. \quad (8)$$

Here $\text{CV} = \sigma/m$ is the 'coefficient of variation' of the partner-acquisition distribution. We see that, even with the average rate of partner change, m , remaining constant, along with βD also remaining constant, we can change from a situation in which no epidemic is possible

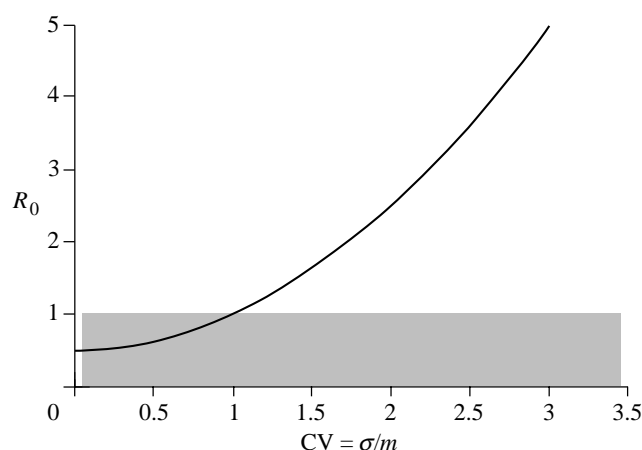


Figure 1. This figure shows how the basic reproductive number, R_0 , depends on the coefficient of variation, $\text{CV} = \sigma/m$, of the distribution in rate of partner acquisition, $p(i)$, for fixed values of β , D , and the average rate of acquiring partners, m . Specifically, $\beta = 0.1$, $D = 5$ yr, $m = 1 \text{ yr}^{-1}$, so that $\beta D m = 0.5$. The shaded area represents $R_0 < 1$, where no epidemic is possible.

(if $\beta D m < 1$), to one where R_0 is much greater than unity, simply by behaviour which changes the CV of the partner-acquisition distribution. Figure 1 illustrates this, using the negative binomial, as familiar in this context, to characterize the form of the partner-change distribution. We believe that such marked changes in R_0 , produced by relatively modest secular changes in patterns of partner acquisition, are likely to play a major part in explaining the different epidemiological trends seen in different parts of Africa, and in the world more generally.

As first emphasized by Yorke *et al.* (1978) in connection with gonorrhoea, such behavioural heterogeneity has important implications for public health policies. If there is marked variation in the patterns of sexual activity within a community, then control efforts—which essentially aim to reduce R_0 —will have much more effect if targeted at the most active individuals and core groups. Again, the above analysis makes plain the essential dynamic mechanisms underlying these insights. Interestingly, there are parallels with other recent work on the relationship between the structure of networks and their vulnerability to disturbance (Albert *et al.* 2000). This work shows that for networks with high CV in the distribution of connections among nodes, as is the case for the Internet and the World Wide Web, the system is robust to random removal of links, but very highly vulnerable to attacks deliberately targeted at the most highly connected nodes. Much of this work translates directly to the dynamics of STDs, and to their control.

There are, of course, many other refinements and complications which are important for detailed understanding of the emergence of HIV in particular places and groups of people. Some of these are discussed below.

5. COMPLICATIONS IN THE EMERGENCE OF HIV

One of the most important questions is, if the strains of HIV currently found in human populations suggest an original 'emergence' back around 1930, why was AIDS

not seen sooner? The explanation that its incidence simply grew slowly, hidden amongst other infections and malnourishment in rural areas of Africa, for 40 years or more, has an air of plausibility. But it does not sit well with epidemiological analysis. In the early stages of exponential growth, an epidemic like HIV and AIDS will grow with a doubling time given roughly by

$$t_d \sim (0.7D)/(R_0 - 1). \quad (9)$$

Observed doubling times in Africa over the past two decades have varied from place to place and among groups, but have generally been of the order of a few years and often shorter. These are consistent with equation (9), given corresponding current estimates of R_0 and D . Given the awesome power of exponential growth, it is hard to see how HIV and AIDS could have escaped attention if it had first appeared around 1930, and doubled at anything like currently observed rates.

Obviously, one explanation is that early R_0 -values were low, possibly mainly because epidemiologically effective partner-acquisition rates, c , were lower. But then one begins to run into the problem of accepting R_0 -values above one, but very close to it (so that $R_0 - 1$ is small). Of course this is possible, but such contrivance is *a priori* unlikely.

This problem arose earlier in sharper form, when work on molecular phylogenies of HIV suggested dates of first emergence around 140 years ago (e.g. Sharp & Li 1988). May & Anderson (1990) suggested two effects, either of which could, in principle, make for longer doubling times in the earlier phases of the epidemic.

(a) Possible effects of two phases of infection

This invoked the possibility that infected individuals had two distinct intervals of infectiousness, one shortly after first becoming infected and another, significantly later, upon entering the AIDS phase. The estimation of the initial doubling rate for HIV seroprevalence now becomes messier, and very tiny secular changes in partner-acquisition rates can reconcile rather slow initial doubling rates with the faster ones currently seen, without invoking implausible values of R_0 only just in excess of unity. We shall not discuss this further here, but refer the interested reader to May & Anderson (1990).

(b) Possible effects of geographical heterogeneity

Perhaps more relevant is the way in which geographical heterogeneity or spatial structure can affect the early spread of a newly emerged pathogen. The following illustration is based on a deliberately oversimplified 'interlinked villages' model, but it makes some basic ideas clear. Details of the analysis, presented here rather telegraphically, are in May & Anderson (1990).

Consider a rural population made up of a large number, n ($n \gg 1$), of roughly equally sized villages. For simplicity, we treat this population as behaviourally homogeneous, with all members of the population acquiring new sexual partners at the average rate c (i.e. zero variance). The probability that an individual in village i will choose any one partner from village j , p_{ij} , is assumed to be $p_{ij} = K(\delta_{ij} + \epsilon)$, where δ_{ij} is the Kronecker delta function ($\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise) and K

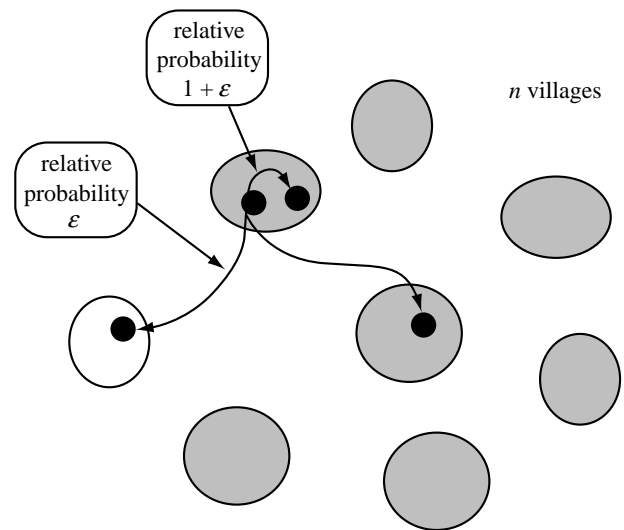


Figure 2. Schematic illustration of the 'loosely interlinked villages' model described in the text, where individuals choose sexual partners from within their own village with probability $K(1 + \epsilon)$, and from any one of the $n - 1$ other villages with probability $K\epsilon$; K is the normalization constant, $K = 1/(1 + n\epsilon)$.

is the normalization constant $K = 1/(1 + n\epsilon)$, ensuring that $\sum_j p_{ij} = 1$. This corresponds to assuming the probability that an individual will choose a partner from any given village other than his or her own, relative to the probability of choosing the 'girl or boy next door' is $\epsilon/(1 + \epsilon)$; we assume $\epsilon \ll 1$. These assumptions are depicted schematically in figure 2. Finally, we assume that HIV appears in one and only one village at $t = 0$, infecting some small initial fraction, Δ , of that village: $y_1(0) = \Delta$. Thereafter HIV spreads, as described by the basic models defined and discussed by May & Anderson (1987, 1988) (see also Anderson & May 1991).

In the early stages of emergence and spread of HIV in this simple model, the fraction of the sexually active population infected at time t , in the village where HIV was first seeded, relative to the initial value Δ , is

$$y_1(t)/\Delta = (1 - 1/n)e^{\rho t} + (1/n)e^{\rho' t}. \quad (10)$$

On average, the corresponding fraction infected in any one of the other $n - 1$ remaining villages, $y_i(t)$ with $i \neq 1$, is

$$y_i(t)/\Delta = (1 - 1/n)(e^{\rho t} - e^{\rho' t}). \quad (11)$$

Here the rate coefficients ρ and ρ' are defined to be

$$\rho = (\beta c D - 1)/D, \quad (12)$$

$$\rho' = [\beta c D/(1 + n\epsilon) - 1]/D. \quad (13)$$

As usual, β is the transmission probability and D the average duration of infectiousness (for details, including the appropriate formula for D when there is overall growth in the population, see May & Anderson (1990)).

These expressions can be understood intuitively, in biological terms. For the overall system of interlinked villages, the basic reproductive number for HIV infections is $R_0 = \beta c D$. This overall value may be partitioned into a component arising from contacts within the same

village, $R_{01} = R_0/(1 + n\epsilon)$, and a component from contacts between villages, $R_{02} = n\epsilon R_0/(1 + n\epsilon)$.

If the basic reproductive rate for HIV exceeds unity for intra-village contacts alone, $R_{01} > 1$, things are relatively straightforward. In this case, we see from equations (12) and (13) that both $\rho' > 0$ and $\rho > 0$, so that both $y_1(t)$ and $y_i(t)$ (with $i \neq 1$) grow exponentially once the infection is seeded in the village labelled 1. Unless R_{01} is only fractionally above unity, the seroprevalence doubling times will not be very different from those estimated by treating this concatenation of loosely coupled subpopulations as if it were a homogeneous population.

But if $R_{01} < 1$, although overall $R_0 = R_{01} + R_{02} > 1$, a more complicated situation arises. Now $\rho' < 0$, although $\rho > 0$. It follows from equation (10) that $y_1(t)$ will at first decrease below the initial value Δ . Only after some time (roughly $\tau \sim (\ln n)/\rho$) will the second term in equation (10) assert itself, and lead to exponential growth of $y_1(t)/\Delta$ above unity. While this is happening, seroprevalence levels in other villages are increasing, on average, from values that are very small early on.

Figure 3 illustrates this situation, for a set of parameter values that are a bit extreme—but not unrealistic—as a metaphor for HIV in Africa. What is happening here is that levels of infection are initially falling in the focal village 1, because the basic reproductive number is below unity for solely intra-village contacts ($R_{01} < 1$).

At the same time, infection is trickling out to other villages, such that overall $R_0 > 1$. In each newly infected village, infection will at first tend to die out ($R_{01} < 1$), but each local decrease must be seen against a rising tide of diffusion of infection to more and more villages. Eventually, this diffuse background rises to levels where the combination of infections originating within a village plus infections imported from other villages makes the process self-sustaining in each individual village, and the seroprevalence levels really take off—as illustrated in figure 2.

The situation illustrated in figure 3 does require that the epidemiological and demographic parameters be such that $R_{01} < 1$ while $R_{01} + R_{02} > 1$ (which can fairly easily happen if $n\epsilon$ exceeds unity). Although restrictive, this constraint is nothing like the exquisitely delicate requirement that R_0 lies marginally above unity, which is essentially the only way to produce sufficiently long doubling times in a homogeneous system.

(c) *Effects of demographic stochasticity*

Although indicative, the above model is ultimately a nonsense. It treats infection levels as continuous variables, even though $y_1(t)$ is initially small and at first gets even smaller, while all other $y_i(t)$ are initially zero and become finite when the first infected individual arrives. Any accurate study must necessarily be based on demographically stochastic models.

To get some feeling for these stochastic effects, let us go back to our original statement that if a new infection is 'seeded' into a large population, it will establish an epidemic if $R_0 > 1$, and die out otherwise. This is true if the seed, although a very tiny fraction of the total population, nevertheless comprises a sufficiently large number of individuals to make statistical fluctuations unimportant. But suppose the seed is only a few, or even one, infected individuals. Even for $R_0 > 1$, and especially if R_0 only just

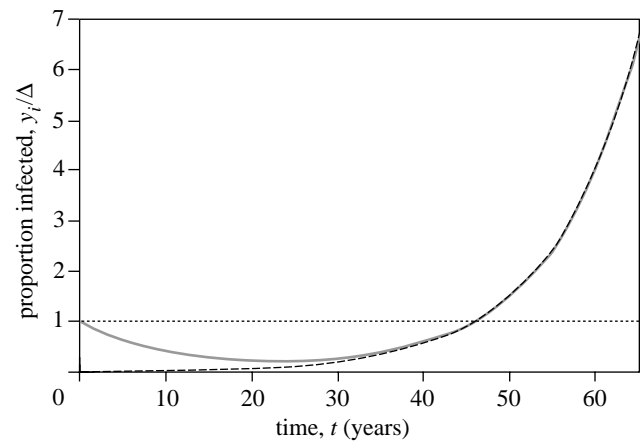


Figure 3. The solid curve shows the fraction infected in the village where HIV first appeared (as a ratio to the initial fraction infected, Δ), y_1/Δ , as a function of the time, t , since the infection first appeared. The dashed curve shows the overall average value of the corresponding fraction infected, y_i/Δ ($i \neq 1$), in any other village. This illustration is for the model defined in the text with the parameters having the values $\beta = 0.1$, $c = 3 \text{ yr}^{-1}$ and $D = 5 \text{ yr}$ (which are not unreasonable estimates of these parameters in the early stages of the epidemic); also $n = 100$ and $\epsilon = 0.02$. Thus, for this parameter combination, $\rho = 0.1 \text{ yr}^{-1}$ and $\rho' = -0.1 \text{ yr}^{-1}$, corresponding to the infection's basic reproductive number being below unity for within-village transmission alone ($R_{01} < 1$), although above unity overall ($R_0 > 1$). The consequent initial decline in the levels of infection within any one 'seeded' village, against a rising tide of diffusion of infection among villages, is discussed in the text.

exceeds unity, early rolls of the dice can see the infection die out if the initial seed is small.

Specifically, if the initial seed is n infected individuals, who mix homogeneously within the susceptible population, the probability that the infection will die out is R_0^{-n} (this is the standard Yule–Kendall process, applying to surnames and all other sorts of things; see Karlin & Taylor 1975). That is, the probability, P , for an infection to establish itself in a large population, if seeded with n infected individuals, is

$$P = 1 - (1/R_0)^n. \quad (14)$$

This is illustrated in figure 4, which shows P as a function of R_0 , for various n . We see that such effects of demographic stochasticity become negligible if n is above 10 or so (unless R_0 is very close to 1), but that they are quite significant if the initial seed is one or two people (unless $R_0 \gg 1$).

Incorporating such stochastic effects into the 'villages' model will clearly lead to even longer times, compared with those illustrated in figure 3 for the deterministic version, before the overall epidemic enters the simple exponential growth phase (growing as $\exp \rho t$) seen at the right of the figure. Such stochastic effects can also result in the infection dying out, in circumstances with $R_0 > 1$ where the deterministic version discussed above would have the epidemic establishing itself.

This being acknowledged, the rough analysis embodied in equations (10)–(13) suggests that more rigorous

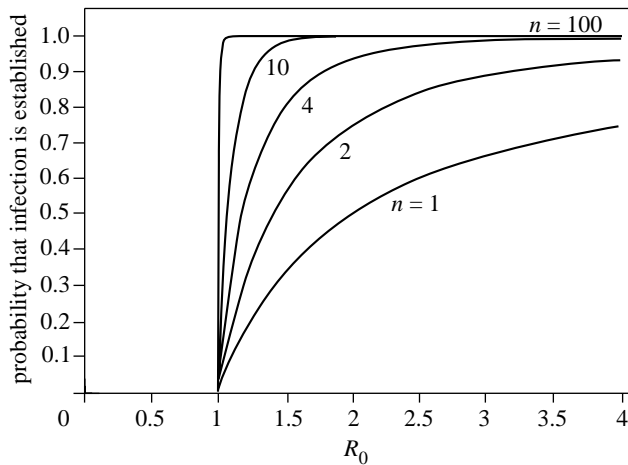


Figure 4. The probability that a newly introduced infection will establish itself in a large population, as a function of its basic reproductive number R_0 , when initially introduced by a 'seed' of n infected individuals (here $n = 1, 2, 4, 10, 100$).

stochastic studies will tend to see initial chains of infection stuttering to extinction in any one village, while at the same time throwing off sparks to ignite new chains in other villages. Although each newly sparked chain will tend to extinguish itself ($R_{01} < 1$), the overall number of ignition points will steadily increase over time ($R_0 > 1$), until eventually a conflagration is lit.

The 'villages' model is oversimplified in other important ways. For one thing, it assumes homogeneity in partner-acquisition rates, whereas we have seen that real heterogeneities can have big effects on R_0 . For another, the above model has random links, independent of details of spatial configuration, among n essentially identical villages; African and other reality is much more complex (involving movement between urban and rural areas; seasonal and other movement of workers; sexual contacts along truck routes; and much else). In short, we recognize that the pattern seen in figure 3 is based on a crude caricature of the possible real effects of spatial structure upon the early emergence of HIV. At the same time, we believe these ideas give a qualitatively reliable account of some of the epidemiological complexities inherent in the kinds of spatial heterogeneities that arise in rural regions of developing countries.

6. CROSS-IMMUNITY AND OTHER CONSIDERATIONS

Up to this point, we have focused on R_0 , which informs us of the chances of successful invasion into a totally susceptible population. However, many invading infectious agents may find themselves within a host population that is not totally susceptible, by virtue of cross-immunity with other infectious agents, notably those from which the new agent may have been derived or to which it may be related. Resulting immunological interference from pre-existing pathogens will limit the resources available to the invading pathogen, and can prevent its spread even when it has R_0 well in excess of unity. This is because the effective reproductive number—the number of secondary cases produced, given that a certain fraction of the population is already immune—may be lower than unity under these circumstances.

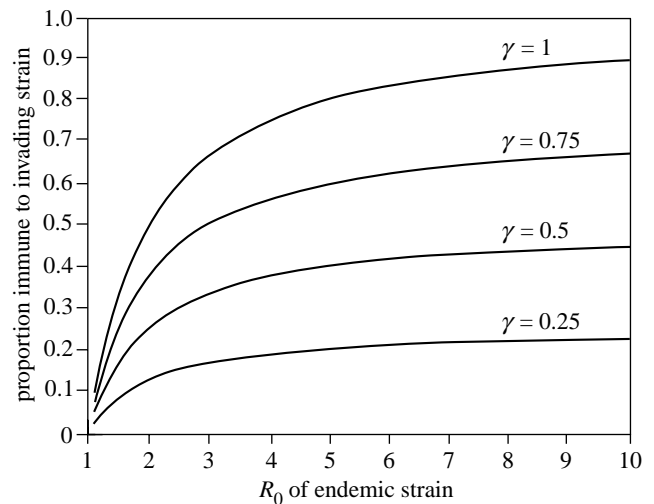


Figure 5. The proportion of a host population immune to an invading strain, by virtue of cross-infection from an endemic strain with basic reproductive number R_0 (the R_0^A of equations (15) and (16)). The figure shows the immune proportion for various values of the cross-immunity parameter, γ .

In essence, the invading pathogen must be able to find for itself a host population with a susceptible fraction of at least $1/R_0$. Suppose there is immunological interference, caused by some other pathogen, A, which is already established in the host population and whose basic reproductive number is $R_0^A > 1$. The fraction of the host population susceptible to the invading infection is then $1 - \gamma(1 - 1/R_0^A)$, where γ is the degree of cross-immunity between the two infectious agents; $(1 - 1/R_0^A)$ is the equilibrium fraction of the host population to have experienced infection with the pathogen A. The invasion criterion thus becomes

$$R_0[1 - \gamma(1 - 1/R_0^A)] > 1. \quad (15)$$

This restriction on R_0 is shown, as a function of γ and R_0^A , in figure 5. Equation (15) can alternatively be expressed as

$$\gamma < (1 - 1/R_0)/(1 - 1/R_0^A). \quad (16)$$

Notice that if $R_0 > R_0^A$, the inequality of equation (15) or (16) is automatically satisfied (remember, $\gamma < 1$) and cross-immunity has no effect. But if $R_0 < R_0^A$ equation (16) can be violated if γ is sufficiently large; in this case, establishment of the new infection is prevented by cross-immunity.

More generally, the invading infection may face competition, in the form of cross-immunity, from several other pathogens which are endemic within the host population. The analogues of equation (15) will then be more complex, depending on the magnitudes of the cross-immunities the invader faces from other established pathogens, as well as on the networks of immunological interactions among these pathogens themselves. The effect will, however, broadly be as above, only more so.

In addition, the discussion epitomized by equations (15) and (16) pertains to competitive interactions between pathogens at essentially a single-locus level.

Where multiple loci or multiple epitopes are involved, the situation is more complex. The pathogen is then subject to constraints arising from a network of interactions among infectious agents, with varying degrees of overlap between their antigenic determinants. It has been shown that, under strong immune selection, such populations of pathogens may self-organize non-overlapping combinations of antigenic variants. In other words, instead of all possible combinations of alleles or epitopes, a set of discrete non-overlapping combinations of alleles or epitopes will dominate at any given time, as these types will suppress all but each other through cross-immunity. For an invading pathogen to breach this asymmetrical equilibrium, it must compete effectively with all the dominant strains with which it shares antigenic variants. This is effectively impossible for new strains arising through recombination, even when they have R_0 -values very much higher than the dominant strains. Mutations arising at the relevant loci may alter the balance among the circulating strains. To occupy a unique immunological niche, however, an emerging pathogen must accrue new variants at all the relevant loci. For a more detailed discussion of these considerations, with some data bearing upon possible examples, see Gupta *et al.* (1996, 1998).

Finally, within a wide range of intermediate levels of immune selection, such multilocus or multiple-epitope systems appear to exhibit cyclical or chaotic fluctuations in strain frequency. The potential consequences for an invading strain are twofold. On the one hand, the invader may be more prone to elimination by stochastic events, because it will itself be subject to the fluctuations inherent in the system. On the other hand, the invader may find itself an 'immunological' niche within this mayhem, when the composition of the dominant strains suddenly changes in its favour. Again, see Gupta *et al.* (1996, 1998).

Although often anecdotal, the medical literature contains several discussions and conjectures about the effects of such cross-immunity on the emergence of new diseases.

For example, the popular impression is that an earlier STD, syphilis, emerged in Europe when Columbus's voyage brought it from the New World. But Cockburn (1961) and others have argued that treponeme infections have long been established in the Old World. Cockburn further argued that, before the renaissance, when hygiene generally was bad, skin-to-skin transmission was favoured, and yaws was the dominant treponeme infection in Europe. Moreover, he suggested, cross-immunity resulted in yaws effectively excluding syphilis. As more hygienic environments began to prevail in the late 15th century, skin-to-skin transmission diminished, yaws waned, and the door was opened for syphilis to emerge, indigenously, in Europe. Such a view puts cross-immunity centre stage in the emergence of syphilis—not introduced by Columbus, but a result of better hygiene, which, like Columbus's voyages themselves, arose from the new spirit of the renaissance. For further discussion, see Hackett (1963) and Wilcox (1973).

Slomka (1996) has discussed the possibility that cross-immunity may help explain the changing patterns in genital herpes simplex virus, HSV. In Europe, there appears to be a change from HSV-2 to HSV-1 (e.g. Nilsen

& Myrnel 2000; Ross *et al.* 1993), although puzzlingly with an opposite trend in the USA (Halioua & Malkin 1999; Kinghorn 1994). Directly relating to patterns in the emergence of HIV, Anderson & May (1996) have discussed how competition between HIV-1 and HIV-2, mediated by cross-immunity, can explain some of the observed epidemiological data.

Potentially important for understanding the role of cross-immunity in the transmission dynamics and aetiology of HIV is the recent finding that co-infection with scrub typhus, *Orientia tsutsugamushi* (formerly *Rickettsia*), may suppress HIV (Watt *et al.* 2000; J. Cohen 2000). Watt *et al.* report that ten AIDS patients in Thailand who became infected with scrub typhus, but had no other infections, had markedly lower HIV loads than a control group with AIDS and malaria or leptospirosis. Preliminary data from test tube and mouse experiments with human sera suggest that antibodies against *O. tsutsugamushi* also bind HIV, being particularly effective in preventing formation of syncytia. This work is in its earliest stages, but it hints at practical realization of the effects of cross-immunity, discussed abstractly in equations (15) and (16).

Moving beyond STDs, and into even more anecdotal territory, it has been speculated that the decline of leprosy in most countries in Europe (Norway being a conspicuous exception until recently) from the 15th century onwards resulted, in part, from cross-immunity acquired from exposure to tuberculosis (see Lietman *et al.* (1997), and references therein). It has also been speculated that the reason why yellow fever is not found in Asia, even though its main vector (*Aedes aegypti*) is found there, is that the high incidence of another haemorrhagic fever, dengue (also spread mainly by *A. aegypti*) in Asia has excluded it by cross-immunity.

In summary, cross-immunity, both among strains of a specific pathogen and between different pathogens, can influence—and arguably has influenced—the emergence of new infections. Clearly, such considerations, possibly carried down to the level of epitopes of a virus (Nowak & May 2000), can also be relevant to the design of vaccines.

Many other kinds of complications can arise in more detailed and realistic exploration of the themes developed in this paper. For instance, although we emphasize the importance of behavioural heterogeneities in rates of acquiring new sexual partners, we assumed partners were selected randomly within the proportional constraints set by the activity distribution. But it could be that choice is more 'assortative', with highly active individuals mixing essentially only with others in their activity class; conversely, there could be 'disassortative' patterns, with very active individuals differentially favouring less active partners (Gupta *et al.* 1989). Many other complications can be relevant, associated with such factors as: age-specific patterns of sexual partner choice, possibly different for males and females; geographical location, and/or patterns of movement; socio-economic status; genetic heterogeneity within the population which affects susceptibility; and a host of other effects relating to aspects of human behaviour. Many of these issues are reviewed in Anderson & May (1991), but others still await attention.

7. DIFFERENT KINDS OF EMERGING DISEASES

Emerging diseases are nothing new in human history. Indeed, our history can be written largely in terms of newly emergent diseases (e.g. McNeill 1976; Diamond 1997). For most such diseases, an important concept is the 'threshold population', N_T , below which host numbers are too small to maintain the infection with the population. Such threshold population sizes can be set by the criterion $R_0 > 1$, via the effect of population size on the transmission rate $T(\alpha, N)$ of equation (1), but they more usually derive from 'fade out' considerations. Directly communicated infections with short duration of infectiousness—as is the case for most of today's childhood infections—require a very large population, even when $R_0 \gg 1$, to maintain a continuing supply of infected individuals at all times. Failing this, the infection will 'fade out'. For measles, for example, the threshold population size is around 300 000 (Bartlett 1957; Black 1966; Keeling & Grenfell 1997). Infections such as smallpox, measles, rubella, pertussis, and the like first emerged around 10 000 years ago, when, following the advent of agriculture and associated settlements, human numbers exceeded the levels for such infections to be maintained. Infections which involve intermediate vectors, and/or which involve long periods of infectiousness, can be maintained in populations as small as those characteristic of pre-agricultural bands of hunter-gatherers. Even here, the population size and structure can be important. Returning to HIV and other STDs, equation (2) makes it clear that population size as such is unimportant; what matters is epidemiologically effective rates of partner change, c .

We conclude by re-emphasizing the differences between a disease's ability to jump from one species into a new host species, and its ability autonomously to establish itself and spread within the new population.

Many ecologically important diseases of non-human animals arise sporadically, or even continuously, even though $R_0 < 1$ for that species, as a result of infection from a reservoir in another host species. For example, avian malaria has extinguished several species of Hawaiian birds. It is, however, ordinarily very difficult for a disease, however nasty it is and however large R_0 is, to extinguish its host population, because—before reaching the extinction point—the host population becomes too small to maintain the infection. However, if the disease arises in a given population, where it is not self-sustaining ($R_0 < 1$), by riding on the shoulders of some primary reservoir of infection sustained in another population, then extinction is possible (requiring only that the disease-associated death rate exceeds the pre-infection basic reproductive rate in this population). This appears to have been what happened to the Hawaiian birds in question, with the reservoir of avian malaria being maintained by migrating wildfowl.

It follows that for newly emerging infections of humans, we need always to ask about the disease ecology. If the infection has $R_0 < 1$ in its human host, then the focus should be on identifying the source and potential reservoir, and reducing or eliminating infective contacts. Quarantine and other control measures are of course also important (especially if we are unsure of the magnitude

of R_0 , as we often will be), but are much less important if $R_0 < 1$ than if $R_0 > 1$. Thus the few cases of bubonic plague seen each year (as a result of contacts with rodents which maintain *Yersinia pestis*) in the south-western USA excite relatively little attention, but this would be vastly different if we had a medieval $R_0 > 1$ for subsequent transmission within the human population.

If $R_0 > 1$ for a newly emerging infection, then the origin of the infection, although important to understand, is less important than intervention aimed at stopping a self-sustaining epidemic.

For HIV, we have, in effect, a mixture of both kinds of situation, with $R_0 > 1$ in some groups and places, and $R_0 < 1$ in others. It is very clear that R_0 exceeded unity among homosexual males in many large cities in the developed world in the early stages of the epidemic, and that R_0 currently exceeds unity for transmission among needle-sharing drug users in many places and, most notably, for heterosexual transmission in many parts of Africa and other countries. On the other hand, for heterosexual transmission within many groups in developed countries, it remains unclear whether we have self-sustaining although slow-growing epidemics ($R_0 > 1$), or whether we are seeing decaying chains of infection within these groups ($R_0 < 1$), with infection persistently sustained by contacts with members of needle-sharing or bisexual groups where $R_0 > 1$.

In short, questions about the emergence of new diseases involve a lot more than where the infection came from. And for a self-sustaining epidemic, such as HIV, questions of its origin—although interesting for the lessons to be learned—are of secondary importance beside those of effective therapies and ultimate control.

REFERENCES

- Albert, R., Jeong, H. & Bacabasi, A.-L. 2000 Error and attack tolerance of complex networks. *Nature* **406**, 378–382.
- Anderson, R. M. & May, R. M. 1979 Population biology of infectious diseases. Part I. *Nature* **280**, 361–367.
- Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans: dynamics and control*. Oxford University Press.
- Anderson, R. M. & May, R. M. 1996 The population biology of the interaction between HIV-1 and HIV-2: coexistence or competitive exclusion? *AIDS* **10**, 1663–1673.
- Anderson, R. M., Gupta, S. & May, R. M. 1991 Potential of community-wide chemotherapy or immunotherapy to control the spread of HIV-1. *Nature* **350**, 356–359.
- Bartlett, M. S. 1957 Measles periodicity and community size. *J. R. Statist. Soc. A* **120**, 48–70.
- Black, F. L. 1966 Measles endemicity in insular populations: critical community size and its evolutionary implication. *J. Theor. Biol.* **11**, 207–211.
- Cockburn, T. A. 1961 The origin of the treponematoses. *Bull. WHO* **24**, 221–228.
- Cohen, A. N. & Carlton, T. L. 1995 *Nonindigenous aquatic species in a United States estuary: a case study of the biological invasions of the San Francisco bay and delta*. Washington, DC: US Fish and Wildlife Service.
- Cohen, J. 2000 The odd intersection of HIV and scrub typhus. *Science* **289**, 1278.
- Cohen, M. L. 2000 Changing patterns of infectious disease. *Nature* **406**, 762–767.
- Crawley, M. J. 1986 The population biology of invaders. *Phil. Trans. R. Soc. Lond. B* **314**, 711–731.

- Diamond, J. M. 1997 *Guns, germs and steel*. New York: Random House.
- Elton, C. S. 1958 *The ecology of invasions by animals and plants*. London: Methuen.
- Enserink, M. 2000a Malaysian researchers trace Nipah virus outbreak to bats. *Science* **289**, 518–519.
- Enserink, M. 2000b New arenavirus blamed for recent deaths in California. *Science* **289**, 842–843.
- Fisher, R. A. 1930 *The genetical theory of natural selection*. New York: Dover.
- Gupta, S., Anderson, R. M. & May, R. M. 1989 Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS* **3**, 807–817.
- Gupta, S., Maiden, M. C., Feavers, I. M., Nee, S., May, R. M. & Anderson, R. M. 1996 The maintenance of strain structure in populations of recombining infectious agents. *Nat. Med.* **2**, 437–442.
- Gupta, S., Ferguson, N. & Anderson, R. M. 1998 Chaos, persistence, and evolution of strain structure in antigenetically diverse infectious agents. *Science* **240**, 912–915.
- Hackett, C. J. 1963 On the origin of the human treponematoses. *Bull. WHO* **29**, 7–41.
- Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. 2000 AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614.
- Halioua, B. & Malkin, J. E. 1999 Epidemiology of genital herpes—recent advances. *Eur. J. Dermatol.* **9**, 177–184.
- Karlin, S. & Taylor, H. M. 1975 *A first course in stochastic processes*, 2nd edn. New York: Academic Press.
- Keeling, M. J. & Grenfell, B. T. 1997 Disease extinction and community size: modelling the persistence of measles. *Science* **275**, 65–67.
- Kinghorn, G. R. 1994 Epidemiology of genital herpes. *J. Int. Med. Res.* **22**, 14A–23A.
- Kornberg, H. & Williamson, M. H. (eds) 1987 *Quantitative aspects of the ecology of biological invasions*. London: The Royal Society.
- Lietman, T., Porco, T. & Blower, S. 1997 Leprosy and tuberculosis: the epidemiological consequences of cross-immunity. *Am. J. Publ. Hlth* **87**, 1923–1927.
- May, R. M. & Anderson, R. M. 1987 Transmission dynamics of HIV infection. *Nature* **326**, 137–142.
- May, R. M. & Anderson, R. M. 1988 The transmission dynamics of human immunodeficiency virus HIV. *Phil. Trans. R. Soc. Lond. B* **321**, 565–607.
- May, R. M. & Anderson, R. M. 1990 Parasite–host coevolution. *Parasitology* **100**, S89–S101.
- McKussick, L., Harlsman, W. & Coates, T. J. 1985 AIDS and sexual behaviour reported by gay men in San Francisco. *Am. J. Publ. Hlth* **75**, 493–496.
- McNeill, W. H. 1976 *Plagues and peoples*. New York: Doubleday.
- Mooney, H. A. & Drake, J. A. (eds) 1986 *Ecology of biological invasions of North America and Hawaii*. New York: Springer.
- Morse, S. S. (ed.) 1993 *Emerging viruses*. Oxford University Press.
- Nilsen, A. & Myrnel, H. 2000 Changing trends in genital *Herpes simplex* virus infection in Bergen, Norway. *Acta Obstet. Gynecol. Scand.* **79**, 693–696.
- Nowak, M. A. & May, R. M. 2000 *Virus dynamics*. Oxford University Press.
- Peterman, T. A., Stoneburner, R. L., Allen, J. R., Jaffe, H. W. & Curran, J. W. 1988 Risk of HIV transmission from heterosexual adults with transfusion-associated infections. *J. Am. Med. Assoc.* **259**, 55–63.
- Ross, J. D., Smith, I. W. & Elton, R. A. 1993 The epidemiology of *Herpes simplex* types 1 and 2 infection in Edinburgh 1978–1991. *Genitourin. Med.* **69**, 381–383.
- Sharp, P. M. & Li, W. H. 1988 Understanding the origins of AIDS viruses. *Nature* **336**, 315.
- Slomka, M. J. 1996 Seroepidemiology and control of genital herpes: the value of type specific antibodies to *Herpes simplex* virus. *Bur* **6**, 41–45.
- Watt, G. (and 11 others) 2000 HIV-1 suppression during acute scrub-typhus infection. *Lancet* **356**, 475–479.
- Wilcox, R. R. 1973 The treponemal evolution. In *Sexually transmitted diseases* (ed. L. Nicholas), pp. 3–25. Springfield, IL: Charles Thomas.
- Yorke, J. A., Hethcote, H. W. & Nold, A. 1978 Dynamics and control of the transmission of gonorrhoea. *J. Sex. Trans. Dis.* **5**, 51–56.