Bulletin of Mathematical Biology Vol. 53, No. 1/2, pp. 3–32, 1991.
Printed in Great Britain.

0092–8240/91$3.00+0.00
Pergamon Press plc

DISCUSSION: THE KERMACK–McKENDRICK EPIDEMIC THRESHOLD THEOREM

The paper reviews the work of Kermack and McKendrick on the development of simple mathematical models of the transmission dynamics of viral and bacterial infectious agents within population of hosts. The focus of attention is centred on the notion of a threshold density of susceptible hosts to trigger an epidemic and recent extensions of this idea as expressed in the definition of a basic or case reproductive rate of infection. The main body of the paper examines recent developments of the basic Kermack–McKendrick model with an emphasis on deterministic models that describe various types of heterogeneity in the processes that determine transmission between infected and susceptible persons. Particular attention is given to the role of behavioural heterogeneity within the framework of a contact or mixing matrix which defines "who acquires infection from whom".

Introduction. Human fascination with epidemics of infectious diseases and the associated patterns of mortality has a long history. The lists of epidemics compiled by the Chinese scholar, Ssu Kwong, who lived during the Sing Dynasty (AD 960–1279) in China, the "epidemics" of the Greek scholar Hippocrates (458–377 BC) and the rudimentary medical statistics of John Grant (1620–74) and William Petty (1623–87), who studied the London Bills of Mortality in the seventeenth century, well illustrate this point. However, the scientific study of the epidemiology of infectious diseases did not begin in earnest until the development of the "germ theory of disease".

In the earliest medical literature there are expressions of the idea that invisible living creatures might be responsible for disease. Such reference is found, for example, in the writings of Aristotle (387 BC) and of the Arabian physician, Rhazes (AD 860–938). However, reliable methods for the isolation and identification of such creatures ("microbes") were not developed until the nineteenth century. The work of three outstanding scientists, Pasteur (1827–75), Lister (1827–1912) and Koch (1843–1910), laid the foundations of modern microbiology and established a set of principles (often referred to as "Koch's postulates") for establishing the relationship between disease and the presence of an infectious agent within the host.

The application of mathematics to the study of infectious disease appears to have been initiated by Daniel Bernoulli in 1760 by his use of a simple mathematical method to evaluate the effectiveness (in terms of an improvement in life expectancy) of the technique of variolation to protect against smallpox infection (Bernoulli, 1760). Since this early beginning developments have been many and varied, although it is probably fair to say that their impact on public health policy and planning for the prevention of infection and associated disease has been rather limited (Fine, 1979; Anderson and May, 1982; Dietz and Schenzle, 1985). In part this is a consequence of the tendency for theory to have become rather detached from its empirical base.

The origins of modern theoretical epidemiology owe much to the work of En'ko (1889), Hamer (1906), Ross (1908) and Kermack and McKendrick (1927). Hamer (1906) postulated that the course of an epidemic depends on the rate of contact between susceptibles and infectious individuals. This notion has become one of the most important concepts in mathematical epidemiology: it is the so called "mass action" principle of transmission for directly transmitted viral and bacterial infections. The principle is based on the idea that the net rate of spread of infection is proportional to the product of the densities of susceptible and infectious persons. The idea was originally formulated in a discrete-time model, but in 1908 Ronald Ross (celebrated as the discoverer of malaria transmission by mosquitoes) translated the problem into a continuous time framework in his pioneering work on the transmission dynamics of malaria (Ross, 1915; Ross and Hudson, 1917). The ideas of Hamer and Ross were extended and explored in more detail by Kermack and McKendrick (1927) and Soper (1928).

The problem addressed by Kermack and McKendrick in their classic 1927 paper was of great topical interest and the source of much controversy at that time. The "bell shaped geometry" of the epidemic curve was well understood (Fig. 1) but the controversy centred on the factor or factors that determined both the magnitude of the epidemic and its termination within a given population.

Two explanations for the termination of an epidemic were most in favour amongst medical circles at that time, namely: (1) that the supply of susceptible people had been exhausted and (2) that during the course of the epidemic the virulence of the infectious agent had gradually (or rapidly) decreased. Kermack and McKendrick (1927) addressed this problem by formulating a simple deterministic model of the transmission of a directly (= contact) transmitted viral or bacterial agent in a closed population (no birth, death, immigration or emigration). In this model, they decribed the course of "sickness" in an individual and recovery or death by a set of "vital rates" denoting infectivity, recovery and death. The conclusions they arrived at are best summarized in their own words.

> "... the course of an epidemic is not necessarily terminated by the exhaustion of the susceptible members of the community. It will appear that for each particular set of infectivity, recovery and death rates, there exists a critical or threshold density of population. If the actual population density be equal to (or below) this threshold value the introduction of one (or more) infected persons does not give rise to an epidemic, whereas if the population be only slightly more dense a small epidemic occurs. It will appear also that the size of the epidemic increases rapidly as the threshold density is exceeded, and in such a

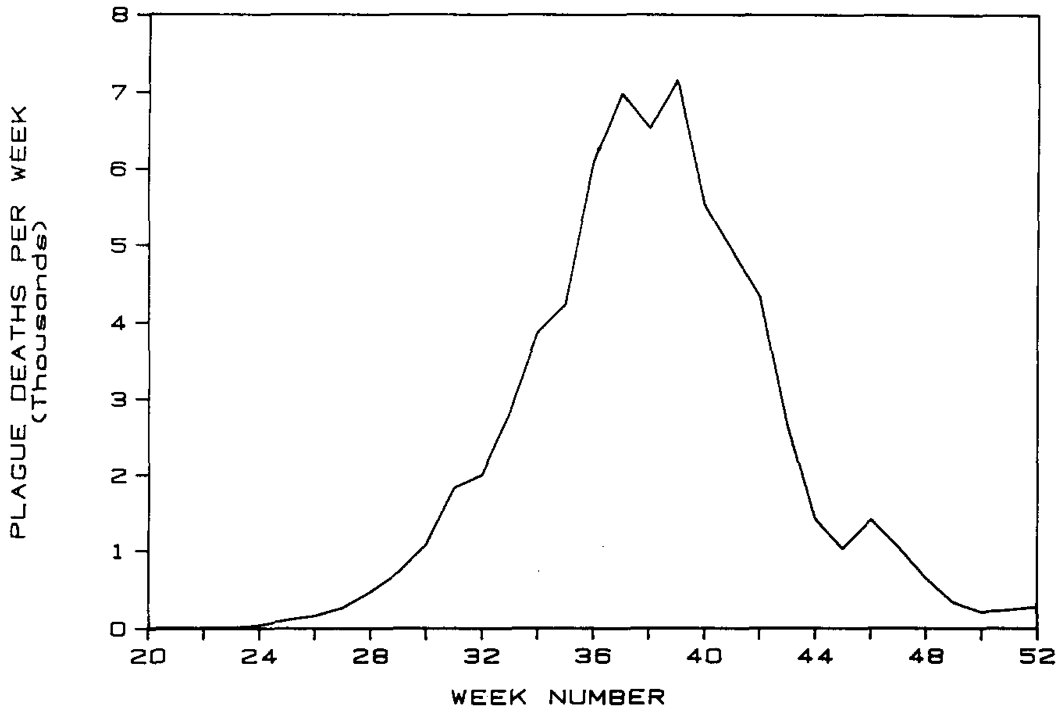


Figure 1. The 1665 Plague epidemic in London. Weekly reports of deaths due to the plague are recorded in Daniel Defoe's journal (Brayley, 1722).

manner that the greater the population density (of susceptibles) at the beginning of the epidemic, the smaller will it be at the end of the epidemic. In such a case the epidemic continues to increase so long as the density of the unaffected population is greater than the threshold density, but when this critical point is approximately reached the epidemic begins to wane and ultimately die out. This point may be reached when only a small proportion of the susceptible members of the community have been affected."

Today this concept plays a central role in much of the mathematical theory of infectious disease epidemiology and it has very important implications for the design of control programmes based, for example, on mass immunization (see Anderson and May, 1983). In this paper, a brief review is presented of the published work of Kermack and McKendrick from 1927 to 1939. The major body of the paper addresses recent developments in this area with particular focus on the significance of heterogeneity in transmission induced by a wide variety of factors including genetic, spatial and behavioural processes. The article is written with the epidemiologist and ecologist in mind and technical details are kept to a minimum. Those interested in the techniques of model formulation and analysis are referred to the appropriate source references.

The Kermack–McKendrick Model. In their 1927 paper, Kermack and McKendrick considered the following problem with directly (=contact) transmitted viral and bacterial agents in mind. One or more infected persons is/are introduced into a closed population (no birth, emigration or immigration) of susceptible individuals and the infectious disease spreads from the affected to the unaffected by contact infection. Each infected individuals runs through the course of his or her sickness and is finally removed by recovery or death. They considered the general case in which the rate of infectiousness, recovery or death depend on the duration (=stage) of infection in an individual. They assumed that the total population remained constant in size during the course of the epidemic, excepting modifications in size caused by deaths due to the disease, and that on recovery complete and lasting immunity was induced to the infectious agent under consideration. Their model was compartmental and deterministic in structure, where the total population at time t, $N(t)$, was divided into susceptibles, infecteds (=infectious people) and recovereds denoted respectively by $X(t)$, $Y(t)$ and $Z(t)$ at time t. Under the assumption that the "vital" rates of transmission (=infectivity) and recovery/death are constant and independent of the duration of infection in an individual, the basic differential equations are of the form:

$$\mathrm{d}X/\mathrm{d}t = -\beta XY \tag{1}$$

$$\mathrm{d}Y/\mathrm{d}t = \beta XY - vY \tag{2}$$

$$\mathrm{d}Z/\mathrm{d}t = vY. \tag{3}$$

Here β denotes the transmission coefficient (recording infectiousness and the probability of contact between people) and v records the recovery/death rate in the infected class Y. If the disease does not induce mortality (v is the recovery rate) then the class Z denotes those who have recovered and are immune to reinfection. If v records mortality, then the class Z records the number of deaths induced by the epidemic.

The existence of a critical threshold density of susceptibles for the occurrence of a major epidemic can be deduced in a heuristic manner, via inspection of the right-hand side of equation (2). Following the introduction of a few infecteds [$Y(0)$ at time $t=0$] into a susceptible population [where $X(0) \simeq N(0)$], the rate of increase in the density of infecteds will only be positive provided the term $\beta X > v$, if a major epidemic is to occur. More formally, the critical density of susceptibles, N_T, is given by

$$N_T = v/\beta. \tag{4}$$

No epidemic can occur unless the population of susceptibles exceeds this value (the recovery rate, v, divided by the transmission coefficient, β), and if it does exceed this value then the size of the epidemic (to a first approximation) is roughly equivalent to $2n$ times the degree to which the initial population density of susceptibles exceeds the critical value [$n = N(0) - N_T$]. Thus at the end of the epidemic the density of susceptibles will be as far below the threshold density, as initially it was above (Kermack and McKendrick, 1927).

Embodied within the threshold density concept is a further idea of fundamental importance in epidemiology, namely, that of a basic or case reproductive rate of infection, R_0, which records the average number of secondary cases of infection produced by one primary case in a totally susceptible population. The idea is directly analogous to Fisher's concept of a net reproductive rate which is widely used in the disciplines of population genetics, ecology and demography (Fisher, 1930). For an infectious agent to spread in a population it is intuitively obvious that

$$R_0 \geqslant 1. \tag{5}$$

If this condition is not satisfied the infection will die out (see Ross, 1915; Macdonald, 1957; Dietz, 1974; Anderson and May, 1979; May and Anderson, 1979; Anderson, 1982).

In the context of the simple Kermack–McKendrick model [equations (1)–(3)] R_0 is defined as

$$R_0 = \beta N / v. \tag{6}$$

Note that the concept of a threshold density of susceptible (N) for R_0 to exceed unity in value is explicit in this definition. Although Kermack and McKendrick did not use the notion of a basic or case reproductive rate, the definition of R_0 is clearly detailed in their analyses of the behaviour of the simple model (i.e. see p. 716 of Kermack and McKendrick, 1927). In the 1927 paper, and subsequent ones, they also addressed the more general case where both infectivity and recovery are functions of the duration of infection in an individual (e.g. distributed infectious and recovery rates). If we denote β and v as functions of the time s since infection then R_0 is defined as

$$R_0 = N \int_0^\infty \beta(s) \exp\left[-\int_0^s v(a)\,\mathrm{d}a\right] \mathrm{d}s. \tag{7}$$

Hence the threshold density, N_T, is given by

$$N_T = 1 \Bigg/ \left[\int_0^\infty \beta(s) \exp\left[-\int_0^s v(a)\,\mathrm{d}a\right] \mathrm{d}s\right]. \tag{8}$$

A further problem addressed by Kermack and McKendrick was the issue of what fraction of the population will be infected during the course of an epidemic in a closed population. If we denote this fraction as I then in the general case, where a small fraction $I(0)$ is infected at time $t = 0$,

$$I = 1 - (1 - I(0)) \exp(-R_0 I). \tag{9}$$

A very clear discussion of the properties of this equation has recently been presented by May (1990). It has one, and only one, solution corresponding to a finite value of I (Fig. 2). As explained by May (1990), this contrasts with the crisp distinction made earlier between the case $R_0 < 1$ (no epidemic) and $R_0 > 1$ (an epidemic). This issue has caused some confusion, particularly in the epidemiological literature concerned with plant pathogens (Jeger, 1986).

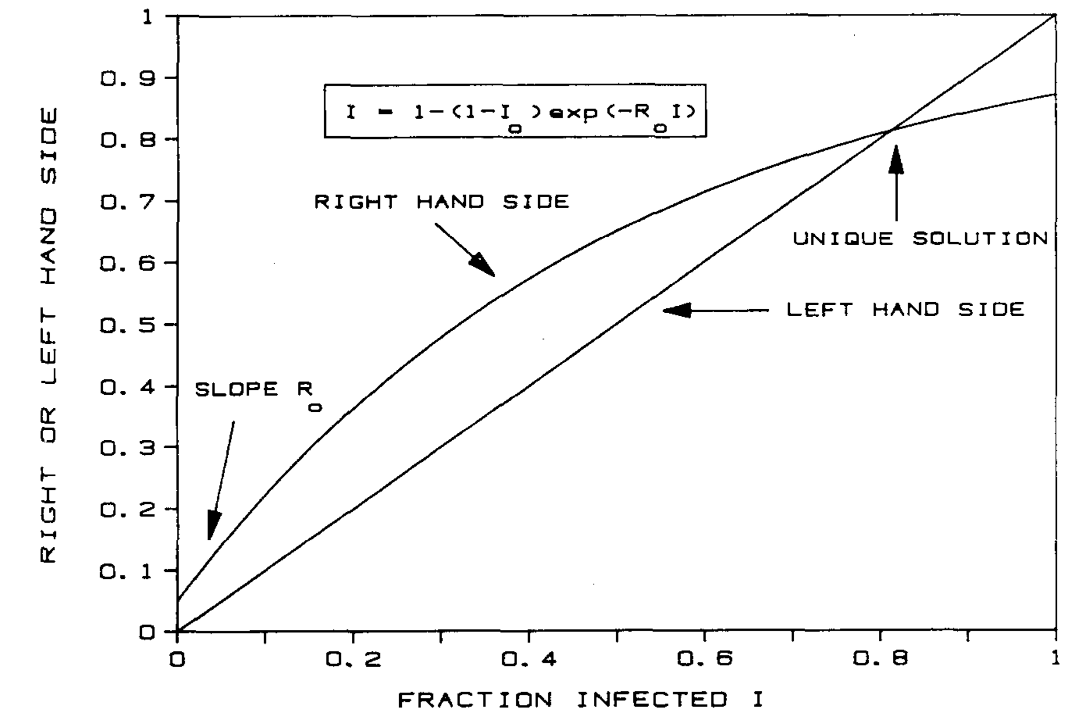


Figure 2. The fraction infected, I, in a closed epidemic in a population of size N where a proportion $I(0)$ are infected at time $t=0$. The graph records plots of the right and left hand sides of equation (9) in the main text for various values of I with $R_0=2.0$ and $I(0)=0.05$. See text for explanation.

However, there is a very straightforward explanation of this situation which rests on the distinction between a small trickle of secondary cases ensuing from the introduction of infection into a susceptible population, and the occurrence of a major epidemic. Even if $R_0 < 1$ (and $N < N_T$) there will always be some subsequent infection leading to the total infected exceeding the original seed of infection, but this will be a decaying chain of infection events with no "runaway" chain reaction or epidemic. (It is interesting to note the similarity of the concept of chain infection events and R_0 with those employed in chemical chain reactions and, in particular, nuclear chain reactions.) Conversely if $R_0 > 1$, the chain of subsequent infection events expands rapidly to generate what is commonly called an "epidemic". The arguments above, of course, apply to the situation where $I(0)$ is small—in other words when we are considering the introduction of a few infecteds into a largely susceptible population.

In subsequent papers, Kermack and McKendrick (1927, 1932, 1933, 1937, 1939) expand their simple theory to consider the question of endemicity of infection, where the assumption of a closed population is relaxed, and to compare theoretical predictions with the patterns of infection recorded in a classic series of experiments on viral and bacterial diseases in mouse colonies carried out by Greenwood *et al.* (1936). The elegance of their work lies in part on their efforts to estimate parameters from data and compare prediction with observed pattern.

The issue of what factors influence the endemic persistence of an infection has received much attention since the work of Kermack and McKendrick (e.g. for reviews see Anderson and May, 1979, 1985a; May and Anderson, 1979). For example, if we extend the simple closed epidemic model [equations (1)–(3)] by the addition of a net immigration rate of new susceptibles, Λ, to the right-hand side of equation (1), and subtract mortality terms from each equation under the assumption of a constant per capita mortality rate, μ, then it can be shown that an endemic equilibrium is attained with the infection persisting in the population, provided

$$\Lambda/\mu > N_T \tag{10}$$

where N_T is as defined in equation (4) with the addition of a term μ to the denominator. Equation (10) simply states that the equilibrium population density in the absence of infection, Λ/μ, must exceed the critical density of susceptibles required to initiate an epidemic (Anderson and May, 1979). These notions can be extended to consider the persistence of infection in a population subject to births and deaths (as opposed to immigration and death) and the ability of infectious agents to regulate the abundance of their host population (Anderson, 1979).

As noted earlier, the simple model of Kermack and McKendrick and the threshold theorem derived from this model has played a pivotal role in subsequent developments in the study of the transmission dynamics of infectious diseases. In the following sections a series of recent developments are reviewed with an emphasis on the implications of theory for the control of infections by mass vaccination and the significance of various types of heterogeneity in the transmission process on patterns of infection within communities.

Control by Mass Vaccination. Two central questions in the design of mass vaccination programmes for the control of childhood viral and bacterial infections are what proportion of the community must be immunized to interrupt or block transmission and what is the optimum age at which to administer the vaccine? These questions can be addressed via the use of simple mathematical models which are extensions of the basic Kermack–McKendrick equations. The developments necessary to crudely mirror the known properties of many common childhood infections, such as measles and rubella, include further compartmentalization of the population to include a class of infants who are protected from infection by maternally derived antibodies and a latent class who are infected but not yet infectious, plus the inclusion of age structure (see Anderson and May, 1983, 1985a; Dietz and Schenzle, 1985).

The central issue underlying control by mass vaccination is that of the threshold density of susceptibles necessary to ensure that the case reproductive rate, R_0, is greater than unity in value. In broad terms the objective of mass vaccination is to reduce the density of susceptibles below this critical value ($R_0 < 1$). If infants are vaccinated at, or close to birth, then the critical proportion, p, that must be immunized to block transmission (i.e. the level of "herd immunity") is given by

$$p > [1 - 1/R_0] \tag{11}$$

under the assumption that the "vital" rates (of infection, mortality recovery, etc.) are constant and independent of age (Dietz, 1974). In practical terms the magnitude of R_0 can be estimated approximately from age-stratified serological data (recording the proportion, by age class, who have experienced infection as judged by the presence of antibodies specific to the antigens of the infectious agent) by reference to the average age of infection A, where

$$R_0 \simeq L/A. \tag{12}$$

Here L represents life expectancy. Note that this relationship only holds for developed countries where births approximately balance deaths in a population of constant size (May and Anderson, 1984). At equilibrium R_0 is also related to the total fraction susceptible in the population, x, where

$$x = 1/R_0. \tag{13}$$

More generally, if maternally derived antibodies protect against infection and against effective vaccination for an average period of D years, and vaccination takes place at an average age of V years, then the critical proportion to be immunized, p, is given by (see Anderson and May, 1983)

$$p > [1 - 1/R_0][1 - V/L] \tag{14}$$

where R_0 is now defined as

$$R_0 = L/(A - D). \tag{15}$$

These simple expressions [equations (14) and (15)] have enabled estimates to be derived of the levels of cohort immunization required to interrupt the transmission of a number of common childhood infections in both developed (Nokes and Anderson, 1988) and developing (May and Anderson, 1985; McLean and Anderson, 1988a,b) countries. In the United Kingdom, for example, the use of these equations in conjunction with parameter estimates derived from serological and demographic data suggest that between 90 and 95% immunization of children by the age of 1–1.5 years (applied uniformly in all health districts) is required to block the transmission of the measles, mumps and rubella viruses and the bacterial agent responsible for pertussis (i.e. whooping cough).

With respect to the optimum age at which to vaccinate, this depends on the net rate of infection (inversely related to the average age at infection: $\lambda = 1/A$), and the duration of protection provided by maternal antibodies. In developing countries A may be low, such that with an average duration of maternal protection of around 3–6 months, the "window" of susceptibility in which vaccine can be administered is very narrow. For example, Katzmann and Dietz (1984) have shown that for a one-stage vaccination programme in a community with intense transmission prior to control, the optimum age at which to vaccinate, T, is approximately given by

$$T \simeq [\ln(1/D) - \ln(1/A)]/[1/D - 1/A]. \tag{16}$$

In many urban centres in developing countries A for measles is often as low as 1.5 years while D is around 0.5 years. In these circumstances, equation (16) suggests that T is around 10 months of age. The World Health Organization recommends immunization against measles at around 10 months to 1 year of age.

Heterogeneity in Transmission. Much of the research in mathematical epidemiology in recent years has centred on the treatment of various types of heterogeneity in transmission of infectious agents within and between communities of hosts. Such heterogeneities may arise, for example, via spatial factors, genetic variability (in both host and infectious agent) and differences in behaviour in various strata or groups of the host population. In this section a brief review is presented of the major areas of development over the past decade.

1. Age-dependent mixing in human communities. For many (if not most) directly transmitted viral and bacterial infections, the per capita force or rate of infection tends to vary systematically with age. An example is presented in Fig. 3, which records the rate of infection with the measles virus in different age groups in England and Wales (Anderson and May, 1985b).

Age dependency in transmission can be mirrored by one simple modification to the basic model of Kermack and McKendrick [equations (1)–(3)], namely, the replacement of the force of infection term $\lambda(t)$ [where $\lambda(t) = \beta Y(t)$] in equations (1) and (2) by a more general function $\lambda(a, t)$. Under the assumption of a "mass action" form of transmission (see May and Anderson, 1984;

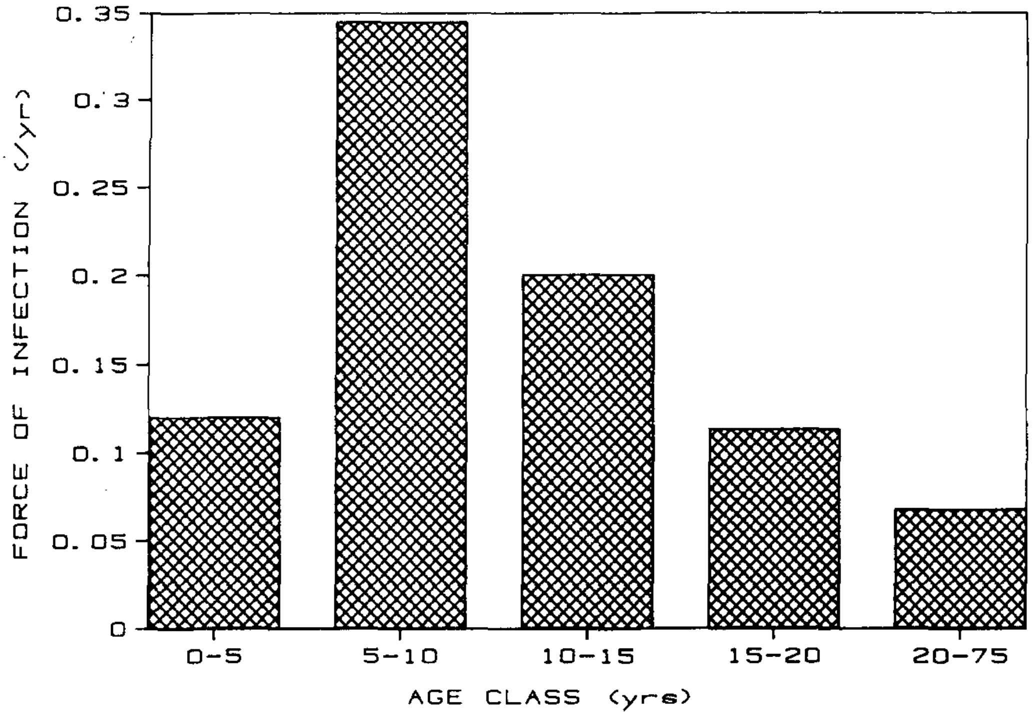


Figure 3. Age-dependent forces of infection [$\lambda(a)$] for the measles virus in developed countries. Data from Table 2 in Anderson and May (1985a).

Anderson and May, 1985b; Dietz and Schenzle, 1985) then the modification can be expressed as

$$\lambda(a, t) = \int_0^L \beta(a', a) Y(a', t) \, da'. \tag{17}$$

Here L denotes life expectancy and the term $\beta(a', a)$ denotes the transmission coefficient ensuing from the contact of susceptibles of age a, with infectious individuals of age a'. The force of infection is therefore a composite parameter denoting the sum rates of contact of susceptibles within a given age class with infectious people of all age classes. Any given value of the transmission coefficient $\beta(a, a')$ presents two components; namely, contact between two age classes and the likelihood that such contact (if it is between susceptible and infectious persons) will give rise to a new case of infection. Both components may be functionally related to age in different ways. However, in much of the published research in this area it has been assumed that behavioural contact patterns are the most important factor (empirical evidence is very limited with respect to both components). If this is the case the term $\beta(a, a')$ effectively denotes "who acquires infection from whom". If the age span from birth to life expectancy, L, is divided into a series of discrete age classes within which the values of $\beta(a, a')$ are constant, for each pair of age classes the transmission coefficients, the β_{ij}'s, form a "who acquires infection from whom" matrix of values ("WAIFW" matrix).

When mixing and contact depends on age the concept of a reproductive rate (and hence of a threshold density of susceptibles) requires modification. In essence we must now define an age-specific basic reproductive rate, R_{0i}, for an infective in age class i where

$$R_{0i} = \sum_{j=1}^{n} B_{ij}(a_j - a_{j-1}). \tag{18}$$

Here $a_j - a_{j-1}$ defines the age interval of a given class (n is the total number of age classes) and B_{ij} is defined as

$$B_{ij} = (N/vL)\beta_{ij} \tag{19}$$

where $1/v$ is the average duration of infectiousness, L is life expectancy and N is population size (assumed constant). The term R_{0i} defines the average number of secondary cases generated by a primary case in age class i in all other age classes in a "susceptible" population (Anderson and May, 1985b). For vaccination close to birth, the criterion for eradication by mass immunization now becomes

$$P > 1 - 1/Q \tag{20}$$

where Q is the dominant eigenvalue of the matrix whose elements are $B_{ij}(a_j - a_{j-1})$. If we assume that the mixing matrix is symmetric ($\beta_{ij} = \beta_{ji}$) then Q is the dominant eigenvalue of the matrix whose elements are given by the R_{0i} of equation (18).

Extensive studies have been carried out on the practical implications of age dependency in transmission of a number of common childhood infections (see Anderson and May, 1985b; Dietz and Schenzle, 1985; Nokes *et al.*, 1986; Anderson *et al.*, 1987; Grenfell and Anderson, 1989; McLean and Anderson, 1988a,b). In all these studies, crude assumptions have had to be made on the structure of the mixing matrix, since empirical data on mixing patterns stratified by age are very limited. The only data available are those on age-related changes in the force or rate of infection (see Fig. 3) but the observed patterns could arise under many different assumptions concerning the structure of the mixing matrix.

With these constraints on interpretation, the general conclusion to emerge from the incorporation of age dependency in mixing is that the estimated critical level of what vaccination is required to block transmission is a little less than that calculated on the basis of homogeneous mixing between age classes (see Anderson and May, 1985b). In practical terms the difference appears to be of little significance. However, more research of both an empirical and a theoretical nature is required in this area.

2. Heterogeneity in sexual behaviour. The emergence of AIDS (the Acquired Immune Deficiency Syndrome) over the past decade as an epidemic on a global scale has triggered much research on the dynamics of sexually transmitted diseases (STDs) (May and Anderson, 1987; Anderson and May, 1988). Patterns of sexual behaviour as reflected, for example, by rates of sexual partner change, are very variable both within and between different human communities (Anderson and Johnson, 1990). As such, models of the transmission dynamics of STDs must take account of this observed heterogeneity in the formulation of the transmission term (Fig. 4). In general,

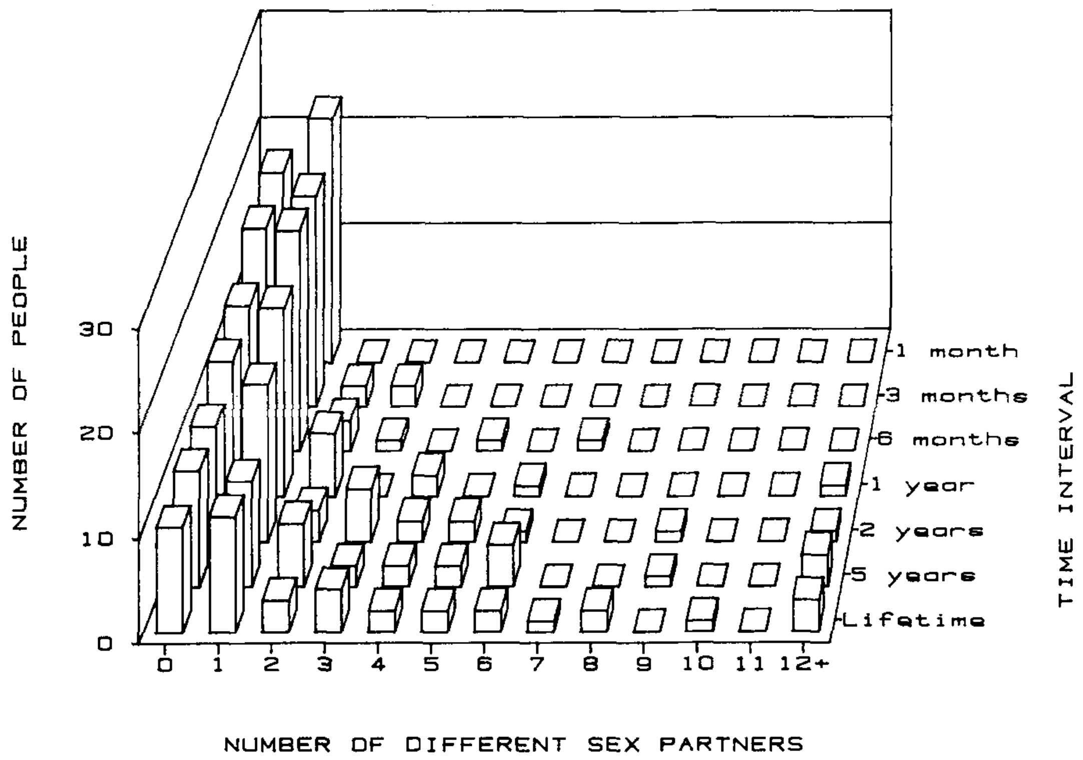


Figure 4. Claimed rate of sexual partner change in a group of 18–22-year-old heterosexual students sampled in England and Wales in 1987 (males and females combined). The graph records the number of students with different rates of sexual partner changes, defined for seven different time intervals (1 month to lifetime).

the early models of STD transmission assume that the per capita rate of infection, λ, is determined by the probability of transmission per partner contact, β, times the mean rate of acquiring new sexual partners, c (the rate of sexual partner change), times the proportion of infectious persons in the sexually active population (Y/N) (Hethcote and York, 1984; Anderson *et al.*, 1986). For example, for the spread of an STD in a male homosexual population the case reproductive rate, R_0, is often defined as

$$R_0 = \beta c V \tag{21}$$

where V is the average duration of infectiousness of an infected person (May and Anderson, 1987). However, this formulation takes no account of heterogeneity in rates of sexual partner change.

A more general approach is based on the stratification of the population on the basis of their rate of sexual partner change, i (Hethcote and York, 1984; Anderson *et al.*, 1986; May and Anderson, 1987, 1988). Consider a closed population of homosexual males [in the context of the transmission of the aetiological agent of AIDS, the human immunodeficiency virus (HIV)], divided into sub-groups, N_i, whose members on average acquire i new sexual partners per unit of time. Initially $N_i(0)=N(0)p(i)$, where $p(i)$ is the initial probability distribution in rates of acquiring partners (Fig. 4). The rates of change in the numbers of susceptible and infected persons in group i, X_i and Y_i respectively, may be expressed as

$$\mathrm{d}X_i/\mathrm{d}t = -i\lambda X_i \tag{22}$$

$$\mathrm{d}Y_i/\mathrm{d}t - i\lambda X_i - vY_i. \tag{23}$$

Here λ is the per capita rate of infection and $1/v$ is the average duration of infectiousness. Under the assumption of "proportional mixing" where the sexual partners are chosen randomly weighted by their sexual activity i, the term λ is

$$\lambda = \beta \sum_i iY_i / \sum_i iN_i \tag{24}$$

where β is as defined for equation (21). For this type of heterogeneous mixing model it can be shown that the basic reproductive rate R_0 is as presented in equation (21) but with c now defined as

$$c = m + \sigma^2/m \tag{25}$$

where c is the mean rate of partner change and σ^2 is the variance of the rate (May and Anderson, 1987, 1988). Note that the variance may dominate the magnitude of R_0 since empirical evidence suggests that it is typically a power function of the mean ($\sigma_2 = am^b$, where a and b are constants) with the power (the value of b) being of the order of 3 (Anderson and May, 1988).

In other words, variability in sexual behaviour is a dominant feature of the pattern of spread and persistence of the infection, since those in the "tail" of the distribution of sexual activity are both more likely to acquire infection and to transmit it. Empirical evidence suggests that more than 70% of the total partnerships formed by a given community are centred on less than 30% of the population of sexually active individuals (Anderson, 1988).

In a closed epidemic (no recruitment of susceptibles) the overall fraction infected, I, is given by (May and Anderson, 1988)

$$I=\sum_i [1-\exp(-i\alpha)]N_i/N \tag{26}$$

where $N=\sum N_i$ and

$$\alpha=-(\beta/v)\ln\left[1-\sum_i iN_i[1-\exp(-i\alpha)]\Big/\sum_i iN_i\right]. \tag{27}$$

It is assumed here that once an individual is infected he remains so for life. In the case where the coefficient of variation is zero, equations (26) and (27) reduce to the Kermack–McKendrick result for a homogeneously mixing population [see equation (9)] when $I(0)$ (the fraction initially infected) is extremely small. When mixing is heterogeneous, the epidemic essentially "burns" itself out in the highly sexually active groups, with the fraction of those in the low activity classes who escape infection being larger and larger as heterogeneity becomes more and more pronounced (the coefficient of variation large). This point is illustrated in Fig. 5, where I is plotted as a function of R_0 for various values of the coefficient of variation (CV) (May and Anderson, 1987).

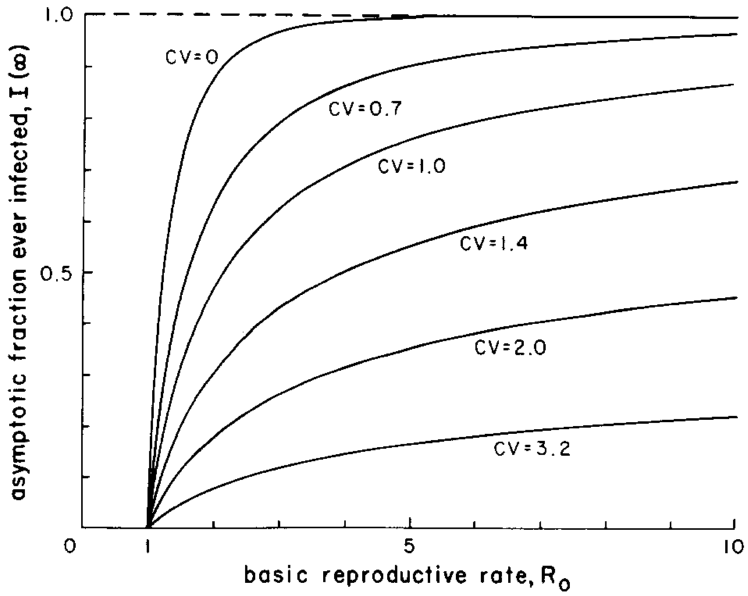


Figure 5. Within a closed population of homosexual males, the fraction infected during an epidemic, I, is shown as a function of the basic reproductive rate, R_0, of the infection. The distribution in rates of acquiring new sexual partners within the population is taken to be a gamma distribution, with the coefficient of variation $CV=\sigma/m$, having the values 0 (classic Kermack–McKendrick epidemic with homogeneous mixing), 0.7, 1, 1.4, 2 and 3.2 as shown. It is assured that 50% of those infected with HIV eventually die from AIDS (see text and May and Anderson, 1988).

This simple model is very illuminating in the sense that it provides a qualitative understanding of how variability in sexual activity influences the magnitude of the epidemic. However, the assumption of proportional mixing (essentially random mixing weighted by sexual activity) is crude, given that it takes no account of "networks" of sexual contacts which define (as in the case of age dependency in transmission—see previous section) "who mixes with whom". Put another way, we ideally need to take account of the proportions of sexual partnerships made by individuals in a given activity group with individuals in the same and other activity classes.

3. Contact networks. In relaxing the assumption of proportional mixing it is necessary to design a choice function which defines the proportion of the contacts of an individual in class i that are made with individuals in class j, $p(i,j)$. In general terms the stratification into classes could be based on spatial location, behaviour or age. However, to continue the discussion in the previous section on HIV spread in male homosexual communities, we first consider stratifications by sexual activity (defined by the rate of acquiring new sexual partners per unit of time). To simplify matters, discrete classes are considered (with a constant rate of partner change within a given class) such that $p(i,j)$ defines a mixing matrix. We retain the notation outlined in the previous section with respect to X_i and Y_i (Nold, 1980).

The net rate of infection of class i, $\lambda(i)$ can be defined as

$$\lambda(i)=c_iX_i\sum_{i=1}^{n}p(i,j)\sum_{r=1}^{m}\beta_rY_{j,r}/(X_j+Y_j). \tag{28}$$

Here n denotes the number of sexual activity classes, c_i records the mean rate of sexual partner change in class i, $Y_{j,r}$ is the number infected in infectious class r and sex activity class j and β_r is the transmission probability associated with infectious class r (the term is written in general form to provide the option of an infected person passing via a series of infectious classes).

There are a series of constants on the elements of the mixing matrix [the $p(i,j)$s] as follows:

$$1<p(i,j)=1 \quad \text{for all } i,j \text{ combinations} \tag{29}$$

$$\sum p(i,j)=1 \tag{30}$$

$$c_i(X_i+Y_i)p(i,j)=c_j(X_j+Y_j)p(j,i). \tag{31}$$

The first two are trivial and obvious but the third [equation (31)] states an important property of the system; namely, that class i cannot in total have more or less sexual contacts with class j than class j can have with class i. This property is of major importance when the infection (such as HIV) is a cause of mortality, since those in high activity classes will acquire infection more rapidly and hence die more rapidly than those in low activity classes.

The distribution of sexual activity will therefore change through time as the epidemic develops and equations (29)–(31) must be satisfied for all values of time t. We will return to this point at a later stage.

A very simple form of $p(i, j)$ is obtained [which satisfies equations (29)–(31)] if all sexual contacts are within group in character [$p(i, i) = 1$, $p(i, j \neq 1) = 0$]. This has been termed *restricted* mixing (Jacquez *et al.*, 1988; Anderson, 1989). More generally, the idea of restricted mixing is analogous to the notion of assortative mating in the fields of behavioural ecology and evolutionary biology (Gupta *et al.*, 1990).

Restricted mixing is unlikely to occur in practice since there are always likely to be cross-linkages between activity classes [or, more generally, different at risk groups such as male homosexuals and heterosexuals (via bisexual men), and intravenous drugs users and the general heterosexual population]. As noted earlier, the most widely employed assumption (because of mathematical convenience, not correspondence with observed pattern) is that of *proportional* mixing. Here the proportion of sexual contacts of people in class i that are made with people in class j is equal to the fraction of total contacts made by the population that are due to people in class j. A third option has been termed *preferred* mixing, which is a linear combination of restricted and proportional mixing. In this case a fraction f_i of the contacts of people in class i are "reserved" for within-class mixing. The remaining contacts are distributed in line with the proportional mixing assumption. Restricted, proportional and preferred mixing satisfy the constraints defined by equations (29)–(31).

A fourth, and final, option is complex choice mixing, which is none of the above (i.e. proportional, preferred, restricted) but which satisfies equations (29)–(31) (Anderson, 1989; Gupta *et al.*, 1990). To illustrate the relationship between a network of sexual contacts and the elements of the associated mixing matrix [the $p(i, j)$s], Fig. 6 records a simple example (hypothetical) of contacts between three sexual activity classes. The classes denote people with 1, 2 or 3 different sexual partners per year and contains 10, 5 and 2 individuals respectively. Under complex choice mixing, one form of the $p(i, j)$ matrix that satisfies equations (29)–(31) is as follows:

$$p(i, j) = \begin{pmatrix} 0.4, & 0.4, & 0.2 \\ 0.4, & 0.4, & 0.2 \\ 0.33, & 0.33, & 0.33 \end{pmatrix}. \qquad (32)$$

Numerical studies of models for HIV transmission in male homosexual communities that incorporate different assumptions concerning the structure

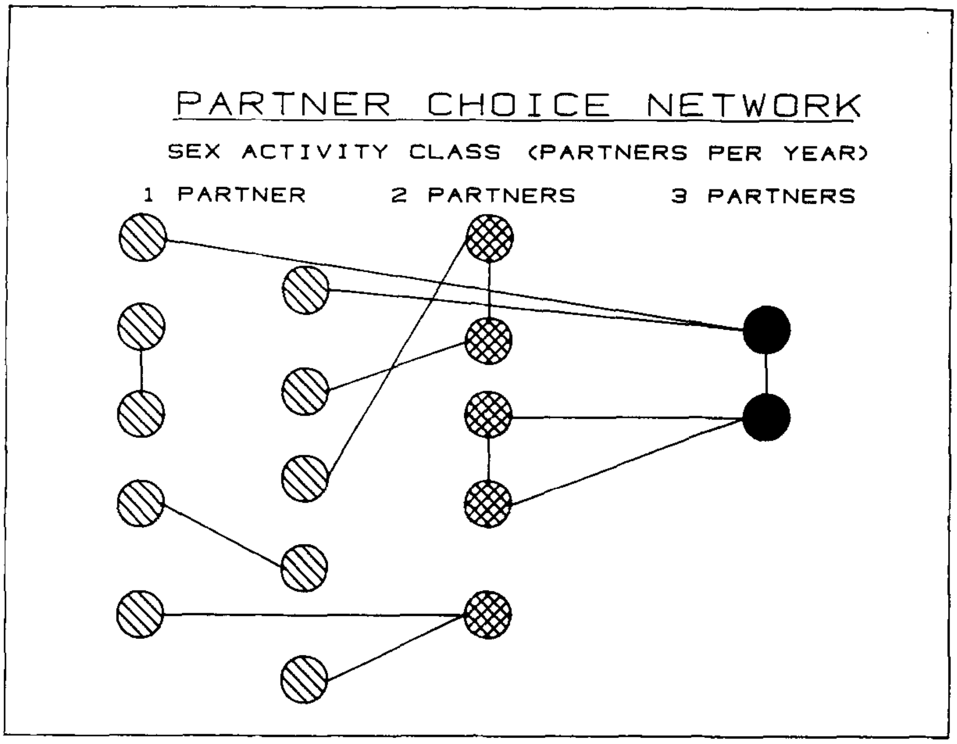


Figure 6. Partner choice network. A simple illustration of a partner choice network in which 17 individuals (male homosexuals) are distributed into three sexual activity classes (one, two or three partners per year). The network generates the choice probability matrix given in equation (32) in the text (Anderson, 1989).

of the mixing matrix generate very different patterns of temporal change in the spread of HIV and the incidence of AIDS (Jacquez *et al.*, 1988; Anderson, 1989; Gupta *et al.*, 1990). Proportional mixing does not necessarily generate the epidemic of the greatest magnitude (as judged, say, by cumulative cases over a defined time period). In general, however, high degrees of within-group mixing tend to reduce the overall magnitude of the epidemic. An illustration of this point is provided in Fig. 7, in which temporal trajectories of the number of male homosexuals infected with HIV, generated by a deterministic transmission model (with six sexual activity classes, recruitment of susceptibles, three infectious classes and a distributed incubation period), with different mixing assumption are displayed. Four trajectories are recorded for proportional mixing, restricted mixing (with only the high activity classes 5 and 6 "seeded" with infection) and two forms of more complex choice. In complex 1, the matrix denotes high within-group contact and very low between-group contact. In complex 2, there is high within-group mixing only in the highest activity class (6) (see Gupta *et al.*, 1990, for details). Note that complex epidemic curves can be generated with multiple peaks in the incidence of infection (HIV) or disease (AIDS). These multiple peaks reflect the spread of infection from one activity class to another (or risk group, i.e. male homosexuals, heterosexuals and intravenous drug users). High within-group mixing can generate an explosive

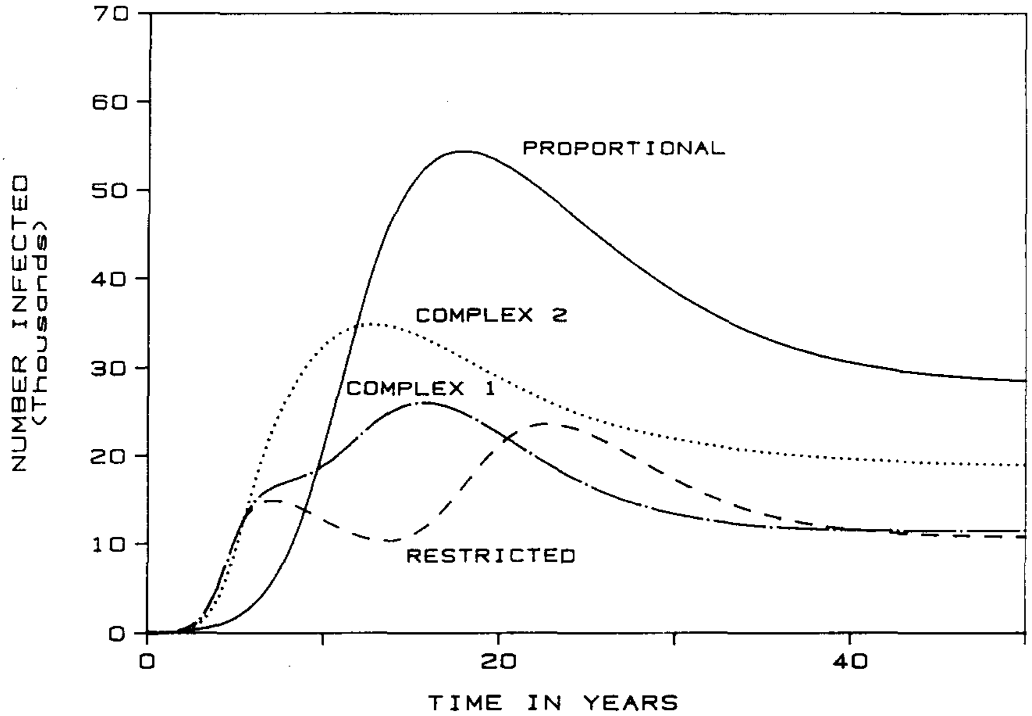
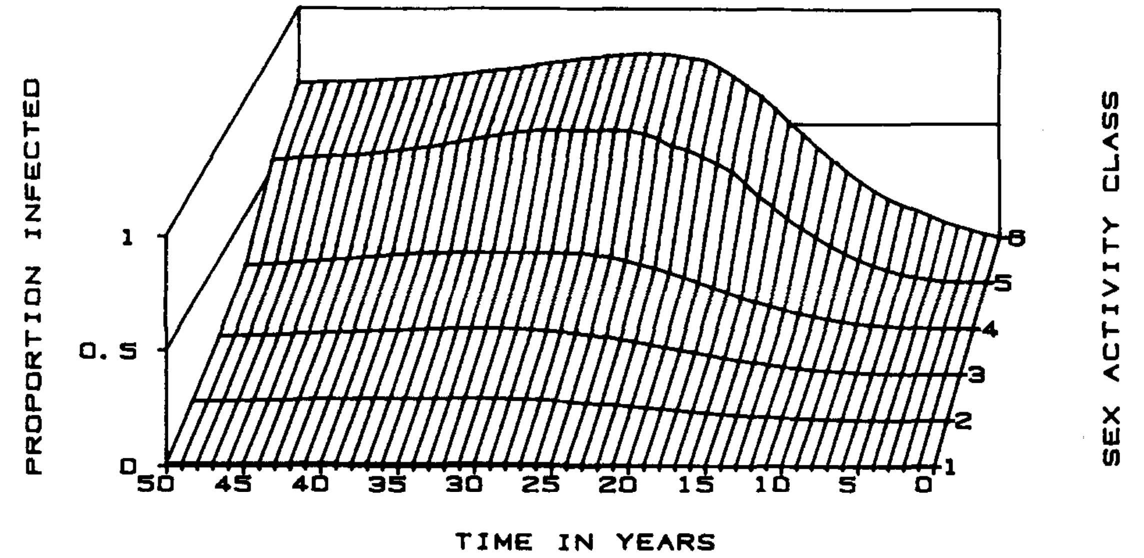


Figure 7. Simulations of temporal trends in the number of male homosexuals infected with HIV (population size of 500 000 at $t = 0$, 4% of the sexually active male population in England and Wales between ages of 16 and 46 years). The model employed to generate the trajectories is as described in Anderson *et al.* (1990b) and Gupta *et al.* (1990). The different trajectories record predictions under different assumptions concerning the mixing matrix. Four simulations are recorded for proportional mixing, restricted mixing and two types of complex mixing (high within-class mixing).

epidemic in the highest sexual activity class where levels of infection rise from close to zero to 80% plus in the space of 4 years (as observed in the case of HIV in male homosexuals in San Francisco in the early 1980s).

This point is illustrated more clearly in Fig. 8, in which temporal changes in the proportion infected with HIV in each activity class are recorded for the proportional mixing simulation and the complex 1 assumption (high within-class mixing) simulation. These patterns are of practical significance with respect to the interpretation of current trends of AIDS and HIV infection in male homosexual populations in many developed countries. They caution against interpreting the current pattern of decline in the rate of reporting new cases of AIDS as necessarily indicating that the epidemic is close to its peak. More precise interpretation of observed trends requires data to guide the choice of a suitable structure of the mixing matrix. This would require detailed information on networks of sexual contact, which will be very difficult to acquire in practice.

The above discussion glosses over one major complication in the formulation of dynamic models containing assumptions that define the structure of the mixing matrix. As noted earlier, if the infectious disease induces



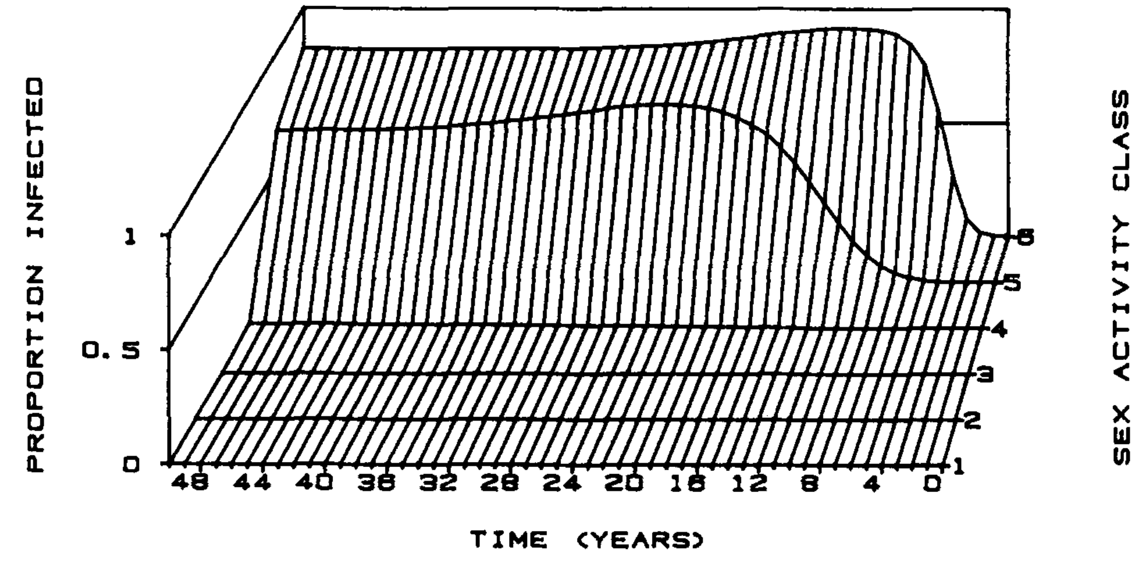


Figure 8. Temporal changes in the proportion of each of six sexual activity classes infected with HIV under the assumption of proportional mixing (top graph) and a complex choice matrix (bottom graph, high within-class mixing with classes 5 and 6 seeded at $t=0$) (Gupta *et al.*, 1990). The six sexual activity classes have mean rates of sexual partner change set as follows: $c(1)=0.45$, $c(2)=3.21$, $c(3)=7.03$, $c(4)=13.85$, $c(5)=43.22$ and $c(6)=81.31$ (yr^{-1}). The proportions in each class at $t=0$ were 0.555, 0.145, 0.105, 0.0825, 0.0725 and 0.04, respectively, for classes 1–6. The overall mean rate per annum at $t=0$ was 8.7 with variance 802.

mortality (or lasting immunity to reinfection in cases other than HIV) then the distribution of sexual activity will change through time as the epidemic spreads and removes more rapidly those in high activity classes when compared with those in low activity groups. In these circumstances, as time progresses, it is likely that imbalances will occur between the number of contacts required by one activity class from another, and what are in fact available. When this occurs a series of assumptions must be made to ensure that the constraints defined in equations (29)–(31) are satisfied at all times during the course of the epidemic. In verbal terms we can define the general nature of these assumptions under the umbrella of a set of *behavioural rules*. These rules may be viewed as hierarchical in structure where a choice or decision at one level leads to a further set of rules to choose between at a lower level. For example, to cope with potential imbalances between sexual contacts required by one group and those available, we have to decide whether to change the structure of the mixing or choice matrix (from its initial state at time $t=0$), or to change the mean rates of sexual partner change of the different groups to accord with availability, or to change both, to meet the constraints of equations (29)–(31).

Once we have decided on one of the above three options we then have to decide precisely how such changes will occur. For instance, if we decide to change the group rates of sexual partner change, do those in high activity classes change their demand, do those in low activity groups change, or does change occur in proportion to the demands of each group? Many possible behavioural scenarios are possible, and unfortunately at present, there is little data available to guide the choice of one option *vs* another. The behavioural rules are clearly qualitative in character and once an assumption has been made, this must be translated into quantitative (algebraic) terms to meet the constraints defined in equations (29)–(31).

Obviously, there are many possible numerical arrangements that will meet the constraints under the qualitative umbrella of a chosen set of behavioural rules. Data is urgently required in this area, but there is also much scope for further theoretical development with respect to the definitions of a spectrum of possible assumptions concerning the manner in which imbalances in supply and demand are catered for, with the help of a few parameters with clear definition in behavioural terms. The literature associated with behavioural ecology and evolutionary biology may be of help in this context.

The problems outlined above are of somewhat greater magnitude and complexity once we move to finer stratification of the sexually active population. For example, in many developing countries HIV is spread by heterosexual contact and, to further complicate matters, there appear to be marked differences in rates of sexual partner change by age and sex, and in the probability of transmission of the virus from males to females and vice versa. In these circumstances we must define a more general sexual partner choice function, $p(k, i, j, a, a')$ to denote the proportion of contacts of an individual of sex k (male or female), activity class i and age a, that are made with individuals of the opposite sex in activity class j and of age a'. Furthermore, the rates of sexual partner change must be formulated as functions of sex, k, and age, a, $-c_{i,k}(a)$. With appropriate notational changes, the constraints defined in equations (29)–(31) still apply, and must be satisfied at all times as the age and sex structure of the population changes under the impact of the AIDS epidemic.

Very little information is available to guide, even in the most qualitative of senses, our choice of a suitable mixing matrix and the assumptions to be made to deal with imbalances in the supply and demand of sexual partners. In one area, however, some information is available. In many societies in sub-Saharan African countries males appear, on average, to choose female sexual partners younger than themselves. Recent analytical and numerical studies have begun to address how such behavioural attitudes of a given population, in conjunction with unequal transmission probabilities between the sexes, might influence the potential long-term demographic impact of AIDS in Africa (Anderson *et al.*, 1990a).

An illustration of one such numerical study is presented in Fig. 9. Two simulations of the projected impact of AIDS in a population of 16.6 million people (at time $t=0$), a 3.8% annual growth rate in the absence of HIV infection and a 1:1 sex ratio are recorded. In one, sexual partner choice is restricted to within an age class, while in the other males, on average, choose female partners younger than themselves. Note that the latter assumption results in the greatest demographic impact (all other parameters being the same) due to the influence of infection and mortality in young females on the net fertility of the population (see Anderson *et al.*, 1990a, for details). As in the case of assessing the importance of mixing matrices on the spread of infection in male homosexual populations, the problems lie not in model formulation but in the choice of behavioural rules and the assignment of values to the elements of the five-dimensional choice function. This area of research is likely to be the focus of much activity in the coming years.

4. *Spatial heterogeneity.* The treatment of spatial heterogeneity in the distribution of people in a defined population is in many senses similar to the problems outlined above in the context of variability in transmission arising from heterogeneity in behaviours that determine contact between susceptible and infected persons. Conventional epidemiological models usually assume homogeneous spatial mixing, with susceptible and infected individuals mingling like the molecules in an ideal gas. In practice, however, spatial factors often play an important role in determining the net rate of transmission, whether at the fine scale of within and between households at a village level or at the broader scale of within and between cities and towns at a countrywide scale (May and Anderson, 1984). Mathematical models embodying such spatial heterogeneity have been explored by several authors in recent years (Hethcote, 1978; Murray and Cliff, 1975; Nold, 1980; Post *et al.*, 1983). A simple approach is to consider n groups, with the population in the ith spatial location being N_i, with births balancing deaths so that all N_i are constant. If we

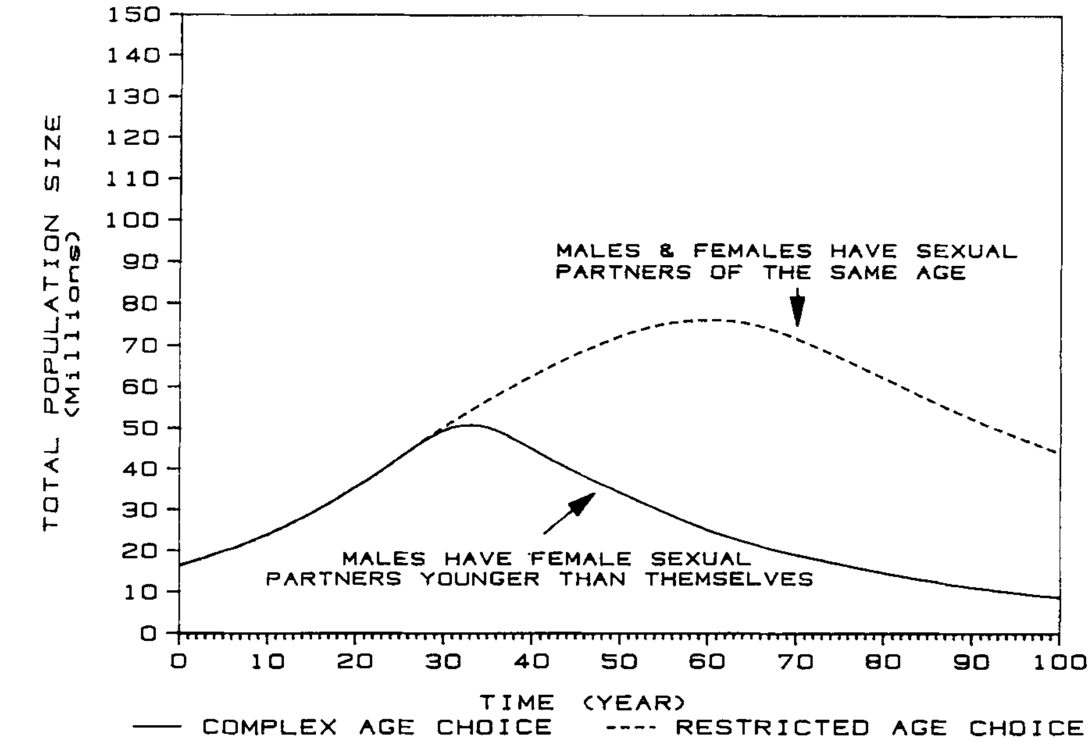


Figure 9. The influence of different patterns of sexual contact between age classes of males and females on the predicted demographic impact of HIV-1 in a developing country (Anderson *et al.*, 1990a). The model incorporates epidemiological and demographic processes and the simulations show changes in population size from the introduction of HIV-1 at $t=0$ into a population of 16.6 million people. In the absence of HIV the population was set to grow at 4% per annum. The doubling time of the epidemic was set at 1.5 years in its yearly stages, the transmission probability from males to females was assumed to equal that from females to males, the efficiency of vertical transmission was set at 50% with a mean incubation period of AIDS of 8 years in adults and 2 years in infants. The two trajectories record predicted trends under the assumption of restricted mixing within age classes and a mixing pattern in which males, on average, have sexual contacts with females younger than themselves (Anderson *et al.*, 1990a).

define the transmission parameter β_{ij} to represent the probability that infectious individuals in group j will infect a susceptible in spatial location i, then the force of infection in location i, λ_i, is given as

$$\lambda_i = \sum_{j=1}^{n} \beta_{ij} Y_j \tag{33}$$

under the Kermack–McKendrick "mass-action" assumption.

One of the practical problems associated with the spread and persistence of directly transmitted infections in spatially heterogeneous populations concerns the question of what is the optimal immunization policy to interrupt transmission in the total population. For example, is it best to vaccinate greater proportions in the densely populated location or is it best to vaccinate equal proportions in all settings? To assess this problem it is important to be clear about the meaning of the word optimal. In most practical senses, it is simplest to define an optimal schedule as one which minimizes the total number of immunizations delivered per unit of time—consistent with the constraint of interrupting transmission in all spatial locations in the total population (May and Anderson, 1984).

The eradication criteria for the Kermack–McKendrick model (with recruitment of susceptibles and stratified by spatial location) is given by the requirement that

$$\det|A_{ij}|=0. \tag{34}$$

Here A_{ij} is the $n \times n$ matrix whose elements a_{ij} are

$$a_{ij}=\left[\frac{\beta_{ij}N_j(1-p_j)}{(v+\mu)}-\delta_{ij}\right] \tag{35}$$

where p_j is the fraction of newly born infants in group i that are effectively immunized (essentially at birth), $1/v$ is the average infectious period, $1/\mu$ is the life expectancy and $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i \neq j$ (May and Anderson, 1984). The fraction of the total population immunized, P, is

$$P=\sum p_i N_i / \sum N_i. \tag{36}$$

The optimal schedule is that which minimizes the value of P subject to the constraints set by equation (34) and the requirement that $1 \geqslant p_i \geqslant 0$.

May and Anderson (1984) have shown that in the case where intragroup transmission rates systematically exceed intergroup transmission rates, the vaccination coverage required under an optimal population-based programme of immunization is lower than would be estimated assuming homogeneous mixing within the total population. The principal message emerging from their analysis (noting the simplicity of the model employed) is that, all other things being equal, attention should be differentially focused on the larger and high density groups. In many ways this conclusion is analogous to the message that emerges from heterogeneous mixing models of the transmission dynamics of sexually transmitted infections. The greatest impact with respect to slowing the spread of infection results if education to change sexual behaviour (e.g. rates of sexual partner change) is targeted at the groups with the highest levels of sexual activity (May and Anderson, 1987; Anderson and May, 1988; Anderson *et al.*, 1990b).

5. *Genetic heterogeneity.* Although age-related, spatial and behavioural heterogeneities in host populations have received increasing attention in recent years, genetic heterogeneity—whereby some hosts may be more or less

resistant to infection than others—is a relatively neglected area in mathematical epidemiology. Those developments that have taken place in recent years tend to fall into two distinct areas.

The first concerns the importance of significantly different degrees of intrinsic susceptibility to infection among different host genotypes to the design of immunization programmes. As shown by Anderson and May (1984), such factors can seriously complicate the estimation of eradication criteria. In some sense the problem is linked to that of age-related changes in the rate of transmission. Observed patterns of change in this rate (see Fig. 3) show a decline in older age classes. This could rise either as a consequence of reduced contact between individuals or reduced susceptibility to infection as a consequence of increased age, or it could alternatively be that such apparent decline is simply because the relatively more susceptible genotypes have largely been infected at a much earlier age. If this is the case, then estimates of p (the eradication criterion) based on pre-immunization values of age-specific rates of infection will be too optimistic. As overall infection rates decline under an immunization programme, an increasing number of relatively more susceptible genotypes will move into the older age classes of susceptibles. A simple example illustrates the significance of this to the estimation of the critical level of vaccination, p, required to block transmission. In the context of the simple Kermack–McKendrick model we assume that the population (of size N) consists of a fixed proportion $(1-f)$ of genotype A which is less resistant to a particular infection than is the remaining fraction f who are of genotype B. Specifically the transmission rate for susceptibles of genotype A is β_A and for genotype B is β_B; we write $\beta_A = \beta$ and $\beta_B = \varepsilon\beta$, where $\varepsilon < 1$.

The eradication criterion for this model is given by

$$p > 1 - v/[\beta N[1 - f(1-\varepsilon)]]. \qquad (37)$$

This condition is similar to that defined in equation (11) for a genetically homogeneous population, and reduces exactly to equation (11) when $f \to 0$ and $\varepsilon \to 1$. The important point with respect to this criterion [equation (37)] is that if the apparent change in the force of infection (apparent in the sense that it is due to genetic heterogeneity as opposed to behavioural factors) is used to estimate the value of p we can arrive at an underestimate of the value that would be derived on the basis of a knowledge of genetic variation in susceptibility to infection (Anderson and May, 1984). The important practical issue to emerge from this type of analysis is that great care should be taken in ascertaining whether age-related changes in the observed rate of infection (derived from age-stratified profiles) arise as a result of genetic factors or the behavioural processes that influence the rate of contact between susceptible and infecteds. One approach to resolving this issue is to try to ascertain (i.e. by HLA typing) if those seronegative (i.e. those who have not apparently experienced infection) in older age classes are genetically different from those who have experienced the infection.

The second area in which the model of Kermack and McKendrick has been developed to assess genetic factors concerns attempts to meld the theories of infectious disease transmission and population genetics to monitor changes in host population abundance (i.e. of susceptibles and infecteds) and gene frequences (see May and Anderson, 1983). Of particular importance in this context is the use of models of transmission dynamics in the derivation of genetic fitness functions that capture both frequency- and density-dependent processes. Much of the recent work has centred only on frequency-dependent effects (see Gillespie, 1975) but the Kermack–McKendrick framework can be developed to consider a system with population abundance and gene frequency changes.

For example, consider a host population in which two alleles (A and a) at a single locus determine susceptibility to different genetic strains of a haploid parasite. For a diploid host there will now be three genotypes of susceptibles, X_i, labelled by the subscript i ($i=1, 2, 3$ for AA, Aa, aa genotypes respectively). If the pathogen is haploid, there will be a population of hosts of genotype i infected with parasite genotype s ($s=1, 2$), which are denoted by $Y_{i,s}$ (assuming that concurrent infections cannot occur). Hence the total host population, $N(t)$, is obtained by summing over these nine subpopulations.

$$N=\sum_{i=1}^{3}\left(X_i+\sum_{s=1}^{2} Y_{i,s}\right). \tag{38}$$

The generalization of the Kermack–McKendrick equations (1)–(3) gives:

$$\mathrm{d}X_i/\mathrm{d}t=aN(1-N/K)_+\theta_i-b_iX_i-X_i\left[\sum_j \sum_s \beta_{i,j,s}Y_{j,s}\right] \tag{39}$$

$$\mathrm{d}Y_{i,s}/\mathrm{d}t=X_i\sum_j \beta_{i,j,s}Y_{j,s}-(b_i+\alpha_{i,s})Y_{i,s}. \tag{40}$$

Here b_i is the disease-free death rate of genotype i, $\alpha_{i,s}$ the disease-induced death rate of genotype i infected with parasite strain s (hosts are assumed not to recover from infection), and $\beta_{i,j,s}$ is the transmission coefficient for infection of host genotype i with parasite strain s by an infected host of genotype j. The birth rate is assumed to be logistic in form with the subscript + denoting that the birth rate is zero for $N>K$. Random mating is assumed, so that newly born susceptibles are approportioned among the three genotyes in the proportions $\theta_i=p^2$, $2pq$, q^2 ($q=1-p$) for $i=1, 2, 3$ respectively. The gene frequency p of gene A is clearly

$$p=[X_1+\tfrac{1}{2}X_2+\textstyle\sum(Y_{1,s}+\tfrac{1}{2}Y_{2,s})]/N. \tag{41}$$

The study of equations (39) and (40) is a formidable task but some progress has been made by May and Anderson (1983), Beck (1984) and Beck *et al.* (1984). In a particularly elegant study, Beck (1984) has shown that to an excellent approximation the system of nine equations can be reduced to just two differential equations for p and q given certain constraints on parameter values. She showed that the system can exhibit fixation of either a particular parasite genotype or a particular host genotype, or both, or can exhibit stable limit cycles. The outcome tends to depend on the trade-offs between transmissibility and virulence for the parasite. Of particular interest is the ability of the system, for certain domains of parameter space, to show widely changing gene frequencies through time in a population of relatively constant size. Recently, similar patterns have been demonstrated for a model of macroparasite transmission in host communities, where selection induced by host immune defences or drug application acts on different parasite genotypes (Anderson *et al.*, 1990b). One moral from these studies is that relatively simple models that incorporate frequency- and density-dependent processes, in a framework that takes account of population abundance and gene frequency changes, can exhibit very complicated patterns of non-linear behaviour. The modelling of population genetic and epidemiological theory is an important priority for future research on infectious disease transmission. The major problem lies in the need to simplify large systems of coupled equations with many parameters to simpler systems (amenable to analytic study) without the loss of biological detail.

Conclusions. The choice of topics covered in this article is broad, but not in any sense comprehensive with respect to the development of epidemic theory since the publication of the Kermack–McKendrick papers. Many areas of great interest have been omitted, not via any prejudice, but simply because of the limitations of space. One in particular, namely the development of stochastic models of epidemic processes, is of obvious importance with respect to the notions of threshold theorems and infectious disease persistence in communities of people. A very thorough treatment of this area is given in the book by Bailey (1975) and more recent developments are discussed by Ball (1983). Chance events in the chain of contacts between susceptibles and infecteds are clearly of importance in determining the likelihood of a major epidemic developing, particularly in small communities. Similarly, with respect to large communities, oscillatory fluctuation in incidence within recurrent epidemic cycles can result in "disease" fade out in communities below a certain size (and with low net birth rates) during the inter-epidemic phases. For example, this factor underpins the observation that the measles virus only tends to persistence endemically (with recurrent cycles in incidence) in communities of a certain total size (roughly, between 300 000 and 500 000 people; see Bartlett, 1957; Black, 1966). In addition to the problems of persistence, stochastic models have an obvious role in the development of methods for parameter estimation (i.e. the incubation and infectious periods) and in understanding the frequency and perpetuation of epidemic cycles in disease incidence (Becker, 1989).

The major aim of this present review has been to highlight recent developments of the deterministic theory, with particular emphasis on the treatment and significance of various forms of heterogeneity in the transmission process. There are many similarities in the way in which the basic Kermack–McKendrick framework is further compartmentalized to handle spatial, behavioural and genetic heterogeneity. This is especially apparent in the definition of the transmission coefficient, β, with respect to contact within and between different groups or classes in the total population (whether based on age, genotype, behavioural activity or spatial location). This coefficient can be partitioned into two separate components—one representing the likelihood of transmission following contact between a susceptible and an infected and the second denoting the probability of contact between individuals in different groups. The latter component is of particular interest at present, in the context of sexually transmitted infections such as HIV. The structure of the choice or contact function, which represents the probability of contact between different groups, has a major impact on the behaviour of any given model. This general observation highlights the need for better empirical studies on human (or host) behaviours that influence transmission events. In the case of HIV this implies sexual behaviour, which is a particulary difficult area for scientific study. Much research of an anthropological nature has examined human sexual behaviour but rather little of the published work is sufficiently quantitative in character to aid in model construction and parameter estimation. It is often argued that such behavioural patterns are impossible to quantify, on the grounds of varying understanding between individuals of the precise meanings of the terms that are often used in survey questionnaires to specify a particular type of behaviour (e.g. penetrative sexual intercourse, oral sex, etc.). In addition there are the obvious problems associated with assessing the accuracy, or truthfulness, of a person's responses to survey questions. However, these difficulties should not detract from the urgent need to acquire such information if a better understanding is required of the processes that shape infectious disease persistence and spread. In any field of research, a start must be made somewhere in data collection. What is required in the early stages of behavioural work in epidemiology is a clear appreciation of the limitations of the information and, concomitantly, great caution in the interpretation of recorded patterns.

I am very indebted to Robert May, Sunetra Gupta, Stephen Blythe and Graham Medley for much advice and help. Financial support from the Medical Research Council and the Overseas Development Agency is gratefully acknowledged.

LITERATURE

Anderson, R. M. 1979. Parasite pathogenicity and the depression of host population equilibrium. *Nature* **279**, 150–152.

Anderson, R. M. (Ed.) 1982. *Population Dynamics of Infectious Diseases*. London: Chapman and Hall.

Anderson, R. M. 1988. The role of mathematical models in the study of HIV transmission and the epidemiology of AIDS. *J. AIDS* **1**, 241–256.

Anderson, R. M. 1989. Mathematical and statistical studies of the epidemiology of HIV. *AIDS* **3**, 333–346.

Anderson, R. M. and A. M. Johnson. 1990. Rates of sexual partner change in homosexual and heterosexual populations in the United Kingdom. In *AIDS and Sex: An Integrated Biomedical and Biobehavioural Approach*, J. Reinisch and B. Voeller (Eds). Oxford: Oxford University Press (in press).

Anderson, R. M. and R. M. May. 1979. Population biology of infectious diseases. I. *Nature* **280**, 361–367.

Anderson, R. M. and R. M. May. 1982. Directly transmitted infectious diseases: control by vaccination. *Science* **215**, 1053–1060.

Anderson, R. M. and R. M. May. 1983. Vaccination against rubella and measles: quantitative investigations of different policies. *J. Hyg. Camb* **90**, 259–325.

Anderson, R. M. and R. M. May. 1984. Spatial, temporal and genetic heterogeneity in host populations and the design of immunization programmes. *IMA J. Math. appl. Med. Biol.* **1**, 223–266.

Anderson, R. M. and R. M. May. 1985a. Vaccination and herd immunity to infectious diseases. *Nature* **318**, 323–328.

Anderson, R. M. and R. M. May. 1985b. Age-related changes in the rate of disease transmission: implications for the design of vaccination programmes. *J. Hyg. Camb.* **94**, 365–436.

Anderson, R. M. and R. M. May. 1988. Epidemiological parameters of HIV transmission. *Nature* **333**, 514–519.

Anderson, R. M., G. F. Medley, R. M. May and A. M. Johnson. 1986. A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. appl. Med. Biol.* **3**, 229–263.

Anderson, R. M., J. A. Crombie and B. T. Grenfell. 1987. The epidemiology of mumps in the UK: a preliminary study of virus transmission, herd immunity and the potential impact of immunization. *Epidem. Inf.* **99**, 65–84.

Anderson, R. M., W. Ng and E. Konnings. 1990a. The influence of different sexual contact patterns between age classes on the predicted demographic impact of AIDS in developing countries. *N.Y. Acad. Sci.* (in press).

Anderson, R. M., S. P. Blythe, S. Gupta and E. Konnings. 1990b. The transmission dynamics of the human immunodeficiency virus type 1 in the male homosexual community in the United Kingdom: the influence of changes in sexual behaviour. *Phil. Trans. R. Soc. Lond. B* (in press).

Bailey, N. J. T. 1975. *The Mathematical Theory of Infectious Diseases*, 2nd edn. New York: Macmillan.

Ball, F. 1983. The threshold behaviour of epidemic models. *J. app. Prob.* **20**, 227–241.

Bartlett, M. S. 1957. Measles periodicity and community size. *J.R. Statist. Soc.* **A120**, 48–70.

Beck, K. 1984. Co-evolution: mathematical analysis of host–parasite interactions. *J. Math. Biol.* **19**, 63–78.

Beck, K., J. P. Keens and P. Ricciardi. 1984. The effect of epidemics on genetic evolution. *J. Math. Biol.* **19**, 79–94.

Becker, N. G. 1989. *Analysis of Infectious Disease Data.* London: Chapman and Hall.

Bernoulli, D. 1760. Essai d'une nouvelle analyse de la mortalite causee par la petite verole et des avantages de l'incubation pout la prevenir. *Mem. Math. Phys. Acad. R. Sci. Paris* 1–45.

Black, F. L. 1966. Measle endemicity in insular populations: critical community size and its evolutionary implications. *J. theor. Biol.* **II**, 207–211.

Dietz, K. 1974. Transmission and control of arbovirus diseases. In: *Epidemiology*, D. Ludwig and K. L. Cooke (Eds), pp. 104–121. Proc. SIAM, Philadelphia.

Dietz, K. and D. Schenzle. 1985. Mathematical models for infectious disease statistics. In: *A Celebration of Statistics*, A. C. Atkinson and S. E. Fienberg (Eds), pp. 167–204. New York: Springer-Verlag.

En'ko, P. D. 1889. On the course of epidemics of some infectious diseases. *Vrach St. Petersburg* **X**, 1008–1010, 1039–1042, 1061–1063.

Fine, P. E. M. 1979. John Brownlea and the measurement of infectiousness: an historical study in epidemic theory. *J. R. Statist. Soc.* **A142**, 347–362.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.

Gillespie, J. H. 1975. Natural selection for resistance to epidemics. *Ecology* **56**, 483–495.

Greenwood, M., A. B. Hill, W. W. C. Topley and J. Wilson. 1936. Experimental epidemiology. *MRC Special Report Series*, 208. London: HMSO.

Grenfell, B. T. and R. M. Anderson. 1989. Pertussis in England and Wales: an investigation of transmission dynamics and control by mass vaccination. *Proc. R. Soc. Lond. B.* **236**, 213–252.

Gupta, S., R. M. Anderson and R. M. May. 1990. The influence of sexual contact networks on the predicted pattern of the AIDS epidemic in male homosexuals in the United Kingdom. *IMA J. Math. appl. Med. Biol.* (in press).

Hamer, W. H. 1906. Epidemic disease in England—the evidence of variability and of persistency of type. *Lancet* **ii**, 733–739.

Hethcote, H. W. 1978. An immunization model for a heterogeneous population. *Theor. Pop. Biol.* **14**, 338–349.

Hethcote, H. W. and J. A. York. 1984. Gonorrhoea; transmission dynamics and control. *Lect. Notes. Biomath.* **56**, 1–105.

Jacquez, J. A., C. P. Simon, J. Koopman, L. Sattenspield and T. Perry. 1988. Modelling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.* **92**, 118–199.

Jeger, M. J. 1986. Asymptotic behaviour and threshold criteria in model plant disease epidemics. *Plant Pathol.* **35**, 355–361.

Katzmann, W. and K. Dietz. 1984. Evaluation of age-specific vaccination strategies. *Theor. Pop. Biol.* **25**, 125–137.

Kermack, W. O. and A. G. McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721.

Kermack, W. O. and A. G. McKendrick. 1932. Contributions to the mathematical theory of epidemics. Part II. *Proc. R. Soc. Lond. A* **138**, 55–83.

Kermack, W. O. and A. G. McKendrick. 1933. Contributions to the mathematical theory of epidemics. Part III. *Proc. R. Soc. Lond. A* **141**, 94–122.

Kermack, W. O. and A. G. McKendrick. 1937. Contributions to the mathematical theory of epidemics. Part IV. *J. Hyg. Camb.* **37**, 172–187.

Kermack, W. O. and A. G. McKendrick. 1939. Contributions to the mathematical theory of epidemics. Part V. *J. Hyg. Camb.* **39**, 271–288.

Macdonald, G. 1957. *The Epidemiology and Control of Malaria.* London: Oxford University Press.

May, R. M. 1990. Population biology and population genetics of plant–pathogen associations. In: *Pests, Pathogens and Plant Communities*, J. J. Burdon and S. R. Leather (Eds). Oxford: Blackwell (in press).

May, R. M. and R. M. Anderson. 1979. Population biology of infectious diseases. II. *Nature* **280**, 455–461.

May, R. M., and R. M. Anderson. 1983. Epidemiology and genetics in the co-evolution of parasites and hosts. *Proc. R. Soc. Lond. B* **218**, 281–313.

May, R. M. and R. M. Anderson. 1984. Spatial heterogeneity and the design of immunization programs. *Math. Biosci.* **72**, 83–111.

May, R. M. and R. M. Anderson. 1985. Endemic infections in growing populations. *Math. Biosci.* **77**, 141–156.

May, R. M. and R. M. Anderson. 1987. Transmission dynamics of HIV infection. *Nature* **326**, 137–142.

May, R. M. and R. M. Anderson. 1988. The transmission of human immunodeficiency virus (HIV). *Phil. Trans. R. Soc. Lond. B* **321**, 565–607.

Murray, G. D. and A. D. Cliff. 1975. A stochastic model for measles epidemics in a multi-region setting. *Inst. Br. Geog.* **2**, 158–174.

McLean, A. R. and R. M. Anderson. 1988a. Measles in developing countries. Part I. Epidemiological parameters and patterns. *Epidem. Inf.* **100**, 111–133.

McLean, A. R. and R. M. Anderson. 1988b. Measles in developing countries. Part II. The predicted impact of mass vaccination. *Epidem. Inf.* **100**, 418–442.

Nokes, D. J. and R. M. Anderson. 1988. The use of mathematical models in the epidemiological study of infectious diseases and in the design of mass immunization programme. *Epidem. Inf.* **101**, 1–20.

Nokes, D. J., R. M. Anderson and M. J. Anderson. 1986. Rubella transmission in South East England: a horizontal seroepidemiological study. *J. Hyg. Camb.* **96**, 291–304.

Nold, A. 1980. Heterogeneity in disease-transmission modelling. *Math. Biosci.* **52**, 227–240.

Post, W. M., D. L. DeAngelis and C. C. Travis. 1983. Endemic disease in environments with spatially heterogeneous host populations. *Math. Biosci.* **63**, 289–302.

Ross, R. 1908. *Report on the Prevention of Malaria in Mauritius.* London: Waterlow and Sons.

Ross, R. 1915. Some *a priori* pathometric equations. *Br. med. J.* **I**, 546–547.

Ross, R. and H. P. Hudson. 1917. An application of the theory of probabilities to the study of *a priori* pathometry. III. *Proc. R. Soc. A* **93**, 225–240.

Soper, M. A. 1928. The interpretation of periodicity in disease prevalence. *J.R. Statist. Soc.* **A92**, 34–61.

ROY M. ANDERSON
Parasite Epidemiology Research Group
Department of Pure and Applied Biology
Imperial College
London University
London SW7 2BB, U.K.