Halloran ME. Concept of Infectious Disease Epidemiology (ch. 27). In: Rothman KJ, Greenland S. (eds). Modern Epidemiology. Philadelphia: Lippincott-Raven. 1998.

27

Concepts of Infectious Disease Epidemiology

M. Elizabeth Halloran

Infectious disease epidemiology is characterized by the presence of at least one other active player in addition to the human population, namely, the infectious agent or parasite. The presence of this additional propagating population sets the stage for aspects specific to infectious disease epidemiology. First and foremost is transmission. Transmission from one host to another is fundamental to the survival strategy of the infectious agent, since any host will eventually either clear the infection or die, even if from an unrelated cause. A consequence of transmission is that, unlike noninfectious diseases, the occurrence of infectious disease events in individuals depends on the occurrence of that disease in other members of the population. Sir Ronald Ross (1916) called this dependence of disease events in infectious diseases "dependent happenings."

Although most methods used in general epidemiology are applicable to the study of infectious diseases, additional concepts are needed to describe the phenomena resulting from the dependence of disease events. These include infectiousness, the transmission probability, contact patterns, and the basic reproductive number. An intervention in infectious diseases can also have several different kinds of effects, including direct effects on a person receiving the intervention as well as indirect effects on other individuals. These different effects require additional parameters and study designs for their evalua-

tion. Exposure to infection plays a special role because exposure to infection is necessary for infection and disease to occur. The components of exposure to infection, such as the contact and mixing patterns of the infective and susceptible hosts, as well as the degree and duration of infectiousness, need to be taken into account in infectious disease epidemiology.

Even when conventional epidemiologic concepts are applicable, they should be used in infectious disease studies only after close examination of the underlying assumptions. Because the temporal evolution of the host population and the disease process under study can be quite rapid compared with the time frame of the study, conventional epidemiologic methods that assume stationarity can produce very biased estimates of effects in infectious disease epidemiology.

Epidemiology of infectious diseases is an extension of ecology and evolution. From our anthropocentric point of view, we say a person with the agent is infected. From the point of view of the infectious agent, however, humans are simply home and lunch, their ecologic niche (see Burnet and White, 1972; McNeill, 1976). Because each infectious agent has its own life cycle, immunology, ecology, evolution, and molecular biology, studies of infectious disease require an understanding of all of these aspects. Most of these considerations are beyond the scope of this chapter. Emerging and reemerging infectious disease and the development of drug resistance will not be considered here.

In this chapter, we introduce a few important concepts of infectious disease epidemiology, focusing on the consequences of the dependent structure of disease events for measures of effect and study design. Some of the ideas about dependent events are applicable beyond the infectious disease setting. An example is drug addiction, in which the number of people becoming addicted depends partly on the number of people already addicted. On the other hand, not all diseases caused by infectious agents result in a dependent happening structure. Examples include Lyme disease and sylvan yellow fever. These are infectious diseases called zoonoses that generally circulate in animal hosts and are occasionally transmitted to humans, so that within the human population the events are not dependent.

TIME LINES OF INFECTION

The time lines of infection within the host can be described with reference to the dynamics of infectiousness and of disease (Fig. 27-1). Both begin with the successful infection of the susceptible host by the parasite. The time line of infectiousness includes the *latent period*, the time interval from infection to development of infectiousness, and the *period of infectiousness* of the host, during which time the host could infect another host. Eventually, the host becomes noninfectious either by recovery from the infection, possibly developing immunity, or by death. The host can also become noninfectious while still alive and still harboring the parasite.

The time line of disease within the host includes the *incubation period*, the time from infection to development of symptomatic disease, and the *symptomatic period*. The probability of developing symptoms or disease after becoming infected is the *pathogenicity* of the interaction of the parasite with the host. Eventually, the host leaves the symptomatic state either by recovering from the symptoms or by death. The host becomes an infectious *carrier* if he recovers from symptoms but remains infectious. For example, people infected with hepatitis B often become infectious carriers. If the parasite has initiated

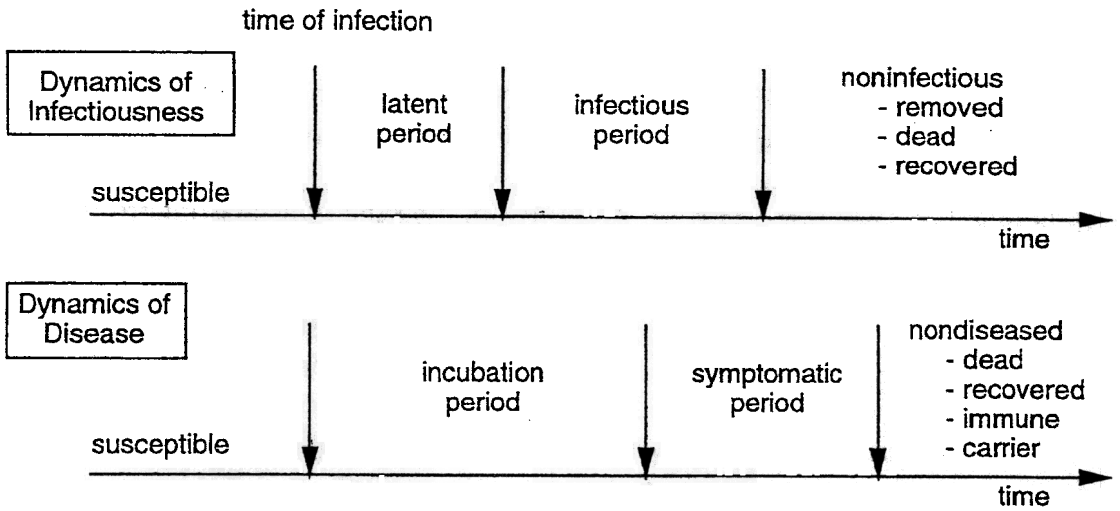


FIG. 27-1. Time lines for infecton and disease.

an autoimmune response, symptoms can continue after the parasite is cleared. Rheumatic heart disease can develop after streptococcus B infection is cleared. An *inapparent case* or *silent infection* is a successful infection that does not develop detected symptoms. Inapparent cases can be infectious.

The terminology differs from that used in noninfectious disease epidemiology. The term *latent period* has an entirely different meaning from that just described, corresponding to the period from development of asymptomatic disease to development of symptoms. The incubation period in infectious disease is a combination of what are called the induction and the latent periods in noninfectious diseases. While the disease process and its associated time line are important to the infected person and to a physician, the dynamics of infectiousness are more important for propagation of the parasite and for public health.

The configuration of the two time lines in Fig. 27-1 and their relation to one another are specific to each parasite and can have important public health consequences and implications for study design. In chickenpox, the latent period is shorter than the incubation period, so that a child with chickenpox becomes infectious to other people before developing symptoms. A study of chickenpox transmission in school children revealed that most transmission occurred before the development of symptoms, thus requiring children with chickenpox to stay home from school does not make much sense from the point of view of trying to reduce transmission. Malaria from *Plasmodium falciparum* has an incubation period of about 14 days in the human host. The stages of the parasite that are infective for mosquitoes occur about 10 days after the development of malaria symptoms. Thus, the latent period is about 10 days longer than the incubation period, so early treatment of symptoms could have an important effect on transmission. Human immunodeficiency virus (HIV) infection is a prime example of a public health nightmare in which the infectious agent has a short latent period and a long incubation period. The latent period is on the order of days to weeks, while the median incubation period to observed symptoms is greater than 10 years. A person infected with HIV can infect other people for a long time before disease is apparent.

TRANSMISSION PROBABILITY

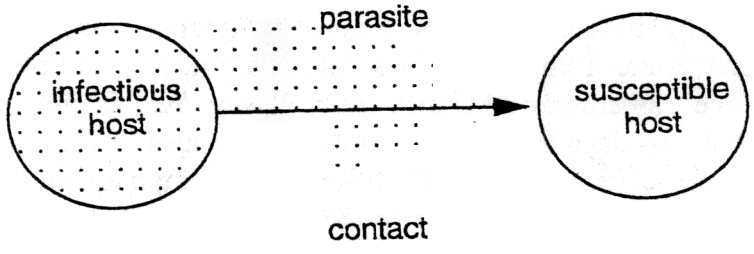
A fundamental parameter of infectious disease epidemiology is the *transmission probability*. The transmission probability is the probability that, given contact between an infective source and a susceptible host, successful transfer of the parasite will occur so that the susceptible host becomes infected (Fig. 27-2). Estimating the transmission probability and its variability in a population is important for understanding the dynamics of infection and the effects of interventions. The transmission probability depends on characteristics of the infective source, the parasite, the susceptible host, and the type and definition of contact. The infectious source could be another person, as in transmission of measles or mumps. It could be an insect vector, such as the mosquito vector of the malaria parasite, or a contaminated inanimate object, such as drinking water containing cholera bacteria or needle syringes infected with hepatitis B virus.

The concept of a *contact* is very broad and must be defined in each particular study. The mode of transmission of a parasite determines what types of contact are potentially infectious. Different definitions of a potentially infective contact for a given parasite, even within the same study, are possible. In a study of whooping cough transmission, a potentially infective contact could be defined as being in the same school on one day with someone with culture-proven whooping cough. Alternatively, it could be defined as living in the same house during the period of presumed infectiousness of the person with whooping cough. In an HIV study, a potentially infective contact could be defined as each sex act between two sexual partners in a steady relationship, one of whom is infected with HIV. Alternatively, the partnership over its entire duration could be defined as one potentially infective contact. In any study or analysis, it is important to use a precise operational definition for a potentially infective contact.

Difficult problems of infectious disease epidemiology are identifying infectious sources and susceptible hosts, quantifying infectiousness and susceptibility, knowing the strain of the parasite, and defining and identifying contacts between infectives and susceptibles.

Estimating the Transmission Probability

There are several methods for estimating the transmission probability. We illustrate two broad approaches here. In the first, infectious individuals are identified and the propor-



Transmission probability depends on
- infectious host
- susceptible host
- contact definition
- parasite

FIG. 27-2. Transmission from an infective to a susceptible host during contact.

tion of contacts that they make with susceptibles that result in transmission is determined. In the second, susceptibles are identified and data gathered on the number of contacts they make with infectives and their infection outcomes. To illustrate the first approach, we present the conventional secondary attack rate. To illustrate the second, we present the binomial model.

Secondary Attack Rate

The general idea of the secondary attack rate or case-contact approach of estimating the transmission probability is to identify infectious persons and then to identify the susceptible people who make contact with them by some definition of contact. The initially identified infectious persons are called the *primary* or *index cases*. The *conventional secondary attack rate* (SAR) is the probability of the occurrence of disease among known (or presumed) susceptible persons following contact with a primary case:

$$\hat{SAR} = \frac{\text{number of persons exposed who develop disease}}{\text{total number of susceptible exposed persons}}. \qquad [27\text{-}1]$$

The SAR is actually a proportion, not a rate. It is often defined for exposure to an infective within some small population unit, such as a household, classroom, or school bus. Within this unit, mixing and exposure to infection are assumed to be homogeneous.

Example. The *household SAR* is the probability that a susceptible individual living within the same household with an infectious person during his or her period of infectiousness will become infected. The household SAR is a commonly used parameter for estimating vaccine efficacy in directly transmitted infections, such as pertussis, mumps, chickenpox, and measles (Fine et al., 1988; Orenstein et al., 1988). The data required are the time of onset of disease for each case in the household, as well as knowledge of who is susceptible. Estimates or assumptions about the minimum and maximum incubation periods, E_1 and E_2, respectively, the latent period, and the maximum time, I, that a person remains infectious are also required (Fig. 27-3) and sometimes obtained from other studies. One sometimes assumes that the onset of symptoms coincides with the onset of infectiousness and that there are no inapparent cases.

The first step in assessing the SAR is to define for the disease under study the time interval after the index case that would include secondary cases. The presumed beginning of infectiousness of the index case is defined as time 0 for each household. *Secondary cases* are those with time of onset between the end of the minimum incubation period E_1 relative to the beginning of infectiousness of the index case ($t = 0$) and the end of the maximum incubation period E_2 relative to the time of the maximum infectious period of the primary case, $t = I$. Thus, secondary cases are those occurring in the interval $(E_1, I + E_2)$. A case with recorded onset time less than one minimum incubation period, E_1, after that of the index case was presumably not infected by the index case and is called a *co-primary case*. Tertiary and higher cases are those occurring after the maximum allowable time interval for the secondary cases.

Example. For an early efficacy study of pertussis vaccines, Kendrick and Eldering (1939) estimated the infectious period for the bacteria from studies of throat cultures, finding that nearly everyone had a positive culture up to 21 days after onset of symptoms. They defined a definite exposure (potentially infective contact) as living in the same house as the index case or being indoors in another house with the index case for at least 30 minutes within $I_d = 21$ days of onset of symptoms of the index case. The mean incubation period of pertussis from two other studies was estimated to be 13 ± 7.6 days and

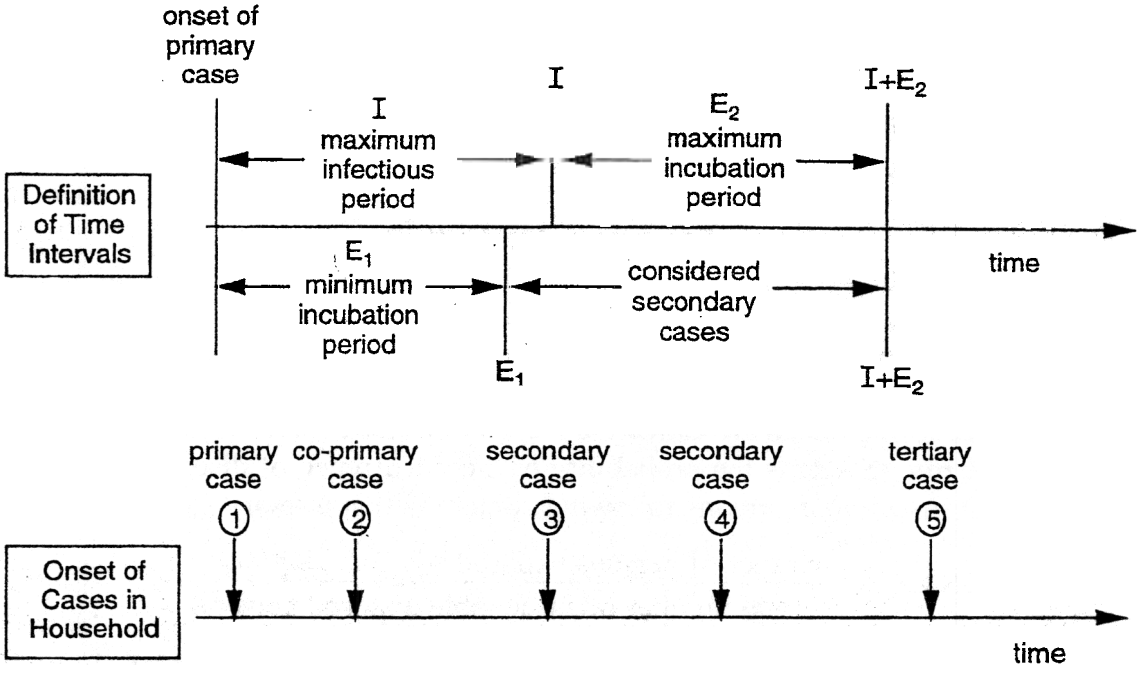


FIG. 27-3. Time periods for estimating the household secondary attach rate.

15.4 ± 1.3 days. Based on this information, Kendrick and Eldering somewhat arbitrarily set the minimum incubation period to $E_1 = 10$ days and the maximum incubation period to $E_2 = 30$ days. Under the definition of definite exposure, secondary cases were those occurring between $E_1 = 10$ and $I_d + E_2 = 21 + 30 = 51$ days after the onset of symptoms in the index case.

Kendrick and Eldering had a second definition for a potentially infective contact called an *indefinite exposure.* Based on the observation that between 21 and 35 days after onset of symptoms, throat cultures were less often positive, someone exposed up to $I_i = 35$ days after onset of symptoms in the index case was defined as an indefinite exposure. The definition of potentially infectious contacts under this less specific definition also included outdoor contacts. Under the less stringent definition of indefinite exposure, secondary cases were those occurring between day $E_1 = 10$ and $I_i + E_2 = 35 + 30 = 65$ days after the onset of symptoms in the index case. This example illustrates that the definition of a potentially infectious contact in a study is somewhat arbitrary and should be made explicit.

The second step in assessing the SAR is to determine for each ascertained case within the minicohort in each household whether it is a coprimary, secondary, tertiary, or higher generation case. The estimated household SAR is the total number of secondary cases in all households divided by the total number of at-risk susceptibles in all households, as in equation 27-1. Coprimary cases are excluded from the denominator. Tertiary or higher cases are excluded from the numerator but included in the denominator.

Example. The case-contact approach is used to estimate the transmission probability of tuberculosis. Upon identification of an infectious case of tuberculosis, public health officials locate people who have made contact with the infectious case and test them for whether they have become infected. The pooled estimate of the proportion who have become infected is an estimate of the transmission probability.

Difficulties in estimating the SAR and case-contact rates include determination of the latent and incubation periods, ascertainment of onset times of cases, and determination of when an exposure to infection has taken place.

Binomial Models of Transmission Probabilities

The binomial model is often used when susceptibles make more than one potentially infectious contact. The probability of transmission during a contact between a susceptible and an infectious person is denoted by p. The probability of the susceptible person's escaping infection during the contact is $q = 1 - p$. Suppose that a person makes n contacts with an infective or with different infectives and that the probability of being infected during any contact is independent of any previous contacts. Then the probability of escaping infection from all n potentially infective contacts is $q^n = (1 - p)^n$. The probability of being infected after n contacts, that is, of not escaping infection from all n contacts, is $1 - q^n = 1 - (1 - p)^n$. The maximum likelihood estimate of the transmission probability under the binomial model is

$$\hat{p} = \frac{\text{number of susceptibles who become infected}}{\text{total number of contacts with infectives}} \qquad [27\text{-}2]$$

Note the similarity between the formula for $\hat{p}$ and the one for $\hat{\text{SAR}}$ in equation 27-1. The difference is in the denominators. In the binomial model, we count the total number of potentially infectious contacts that susceptible individuals make, while in the SAR each susceptible person had just one potentially infectious contact with the infective. The two formulas would be the same if everyone in the binomial model made just one potentially infectious contact. Both p and SAR are measures of the transmission probability.

Example. A study of HIV transmission was conducted in a population of 100 steady sexual couples. At the beginning of the study, one partner in each couple was already infected and the other partner was susceptible. Twenty-five of the 100 susceptible partners became infected during the follow-up period. The total number of sexual encounters reported in the study either up until a person became infected or until the end of the study was 1500. The maximum likelihood estimate of the transmission probability is $\hat{p} = 25/1500 = 0.017$. The probability of becoming infected after two contacts with an infected person is $1 - (1 - \hat{p})^2 = 0.034$.

It is possible to estimate the transmission probability even if the infection status of people making contact with susceptibles is not known, if other information is used. For example, an estimate of the prevalence P of infection in the pool of potential contacts might be available. The probability of becoming infected from a contact with a partner with unknown infection status chosen at random from a population with prevalence P is Pp. Thus, the probability of infection after n total contacts is $1 - (1 - Pp)^n$. The transmission probability can be estimated by solving for p using information on the proportion becoming infected during the study, the total number of contacts, and the prevalence of infection in the pool of contacts.

Sometimes information on the exact number of contacts is not available. Study subjects might give information on the average number of contacts they each make per unit time. From this, the expected number of contacts during the study period can be estimated. The exact form of the binomial model that is used in an analysis depends on the data available and on the assumptions made about the variability of infectiousness of the infectives and the variability of susceptibility of the susceptibles (Kim and Lagakos, 1990). Even when analysis using the binomial model becomes computationally quite demanding, the underlying principle remains simple. There is a probability p of transmission and a probability $q = 1 - p$ of escaping transmission upon contact with an infective.

Other Binomial Models

Chain binomial models are developed from the simple binomial model by assuming that infection spreads within a population in discrete units of time, producing chains of infection governed by the binomial probability distribution. If p is the transmission probability from an infective host to a susceptible host with a contact assumed to last one time unit, then the number of new infections in exposed susceptibles at the end of a time unit is assumed to follow the binomial distribution. The expected distribution of infections after several units of time can be calculated from the chained, that is, sequential, application of the binomial model. The Reed-Frost and Greenwood models are examples of chain binomial models. The Reed-Frost model assumes that exposure to two or more infectious people at the same time are independent exposures. The probability of escaping infection from two infectives is $(1-p)^2$. The Greenwood model assumes that exposure to two or more infectious people at the same time is the equivalent to exposure to one. Under this model, the probability of escaping infection from simultaneous exposure to two infectives is $1-p$. The chain binomial models can be used to estimate the transmission probability from data gathered on each generation of infection or from the final distribution of infections within a collection of households after an epidemic has occurred. Abbey (1952), Bailey (1957), and Becker (1989) discuss chain binomial models.

Transmission in Small Units Within Larger Communities

In the conventional household SAR studies and the HIV partner study described above, the houses and partnerships were assumed to be independent of each other. The susceptibles were assumed to be infected only by the index case, who had somehow become infected. The small units, households, or partnerships could be part of a community, however, so that individuals can become infected either from the index case or in the community at large. If the transmission probability or SAR is estimated without taking into account the opportunity to become infected outside the unit under study, it will overestimate the actual probability of becoming infected per contact. Longini and Koopman (1982) developed a model for transmission in a community of households that takes into account both sources of infection. A similar approach can be used in estimating HIV transmission probabilities, where nonmonogamous partnerships can be thought of as households of size two.

BASIC REPRODUCTIVE NUMBER

A second important parameter in infectious diseases is the basic reproductive number, R_0. Understanding R_0 is important for public health applications in infectious diseases. For *microparasitic diseases,* such as those caused by viruses and bacteria, R_0 is defined as the expected number of new infectious hosts that one infectious host will produce during his or her period of infectiousness in a large population that is completely susceptible. R_0 does not include the new cases produced by the secondary cases, or further down the chain. It also does not include secondary cases who do not become infectious.

Example. If $R_0 = 9$ for measles in a population, then one person with measles introduced to that population would be expected to produce nine new secondary infectious cases before recovering, if the population were completely susceptible. If the person produced two additional cases who did not become infectious, R_0 would still be 9.

In general, for an epidemic to occur in a susceptible population, R_0 must be >1. If $R_0 < 1$, an average case will not reproduce itself, so an epidemic will not spread. Since R_0 is an average, it is possible that a particular infectious person will produce more than one infective case, even when $R_0 < 1$, so there may be a small cluster of cases. We would not, however, expect a self-sustaining outbreak.

For microparasitic infections, R_0 is a composite of three important aspects of infectious diseases: the rate of contacts c, the duration of infectiousness d, and the transmission probability per potentially infective contact p. The average number of contacts made by an infective during the infectious period is the product of the contact rate and the duration of infectiousness—cd. The number of new infections produced by one infective during his infectious period is the product of the number of contacts in that time interval and the transmission probability per contact:

$$R_0 = \begin{matrix}\text{number of} \\ \text{contacts per} \\ \text{unit time}\end{matrix} \times \begin{matrix}\text{transmission} \\ \text{probability} \\ \text{per contact}\end{matrix} \times \begin{matrix}\text{duration} \\ \text{of} \\ \text{infectiousness}\end{matrix} = cpd\,.$$

A term could be included to account for the probability of becoming infectious after infection.

A value of R_0 is not specific to a parasite, but to a parasite population within a particular host population at a particular time. The contact rates in rural areas will be lower than contact rates in urban areas, so we expect the R_0 of measles to be lower in rural than in urban areas. R_0 of malaria may be high during the season of high mosquito density but low during the season when there are few mosquitos. The R_0 of HIV infection in intravenous drug users might be much higher than it is for HIV infection in a heterosexual population.

Because R_0 is the number of new infectious cases per infectious case, it is a dimensionless quantity. Without further information about the magnitude of the parameters that make up R_0, we cannot conclude much about the time frame of an epidemic, the transmissibility of the infectious agent, or the contact rate. R_0 is about 9 for measles in some populations and also about 9 for HIV infection in some populations of intravenous drug users. We know from other sources that measles has a high transmission probability and short duration of infectiousness and moves much faster than HIV, which has a low average transmission probability and longer duration of infectiousness. If we knew only that $R_0 = 9$ for both, then we would know that they both could result in major epidemics, but we would not be able to draw conclusions about the relative time frames of the two.

Indirectly transmitted diseases are those in which a parasite is transmitted between two different host populations. An example is the vector-borne disease malaria, transmitted from humans to mosquitos and back to humans. Another example is heterosexual transmission of sexually transmitted diseases where the infection is transmitted from a man to a woman and back to a man. R_0 for indirectly transmitted diseases depends on the product of the two components of transmission. If a woman infects on average two men and a man infects on average three women, then one infectious case amplifies on average to six infectious cases in the same host population.

By definition, R_0 assumes that all contacts are with susceptibles. In real populations, however, there are often people who are already immune to a parasite. Under these circumstances, the expected number of new cases produced by an infectious person is less than R_0 and is called the *effective reproductive number*, denoted by R. If x is the propor-

tion of a randomly mixing, homogeneous population that is susceptible, R is the product of R_0 times the proportion x of the contacts that are with susceptibles:

$$R = R_0 x \qquad [27\text{-}3]$$

Example. Assume that $R_0 = 9$ for measles in a population and that one-half of the population is immune. Then, the effective reproductive number for measles is $R = 9 \times 0.5 = 4.5$. A case of measles would produce on average only 4.5 new secondary cases in this population.

Simple Insights from R_0

R_0 is a complex parameter that summarizes many of the important aspects of an infectious agent in a host population. It allows us to compare seemingly disparate diseases from the viewpoint of population biology and think about the effects of public health interventions. When a parasite is endemic and over time the average incidence does not change, an infectious case produces on average one new infectious case, and $R = 1$. To reduce transmission so that the parasite dies out, the average number of secondary cases produced by one infective case needs to be <1:

$$R < 1. \qquad [27\text{-}4]$$

Suppose that $R_0 = 5$ for HIV infection in a population. We would have to decrease the contact rate by a factor of five to turn the tide of an epidemic. If condoms reduced the transmission probability by 90%, then R_0 would be reduced to 0.5 if everybody used them. If before intervention, an average case of tuberculosis is infectious for 1 year and produces eight other cases, an intervention strategy emphasizing case-detection and treatment with antibiotics that reduces the period of infectiousness to 2 weeks would reduce R_0 to about 0.3.

If the fraction of susceptibles is low enough, the probability that an infective host comes in contact with a susceptible host before recovering will be very low. The parasite will not be able to persist. If immunization confers complete and lifelong immunity in all of the immunized individuals and a fraction f is immunized before the age of first infection, then $1 - f$ would be the maximum fraction of the population that is susceptible, disregarding immunity from previous disease. Substituting $1 - f$ for x in the formula for R in equation 27-3, theoretically it suffices to make

$$R = R_0(1 - f) < 1, \qquad [27\text{-}5]$$

to eliminate transmission. The fraction that needs to be immunized to eliminate transmission is

$$f > 1 - 1/R_0 . \qquad [27\text{-}6]$$

A higher R_0 requires immunization of a higher fraction to eliminate transmission.

Example. Assume that $R_0 = 9$ for measles in a population. Under the assumption of random mixing, the fraction that needs to be immunized before the age of first infection is $f = 1 - 1/R_0 = 1 - 1/9.0 = 0.89$. R_0 for smallpox before it was eradicated was estimated to be 4–5. For this R_0, the proportion that would need to be immunized before the age of

first infection would be $f = 1 - 1/R_0 = 1 - 1/5.0 = 0.80$. Based on these simple calculations, we would expect it to be harder to eliminate measles than smallpox. This is proving to be the case, although other factors such as heterogeneities in contact rates and susceptibility and measles vaccine failure play important roles.

The fraction of the population that must be vaccinated to make $R_0 < 1$ increases if the vaccine fails or provides only partial protection in some individuals. Suppose that the vaccine fails in some fraction $1 - h$ of the individuals who receive it, while the proportion h are completely protected. Thus, the fraction of the population protected by immunization is hf, and $R = R_0(1 - hf)$. The fraction of the population that needs to be immunized to eliminate transmission is then

$$f = \frac{1 - 1/R_0}{h} \qquad [27\text{-}7]$$

Example. Assume as in the above measles example that $R_0 = 9$. Suppose, however, that there is a failure somewhere along the cold chain required to keep vaccine viable from production to injection. Assume that the vaccine fails completely in 5% of the immunized people while conferring complete and long-lasting protection in the other fraction $h = 0.95$. The fraction f that must be vaccinated to eliminate transmission increases to

$$f = \frac{1 - 1/R_0}{h} = \frac{0.89}{0.95} = 0.94. \qquad [27\text{-}8]$$

If the vaccine fails in 15% of the vaccinated people, then the fraction that must be vaccinated is 0.89/0.85 > 1.0. With this vaccine at this failure rate, it would not be possible to eliminate transmission even if it were possible to vaccinate everyone.

The vaccine might not completely fail but confer only partial protection against infection, resulting in a reduction in the transmission probability to a susceptible from p to bp, where b is the relative susceptibility of a vaccinated susceptible person compared with an unvaccinated susceptible person. If a vaccinated person does becomes infected, the vaccine may still cause a reduction in the degree or duration of infectiousness. Let m be the ratio of the degree of infectiousness and ρ be the ratio of duration of infectiousness in a vaccinated infective compared with an unvaccinated infective. Consider a population in which everyone is immunized with such a partially protective vaccine that reduces infectiousness. The reproductive number R^c for the infectious agent in the presence of this complex vaccine is

$$R^c = c(bpm)(\rho d) = R_0(bm\rho). \qquad [27\text{-}9]$$

A vaccinated person is worth the fraction $bm\rho$, the *immunologically naive equivalent*, of an unvaccinated person from the point of view of transmission. If $bm\rho < 1/R_0$, then R^c will be <1. Thus, the vaccine has to be efficacious enough to reduce the value of an immunologically naive susceptible below $bm\rho = 1/R_0$ to prevent sustained transmission if the agent were introduced (Halloran et al., 1994b).

Example. Suppose that a measles vaccine reduces the transmission probability to the fraction $b = 0.05$ of its value in an unvaccinated person but leaves infectiousness unchanged, so that m = 1 and $\rho = 1$. The protective efficacy is 0.95. If $R_0 = 9.0$ and everyone is vaccinated, then $R^c = 0.05R_0 = 0.45$. If measles were introduced into a population vaccinated with this vaccine, it would not be expected to spread. If, however, $b = 0.12$, so that the protective efficacy is just 0.88, then $R^c > 1$, and we would expect to see an outbreak if the virus were introduced into the fully immunized population.

Herd immunity describes the collective immunologic status of a population of hosts, as opposed to an individual organism, with respect to a given parasite (Anderson and May, 1982). Herd immunity of a population may be high if many people have been immunized or have recovered from infection with immunity or may be low if most people are susceptible. The level of herd immunity has important effects on the transmission of infectious agents. As herd immunity increases, R will decrease. Fine (1993) provides a review of herd immunity.

Estimating R_0

Direct estimation of R_0 is not easy (Dietz, 1993). Two indirect methods can be used when the transmission system is assumed to be in dynamic equilibrium. The first method is based on the concept that when the average incidence rate and prevalence of disease are not changing, an infectious case produces on average one other infectious case, so $R = 1$. From the relation $R = R_0x = 1$, the proportion susceptible at equilibrium would be $x = 1/R_0$. Assuming random mixing, then R_0 is roughly estimated by the reciprocal of the proportion susceptible.

Example. Fine and Clarkson (1982) analyzed the data available for age-specific incidence and immunity levels for measles in England and Wales since 1950. They estimated that about 4–4.5 million individuals were susceptible, or about 9% of the population. May (1982) estimated that this corresponds to an R_0 for measles in those countries of about $1/0.09 = 11$, which is similar to estimates using other methods.

A second method was derived by Dietz (1975). In the simple case in which the incidence rate is assumed to be independent of age, the average age of infection, A, is equal to the reciprocal of the incidence rate, I:

$$A = \frac{1}{I}. \qquad [27\text{-}10]$$

Assuming stationarity, the incidence rate of infection can be estimated either from case reports or from cross-sectional, age-specific surveys of serologic prevalence in the population using the usual methods of estimating incidence from prevalence. If the average life expectancy, L, in a population is known, then R_0 can be estimated from the relation

$$R_0 \approx L/A \qquad [27\text{-}11]$$

If a population is growing substantially, such as in many developing countries, then the life-expectancy L should be replaced by the reciprocal of the per capita birth rate (Anderson and May, 1991).

Example. The higher the average age of infection, the lower the R_0 for any given life expectancy. The average age of measles in the United States in the 1950s was about 5–6, while for rubella it was about 9–10. From this, we would conclude that R_0 for measles was higher than for rubella.

The average age of infection is itself of interest because intervention programs can shift it. If many people are vaccinated, the incidence of infection will decrease, so that the average age of infection in the susceptibles will increase. Some diseases, such as mumps, chickenpox, and rubella, are more serious if acquired at older ages. Thus, the number of total cases could decrease due to a vaccination program at the same time that the number of serious cases would increase. These aspects of infectious disease interventions are studied using dynamic transmission models. Anderson and May (1991) present a com-

pendium of information on R_0 and dynamic transmission models; see also Halloran et al. (1989, 1994c).

Although R_0 is a conceptually useful measure that provides a summary of several aspects of an infectious disease, the simple relations described above usually do not hold. Heterogeneities in the contact rates, transmission probabilities, and duration of infectiousness produce different R_0 values in different subgroups. If individuals of a group of people who live near each other are not immunized, then it is possible for transmission to occur in that group, even when transmission has been eliminated in other segments of the population. The contact rate can increase locally if people move into crowded conditions, such as into college dormitories, military barracks, or refugee camps. Especially when transmission is tenuous or near elimination, heterogeneities can play an important role in determining whether a parasite can persist in a population.

Virulence, R_0, and the Case-Fatality Ratio

R_0 can also be used to study within-host dynamics such as the interaction of infectious agents with the immune system or to quantify evolutionary concepts. *Virulence* is a measure of the speed with which a parasite kills an infected host. Since R_0 is a function of the time spent in the infective state, R_0 could decrease as virulence increases. If the parasite is so highly virulent that it kills its host quickly, then R_0 could be < 1, and the parasite will die out. Viewed in this way, there is evolutionary pressure on parasites to become less virulent and to develop a more benign relation to the host. On the other hand, in some diseases, hosts become more infectious when they become sicker, so the transmission probability increases at the same time virulence increases. Thus, R_0 could increase as virulence increases, putting evolutionary pressure on the parasite to increase virulence. The balance depends on the particular parasite. The *case-fatality ratio* is the probability of dying from a disease before recovering or dying of something else. As virulence increases, the case-fatality ratio increases.

Example. Cholera is a disease that kills very quickly because the infected host becomes dehydrated within hours and dies. Virulence and the case-fatality ratio of untreated cholera are very high. A simple solution of salt and sugar in water given to a person sick with cholera will prevent death from dehydration, so that the person has time to develop immunity and recover from the disease. Use of this oral rehydration method has dramatically reduced the virulence and case-fatality ratio of cholera.

R_0 in Macroparasitic Diseases

The concept of R_0 comes from general population theory and refers to the expected number of reproducing offspring that one reproducing member of the population will produce in the absence of overcrowding. Although R_0 is defined for microparasitic diseases by the number of infective hosts, for larger parasites, called *macroparasites*, such as worms, R_0 is defined as in general population theory to be the expected number of mature female offspring that one female parasite will produce in her lifetime.

Example. The disease schistosomiasis is caused by large, sexually reproducing worms called schistosomes that can live up to 20 years within a human host. If a female schistosome worm has an $R_0 = 2$ in a population of human hosts and an intermediate host population of snails, then the average female schistosome produces two mature female worms from the thousands of eggs that it produces. Most of the eggs are thwarted on their obligatory passage through the environment and the intermediate snail hosts, before be-

ing able to establish themselves in another human host where they can grow into reproducing adults. The two new successful worms could be in one other human host, or in two different hosts. The $R_0 = 2$ refers to the number of worms, not to the number of hosts.

With macroparasites, we are often more interested in the total number of parasites in each host than in the mere prevalence of infection, because the parasite burden in a host can be more relevant than infection *per se* in determining morbidity. In macroparasitic diseases, some hosts can have very heavy infection, that is, many worms, while others have very light infection. This pattern is called *clumping* of infection in *wormy* people. Chemotherapy that targets wormy people could have a greater effect on transmission and morbidity than untargeted therapy. Clumping should be taken into account when designing interventions and their evaluation in helminths (Anderson and May, 1991).

INCIDENCE RATE AS A FUNCTION OF PREVALENCE AND CONTACT RATE

Besides the parameters specific to infectious diseases such as the transmission probability and R_0, the usual epidemiologic measures such as incidence rate, incidence proportion, and prevalence are also used in infectious disease epidemiology. Historically, terminology has differed somewhat in infectious disease epidemiology, with the terms *attack rate* being used for incidence proportion and *force of infection* for incidence or hazard rate. Although the same definitions and methods of estimation hold in infectious disease epidemiology for these usual epidemiologic measures, the dependence of events in infectious diseases results in additional intrinsic relations among the measures. Under the assumption of simple random mixing, constant contact rate c, and transmission probability p, the incidence rate $I(t)$ can be expressed as a function of the prevalence $P(t)$ at time t of infectious persons:

$$I(t) = cpP(t). \qquad [27\text{-}12]$$

This equation represents what Ross pointed out in 1916, namely, that the number of people becoming affected per unit time, the incidence, depends on the number already affected, the prevalence, as well as the contact rate and the transmission probability. The incidence proportion in a given time period depends on the incidence rate in that period; thus, it is also a function of the components of the transmission process.

Equation 27-12 can be used to estimate different quantities, depending on which components have been measured. The product of the contact rate and the transmission probability, cp, equals the more easily estimable ratio of the incidence rate to the prevalence of infectives, $I(t)/P(t)$. Thus, we do not need to observe the underlying contact process and transmission probabilities to obtain some information about their product cp. The transmission probability can be estimated if the other three components are measured.

Example. In malaria, the probability that a human host becomes infected from the bite of a mosquito containing infective stages of the parasite, that is, the transmission probability p, can be estimated from $p = I(t)/cP(t)$. $I(t)$ is the estimated incidence of new malaria infections, c is the estimated number of mosquito bites per person per unit time, and $P(t)$ is the proportion of captured mosquitos with infective stages of the malaria parasite in their salivary glands.

In reality, the incidence rate, contact rate, transmission probability, and prevalence may form a complex relation that is difficult, if not impossible, to measure. It is important in designing and analyzing studies in infectious diseases, however, to make any implicit assumptions about the relation explicit.

Estimation of the transmission probability as described in the preceding sections requires information on actual contacts between infectives and susceptibles, while the usual epidemiologic parameters such as incidence rate and incidence proportion do not. Halloran and Struchiner (1995) classified the parameters as *conditional* and *unconditional* parameters, depending on whether estimation is conditional on knowledge of such contacts or not. Thus, the transmission probability is a conditional parameter, while incidence rate and incidence proportion are unconditional parameters. The parameters transmission probability, incidence rate, and incidence proportion form a hierarchy requiring decreasing amounts of information about the transmission and contact processes (Rhodes et al., 1996).

Comparison of the basic reproductive number, $R_0 = cpd$, and the incidence rate, $I(t) = cpP(t)$, reveals the difference in points of view of the infective and susceptible hosts in infectious diseases. R_0 is the number of new cases that an infectious case is expected to produce, while the incidence rate $I(t)$ reflects the probability that a susceptible person will become infected in a short unit of time. Both quantities contain the product of the contact rate and the transmission probability, cp, the contact being the point at which the susceptibles and infectives meet and transmission being the fundamental event in infectious diseases. The unit composed of the susceptible, the infective, and the contact between them is the irreducible element in the study of transmission.

Contact Rates and Mixing Patterns

Contact patterns in a population play a central role in determining transmission and exposure to infection. There are different ways to think about how individuals in populations make contacts. One is that people behave like gas molecules with the rate of contacts being determined by density. If people were pressed more closely together, as in an urban environment, they would bump into each other more often than if they were less densely distributed, as in a rural environment. For diseases spread through casual contact, such as measles, mumps, or influenza, population density plays a role in determining the value of R_0. Alternatively, contact rates can be determined by choice, such as in sexual contacts or injection of intravenous drugs. In this case, R_0 is determined more by social choice. In many cases, both density and choice will play a role in determining contact rates and mixing patterns.

Regardless of how contacts arise, the simplest assumption about the contact pattern in a population is that of *random mixing*. Under this assumption, every person has an equal chance of making contact with each other person. Consequently, every person also has an equal chance of being exposed to infection because every person is equally likely to make contact with any infectious person. The assumption of equal exposure to infection of people in the comparison groups is important in many studies of interventions and risk factors affecting susceptibility, especially when based on unconditional parameters. As in the discussion above, we denote by c the constant contact rate that does not change over time in a randomly mixing population.

Most populations do not mix randomly but have groups that mix more with their own members than with other groups. The groups could be sexual behavior groups, different age groups within a school, or households in a community. The contact rate of individuals of group j with individuals of group i is denoted by c_{ij}. In a population composed of two mixing groups, group 1 and group 2 (Fig. 27-4), the contact pattern is described by a *mixing matrix* that has the same number of rows and columns as the number of mixing groups. The entries in the matrix represent the rate of contacts of individuals within and between the groups. The mixing pattern of two groups is represented by the matrix:

$$C = \begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \end{bmatrix} \qquad [27\text{-}13]$$

On the diagonals are the rates of contacts within groups, c_{11} and c_{22}. The off-diagonal entries, c_{12} and c_{21}, represent the rates of contacts between the groups corresponding to that row and column.

R_0 will be higher in the group with the higher contact rate, assuming that the transmission probability and duration of infectiousness are the same in both groups. If an epidemic occurs and there is contact between the two groups, the epidemic in the group with the higher contact rates will help drive the epidemic in the group with the lower rates. The group with the higher R_0 would then serve as a *core population* for transmission. A core population is a group with a high R_0, possibly due to a high contact rate, that interacts with a possibly much larger group with a low R_0. The interaction between the two groups helps spread the disease in the population with the lower R_0. The existence of a core group has consequences for intervention programs. It may be easy to reduce the average R_0 for the whole population below 1, while R_0 in the core population remains above 1, so that transmission will persist. In infectious diseases, the chain is only as weak as its strongest link.

Example. Hethcote and Yorke (1984) examined different strategies for reducing gonorrhea, taking into account professional sex workers who acted as a core group and contacts within the general population. They found that an intervention program generally needs to be targeted at the subpopulation with the higher R_0, in this case, the core population of sex workers, to have the greatest effect. Alternatively, a program could try to interrupt the contacts between the two groups.

Unfortunately, like much else in infectious diseases, these contact patterns are often difficult to determine and usually are not measured (Ghani et al., 1997). When conducting studies in infectious diseases where transmission plays a role, it is important to formulate explicitly the underlying assumptions that are being made with respect to contact patterns and exposure to infection. Since groups with different contact rates and mixing patterns could have different exposure to infection, consideration of the contact patterns could be

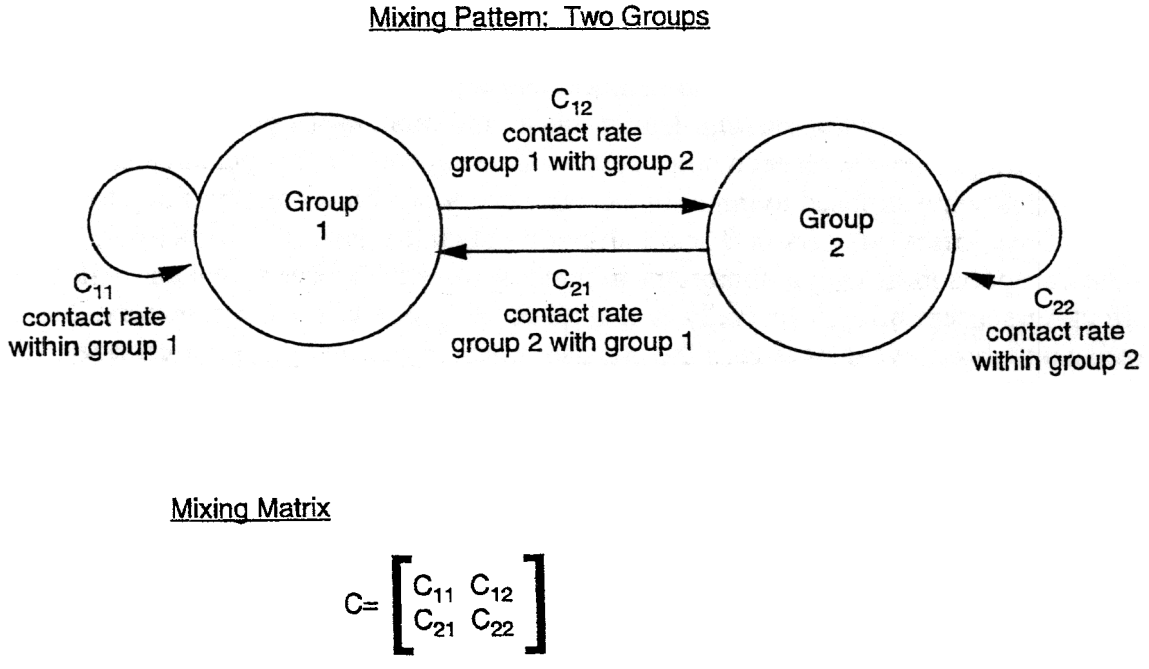


FIG. 27-4. Mixing patterns of two groups.

important for interpreting measures of effect. Failure to take into account unequal exposure to infection in the groups being compared can produce biased estimates of effect.

Dynamic Epidemic Process in a Closed Population

We tie the concepts of epidemiology, population biology, and transmission together with a simple example of an epidemic process in a closed population. Consider an infectious disease in which individuals can go through three states (Fig. 27-5a). They start out susceptible, X, then become infected and infectious, Y, after which they recover with immunity, Z. Models of this type of infection process are called *SIR* models for susceptible, infected, recovered. We opt for the other commonly used *XYZ* notation because I is used throughout this book for incidence rate. If these are the only three states possible, then each person in a population of N individuals is in one of these three states, where $X(t)$ is the number of susceptible people at time t, $Y(t)$ is the number of infectives, and $Z(t)$ is the number of immunes. This simple model ignores the latent and incubation periods and assumes that infection, disease, and infectiousness occur simultaneously. This model could be a simplified representation of measles, mumps, rubella, or chickenpox.

If the population is closed, then there are no births, immigration, deaths, or emigration. Therefore, a closed population is analogous to a closed cohort of people in an epidemiologic study. In a typical cohort study, we would not necessarily be concerned with how the individual people interact. In a study of an infectious disease, however, the underlying contact and transmission processes are important, so we need to think about these processes in our study. We consider a closed population of N initially susceptible people who are assumed to be mixing randomly with contact rate c. Thus, initially everyone in the closed population is in state X at time $t = 0$.

Suppose a parasite such as a measles virus is introduced into this population, so that one person enters the infectious state Y. If $R_0 > 1$, the epidemic is expected to spread. The process in a closed population is illustrated in the two top graphs in Fig. 27-6. The infec-

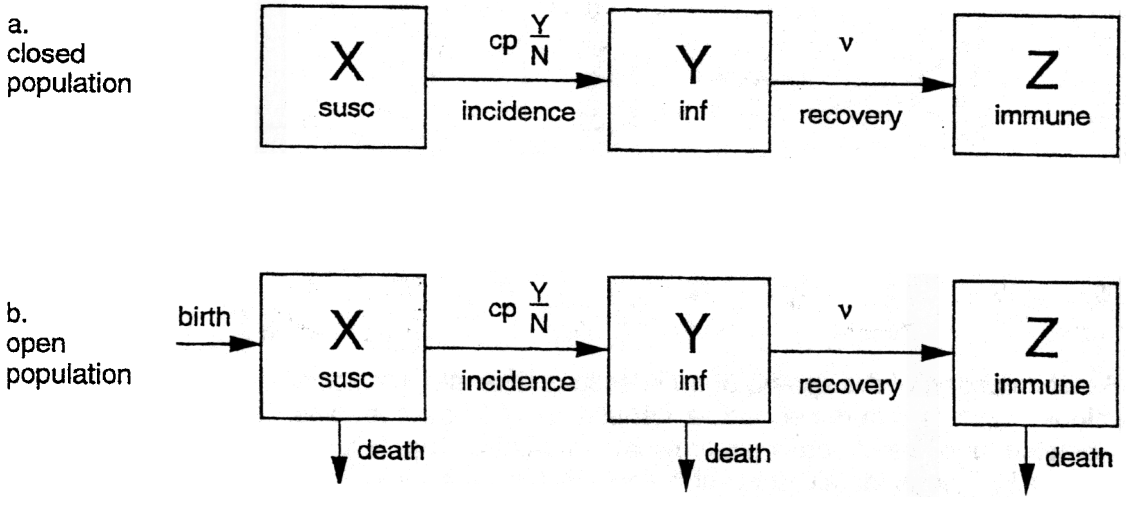


FIG. 27-5. Transmission model for an infectious disease in a host population. The three compartments represent susceptible (X), infective (Y), and immune (Z) hosts at time t. The total host population is of size $N = X + Y + Z$. Susceptible hosts become infected with an incidence rate (force of infection) of cpY/N, where c is the contact rate, p is the transmission probability, and Y/N is the prevalence of infective hosts at time t. The rate of recovery is v. (a) closed population; (b) open population. Arrows represent transitions in and out of compartments.

tion spreads from the first infective to the average number R_0 of susceptibles, depending on the rate of contact c, the transmission probability p, and how long the person is infectious d. If people recover at the rate ν, then they are infectious on average for the time period $d = 1/\nu$. The first infective eventually recovers with immunity into state Z, while the infection spreads from those people he or she infected to more susceptibles. The number of infectives Y initially increases. As the epidemic spreads, the number of susceptibles X decreases, while the number of people with immunity in Z begins to increase. Incidence and prevalence of infection will increase until the number of susceptibles available becomes a limiting factor. Then the number of new cases and prevalence of infectives begin to decrease until the parasite dies out and no people are left in the infective compartment Y. A parasite in a closed cohort where people recover with long-lasting immunity will inevitably die out, as in the top graphs in Fig. 27-6, because the key to parasite *persistence* in a host population is a continuous supply of susceptibles. The susceptibles can be produced either by births, immi-

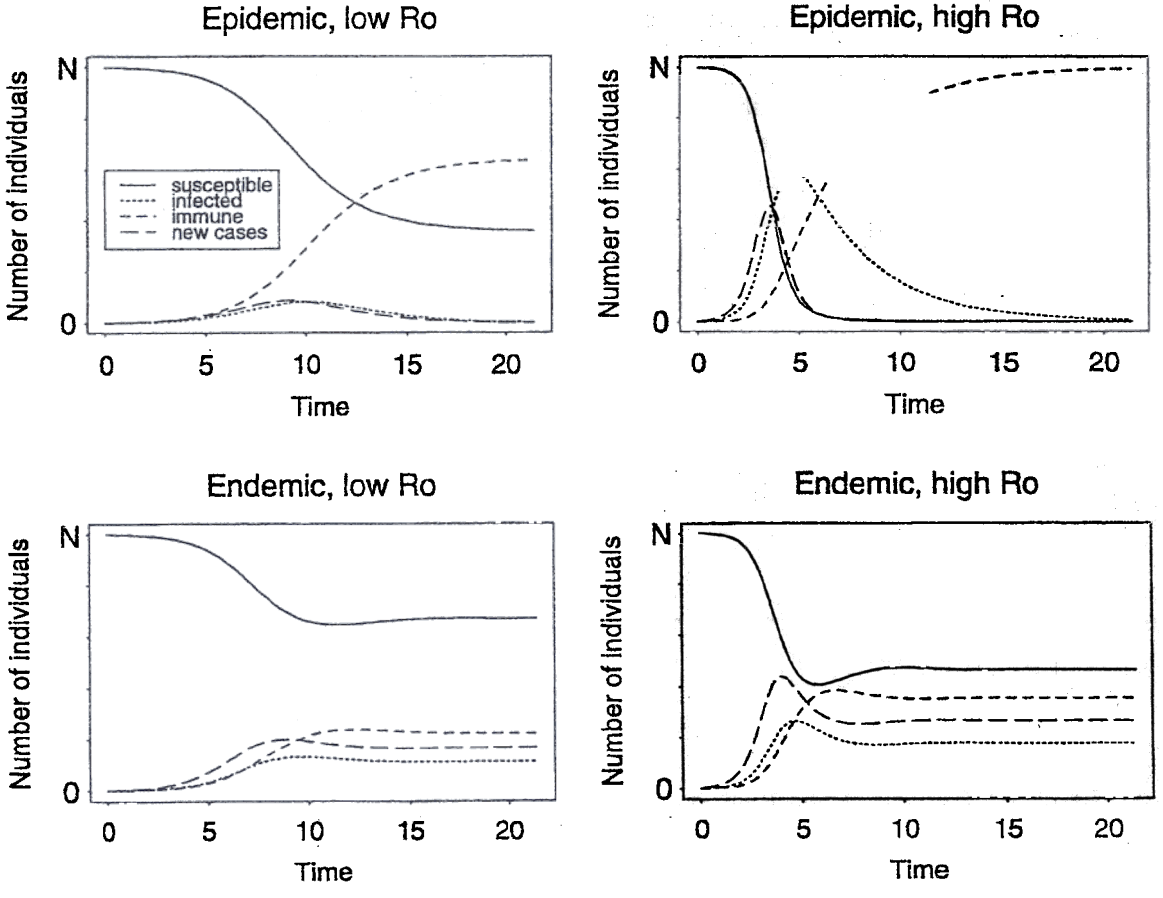


FIG. 27-6. Comparison of the spread of an infectious disease in a closed or open population. The infectious agent is introduced into a population of N susceptibles. Susceptible people become infected and infectious, then develop immunity. **Top left:** Epidemic in a closed population, low R_0. The epidemic dies out before all susceptibles become infected. **Top right:** Epidemic in a closed population, higher R_0. Everyone becomes infected during the epidemic. There are no infectives left as the epidemic dies out. **Bottom left:** Epidemic followed by endemic persistence in an open population, low R_0. The infectious agent does not die out due to the supply of new susceptibles. Prevalence of susceptibles, infectives, and immune people is in dynamic equilibrium. The number of new incident cases is steady. **Bottom right:** Epidemic followed by endemic persistence in an open population, high R_0.

gration into the population, recovery without immunity, or waning of immunity after it is acquired. In this example, however, no new susceptibles are produced.

The dynamics of the epidemic are described by three differential or difference equations that express the rate of change of the number of people in each of the three states. The rate at which people leave the susceptible compartment X and become infected is simply the incidence rate. Prevalence of infectives at time t, $P(t)$, is the number of infectious people $Y(t)$ divided by the size of the population N, or $Y(t)/N$. The formula for incidence as a function of prevalence in the epidemic is

$$I(t) = cpP(t) = cp\frac{Y(t)}{N}. \qquad [27\text{-}14]$$

The change in the number of susceptibles, the population-at-risk, $\Delta X(t)$, per small interval of time Δt at time t equals the incidence rate $I(t)$ times the size of the population-at-risk $X(t)$. The change in the number of infectives, $\Delta Y(t)$, is the difference between the number of new infections and the number of infectives developing immunity. The number of infectives developing immunity in that time interval is the change in the number of immunes $\Delta Z(t)$. The three difference equations for the epidemic model are then

$$\frac{\Delta X(t)}{\Delta t} = -I(t)X(t) = -cp\frac{Y(t)}{N}X(t),$$

$$\frac{\Delta Y(t)}{\Delta t} = cp\frac{Y(t)}{N}X(t) - \nu Y(t),$$

$$\frac{\Delta Z(t)}{\Delta t} = \nu Y(t).$$

We can associate aspects of the epidemic process with the usual epidemiologic measures. An estimate of the incidence rate $I(t)$ estimates $cpY(t)/N$. A cross-sectional study to estimate prevalence $P(t)$ of current infection would yield an estimate of $Y(t)/N$. The number of new infections in an interval of time estimates $[cpY(t)/N]X(t)\Delta t$, the incidence rate times the number at risk for the event times the time interval. The epidemic process of a disease producing long-lasting immunity in a closed population is always either increasing or decreasing. An important consequence for conducting studies in epidemics in closed populations is that there is no stationary state of the disease process. Thus, epidemiologic methods, study designs, or analytic methods that assume stationarity of the disease process are not applicable under epidemic conditions.

The epidemic process also depends on the population biology. Since R_0 is the product of the contact rate, the transmission probability, and the duration of infectiousness, in this model, $R_0 = cp/\nu$. The expected number of new cases per infective host decreases from R_0 to $R = R_0x$, where $x = X(t)/N$, the proportion still susceptible at time t. The epidemic peaks and begins to decrease when $R < 1$, so that $X(t)/N < 1/R_0$, that is, when the proportion of the population still susceptible becomes less than the reciprocal of the basic reproductive number. The greater R_0, the fewer susceptibles will be left when the epidemic peaks (compare the two top graphs in Fig. 27-6). Not all the susceptibles need to become infected before the parasite dies out. The higher R_0 is, the fewer susceptibles will be left at the end of the epidemic. Thus, the incidence proportion after an epidemic provides information on R_0. If an intervention reduced some aspect of R_0, then the intervention would result in the epidemic peaking when a higher proportion of the population was still susceptible, and fewer people would become infected before the epidemic died out.

Transmission in an Open Population and Dynamic Cohort

If the population is open so that birth or immigration and death or emigration can take place, then the susceptibles form a dynamic cohort with the population-at-risk changing over time (see Fig. 27-5b). This open population with the dynamic cohort at risk for infection is amenable to many of the study designs standardly used in dynamic cohorts. In the open population, if the replenishment of susceptibles is fast enough compared with the dynamics of the parasite, then the parasite does not necessarily die out, but can persist and become *endemic* (Fig. 27-6, lower two graphs). The parasite *invades* the population, *establishes* itself, and persists. In the lower two graphs of Fig. 27-6, the prevalence of infectives and number of new cases remain >0, indicating that the infection persists in the population. The prevalence of susceptible, infected, and immune people at equilibrium will depend on R_0 (compare the two lower graphs in Fig. 27-6).

When a disease is first introduced into a population, the dynamics will resemble an epidemic and are not stationary. As stated in the previous section, epidemiologic methods that assume stationarity of the disease process cannot be used during the epidemic phase. If the parasite has achieved a dynamic equilibrium, however, then some relations might be applicable. In choosing study designs and methods of analysis, we need to consider whether the dynamics of transmission are at equilibrium or are changing over time.

We can also define a fixed cohort of susceptibles within a dynamic population and follow them just as in a usual fixed cohort study. An important difference in infectious diseases, however, is that the disease process outside the cohort under study can affect incidence within the cohort, so it is important to think about how the fixed cohort interacts within itself and with the population at large. What we will observe if viewing only the study cohort of susceptibles is an epidemic within the cohort. If the contacts are predominantly with people outside the cohort, the prevalence of infection in the contacts will be similar to the population at large. This value may be changing rapidly over time if there is an epidemic in the larger population, or it may be fairly constant if prevalence is not changing rapidly. If the contacts are predominantly with other members of the initially susceptible study cohort, then initially there will be few infectious contacts, but the number of infectious contacts will increase as the epidemic within the fixed cohort spreads as in the epidemic process described above.

Recurrent Infections

In many infectious diseases, people can have recurrent infections. This possibility has important implications both for the dynamics of disease and for study design. Regarding dynamics, if people can have recurrent infections, then the pool of susceptibles can be replenished by the people recovering. In this case, the parasite might persist in closed populations. In the design and analysis of recurrent infections, methods allowing repeated outcomes in the same person need to be employed.

Example. In this example, the concepts of incidence, dynamics, and prevalence are tied together. Two investigators who have just conducted separate studies of gonorrhea in a heterosexual population of men and women come to different conclusions. The first investigator conducted a study in clinics using a sound sampling scheme with good ascertainment, found that the incidence rate of gonorrhea is much higher in men than women, and so concluded that gonorrhea is a greater problem in men than women. The second investigator conducted a population-based study that was also well designed, found that the

prevalence of gonorrhea is much higher in women than in men, and so concluded that the problem is greater in women. How do we resolve this paradox?

Assume that gonorrhea transmission has been fairly constant in this population and thus is at equilibrium. Women can be infected with gonorrhea for a long time before they develop symptoms, whereas men develop symptoms quickly and go for treatment. Thus, the duration of infectiousness in men is much shorter than in women. Generally, the transmission probability from females to males is lower than that from males to females; however, to make this point as simply as possible, we assume here that they are equal. Assume that the population has an equal number of men and women, that the rate of new partners (contact rate) is the same in both, and that men and women mix randomly with the opposite sex.

Prevalence of infection in the women is higher than in men partly because the duration is longer, so there are a greater number of susceptible men than women who are at risk of becoming new cases. The susceptible men make the same number of contacts and have the same transmission probability as the women, but their contact pool, the women, has a higher prevalence, so the incidence rate is higher in the men. The combined effect in the men of higher incidence rate and greater proportion susceptible results in a higher rate of new cases in men than in women. If we conducted a study in a clinic based on incidence rate or number of new cases, we would conclude that the problem was more serious in men.

If we conducted a prevalence study, we might think the problem mainly was in women. The important point is that they are related through the dynamic process, and understanding the relation resolves the paradox.

If we can reduce prevalence in the women, it will reduce the incidence rate and, consequently, prevalence in the men. This in turn will reduce the incidence rate in women and, consequently, the prevalence in women. The dependence of events in infectious diseases results in interventions having greater overall effects than would be expected from just the direct effects in the individuals receiving the intervention.

MEASURES OF EFFECT

The different kinds of effects in infectious diseases require more measures of effect than in noninfectious diseases. In addition to the usual effect measures of epidemiology, such as incidence rate ratio and incidence proportion ratio, the transmission probability ratio is an effect measure specific to infectious disease epidemiology. As described above, the former are unconditional effect measures, while the transmission probability ratio is a conditional effect measure because it conditions the contact between infective and susceptible. The choice of parameter and the choice of comparison populations in a study depend on whether we are interested in estimating changes in susceptibility or infectiousness, conditional or unconditional effects, or direct, indirect, or overall effects, as is discussed below (Halloran et al., 1997).

Transmission Probability Ratio

The *transmission probability ratio* (TPR) is a measure of the relative risk of transmission from infectives to susceptibles during a contact. For any given type of contact and infectious agent, we can estimate the effect of a covariate on susceptibility, infectiousness, or their combination by our choice of comparison pairs in the TPR. We can also estimate the TPR of differing types of contacts, infectious agents, routes of infection, or strains of an infectious agent.

$$\mathrm{VE}_S = 1 - \frac{\mathrm{SAR}_{10}}{\mathrm{SAR}_{00}} = 0.82,$$

$$\mathrm{VE}_I = 1 - \frac{\mathrm{SAR}_{01}}{\mathrm{SAR}_{00}} = 0.41,$$

$$\mathrm{VE}_T = 1 - \frac{\mathrm{SAR}_{11}}{\mathrm{SAR}_{00}} = 0.89.$$

The interpretation of these estimates is that the vaccine reduces the susceptibility by 82% and infectiousness by 41% and the combined reduction in susceptibility and infectiousness is 89%. Comparing these values with the discussion of the basic reproductive number R^c under partially protective vaccines, the ratio used to estimate VE_S estimates the parameter b, and, given the definition of household SAR, the ratio used to estimate VE_I estimates the parameter $m\rho$. The ratio used to estimate VE_T is closely related to $bm\rho$, the fraction of an immunologically naive equivalent that a vaccinated susceptible person contributes to the basic reproductive number.

Conditional Versus Unconditional Measures

The choice between conditional and unconditional measures of effect in designing studies is important. Although estimating conditional parameters such as transmission probabilities requires more information and is more difficult than estimating unconditional parameters such as incidence rates or incidence proportions, studies based on the transmission probabilities have some advantages. First, they are less easily biased by unmeasured, unequal exposure to infection than studies based on unconditional parameters. As early as 1915, Greenwood and Yule pointed out that to measure the effect of some risk factor on susceptibility to infection, we need equal exposure to infection in the comparison groups. When estimating the effect of a covariate on susceptibility using the transmission probability, we control for exposure to infection by taking into account the actual contacts with infectives. It also has a more obvious biologic interpretation. Second, relative infectiousness can only be estimated from the relative transmission probabilities and cannot be estimated from unconditional parameters. Thus, studies that do not gather information on contacts between infectives and susceptibles cannot in general be used to estimate the effects of covariates or interventions on infectiousness.

Example. Suppose we conduct a study of the efficacy of an HIV vaccine in which we collect information on contacts between infectives and susceptibles. Assume that the vaccine has no effect on infectiousness and that the transmission probability to an unvaccinated person is estimated to be p_0 and to a vaccinated person is estimated to be p_1. Then $\mathrm{VE}_p = 1 - p_1/p_0$. If we estimated $p_1 = 0.2p_0$, then we would estimate the efficacy of the vaccine in reducing susceptibility based on the transmission probabilities to be $\mathrm{VE}_p = 1 - 0.20 = 0.80$.

Alternatively, we could choose to use an unconditional parameter and estimate vaccine efficacy, VE_{1R}, from the incidence rate of the vaccinated, $I_1(t)$, compared with the unvaccinated group, $I_0(t)$. For this, we need only the time of each infection and the person-time at risk. Recall the formula for incidence as a function of the contact rate, transmission probability, and prevalence of infectives is $I(t) = cpP(t)$. If the study is randomized and blinded, then the contact rates and prevalence of infection in the contacts in the two groups might be assumed to be equal. Assuming furthermore that they are simple constants, then

Suppose that there are two types of infectives and susceptibles making a specified type of contact. Then there are four transmission probabilities corresponding to the possible combinations of pairs of infectives and susceptibles. For example, if we are studying the transmission probabilities in a population of vaccinated and unvaccinated people, the infectious person in a contact could be either vaccinated or unvaccinated, and the susceptible person could be either vaccinated or unvaccinated. If we denote being vaccinated by 1 and being unvaccinated by 0, we can distinguish four transmission probabilities: p_{00}, p_{10}, p_{01}, and p_{11}, where, for example, p_{01} denotes the transmission probability from a vaccinated infective to an unvaccinated susceptible. We estimate the effect of the vaccine in reducing susceptibility by the ratio of the transmission probability from unvaccinated infectives to vaccinated susceptibles, p_{10}, compared with the transmission probability from unvaccinated infectives to unvaccinated susceptibles, p_{00}. To estimate the effect of the vaccine on reducing infectiousness, we compare the transmission probabilities from vaccinated and unvaccinated infectives to the unvaccinated susceptibles, p_{01} and p_{00}, respectively. The combined effect of the vaccine on reducing susceptibility and infectiousness is estimated by comparing the transmission probability when both people in the contact are vaccinated, p_{11}, to the transmission probability when both people are unvaccinated, p_{00}. The three TPRs of interest are

$$\begin{aligned} \text{relative susceptibility: } \mathrm{TPR}_S &= \frac{p_{10}}{p_{00}}, \\ \text{relative infectiousness: } \mathrm{TPR}_I &= \frac{p_{01}}{p_{00}}, \\ \text{combined effect: } \mathrm{TPR}_T &= \frac{p_{11}}{p_{00}}. \end{aligned} \qquad [27\text{-}15]$$

Analogously, we could compare the transmission probability of tuberculosis in Caucasians (c) compared with African Americans (a) or between the two groups, with estimates of p_{cc}, p_{ca}, p_{ac}, and p_{aa}, by defining the appropriate pairs of transmission probability ratios. We can compare the male–male, male–female, female–male, and female–female probability of transmission of HIV from the ratios of estimates of p_{mm}, p_{mf}, p_{fm}, and p_{ff}.

The *excess* or *prevented transmission probability fraction in the exposed* is a causal parameter of interest specific to infectious diseases. If the TPR is >1, then the excess transmission probability fraction is TPR – 1. If the TPR is <1, then the prevented transmission probability fraction in one group compared with the other is 1 – TPR.

Example. Vaccine efficacy is usually estimated by 1 – RR, where RR is some measure of relative risk. It thus has the form of the prevented fraction in the exposed. When the vaccine efficacy estimate is based on the transmission probability, VE = 1 – TPR, it is an example of the prevented transmission probability fraction in the exposed.

Example. As described above, the household secondary attack rate (SAR) is an estimator of the transmission probability that is sometimes used to measure vaccine efficacy. Suppose in a study of the efficacy of a pertussis vaccine that the household SARs from an unvaccinated case to unvaccinated and vaccinated susceptibles were estimated to be $\widehat{\mathrm{SAR}}_{00} = 0.85$ and $\widehat{\mathrm{SAR}}_{10} = 0.15$, respectively, and from a vaccinated case to unvaccinated and vaccinated susceptibles were $\widehat{\mathrm{SAR}}_{01} = 0.50$ and $\widehat{\mathrm{SAR}}_{11} = 0.09$, respectively. Then the efficacy of the vaccine in reducing susceptibility, infectiousness, and combined effects on both is estimated by

$$VE_{IR} = 1 - \frac{I_1(t)}{I_0(t)} = 1 - \frac{cp_1P(t)}{cp_0P(t)} \cong 1 - \frac{p_1}{p_0} = 0.80.$$

In this case of equal exposure to infection, as well as simple constants, VE_{IR} approximately equals VE_p. This is not generally the case, even when exposure to infection is equal in the two groups because generally c, p, and $P(t)$ are not simple constants and therefore do not cancel. This simple case is used to illustrate the relation of the conditional to unconditional effect measures. The interpretation of the estimated effect would generally depend on the choice of parameter.

The unconditional effect measures can be easily biased by unequal exposure to infection. Continuing the example of the HIV vaccine trial, assume people know whether they are in the vaccinated or control group. Suppose that the people in the vaccine group believe themselves to be well protected, so they increase their contact rate and it becomes four times higher than in the unvaccinated group, that is, $c_1 = 4c_0$. You, however, do not know this because you did not include collection of information on contact rates, or at least change in contact rates, in your study design. The expected estimate of vaccine efficacy under this situation would be

$$VE_{IR} = 1 - \frac{I_1(t)}{I_0(t)} = 1 - \frac{c_1p_1P(t)}{c_0p_0P(t)} = 1 - \frac{4c_0(0.20p_0)}{c_0p_0} = 0.20.$$

The estimated efficacy of 0.20 would underestimate the actual effect of 0.80 of the vaccine in reducing susceptibility.

Suppose now that the contact rate is the same in the two groups but that the vaccinated group has become incautious about their choice of partners. Assume that the prevalence of infection in the partner pool of the vaccinated group is twice as high as the prevalence of infection in the partner pool of the unvaccinated group. Thus, $P_1(t) = 2P_0(t)$. The expected estimate of vaccine efficacy under this situation would be

$$VE_{IR} = 1 - \frac{I_1(t)}{I_0(t)} = - \frac{cp_1P_1(t)}{cp_0P_0(t)} = - \frac{(0.20p_0)2P_0(t)}{p_0P_0(t)} = 0.60$$

This again underestimates the effect of the vaccine on susceptibility.

The combined effect of the change in exposure to infection and the biologic protective effect of the vaccine is an important public health measure of interest, since an increase in exposure could outweigh the protective efficacy of the vaccine. In conducting studies, however, we generally want to differentiate covariate effects on susceptibility from covariate effects on exposure to infection. In the design and analysis of a study, it is therefore important to distinguish risk factors for susceptibility from risk factors for exposure to infection.

Improvement in estimates of covariate effects on susceptibility based on parameters not accounting for actual exposure can be achieved by stratifying according to some surrogate measure or risk factor for exposure to infection. For instance, children in a vaccine study may be stratified according to whether they attend day school or stay at home. To stratify by surrogates or risk factors for exposure is not the same as conditioning on actual contacts with infectives, however.

Case-control studies in infectious diseases need to satisfy the same assumptions as case-control studies in noninfectious diseases. The assumptions underlying many types of case-control studies may, however, be dramatically violated in studies of infectious diseases. Infectious diseases are often not rare, and stationarity assumptions commonly do not apply (Struchiner et al., 1990). Thus, the underlying assumptions should be examined closely for their applicability. Unequal exposure to infection could also be a practically important factor.

Example. In a case-control study of the bacille Calmette-Guérin (BCG) vaccine against leprosy by Muliyil et al. (1991), exposure to an infectious case within the household was much more strongly associated with the development of leprosy [odds ratio (OR) = 11.74; 95% confidence interval (CI): 3.97–34.71] than was BCG vaccination (OR = 0.80; 95% CI: 0.59–1.10). Unequal exposure to infection that was not taken into account could easily have biased the estimated effect of BCG vaccine.

A nonrandomized intervention could even exacerbate an imbalance in exposure to infection between vaccinated and unvaccinated groups, so that the bias in the estimate becomes worse due to indirect effects of intervention. For example, consider an observational study done in a population in which people living in low-transmission areas were those who tended to get vaccinated. In this case, the lower exposure to infection in the vaccinated group, if not taken into account, would produce an overestimate of vaccine efficacy. The bias in the vaccine efficacy estimate could be even further increased by the localized decrease in transmission produced by vaccination. Since this increased bias results from the indirect effects of the intervention due to the dependence of events in infectious diseases, this bias is called *dependent confounding*.

Exposure and Contact Efficacy

An intervention could alter the contact rates or contact pattern among persons receiving the intervention or in a population receiving the intervention program. *Contact rate efficacy* is the relative change in the contact rates due to an intervention program. *Exposure* or *behavior efficacy* is the relative increase or decrease in exposure to infection in the person receiving the intervention, or the relative change in the rate of infection or disease due to the change in exposure to the infectious agent, depending on the outcome measure chosen (Halloran et al., 1994a). The change in exposure to infection can occur in intervention studies either as the primary goal of the intervention or secondary to belief in the prophylactic efficacy of a measure. In the above example of an HIV vaccine study, the vaccine had the effect of increasing the rate of contacts in the vaccinated group by a factor of four.

STUDY DESIGNS FOR DEPENDENT HAPPENINGS

In the preceding discussion, we were interested in estimating effects of covariates or interventions on susceptibility, infectiousness, or some aspect of the contacts and contact process. These effects are defined for individuals or pairs of individuals. In addition, however, we might want to estimate the indirect, total, or overall effects of an intervention program in a population. Indirect effects are benefits or detriments from an intervention program in a population to individuals not directly receiving the intervention, compared to the situation in which the population had not had the intervention program. An example would be a reduction in incidence in unvaccinated people due to widespread vaccination in a population. Total effects are the combined direct effect in individuals actually receiving the intervention and the benefits due to the indirect effects of the intervention program as a whole. An example would be the reduction in incidence in vaccinated people due to widespread vaccination in a population. The overall effect of an intervention program is the effect on the population as a whole, including both those receiving and those not receiving the intervention. An example would be the overall reduction in incidence produced by a vaccination program. Some types of interventions, such as environmental interventions, have only indirect or overall effects, since individuals do not receive the intervention.

Although outcomes will still be measured in individuals, indirect, total, and overall effects are defined for the distribution of the intervention in the whole population, so the actual unit of interest is the population. Common to these effects is the need to imagine a population A that received the intervention program and another population B that did not receive the intervention program. The different kinds of effects are measured by choosing different subpopulations from A to compare with population B. For indirect effects, the subpopulation in A composed of individuals not receiving the intervention is compared with population B. To measure total effects, the subpopulation in A composed of individuals receiving the intervention is compared to population B. For the overall public health benefits, the entire population A is the population of interest, compared with B. The comparison of these different subgroups from population A to population B are designated study designs IIA, IIB, and III, respectively. Included here for completeness, study design I measures direct effects with unconditional parameters and compares people receiving the intervention with people not receiving the intervention within the same population A. Struchiner et al. (1990) and Halloran and Struchiner (1991, 1995) discuss the study designs for dependent happenings, including difficulties of causal inference. These study designs face the same problems as study designs in noninfectious diseases that use separate populations as comparisons, such as ecologic studies or studies using historical controls.

Owing to the indirect effects of infectious disease interventions, to measure the excess or prevented number of cases in the exposed, the comparison needs to be made between the incidence proportion in the vaccinated group (population A) and what the incidence proportion would have been in the unvaccinated group if no vaccination had taken place (population B). If the comparison is made between the number of cases in the vaccinated and the unvaccinated groups in the same population, as would be usual in noninfectious diseases, both groups will have experienced a reduction in incidence. Thus, the prevented fraction in the exposed (vaccinated) times the number of people exposed (vaccinated) will not yield a good estimate of the number of cases that were prevented by the vaccine. The comparison figure in population A, the number of cases in the unvaccinated group, does not represent the number of cases that would have occurred in the unvaccinated population had the vaccination program not taken place (population B). It is often not possible to know what the incidence proportion would have been in the absence of the intervention program. If estimates of the prevented fraction or number of cases prevented are made under these circumstances, it is important to note that they were calculated ignoring possible indirect effects.

SUMMARY

Because of the fundamental role of transmission of the infectious agent and dependent happenings, epidemiologic measures of interest in infectious disease epidemiology include the transmission probability, the contact rate, infectiousness, and the basic reproductive number (R_0), as well as direct and indirect effect measures. Measures such as the transmission probability that condition on contact between infectives and susceptibles are called *conditional* parameters, while those that do not, such as incidence rate and incidence proportion, are *unconditional* measures. The incidence rate is a function of the contact rate, the transmission probability, and the prevalence of infectives in the population. The dynamics of infection within a population need to be taken into account in the design and interpretation of studies. It is important to distinguish risk factors for susceptibility from risk factors for exposure to infection. Changes in contact rates and exposure efficacy are additional measures of interest.