A RAND NOTE

Modeling Heterogeneity in Susceptibility and Infectivity for HIV Infection

N. Scott Cardell, David E. Kanouse

RAND

The research described in this report was supported by RAND as part of its program of public service.

Published 1991 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

A RAND NOTE

N-3134-RC

Modeling Heterogeneity in Susceptibility and Infectivity for HIV Infection

N. Scott Cardell, David E. Kanouse

RAND

MODELING HETEROGENEITY IN SUSCEPTIBILITY AND INFECTIVITY FOR HIV INFECTION

N. Scott Cardell
Washington State University
Pullman, WA 99164

David E. Kanouse
The RAND Corporation
Santa Monica, CA 90406

Abstract

Models of the spread of human immunodeficiency virus (HIV) infection must deal with substantial heterogeneity in the populations at risk. The virus is spread by behaviors that are far from uniformly distributed in the population, and substantial variations in biological aspects of susceptibility and infectivity are also likely. How adequately a model represents this heterogeneity will substantially determine its accuracy and usefulness for capturing the dynamics of the epidemic, for making forecasts of future spread, and for answering questions of policy interest.

There are two main ways in which a model may handle heterogeneity: by partitioning the population into discrete risk groups that are in some respect homogeneous within group but heterogeneous between groups, and by introducing model parameters to capture the effects of heterogeneity in a group or in the population as a whole. This paper discusses the dynamics of heterogeneity in HIV spread and develops a theory of heterogeneity in susceptibility and infectivity within a population that allows a simple representation of key phenomena within an epidemic model. It is suggested that the effects of heterogeneity-related phenomena can be captured by letting two key parameters, the mean susceptibility over time of the uninfected and the mean infectivity of the infected, depend upon $\frac{X}{P}$, the proportion of the population that is uninfected. (The mean infectivity may also depend on the cumulative proportion of the population that is removed through death or other causes). Because $\frac{X}{P}$, as we define it, is monotonic over time, this approach is general, and it allows considerable flexibility in the choice of functional form to fit available data.

1. Introduction

Models of the spread of human immunodeficiency virus (HIV) infection must deal with substantial heterogeneity in the populations at risk. The virus is spread within and between populations that differ in both type and frequency of behaviors that are

Reprinted from *Mathematical and Statistical Approaches to AIDS Epidemiology*, Vol. 83, Lecture Notes in Biomathematics, Carlos Castillo-Chavez (ed.), pp. 138–156, ©1989 by Springer-Verlag. Reprinted by permission.

epidemiologically linked to HIV transmission, as well as in other ways that may be relevant to the goals underlying development of a model. How adequately a model represents this heterogeneity will substantially determine its accuracy and usefulness for capturing the dynamics of the epidemic, for making forecasts of future spread, and for answering questions of policy interest.

The heterogeneity dealt with in this paper concerns the extent of variation within a population in individuals' susceptibility to infection or infectivity to others. We define susceptibility as proportional to the probability per unit time that an uninfected individual will become infected, holding factors external to the individual constant. Susceptibility has both biological and behavioral components; that is, it will depend both on the nature and frequency of epidemiologically risky behaviors that the individual engages in and on biological resistance to becoming infected. This definition is broader than the usual definition, which is restricted to the biological component.[1] We define infectivity similarly, as proportional to the probability per unit time that an infected individual will infect another (uninfected) person, again holding factors external to the individual constant. Like susceptibility, infectivity has both biological and behavioral components.

Within a population, the extent of variation in susceptibility and infectivity will reflect the extent of variation in relevant biological and behavioral factors. If all individuals behave in the same way and all have the same biological propensity to infect or acquire infection, then individual susceptibility and infectivity at any given time will reflect only the individual's current status (infected or uninfected) and the population mean for individuals with that status. If behavioral or biological factors vary, however, that variation introduces selection dynamics that alter the course of an epidemic over time. Those dynamics should be taken into account in epidemiological modeling.

The effects of population heterogeneity on the spread of an infectious agent are greatest when modes of transmission are relatively inefficient, as is the case with HIV (Friedman and Klein, 1987; Wiley, 1987). In these circumstances, the effects of variability in the probability of transmission per exposure or in the frequency of exposure are much greater than when transmission is relatively efficient. Growth is expected to occur at a constant exponential rate in the early phases of an epidemic. As we shall show, however, heterogeneity may first substantially accelerate and then moderate the rate of growth that would occur in the absence of heterogeneity.

Although the HIV epidemic is in some ways quite complex, its essential features are those of a relatively simple class of epidemics that may be described by an SI model, in which a constant-sized population consists entirely of those susceptible and those infected; no one is immune and no one recovers (Hethcote, 1976). The early growth of such an epidemic within a homogeneous population is at a constant exponential rate, gradually slowing to half this initial rate when the proportion infected reaches 50 percent. Manifestly, the HIV epidemic has not followed this pattern but has instead exhibited a steady increase in the doubling time (Curran et al., 1988). The existence of biological or behavioral heterogeneity in the populations in which HIV is spreading is one possible explanation.

---

[1] We have adopted this broader definition of susceptibility because it offers distinctive advantages over the more usual definition. In particular, it allows unitary representation of the effects of sources of individual variation that are functionally equivalent. For instance, the effects of biological resistance may be difficult to distinguish from the effects of using condoms.

Because of the important role that heterogeneity may play in the dynamics of the HIV epidemic, an AIDS epidemiological model needs to account for its effects. A model may handle heterogeneity in either or both of two ways: by partitioning the population into discrete risk groups that are in some respect homogeneous within group but heterogeneous between groups (e.g., Jacquez et al., 1988; Sattenspiel and Simon, 1988), and/or by introducing model parameters to capture the effects of heterogeneity in a group or in the population as a whole. In Section 2 of this paper, we discuss the qualitative effects of within-group heterogeneity on the course of an epidemic. In Section 3, we develop a theory of heterogeneity in susceptibility and infectivity that allows a simple representation of the key phenomena associated with this heterogeneity in the spread of HIV. In Section 4, we illustrate the dynamics involved using a simple one-population simulation, and show how these dynamics may be represented in a model using an approach to parameterization adopted in our own deterministic simulation model of the spread of HIV infection in the United States (Cardell et al., 1987).

2. Effects of Heterogeneity

In the standard epidemiological model, all individuals are equally susceptible to infection and equally infectious once infected. If we allow for individual variation in susceptibility, the most susceptible individuals become infected first. As a result, the average susceptibility of those who remain uninfected declines over time and is greatest when no one is infected. Because the epidemic growth rate is a function of the average susceptibility of those who remain uninfected, variation in susceptibility results in slower growth for the same average susceptibility for the total population.

Individual variation in susceptibility can occur without individual variation in infectivity. If infectivity varies as well, the consequences depend on the relationship between infectivity and susceptibility. If infectivity varies independently of susceptibility, infectivity variation will have no effect on the epidemic, because there will be no reason for the most infectious to become infected first and therefore the mean infectivity of those who are infected at any given time will remain at the population average. If, however, both susceptibility and infectivity vary and are positively correlated (e.g., if both represent a tendency to engage in behaviors that are epidemiologically linked to HIV infection), then the most infectious, being more susceptible on average, will tend to become infected first. As a result, the mean infectivity of the infected will exceed the mean infectivity that the population would have if all its members were infected.

When susceptibility and infectivity are positively correlated, the mean infectivity of the infected will be greatest when the smallest fraction is infected. When this fraction is very small, the mean susceptibility of those uninfected will also be near its maximum value (the mean value for the total population). Therefore, introducing both susceptibility variation and a correlated infectivity variation will increase the rate at which the epidemic spreads in its earliest stages. As the epidemic progresses, however, both the mean susceptibility of the uninfected and the mean infectivity of the infected will decline, and the rate of spread will fall to a lower level than would prevail at the same level of infection if there were no variation. In fact, it can be shown that the growth rate will fall below the "no variation" rate before one half the population is infected.

Although our larger model (Cardell et al., 1987) includes many additional factors, our focus in this paper is on heterogeneity among individuals in susceptibility and infectivity. We do not deal with stages of HIV infection or with variation in infectivity over the course of the disease (Blythe and Anderson, 1988; Hyman and Stanley, 1988). We treat the population as constant in size, ignoring vital dynamics. We also ignore behavioral change over time and heterogeneity of other types than that considered here, such as in mixing.

Susceptibility when uninfected and infectivity when infected are constants associated with an individual. Since there is no recruitment, and we assume that no one recovers, we continue to include those removed through death in our definition of the population. An important consequence of this is that the proportion uninfected can only be monotonically nonincreasing. These assumptions can reasonably be applied to cohorts over time, but would require modification to apply to fixed age groups.

3. Accounting for Heterogeneity in Epidemiologic Models

Given that variations in susceptibility and infectivity can potentially have important effects, what can be done to deal with these effects in epidemiological models? In the discussion that follows, we will begin with the simplified one-population SI model but will extend the model to consider some of the particulars of the HIV epidemic and what heterogeneity entails in that context.

Consider a simplified case where we have only one population and include only subgroups for uninfected, infected, and dead from HIV-related causes. Further let the latency between infection and death be a fixed time period, τ. (For discussions of models in which the latency between infection and death is not fixed, see for example, Cardell et al., 1987 and Castillo-Chavez et al., 1989.) Let $X(t)$ denote the uninfected population, $Y(t)$ the infected and living population, and $Z(t)$ those dead of HIV-related causes. To allow for individual variation in susceptibility and infectivity, characterize each individual by what his susceptibility to infection is when uninfected and what his infectivity to others is when infected. Denote susceptibility by s and infectivity by h. Let $x(s,h,t,)$, $y(s,h,t)$, $z(s,h,t)$ denote the distribution of individuals over s and h at time t for the uninfected, infected and dead groups, respectively. Thus $X(t) = \int_0^\infty \int_0^\infty x(s,h,t)dsdh$, $Y(t) = \int_0^\infty \int_0^\infty y(s,h,t)dsdh$, and $Z(t) = \int_0^\infty \int_0^\infty z(s,h,t)dsdh$. Let $P = X(t) + Y(t) + Z(t)$, and $p(s,h) = x(s,h,t) + y(s,h,t) + z(s,h,t)$. (Note that $Z(t) = P - X(t-\tau)$, $Y(t) = X(t-\tau) - X(t)$). In this simplified model the only processes are infection and death. Because these processes move individuals between the above categories, P and $p(s,h)$ *do not depend on time*. Assuming random contact (or proportionate mixing; see Hethcote and Van Ark, 1987), we can write the probability per unit time (or hazard) that an infected individual characterized by s_1 and h_1 will infect an uninfected individual characterized by s_2 and h_2 as:

$$\alpha s_2 h_1.$$

That is, if person 2 is uninfected at time 0, the probability that infected person 1 will *not* have infected person 2 by time t is (ignoring all other sources of infection):

$$e^{-\alpha s_2 h_1 t}.$$

The constant α is chosen to allow the convenient normalizations

$$\bar{s} \equiv \frac{1}{P}\int_0^\infty \int_0^\infty s \cdot p(s,h)dsdh = 1, \quad \bar{h} \equiv \frac{1}{P}\int_0^\infty \int_0^\infty h \cdot p(s,h)dsdh = 1.$$

Note that the size of α depends on the units of time and that α may exceed one.

We assume that the population groups are large enough to treat without loss of generality all functions as continuous, and hence we can use deterministic equations. (In practice, this means we can only apply these equations after a significant number of people are infected.)

$$\frac{dx(s,h,t)}{dt} = -\alpha s \left(\int_0^\infty \int_0^\infty r \cdot y(q,r,t)dqdr \right) x(s,h,t) \tag{1}$$

Given as initial conditions x, y, z specified for $t \in [\tau, 0]$.

$$\begin{aligned} z(s,h,t) &= y(s,h,t-\tau) + z(s,h,t-\tau) \\ &= p(s,h) - x(s,h,t-\tau) \end{aligned} \tag{2}$$

$$\begin{aligned} y(s,h,t) &= p(s,h) - x(s,h,t) - z(s,h,t) \\ &= x(s,h,t-\tau) - x(s,h,t). \end{aligned} \tag{3}$$

Let $b(t) = \alpha \int_0^t \left(\int_0^\infty \int_0^\infty r \cdot y(q,r,u)dqdr \right) du$; then the solution to equation 1 above is $x(s,h,t) = x(s,h,0)e^{-b(t)\cdot s}$. If we assume that the same infection process held before $t = 0$, $x(s,h,0) = p(s,h)e^{-ks}$. Let $a(t) = b(t) + k$; then:

$$x(s,h,t) = p(s,h)e^{-a(t)\cdot s}, \tag{4}$$

substituting into equation (2)

$$\begin{aligned} z(s,h,t) &= p(s,h) - p(s,h)e^{-a(t-\tau)\cdot s} \\ &= p(s,h) \cdot \left(1 - e^{-a(t-\tau)\cdot s}\right), \end{aligned} \tag{5}$$

substituting into equation (3)

$$\begin{aligned} y(s,h,t) &= p(s,h) - p(s,h)e^{-a(t)\cdot s} - p(s,h) + p(s,h)e^{-a(t-\tau)\cdot s} \\ &= p(s,h) \cdot \left(e^{-a(t-\tau)s} - e^{-a(t)s}\right) \end{aligned} \tag{6}$$

$$X(t) = \int_0^\infty \int_0^\infty x(s,h,t)dsdh. \tag{7}$$

The number of new infections per unit time is:

$$\begin{aligned}\frac{dX(t)}{dt} &= \int_0^\infty \int_0^\infty \frac{dx(s,h,t)}{dt} dsdh \\ &= \int_0^\infty \int_0^\infty -\left(\alpha s x(s,h,t) \int_0^\infty \int_0^\infty r \cdot y(q,r,t) dqdr\right) dsdh \\ &= -\alpha \left(\int_0^\infty \int_0^\infty s x(s,h,t) dsdh\right) \cdot \left(\int_0^\infty \int_0^\infty h \cdot y(s,h,t) dsdh\right). \qquad (8)\end{aligned}$$

Let $S(t)$ denote the mean susceptibility of the uninfected and $I(t)$ the mean infectivity of the infected; then

$$S(t) = \frac{\int_0^\infty \int_0^\infty s x(s,h,t) dsdh}{\int_0^\infty \int_0^\infty x(s,h,t) dsdh} = \frac{\int_0^\infty \int_0^\infty s x(s,h,t) dsdh}{X(t)} \qquad (9)$$

and similarly

$$I(t) = \frac{\int_0^\infty \int_0^\infty h \cdot y(s,h,t) dsdh}{Y(t)}; \qquad (10)$$

thus, multiplying equation (9) by $X(t)$ and equation (10) by $Y(t)$ and substituting the results into equation (8),

$$\frac{dX(t)}{dt} = -\alpha X(t) S(t) Y(t) I(t). \qquad (11)$$

The specific formulae used from here depend on the specific choice assumed for $p(s,h)$. The ratio $\frac{p(s,h)}{P}$ is the bivariate probability density function for s and h in the population. Let $f(s,h) = \frac{p(s,h)}{P}$. Obviously, if s and h are independently distributed, $I(t)$ will be a constant. However, remember that s and h include individual variation in behavior. Epidemiological evidence on the infectivity of different behaviors suggests that, whereas some behaviors may be more likely to transmit the HIV virus in one direction than the reverse (for example, anal intercourse from the insertive to the receptive partner), almost all risky behaviors can transmit in either direction. Further, behavioral data suggest wide variation in the frequency of risky behaviors overall (Turner, Miller, and Moses, 1989). Thus, we can expect substantial variation in s, and h positively correlated with s.

Let us first consider a particularly simple case: let s and h be identical and exponentially distributed, so that (under our normalization) $f(s,h) = f(s) = e^{-s}$ and $h \equiv s$. (Alternatively, we could write $f(s,h) = \delta(h-s)e^{-s}$ where δ is the Dirac delta function.) Thus we have

$$S(t) = \frac{\int_0^\infty s e^{-\left(a(t)+1\right)s} ds}{\int_0^\infty e^{-\left(a(t)+1\right)s} ds} = \frac{\left(\frac{1}{a(t)+1}\right)^2}{\frac{1}{a(t)+1}} \qquad (12)$$

$$= \frac{1}{a(t)+1} \equiv \frac{X(t)}{P}, \qquad (13)$$

the proportion uninfected.

$$I(t) = \frac{\int_0^\infty h\left(e^{-\left(a(t-\tau)+1\right)h} - e^{-\left(a(t)+1\right)h}\right)dh}{\int_0^\infty \left(e^{-\left(a(t-\tau)+1\right)h} - e^{-\left(a(t)+1\right)h}\right)dh}$$

$$= \frac{\left(\frac{1}{a(t-\tau)+1}\right)^2 - \left(\frac{1}{a(t)+1}\right)^2}{\frac{1}{a(t+\tau)+1} - \frac{1}{a(t)+1}}$$

$$= \frac{1}{a(t-\tau)+1} + \frac{1}{a(t)+1}$$

$$\equiv \frac{Y(t)}{P} + 2\frac{X(t)}{P} = 1 + \frac{X(t)}{P} - \frac{Z(t)}{P}. \tag{14}$$

Note that early in the epidemic (i.e., $\frac{X(t)}{P} \approx 1$) $S(t)$ is approximately the population average ($\bar{s} \equiv 1$), while $I(t)$ is approximately twice the population average ($I(t) \approx 2 \equiv 2\bar{h}$); both $S(t)$ and $I(t)$ decline over the course of the epidemic. Obviously, the exponential distribution is a particularly simple case. One convenient generalization is the Γ distribution. The shape parameter can be chosen to give varying degrees of concentration of the distribution in the "low risk" end while maintaining a significant fraction at very high risk. (Exponential is Γ with shape parameter 1).

Let s be Γ with shape parameter c normalized to a mean of 1, i.e.,

$$f(s) = \frac{c^c s^{c-1} e^{-cs}}{\Gamma(c)}.$$

Then,

$$S(t) = \frac{\int_0^\infty c^c s^c e^{-\left(c+a(t)\right)s} ds}{\int_0^\infty c^c s^{c-1} e^{-\left(c+a(t)\right)s} ds}$$

$$= \frac{\left(\frac{c}{c+a(t)}\right)^{c+1} \cdot \Gamma(c)}{\left(\frac{c}{c+a(t)}\right)^{c} \cdot \Gamma(c)} = \frac{c}{c+a(t)}, \tag{15}$$

while

$$\frac{X(t)}{P} = \left(\frac{c}{c+a(t)}\right)^c, \tag{16}$$

hence

$$S(t) = \left(\frac{X(t)}{P}\right)^{1/c}. \tag{17}$$

If c is taken less than one, the bulk of the population has susceptibility less than the (arithmetic) average. We use $c = \frac{1}{3}$ for the base case in our elaborated model (Cardell

et al., 1987). If we kept $h \equiv s$ we would have

$$I(t) = \left(1 + \frac{X(t)}{P} - \frac{Z(t)}{P}\right)\left(\left(1 - \frac{Z(t)}{P}\right)^2 + \left(\frac{X(t)}{P}\right)^2\right). \tag{18}$$

However, $h \equiv s$ is not plausible, particularly when s is assumed to have a wide variation. Let h now denote the common factors that influence both susceptibility and infectivity. We will show that under this altered definition, h plays exactly the same role in our simplified model as under the definition just considered. Susceptibility and infectivity depend primarily on frequencies of risky behaviors. Since transmission can only occur in the presence of risky behaviors, we can conclude that susceptibility and infectivity should vary through some limited range conditional on h. (For instance, the probability per month of an individual becoming infected, or if infected, infecting someone else, can never exceed his probability of engaging in at least one risky behavior). We reflect this conclusion in the structure of equations. Let

$$s = h \cdot u$$
$$i = h \cdot v$$

where

h, u, v are independent random variables,
$h, u, v \geq 0$,
u, v with finite ranges,
i is individual infectivity (temporarily).

Let $f(h)$, $f(u)$ and $f(\nu)$ be the respective marginal distribution functions. Without loss of generality, we normalize h, u, ν so:

$$E(h) = E(u) = E(\nu) = 1.$$

Similarly, as before,

$$S(t) = \frac{\int_0^\infty \int_0^\infty h \cdot u f(h) f(u) e^{-a(t)h \cdot u} dh du}{\frac{X(t)}{P}} \tag{19}$$

$$\left(\frac{Y(t)}{P}\right) \cdot I(t) = \int_0^\infty \int_0^\infty \int_0^\infty h\nu f(h) f(u) f(\nu) \left(e^{-a(t-\tau)hu} - e^{-a(t)hu}\right) dh du d\nu$$

$$= \int_0^\infty \nu f(\nu) d\nu \int_0^\infty \int_0^\infty h f(h) f(u) (e^{-a(t-\tau)hu} - e^{-a(t)hu}) dh du \tag{20}$$

but

$$\int_0^\infty \nu f(\nu) d\nu \equiv E(\nu) \equiv 1.$$

Thus,

$$I(t) = \frac{\int_0^\infty \int_0^\infty hf(h)f(u)(e^{-a(t-\tau)hu} - e^{-a(t)hu})dhdu}{\left(\frac{Y(t)}{P}\right)} \tag{21}$$

$$\frac{X(t)}{P} = \int_0^\infty \int_0^\infty f(u)f(h)e^{-a(t)hu}dhdu \tag{22}$$

$$\frac{Z(t)}{P} = 1 - \int_0^\infty \int_0^\infty f(u)f(h)e^{-a(t-\tau)hu}dhdu \tag{23}$$

$$\frac{Y(t)}{P} = 1 - \frac{X(t)}{P} - \frac{Z(t)}{P}. \tag{24}$$

Note that the ν and functions of ν drop out of all the final computations. That is, random variations in infectivity that are unrelated to susceptibility have no effect on model results. Thus, we can ignore ν, take $i \equiv h$, and consider h to be individual infectivity as before.

Before proceeding, let us consider a simple choice for $f(u)$ and $f(h)$. Let h be gamma with shape parameter 2 and u be uniform $[0, 2]$; then

$$f(h) = 4he^{-2h}$$

$$f(u) = \begin{cases} \frac{1}{2} & 0 \le u \le 2 \\ 0 & u > 2 \end{cases}$$

$$\begin{aligned} \frac{X(t)}{P} &= \int_0^2 \int_0^\infty \left(\frac{1}{2}\right) 4he^{-2h}e^{-a(t)hu}dhdu = \int_0^\infty 2he^{-2h}\frac{1}{a(t)h}\left(1 - e^{-a(t)h2}\right)dh \\ &= \frac{2}{a(t)}\int_0^\infty \left(e^{-2h} - e^{-2h\left(1+a(t)\right)}\right)dh \\ &= \frac{2}{a(t)} \cdot \left(\frac{1}{2} - \frac{1}{2\left(1+a(t)\right)}\right) = \frac{1}{1+a(t)}, \end{aligned} \tag{25}$$

and

$$\begin{aligned} \frac{X(t)}{P}S(t) &= \int_0^\infty \int_0^\infty hu \cdot 2he^{-2h}e^{-a(t)hu}dudh \\ &= \int_0^\infty 2h^2e^{-2h}\left(\frac{1}{a(t)^2h^2} - \frac{e^{-a(t)h\cdot 2}}{a(t)^2h^2} - \frac{2e^{-a(t)h\cdot 2}}{a(t)h}\right)dh \\ &= \frac{1}{a(t)^2} - \frac{1}{a(t)^2} \cdot \frac{1}{1+a(t)} - \frac{1}{a(t)}\frac{1}{\left(1+a(t)\right)^2} \\ &= \frac{1}{\left(1+a(t)\right)^2}. \end{aligned} \tag{26}$$

Thus[2]

$$\frac{X(t)}{P} = \frac{1}{1+a(t)}$$

$$S(t) = \frac{1}{1+a(t)} = \frac{X(t)}{P}. \tag{27}$$

$$\frac{Y(t)}{P} = \frac{1}{1+a(t-\tau)} - \frac{1}{1+a(t)} \tag{28}$$

$$\begin{aligned}
\frac{Y(t)}{P} I(t) &= \int_0^\infty \int_0^2 h \cdot 2he^{-2h} \left(e^{-a(t-\tau)hu} - e^{-a(t)hu} \right) dudh \\
&= \int_0^\infty 2h^2 e^{-2h} \Big(\frac{1}{a(t-\tau)h}(1 - e^{-a(t-\tau)h\cdot 2}) - \frac{1}{a(t)h}(1 - e^{-a(t)h\cdot 2}) \Big) dh \\
&= \int_0^\infty 2h \Big(\frac{1}{a(t-\tau)} e^{-2h} - \frac{1}{a(t-\tau)} e^{-2\left(1+a(t-\tau)\right)h} \\
&\quad - \frac{1}{a(t)} e^{-2h} + \frac{1}{a(t)} e^{-2\left(1+a(t)\right)h} \Big) dh \\
&= \frac{1}{2} \frac{1}{a(t-\tau)} - \frac{1}{2} \frac{1}{a(t-\tau)} \frac{1}{\left(1+a(t-\tau)\right)^2} - \frac{1}{2} \frac{1}{a(t)} + \frac{1}{2} \frac{1}{a(t)} \frac{1}{\left(1+a(t)\right)^2} \\
&= \frac{1 + \frac{a(t-\tau)}{2}}{\left(1+a(t-\tau)\right)^2} - \frac{1 + \frac{a(t)}{2}}{\left(1+a(t)\right)^2} \\
&= \frac{1}{2} \left(\frac{1}{1+a(t-\tau)} - \frac{1}{1+a(t)} + \frac{1}{\left(1+a(t-\tau)\right)^2} - \frac{1}{\left(1+a(t)\right)^2} \right) \qquad (29) \\
\Longrightarrow I(t) &= \frac{1}{2} + \frac{1}{2} \left(\frac{1}{1+a(t-\tau)} + \frac{1}{1+a(t)} \right) \\
&= 1 + \frac{1}{2} \left(\frac{X(t)}{P} - \frac{Z(t)}{P} \right). \qquad (30)
\end{aligned}$$

Note that by comparison to the case where s is exponential (as it is here) and $h \equiv s$, the selection effect on $I(t)$ has been halved: that is, the effect of the fact that the most susceptible tend to become infected first on the average infectivity of the infected is one half of the earlier result.

[2]It is no coincidence that these results are the same as for the case where s is exponentially distributed. $f(h)$ and $f(u)$ are convenient in part because they result in an exponentially distributed s.

We have normalized h and u so that

$$1 \equiv \int_0^\infty \int_0^\infty f(h)f(u)hu \; dhdu$$
$$= \int_0^\infty \int_0^\infty \frac{x(h,u,t)}{P} hu \; dhdu + \int_0^\infty \int_0^\infty \frac{y(h,u,t)}{P} hu \; dhdu$$
$$+ \int_0^\infty \int_0^\infty \frac{z(h,u,t)}{P} hu \; dhdu$$

and

$$1 \equiv \int_0^\infty \int_0^\infty f(h)f(u)h \cdot dhdu$$
$$= \int_0^\infty \int_0^\infty \frac{x(h,u,t)}{P} h \cdot dhdu + \int_0^\infty \int_0^\infty \frac{y(h,u,t)}{P} h \cdot dhdu$$
$$+ \int_0^\infty \int_0^\infty \frac{z(h,u,t)}{P} h \cdot dhdu.$$

For the following discussion it is convenient to think of susceptibility and infectivity as aggregate quantities. Thus, the aggregate susceptibility of the uninfected population is $S(t)X(t)$. Let $Q(t)$ be the proportion of the initial aggregate susceptibility that remains in the uninfected population, then:

$$Q(t) = \frac{1}{P} X(t)S(t) = \int_0^\infty \int_0^\infty h \cdot u \cdot f(h)f(u)e^{-a(t)hu} dhdu. \tag{31}$$

Similarly, recall that h characterizes an individual, so we can consider the aggregate latent infectivity of the uninfected population. Let $R(t)$ be the proportion of this aggregate latent infectivity that remains in the uninfected population, then:

$$R(t) = \int_0^\infty \int_0^\infty hf(h)f(u)e^{-a(t)hu} dhdu. \tag{32}$$

We can now write:

$$S(t) = \frac{Q(t)}{\left(\frac{X(t)}{P}\right)} \tag{33}$$

$$I(t) = \frac{R(t-\tau) - R(t)}{\frac{X(t-\tau)}{P} - \frac{X(t)}{P}}. \tag{34}$$

In order to allow generalizations to a multiple-population model and to allow for convenient computer implementation, it is desirable to limit ourselves to functional form choices that allow Q and R to be solved for as simple functions of $\frac{X}{P}$. For the base case in our elaborated model (Cardell et al., 1987), as in the simplified simulation presented ahead, $Q(t) = \left(\frac{X(t)}{P}\right)^4$ and $R(t) = \left(\frac{X(t)}{P}\right)^2$. Note that $R(t)$ declines more slowly than $Q(t)$ over the course of the epidemic. This is a natural consequence of the fact that the infection process directly selects out the most susceptible, but only indirectly

selects out the most infectious (through the relationship between infectivity and susceptibility). We now demonstrate that $R(t)$ must be between $\frac{X(t)}{P}$ and $Q(t)$. That is, $Q(t) \leq R(t) \leq \frac{X(t)}{P}$. For simplicity, assume the moments of the h distribution are all finite.

Let
$$B(u,t) = \int_0^\infty h f(h) e^{-a(t)hu} dh. \tag{35}$$

Then
$$R(t) = \int_0^\infty f(u) B(u,t) du, \tag{36}$$

and
$$Q(t) = \int_0^\infty u f(u) B(u,t) du \tag{37}$$

$$\frac{\partial B(u,t)}{\partial u} = -\int_0^\infty a(t) h^2 f(h) e^{-a(t)hu} dh < 0. \tag{38}$$

For all $u > 0$, all t, all $B(u,t) > 0$, and all $f(u) > 0$ then $u \leq 1$ implies $uf(u) \leq f(u)$, and $u \geq 1$ implies $uf(u) \geq f(u)$. Recall

$$\int_0^\infty f(u) du \equiv 1 \equiv \int_0^\infty u f(u) du$$

$$\Longrightarrow \int_0^1 (1-u) f(u) du = \int_1^\infty (u-1) f(u) du \tag{39}$$

$$R(t) = \int_0^1 (1-u) f(u) B(u,t) du + \int_0^1 u f(u) B(u,t) du + \int_1^\infty f(u) B(u,t) du \tag{40}$$

$$Q(t) = \int_1^\infty (u-1) f(u) B(u,t) du + \int_0^1 u f(u) B(u,t) du + \int_1^\infty f(u) B(u,t) du \tag{41}$$

$$\begin{aligned} R(t) &\geq \left(\int_0^1 (1-u) f(u) du \right) \cdot B(1,t) + \int_0^\infty \min\left(f(u), uf(u)\right) B(u,t) du \\ &= \left(\int_1^\infty (u-1) f(u) du \right) \cdot B(1,t) + \int_0^\infty \min\left(f(u), uf(u)\right) B(u,t) du \qquad (42) \\ &\geq Q(t), \end{aligned}$$

since for $0 \leq u \leq 1$, it follows that $(1-u) \geq 0$ and $B(u,t) \geq B(1,t)$ while for $u \geq 1$, it follows that $(u-1) \geq 0$ and $B(u,t) \leq B(1,t)$. Thus $R(t) \geq Q(t)$ and since $\frac{\partial B(u,t)}{\partial u} < 0$, for any nondegenerate distribution of u (that is, $\int_0^{1-e} f(u) du > 0$, or equivalently, $\int_{1+e}^\infty f(u) du > 0$, some $e > 0$), $R(t) > Q(t)$.

Similar to the above argument, let

$$B(h,t) = \int_0^\infty f(u) e^{-a(t)hu} du. \tag{43}$$

Then

$$\frac{\partial B(h,t)}{\partial h} = -\int_0^\infty a(t)uf(u)e^{-a(t)hu}du < 0, \tag{44}$$

$$\frac{X(t)}{P} = \int_0^\infty f(h)B(h,t)dh, \tag{45}$$

$$R(t) = \int_0^\infty hf(h)B(h,t)dh. \tag{46}$$

Recall $\int_0^\infty f(h)dh \equiv 1 \equiv \int_0^\infty hf(h)dh$. Thus

$$\begin{aligned}
\frac{X(t)}{P} &= \int_0^1 (1-h)f(h)B(h,t)dh + \int_0^\infty \min\big(f(h), hf(h)\big)B(h,t)dh \\
&\geq \left(\int_0^1 (1-h)f(h)dh\right)\cdot B(1,t) + \int_0^\infty \min\big(f(h), hf(h)\big)B(h,t)dh \\
&= \left(\int_1^\infty (h-1)f(h)dh\right)\cdot B(1,t) + \int_0^\infty \min\big(f(h), hf(h)\big)B(h,t)dt \\
&\geq R(t),
\end{aligned} \tag{47}$$

thus $\frac{X(t)}{P} \geq R(t)$ and for $f(h)$ nondegenerate $\frac{X(t)}{P} > R(t)$.

Note that the above gives us four cases:

1) No susceptibility variation (h and u degenerate) $\frac{X(t)}{P} \equiv R(t) \equiv Q(t)$, $S(t) \equiv I(t) \equiv 1$.
2) Susceptibility varies and is identical to infectivity (h nondegenerate, u degenerate) $\frac{X(t)}{P} > R(t) \equiv Q(t)$.
3) Susceptibility variation only (h degenerate, u nondegenerate) $\frac{X(t)}{P} \equiv R(t) > Q(t)$, $I(t) \equiv 1$.
4) Susceptibility and infectivity vary and are related but not identical (h and u nondegenerate, $\frac{X(t)}{P} > R(t) > Q(t)$.

Obviously, the one-population model above is overly simplified. However, some generalizations are direct. Assume that there are J interacting groups, and that individuals in group j interact with individuals in group k for a proportion of their total activity θ_{jk} $\left(\sum_{k=1}^{J} \theta_{jk} \equiv 1\right)$. Then, equation 11 generalizes directly to:

$$\frac{dX(t)}{dt} = -\alpha_j X_j(t) S_j(t) \cdot \sum_k \theta_{jk} Y_k(t) I_k(t), \tag{48}$$

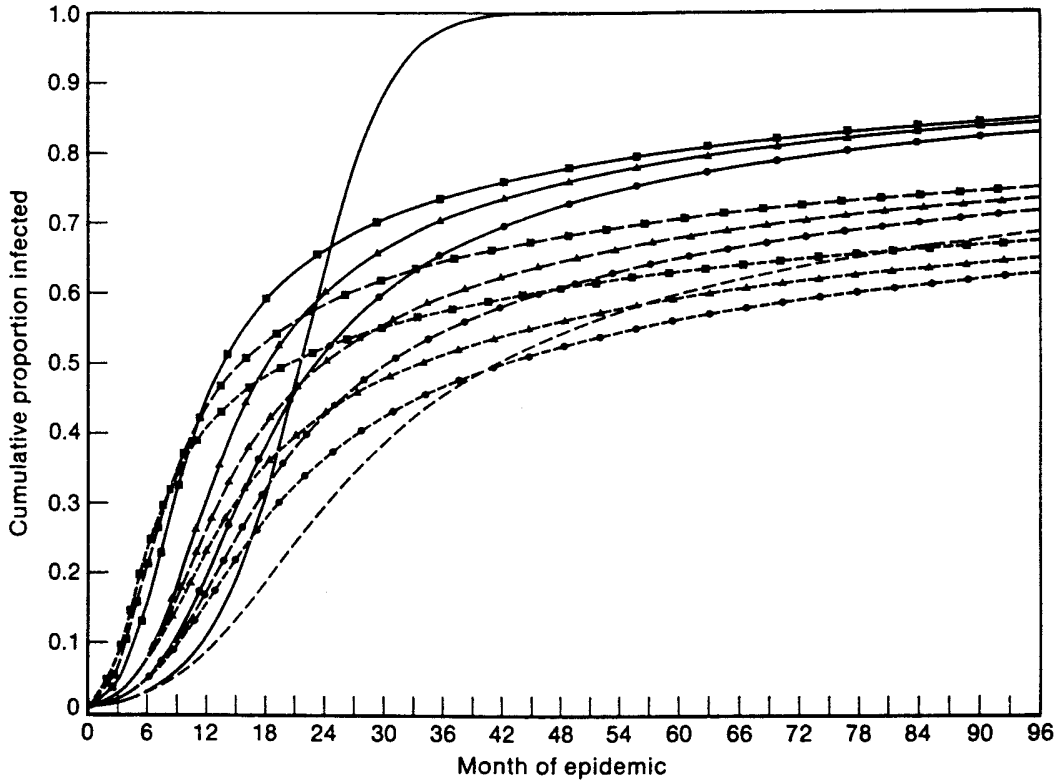
where S and I can be computed from Q, R and X above. This is basically the form used in the RAND HIV model (Cardell et al., 1987). If HIV-transmitting behavior is treated as symmetric, then it is necessary to impose constraints to ensure that the total amount of intergroup activity is the same for each group in the interacting pair jk (Blythe and Castillo-Chavez, 1989; Hethcote and Yorke, 1984; Nold, 1980).

In the epidemic of HIV infection, it is clear that susceptibility and infectivity vary considerably among individuals within identifiable groups (Padian et al., 1987). These variations certainly have behavioral components, which are positively correlated, resulting in a positive correlation between susceptibility and infectivity. There may also be a biological component (Wiley, Herschkorn, and Padian, 1989). The consequences of heterogeneity are that the most susceptible tend to become infected first; the mean susceptibility of the uninfected declines over time; the mean infectivity of those first infected exceeds the mean infectivity of the population that would occur if all were infected; and the mean infectivity of the infected declines over time. These consequences in turn result in an epidemic curve that initially (i.e., when the proportion infected is small) grows faster than the corresponding logistic curve (i.e., the epidemic curve when there is no variation in susceptibility or infectivity), and that later grows more slowly than the corresponding logistic curve.

We have shown that the simple theory developed above yields a practical and plausible representation of the effects of heterogeneity in epidemic models. This representation is an appropriate choice in a variety of situations. For instance, if the interaction between individuals of different susceptibility/infectivity is not random, the interpretation of $S(t)$ and $I(t)$ is more complex, but the same basic phenomena hold, and one can apply formulae 11, 33 and 34 with Q and R chosen as appropriate functions of $\frac{X(t)}{P}$. To apply the above technique the modeler simply chooses appropriate functions $R(t) = f_1\left(\frac{X(t)}{P}\right)$ and $Q(t) = f_2\left(\frac{X(t)}{P}\right)$ such that $\frac{X(t)}{P} > R(t) > Q(t)$. These choices are then used in formulae 33 and 34, and the results used in formulae 11 or 48. The choice $R(t) = \left(\frac{X(t)}{P}\right)^2$ and $Q(t) = \left(\frac{X(t)}{P}\right)^4$ has worked well in the RAND HIV model. We expect that this or a similar choice would work well in other diseases, including conventional STDs, where the behavior that makes one susceptible varies substantially at the individual level.

4. A One-Population Simulation

These dynamics and their consequences can be readily seen in the results of a simulation that introduces heterogeneity into a simple one-population SI model. In this simulation, the starting condition is a 1% prevalence of infection and a value of α that leads to a 23% growth rate for the no-variation case in the first month. Figure 1 shows the cumulative proportion infected at various points over an eight-year period for populations with differing amounts of variation in susceptibility and infectivity. The first column sketches the epidemic growth curve in a population that is homogeneous in susceptibility and infectivity; in this case, the high growth rate results in nearly universal infection in about three years. The second column shows the consequences of introducing variation in susceptibility. (The parameterization chosen for this case is the one used in the "base case" of our epidemiological model, and is described more fully below). In this case, the epidemic grows much more slowly than in the homogeneous case, so that after eight years nearly a third of the population remains uninfected.



NN = no infectivity or susceptibility variation
NM = no infectivity variation, medium susceptibility variation
LL = low infectivity variation, low susceptibility variation
LM = low infectivity variation, medium susceptibility variation
LH = low infectivity variation, high susceptibility variation
ML = medium infectivity variation, low susceptibility variation
MM = medium infectivity variation, medium susceptibility variation
MH = medium infectivity variation, high susceptibility variation
HL = high infectivity variation, low susceptibility variation
HM = high infectivity variation, medium susceptibility variation
HH = high infectivity variation, high susceptibility variation

NOTE: Low, medium and high susceptibility variation are identified respectively as mean relative susceptibility of the uninfected equals the square, cube or fourth power of the proportion uninfected. Low and medium infectivity variation are defined respectively as mean relative infectivity of the infected equals one plus one half the proportion uninfected or one plus the proportion uninfected. High infectivity variation is defined as infectivity perfectly correlated with susceptibility.

Figure 1
Effects of Variation in Infectivity and Susceptibility
in a Single High Growth Rate Population

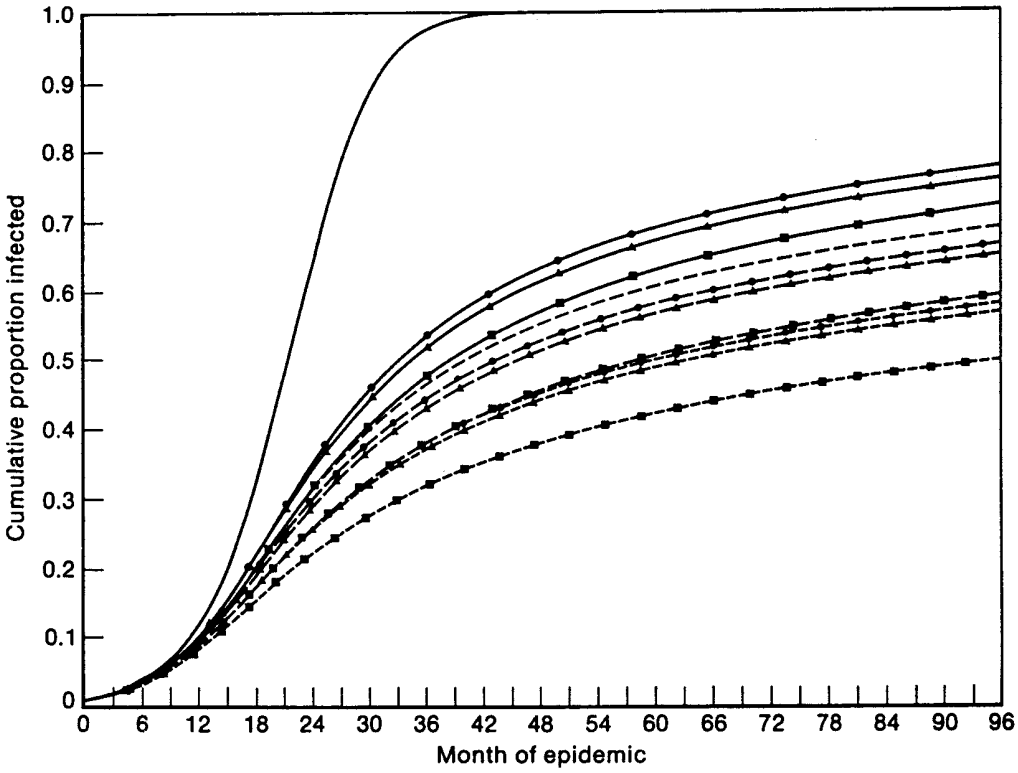
The rightmost three columns at the top of Figure 1 show epidemic growth scenarios when both susceptibility and infectivity are allowed to vary (and are positively correlated). The effects of a low level of variation in infectivity are shown for three different levels of variation in susceptibility. Note that the growth rates in the first 12 to 18 months are higher than when there is no variation. These initially higher growth rates soon moderate, however, so that by 24 months, the epidemic is growing more slowly than it would in the absence of variation. Note also that the effect of introducing infectivity variation along with a given level of variation in susceptibility is to increase the epidemic growth rate. The columns at the bottom of the table provide similar results in cases of "medium" or "high" variation in infectivity. The "high" infectivity variation cases are especially interesting, because in these cases infectivity is defined as perfectly correlated with susceptibility. The effect of their variation is first to increase and then to moderate the epidemic growth rate relative to the case of no variation.

One way of gauging how important heterogeneity may be in modeling the spread of HIV infection is to assess how much difference alternative assumptions about heterogeneity make in the range of future projections that are consistent with a particular observed history. As a simple illustration of this, Figure 2 shows the results of a simulation identical to that presented in Figure 1, except that all cases are initialized to have the same growth rate assumed in the "no variation" case in Figure 1 (23% in the first month). Note that infectivity variation now reduces the rate of epidemic growth for all periods after the initial "observed history". The reason is that variation in infectivity results in a decline over time in the mean infectivity of the infected, and we have now adjusted to start at the same level.

The main point of Figure 2, of course, is that variations in infectivity and susceptibility make a substantial difference in the subsequent epidemic growth rate, even after one has taken the initial growth rate into account. A pragmatic implication is that the absence of good information about the extent of population heterogeneity can be an important source of uncertainty in fitting a model to an observed epidemic growth curve. This will be especially true in the early stages of the epidemic, when there is little indication as to whether and how rapidly the growth rate will moderate as the mean levels of infectivity of the infected and susceptibility of the uninfected decline over time.

5. Discussion

Although our focus in this paper has been on parametric representation of heterogeneity within a single population, we noted at the outset that models may also deal with heterogeneity by partitioning the population into risk groups that differ in their behavior. Indeed, for addressing certain types of heterogeneity, partitioning is the preferred approach. Consider four ways in which subgroups of a population may differ: in the *type* of behavior in which they engage, in their *patterns of interaction* with other subgroups, in the *frequency* of their relevant behaviors and interactions, and in the *consequences* attendant on those behaviors. The first two sources of heterogeneity, type of behavior and patterns of interaction, can best be addressed in a model by partitioning the population into discrete groups rather than by parameterization. For example, group boundaries in the RAND HIV model are defined by participation in



NN = no infectivity or susceptibility variation
NM = no infectivity variation, medium susceptibility variation
LL = low infectivity variation, low susceptibility variation
LM = low infectivity variation, medium susceptibility variation
LH = low infectivity variation, high susceptibility variation
ML = medium infectivity variation, low susceptibility variation
MM = medium infectivity variation, medium susceptibility variation
MH = medium infectivity variation, high susceptibility variation
HL = high infectivity variation, low susceptibility variation
HM = high infectivity variation, medium susceptibility variation
HH = high infectivity variation, high susceptibility variation

NOTE: Susceptibility and infectivity are defined as in Figure 1. Initial prevalence of infection (1%) and initial growth rate in new infections (23% in first month) are defined as the same in all cases. Differences in growth rates thereafter reflect differences in population heterogeneity.

Figure 2
Effects of Variation in Infectivity and Susceptibility
After Initial Growth Rate Has Been Normalized

each of the key risk behaviors (homosexual contact, heterosexual contact, and needle sharing). Heterogeneity in behavioral frequency and epidemiological consequences can be addressed either through categorization or parameterization, but most effectively through both. If we assume proportionate mixing within groups, groups should be defined in such a way that selection of partners for risky behaviors tends to occur within rather than between groups. For that reason, age and geographic location were the other primary determinants of group boundaries in the RAND HIV model.

A full model of HIV spread within the overall population must address many complexities besides heterogeneity. For example, the infected stage is not uniform and the induction time from infection to death is not a constant as assumed above. Most of these complexities, however, need not affect the way that within-group heterogeneity is handled within a model. The technique we have described in this paper allows a simple and plausible representation of the effects of heterogeneity in epidemiological models, including a full model of HIV spread.

Our analysis suggests a number of conclusions. First, the extent of within-group variation in infectivity and susceptibility can have a substantial effect on the dynamics of the growth rate in epidemics of this type. Second, epidemiological projections for HIV are quite sensitive to the amount of variation that is explicitly assumed, and therefore to the amount implicitly assumed where such variation is not explicitly considered. Third, heterogeneity in susceptibility and infectivity naturally slows the epidemic growth rate over time. Such slowing could rather easily be confused with the effects of behavior change, which are similar. The implications of the two processes are quite different, however, and it is important to distinguish them both in modeling work and in monitoring and interpreting actual incidence data. In modeling work, the use of a disaggregated model makes it possible to distinguish the effects of behavior change from the selection effects that result from heterogeneity. Fourth, the importance of heterogeneity in modeling transmission dynamics suggests the importance of gathering data that would permit empirically-based estimation of these parameters. At present, it is difficult to find relevant data for this purpose, but it is possible to describe the types of studies that would be useful. These include studies of the distribution and patterning of risk behaviors in populations and studies that seek to identify and quantify biological markers of infectivity or susceptibility in individuals. Fifth, it is important to consider the potential effects of policy options on the variation in susceptibility and infectivity as well as on their mean levels.

Acknowledgments

This paper is based in part on a presentation at the annual meetings of the Population Association of America, April 1988. This research was supported by RAND corporate funds. We are grateful to Audrey Cardell, James Hammitt, Albert Williams, and two anonymous reviewers for comments on an earlier draft.

REFERENCES

Blythe, S.P. and R.M. Anderson. (1988). Distributed incubation and infectious periods in models of the transmission dynamics of the human immunodeficiency virus (HIV). *IMA J. Math. Biol. Med.*, 5, 1-19.

Blythe, S.P. and C. Castillo-Chavez. (1989). Like-with-like preference and sexual mixing models. Submitted, *Math. Biosci.*

Cardell, N.S., D.E. Kanouse, E.M. Gorman, C. Serrato, P.H. Reuter, and A.P. Williams. (1987). Modeling the spread of human immunodeficiency virus in the United States. Presented to III International Conference on AIDS, Washington, D.C.

Castillo-Chavez, C., K. Cooke, W. Huang, and S.A. Levin. (1989). The role of long periods of infectiousness in the dynamics of acquired immunodeficiency syndrome (AIDS). In *Mathematical Approaches to Resource Management and Epidemiology.* In press, Lecture Notes in Biomathematics, Springer-Verlag.

Curran, J.W., H.W. Jaffe, A.M. Hardy, M. Morgan, R.M. Selik, and T.J. Dondero. (1988). Epidemiology of HIV infection and AIDS in the United States. *Science*, 239, 610-616.

Friedland, G.H, and R. S. Klein. (1987). Transmission of the human immunodeficiency virus. *N. Engl. J. Med.*, 317, 1125-1135.

Hethcote, H.W. (1976). Qualitative analysis of communicable disease models. *Math Biosciences*, 28, 335-356.

Hethcote, H.W. and J.W. Van Ark. (1987). Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. *Math. Biosci.*, 84, 85-118.

Hethcote, H.W. and J.A. Yorke. (1984). *Gonorrhea, transmission dynamics and control.* Lecture Notes in Biomathematics 56, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

Hyman, J.M. and E.A. Stanley. (1988). Using mathematical models to understand the AIDS epidemic. *Math. Biosci.*, 90, 415-473.

Jacquez, J.A., C.P. Simon, J. Koopman, L. Sattenspiel, and T. Perry. (1988). Modeling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.*, 92, 119-199.

Nold, A. (1980). Heterogeneity in diseases-transmission modeling. *Math. Biosci.*, 52, 227-240.

Padian, N., J. Wiley, and W. Winkelstein. (1987). Male to female transmission of human immunodeficiency virus (HIV): Current results, infectivity estimates, and San Francisco population seroprevalence estimates. Presented to III International Conference on AIDS, Washington, D.C.

Sattenspiel, L. and C.P. Simon. (1988). The spread and persistence of infectious diseases in structured populations. *Math. Biosci.*, 90, 341-366.

Turner, C.F., H.G. Miller, and L.E. Moses (eds.). (1989). *AIDS: Sexual Behavior and Intravenous Drug Use.* National Academy Press, Washington, D.C.

Wiley, J. (1987). Models for estimation of transmission probabilities of HIV in epidemiologic studies. Presented at Conference on Statistical and Mathematical Modeling of the AIDS Epidemic, Johns Hopkins University, Baltimore, MD.

Wiley, J.A., S.J. Herschkorn, and N.S. Padian. (1989). Heterogeneity in the probability of HIV transmission per sexual contact: The case of male-to-female transmission in penile-vaginal intercourse, *Stat. Med.*, 8, 93-102.