6 Differential Equation Models for Infectious Disease

> To those familiar with the manifold complexities of real infections in real populations, our "basic models" may seem oversimplified to the point of lunacy.
>
> *Roy Anderson and Robert May (1992, p. 9)*

The aims of this chapter are the same as those stated by Anderson and May (1992) for their monograph *Infectious Diseases of Humans*: to show how relatively simple models can help to interpret data on infectious diseases, and to help design programs for controlling them. We begin with some classic examples, but end with some very recent developments: models for the emergence and management of drug-resistant disease strains, and models for disease progression within the body with particular reference to HIV/AIDS. The study of infectious diseases is one of the most mature applications of dynamic models in biology, so we can present some real and important "success stories" for simple dynamic models. We limit ourselves here to human diseases, but very similar models are also used for animal and plant disease dynamics (see, e.g., Hudson et al. 2002, Campbell and Madden 1990, and the papers discussed in the Preface), an issue of increasing importance as climate change and other anthropogenic stressors render natural populations increasingly susceptible to disease.

This chapter introduces no new mathematics. Rather, following our general approach, it serves to indicate the enormous scope of potential applications for differential equation models. Previous chapters have examined differential equation models at the within-cell (enzyme kinetics, gene regulation) and whole-cell (neuron excitation) levels. Now we go up to the level of human populations, but the model structures and the tools for their analysis—rescaling, eigenvalues, bifurcations, and so on—are the same.

6.1 Sir Ronald Ross and the Epidemic Curve

Sir Ronald Ross (1857–1932) received the 1902 Nobel Prize in medicine for determining the life cycle of the malaria parasite, in particular the role of mosquitos in the parasite life cycle and as vectors for its transmission between humans. From

that humble beginning he went on to found the modern application of dynamic models to the study of infectious diseases.

Ross (1916) gave two motivations for modeling epidemic dynamics. First, he noted that infectious diseases could display three different temporal patterns:

1. *Endemic*: relatively small fluctuations in monthly case counts, and only slow increase or decrease over the course of years (Ross listed leprosy and tuberculosis in this category)
2. *Outbreak*: constantly present but flaring up in epidemic outbreaks at frequent intervals (measles, malaria, dysentery)
3. *Epidemic*: Intense outbreaks followed by disappearance (plague, cholera)

Ross (1916, p. 205) asked "To what are these differences due? Why, indeed, should epidemics occur at all, and why should not all infectious diseases belong to the first group and remain at an almost flat rate?"

Ross's second motivation was to explain the characteristic shape of the *epidemic curve* for diseases in the third class. The epidemic curve is the time course of disease *incidence*, the number of new cases per unit time. Figures 6.1 and 6.2 show a few examples. The characteristic features are a symmetric or nearly symmetric rise and fall, with the outbreak terminating before all individuals susceptible to the disease have become infected. Because susceptibles still remain in the population when outbreaks terminate, it was argued by some at the time that outbreaks terminate because the pathogen loses infectivity; others hypothesized that the uninfected individuals must have been less susceptible to the disease.

Ross's (1916) model was a partial success, allowing him to show that the shape of epidemic curves could be explained without either of these hypotheses. His other goal, to explain different patterns of disease dynamics, was tackled a decade later by Kermack and McKendrick (1927). Current models are largely based on Kermack and McKendrick's modified versions of Ross's models, so we will consider those here. The models (SIR models) are formulated at the level of the available data: the numbers of individuals reported to contract the disease. Individuals are classified as being either *S*usceptible to the disease, *I*nfected by it, or *R*ecovered or *R*emoved. *R*-stage individuals are neither infectious nor infectable: either dead, or having immunity (permanent or temporary) against the disease.

The first of Kermack and McKendrick's basic models described a disease outbreak in a closed population of constant size:

$$\begin{aligned} dS/dt &= -\beta SI \\ dI/dt &= \beta SI - \gamma I \\ dR/dt &= \gamma I. \end{aligned} \qquad [6.1]$$

Initial conditions are $S(0) = S_0 \approx N$, $I(0) = N - S_0 \approx 0$, $R(0) = 0$ where N is the total population size. Since $dS/dt + dI/dt + dR/dt = 0$ the total population size

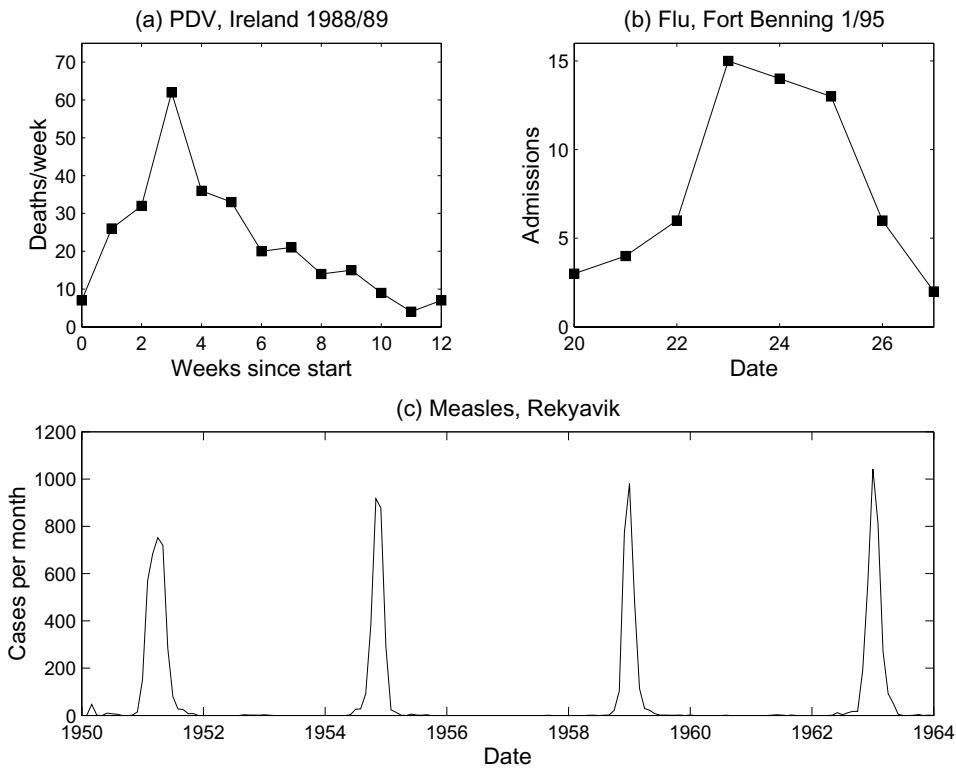

Figure 6.1 Examples of epidemic curves. (a) Phocine distemper virus in Northern Ireland 1988/89 (data from Figure 4 of Hall et al. 1992, provided by John Harwood). (b) An outbreak of influenza in Fort Benning, Georgia in 1995 (data from Davidson 1995). (c) Recurrent outbreaks of measles in Rekyavik, Iceland (data provided by Andrew Cliff, Department of Geography, University of Cambridge).

remains constant at N. The population is closed in the sense that no new susceptibles are added by births or immigration, and so long as R individuals are counted the population size is constant. Thus the assumption of constant population size is really that the only changes in population size are disease-induced deaths. The fraction of infected individuals, $I(t)/N$, is called the *prevalence* of the disease.

The first equation in model (6.1) is disease transmission resulting from contact between susceptibles and infectives. Ross (1916) justified this transmission rate as follows. Each infected individual transmits the pathogen to b individuals per unit time, but new cases arise only if the recipient individual is susceptible. Assuming a constant population of size N, the number of new cases per unit time is therefore $bI(S/N) = \beta SI$ where $\beta = b/N$. This form of transmission is called "mass action" (by analogy with the Law of Mass Action in chemical reactions) or "proportional mixing" (Anderson and May 1992). Mass action has been and still remains the

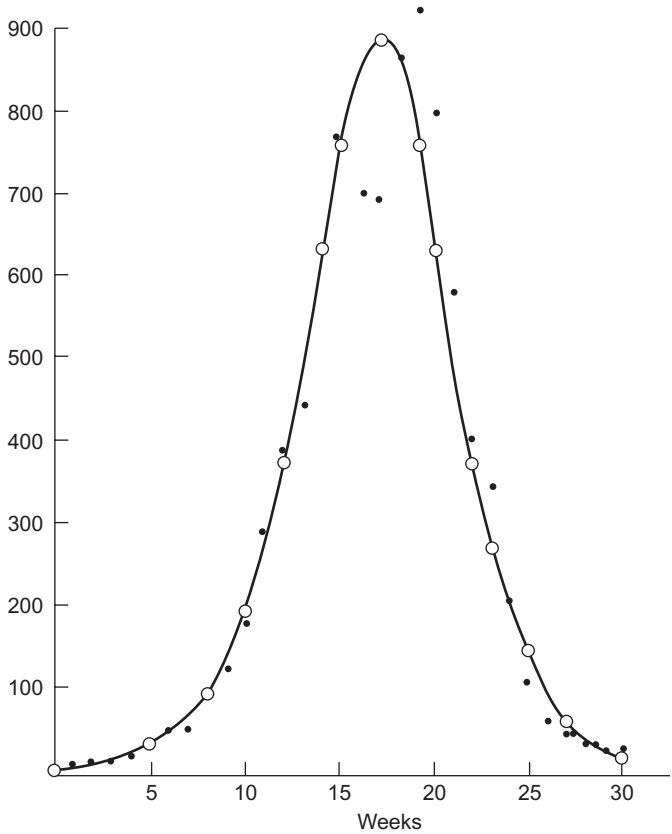

Figure 6.2 Deaths per week from plague in the island of Bombay from December 17, 1905 to July 21, 1906 (from Kermack and McKendrick 1927). The solid line is an approximate solution to their model for a disease with permanent removal—death or immunity—in the rat population on the island. It is compared with data on the human death toll on the assumption that "plague in man is a reflection of plague in rats."

most widely used transmission model; McCallum et al. (2002) review alternative models and empirical studies about the validity of the mass action model.

In the second equation, γ is the rate at which infected individuals recover from the disease (or die), at which point they transfer to the *R*ecovered class. The exit rate γ can be interpreted biologically as the inverse of the mean residence time in the compartment; this interpretation is very important for fitting these models to empirical data.

Whether *R*ecovered individuals are dead versus alive and immune is in one sense irrelevant for the future course of the epidemic, because in either case they have no impact on future infections. However, change in the number of living individuals invalidates Ross's (1916) derivation of the disease transmission rate.

If the living population is not constant, then to justify the βSI in [6.1] we have to assume that the rate of contacts per individual is proportional to the population size—if you double the number of people on the subway, then the kid with the runny nose infects twice as many people. If so, then the rate of new infections is $(bN)I(S/N) = \beta SI$ with $b = \beta$.

6.2 Rescaling the Model

What is the resulting shape of the epidemic curve? At first sight, it appears that we would need to see how the shape and behavior of solutions depend on three parameters: β, γ, and N. However, by rescaling the model (as in Chapter 4) we can reduce to a single parameter.

The benefit of rescaling is that the model becomes simpler just by changing the units of measurement for time and state variables. Usually the most effective rescalings are ones that render all variables in the rescaled model *dimensionless*. For example, $S, I, R,$ and N are all "population size," measured in units like individuals/km^2 or individuals/m^2. The numerical values of these variables depend on the choice of units. However, if we look at the ratios $X = S/N, Y = I/N, Z = R/N$, their values will be the same regardless of the units used for population size. X, Y, Z are called *dimensionless* variables because their numerical values do not depend on the units of measurement.

The dynamic equations for our rescaled variables are easily derived:

$$\begin{aligned} dX/dt = (1/N)dS/dt &= -\beta SI/N \\ &= -(\beta N)(S/N)(I/N) \\ &= -bXY, \end{aligned}$$

and similarly

$$dY/dt = bXY - \gamma Y$$

$$dZ/dt = \gamma Y.$$

This gets us down to two parameters: γ and the new composite parameter $b = \beta N$. We can get rid of one more parameter by defining a rescaled time variable $\tau = \gamma t$. Recall that the mean duration of infection is $1/\gamma$, so a unit increase in τ corresponds to real elapsed time equal to the mean duration of infection. We then have

$$\begin{aligned} dX/d\tau = dX/(\gamma dt) &= (1/\gamma)dX/dt \\ &= -(\beta N/\gamma)X. \end{aligned}$$

The step from $dX/d\tau$ to $(1/\gamma)dX/dt$ follows from the chain rule, $(dX/d\tau) \times (d\tau/dt) = dX/dt$, but the heuristic calculation in the last equation gets the right answer.

The conclusion is that $dx/d\tau$ depends on the single parameter combination $R_0 = \beta N/\gamma$. Doing the same with the other state variables we get the rescaled model

$$\begin{aligned} dX/d\tau &= -R_0XY \\ dY/d\tau &= R_0XY - Y \\ dZ/d\tau &= Y \end{aligned} \qquad [6.2]$$

with initial conditions $X(0) = X_0 \approx 1, Y(0) = Y_0 \approx 0, Z(0) = 0$.

An immediate prediction from this model is a threshold condition for an epidemic to occur. At time 0, $dY/d\tau = Y(R_0X_0 - 1) \approx Y(R_0 - 1)$, for a disease introduced at low incidence into the populations. Consequently, the disease prevalence increases if and only if $R_0 > 1$. Since $X(\tau)$ can only decrease over time, if Y is not increasing at time 0 it can never increase later, so the disease must die out.

The quantity

$$R_0 = \beta N/\gamma \qquad [6.3]$$

is called the "basic reproductive rate" of the disease, and can be interpreted as the expected number of new infections produced by a single infected individual introduced into a population of N susceptibles: βN infections per unit time, multiplied by the expected time $1/\gamma$ in the infectious stage. It therefore should be (and is) a very general property of epidemic models that a disease can be maintained in a population only if its R_0 (defined appropriately for the model) is greater than 1. Measures that reduced R_0 below 1 would then eradicate the disease, such as quarantine of infectives to reduce β or vaccination to reduce the number of susceptibles.

We can also show that epidemics in the model terminate before all susceptibles have become infected—thus achieving one of Ross's goals—and determine how many susceptibles remain.

In model [6.2] any individual who contracts the disease winds up eventually in Z. Since Z can only increase over time but can never go above 1, it must approach some limiting value $Z_\infty = \lim_{t\to\infty} Z(t)$, which is therefore the fraction of all individuals who contract the disease before it dies out. Z_∞ can be found by deriving a one-dimensional differential equation for $Z(t)$. By the chain rule $dX/d\tau = (dX/dZ)dZ/d\tau$ so

$$dX/dZ = (dX/d\tau)/(dZ/d\tau) = -R_0X;$$

hence $X(Z) = X(0)e^{-R_0Z}$. Using this expression for X and $Y = 1 - X - Z$, the third line of [6.2] becomes

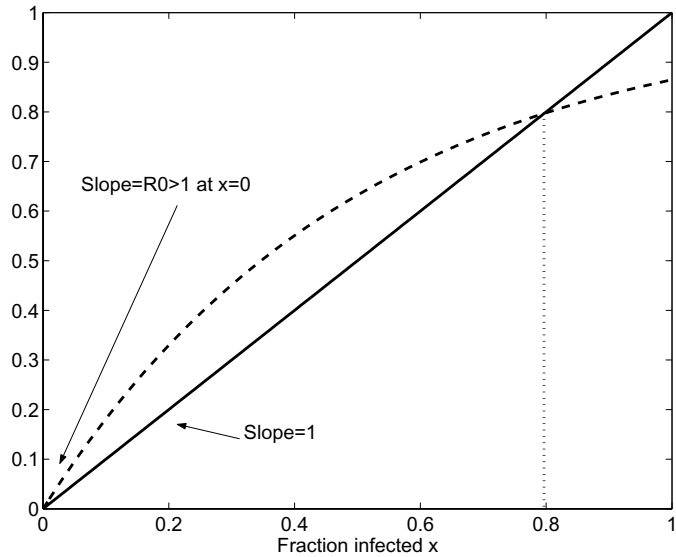

Figure 6.3 Graphical illustration that equation [6.5] has a unique solution between $x = 0$ and $x = 1$. The solution gives the approximate fraction of the population that contracted the disease over the course of an epidemic (Z_∞) in the Kermack-McKendrick SIR model, when the epidemic starts with a small number of infectives and the rest of the population susceptible.

$$dZ/d\tau = 1 - X(0)e^{-R_0 Z} - Z. \qquad [6.4]$$

As $Z(t) \to Z_\infty$, $dZ/d\tau$ decreases to 0, marking the end of the outbreak. When that occurs, since $X(0) \approx 1$ we must have (approximately)

$$Z_\infty = 1 - e^{-R_0 Z_\infty}.$$

Thus Z_∞ is the positive solution of the equation

$$x = 1 - e^{-R_0 x}. \qquad [6.5]$$

We see graphically that [6.5] has a unique solution between 0 and 1 so long as $R_0 > 1$ (see Figure 6.3), representing the fraction of the population that contract the disease before the outbreak collapses. The line $y = x$ has slope 1 and increases without limit. The curve $y = 1 - e^{-R_0 x}$ has slope R_0 at $x = 0$ but saturates to a limiting value of 1 as x increases. Thus the curves must intersect at some point Z_∞ between 0 and 1.

The relationship between Z_∞ and R_0 can be obtained by solving [6.5] for the inverse function, giving

$$R_0 = -\frac{1}{Z_\infty} \ln(1 - Z_\infty).$$

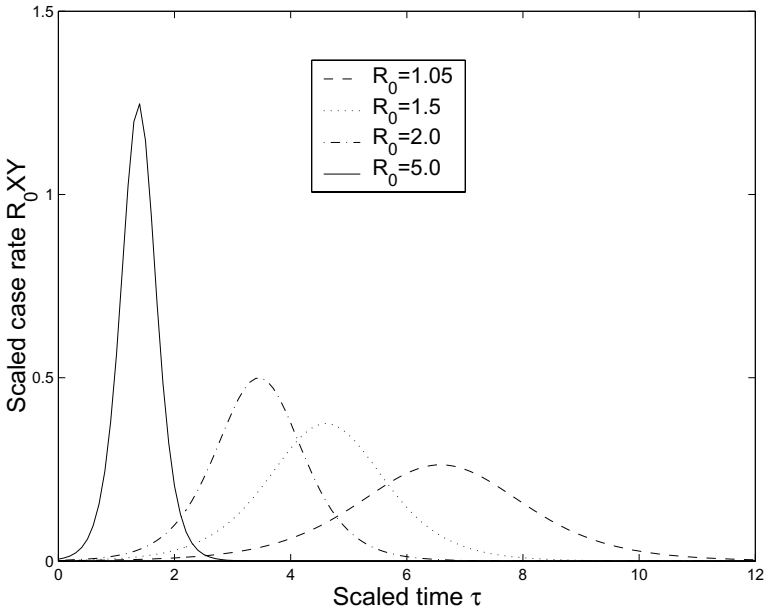

Figure 6.4 Epidemic curves (number of new cases per unit time) calculated from numerical solution of the Kermack-McKendrick model for a disease without recovery in a closed population of constant size. The plot shows $R_0X(\tau)Y(\tau)$ as a function of scaled time τ.

There is also an interesting approximation to [6.5] for R_0 near 1. In that case x is small, and we can use a two-term Taylor expansion $x \approx 1 - (1 - R_0x + (R_0x)^2/2)$ to obtain

$$Z_\infty \approx 2(R_0 - 1)/{R_0}^2 \approx 2(R_0 - 1) \text{ when } R_0 \approx 1.$$

This provides a possible explanation for occasional large outbreaks of a disease that is usually rare: small changes in infectivity as a result of conditions becoming more favorable for disease transmission can produce an epidemic affecting a significant fraction of the population. The approximation is actually an overestimate, but not by much. For example, the exact solution for $R_0 = 1.2$ is for 31% of the population to be infected before the epidemic burns out.

Finally, numerical solution can be used to find the shape of the epidemic curve predicted by the model [6.2], as a function of the single parameter combination R_0, for initial conditions $X(0) \approx 1$, $Y(0) \approx 0$, $Z(0) = 0$. The epidemic curve is defined as the rate at which new cases appear, that is, $R_0X(\tau)Y(\tau)$. Figure 6.4 shows that the model does indeed produce reasonable-looking epidemic curves. Higher values of R_0 naturally lead to a shorter and more intense epidemic, in addition to a higher final infected fraction Z_∞. Recall that τ measures elapsed time in units of the mean duration of infection, so the solutions show that if $R_0 = 5$ the epidemic

only lasts about twice as long as the duration of the infection, while if $R_0 = 1.05$ it last for over ten times the duration of infection.

Exercise 6.1. An isolated village in Iceland experiences an outbreak of influenza in which 812 of the 1100 residents contract the infection. Estimate R_0 assuming that the outbreak started with a single case contracted from outside the village, with all others susceptible at the start of the outbreak.

Exercise 6.2. Find Z_∞ for [6.2] when $R_0 = 10, 20$, and 100.

6.3 Endemic Diseases and Oscillations

We turn now to Ross's second goal, understanding differences in dynamic patterns of incidence among endemic diseases. Consider, for example, two childhood diseases in New York City (measles and chickenpox) prior to the availability of vaccine (Figure 6.5). Both diseases show a pronounced annual cycle, most likely reflecting the higher transmission among children when schools are in session. However, statistical analysis confirms the presence of roughly two-year and three-year periodicities in measles, while the only significant periodicity in chickenpox is the annual cycle. What accounts for this difference?

In order for a disease to persist indefinitely there must be a supply of fresh susceptibles, either through recovery without immunity or through births. The simplest example is an SIS model with constant population size:

$$\begin{aligned} dS/dt &= -\beta SI + \gamma I \\ dI/dt &= \beta SI - \gamma I. \end{aligned} \qquad [6.6]$$

The acronym "SIS" indicates that when infected individuals recover they return to the susceptible class: there is no immunity conferred by infection. Gonorrhea, which we consider below, is a disease of this type.

This model is simple enough to solve. First, we can rescale it in the same way as the SIR model, getting

$$\begin{aligned} dX/d\tau &= -R_0XY + Y \\ dY/d\tau &= R_0XY - Y. \end{aligned} \qquad [6.7]$$

Second, since the $X + Y = 1$ we can replace X by $1 - Y$ in dI/dt to obtain $dY/d\tau = R_0Y(1 - Y) - Y$. Then with a bit of algebra this re-arranges to

$$dY/d\tau = rY(1 - Y/K) \qquad [6.8]$$

where $r = R_0 - 1$, $K = (R_0 - 1)/R_0$. This is the well-known *logistic equation*. If $R_0 < 1$ the disease dies out. For $R_0 > 1$ the qualitative behavior of solutions is easy to determine by graphing $dY/d\tau$ as a function of Y: a parabola with its peak

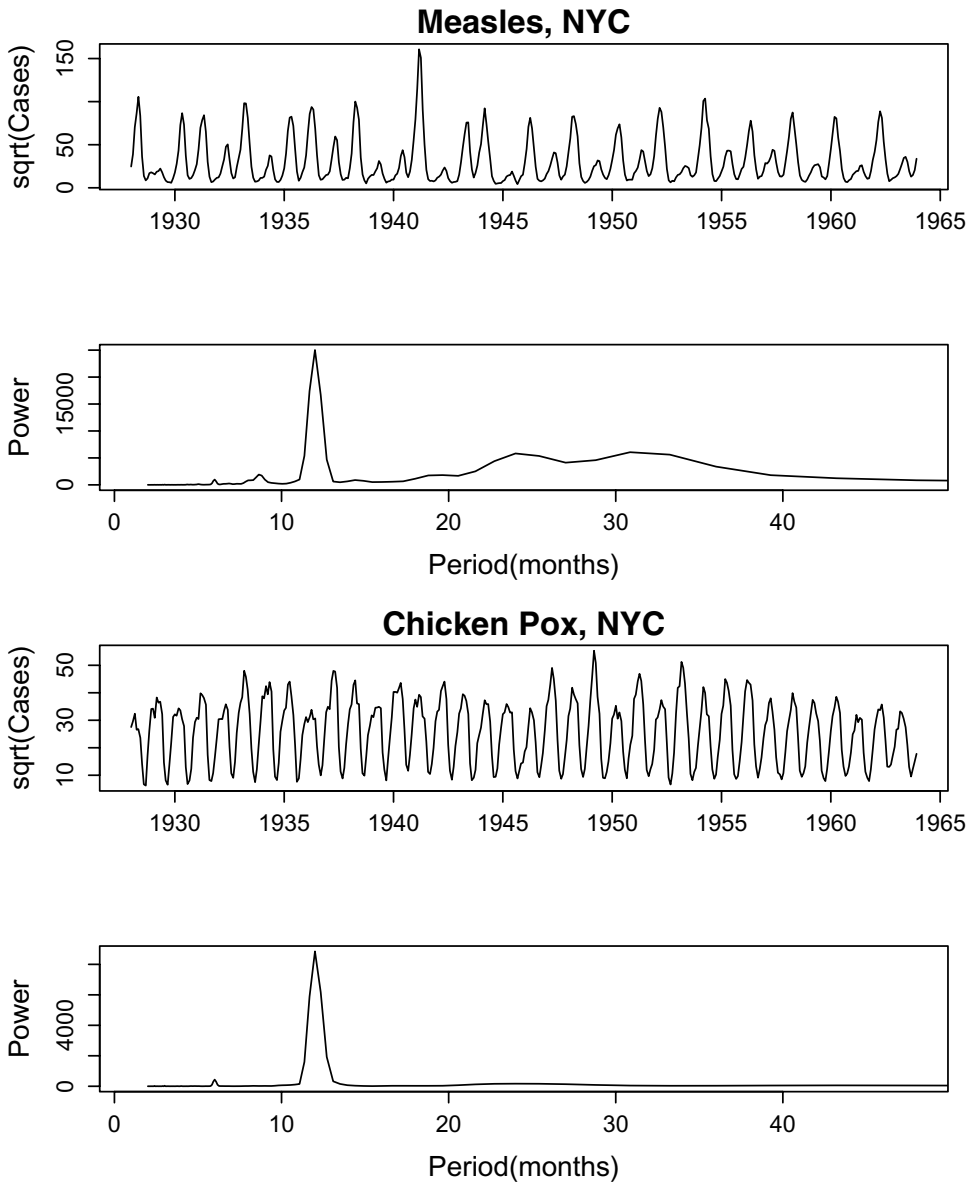

Figure 6.5 Monthly case report totals for measles and chickenpox in New York City prior to vaccination, plotted on square-root scale, and power spectra of the case report time series. The *power spectrum* of a time series represents the relative importance of different oscillation frequencies in the data. The spectra shown here confirm that chickenpox is dominated by a simple annual cycle, while measles shows a mix of annual, two-year, and three-year cycles.

at $Y = K/2$. From this we see that $Y(t) \to K$ from any initial value $Y(0) > 0$. We can also infer the qualitative shape of solutions. When $x(t)$ is below $K/2$, as x increases, dx/dt also increases: the second derivative is positive, hence $x(t)$ is concave up. Between $K/2$ and K, further increases in x leads to a decrease in dx/dt so $x(t)$ is concave down.

Because it reduces to a one-variable model, [6.7] always has a monotonic approach to steady state and cannot account for persistent oscillations in endemic diseases. Oscillations require a model in which the population passes through more disease states. The simplest example is an SIR model in which constant population size is maintained by a balance between births and deaths:

$$\begin{aligned} dS/dt &= \mu N - \beta SI - \mu S \\ dI/dt &= \beta SI - (\gamma + \mu)I \\ dR/dt &= \gamma I - \mu R. \end{aligned} \qquad [6.9]$$

Because population size is constant we only need the first two equations in (6.9). Following Anderson and May (1992) we rescale the variables to $X = S/N$ and $L = \beta I$. X is the fraction of susceptible individuals in the population, and L is the "force of infection"—the probability per unit time of becoming infected, for a susceptible individual. Because we are interested in the period of oscillations, we do not rescale time. The rescaled model is then

$$\begin{aligned} dX/dt &= \mu(1 - X) - LX \\ dL/dt &= (\gamma + \mu)L(R_0X - 1) \end{aligned} \qquad [6.10]$$

where $R_0 = \beta N/(\gamma + \mu)$. R_0 has the same meaning as before: the mean number of new cases produced by a single newly infected individual added to a population of N susceptibles.

Exercise 6.3. What is the qualitative behavior of solutions to [6.8] starting from $x_0 > K$?

6.3.1 Analysis of the SIR Model with Births

The dynamics predicted by the SIR model with births are derived from linear stability analysis of steady states $(\bar{X}, \bar{L})$. There is always a disease-free steady state $(\bar{X}_0, \bar{L}_0) = (1, 0)$. An endemic steady state $(\bar{L}_1 > 0)$ requires $R_0\bar{X} = 1$ (from setting $dL/dt = 0$); hence

$$\bar{X}_1 = 1/R_0. \qquad [6.11]$$

Then setting $dX/dt = 0$ we find

$$\bar{L}_1 = \mu(1 - \bar{X}_1)/\bar{X}_1 = \mu(R_0 - 1). \qquad [6.12]$$

Since $X < 1$ holds at all times (why?), the endemic steady state is biologically meaningful only when $R_0 > 1$, as expected from the meaning of R_0.

The endemic susceptible fraction $\bar{X}_1 = 1/R_0$ is a very general prediction, and will be true whenever an infective's rate of disease transmission is proportional to the susceptible fraction. By definition, a single newly infected individual dropped into a population with $X = 1$ (all susceptible) directly produces R_0 new infections,

on average. At an endemic steady state, each newly infected individual must be exactly replacing itself, that is, producing 1 new infection rather than R_0—hence the susceptible fraction must be $1/R_0$.

The Jacobian matrix for [6.10] is

$$J(X,Y) = \begin{bmatrix} -\mu - L & -X \\ (\gamma+\mu)\,LR_0 & (\gamma+\mu)\,(R_0X-1) \end{bmatrix}. \qquad [6.13]$$

For the disease-free steady state we have

$$J(1,0) = \begin{bmatrix} -\mu & -1 \\ 0 & (\gamma+\mu)\,(R_0-1) \end{bmatrix} \qquad [6.14]$$

with eigenvalues $-\mu$ and $(\gamma+\mu)(R_0-1)$ (using the fact that the eigenvalues of a triangular matrix are the diagonal entries in the matrix). The disease-free steady state is therefore stable if $R_0 < 1$ and an unstable saddle if $R_0 > 1$. The stable eigenvector in the latter case is $(1,0)$—the X axis, which is also the stable manifold. If the population initially consists entirely of susceptible and recovered individuals, it converges to the disease-free steady state as recovered individuals die and are replaced by susceptibles.

The Jacobian for the endemic steady state is

$$J(\bar{X}_1,\bar{Y}_1) = \begin{bmatrix} -\mu\,R_0 & -{R_0}^{-1} \\ (\gamma+\mu)\,\mu\;(R_0-1)\,R_0 & 0 \end{bmatrix}. \qquad [6.15]$$

Since $R_0 > 1$ is necessary for this steady state to exist, the determinant is positive and the trace is negative, implying that the steady state is stable whenever there is an endemic steady state. To see if it is a spiral or a node, we compute the eigenvalues using the formula for a 2×2 matrix from Chapter 2:

$$\lambda = \frac{T \pm \sqrt{T^2 - 4D}}{2}$$

where $T = \text{trace}(J)$ and $D = \det(J)$. Applying this to [6.15] we get eigenvalues

$$\lambda_{1,2} = -\frac{1}{2}\mu R_0 \pm \frac{1}{2}\sqrt{\mu^2 {R_0}^2 - 4\;(\gamma+\mu)\,\mu\;(R_0-1)}. \qquad [6.16]$$

This expression can be understood using two approximations. First, if R_0 is just slightly above 1, the (R_0-1) term within the square root will be dominated by $\mu^2 R_0^2$ so both eigenvalues are real, implying that the steady state is a node. However, endemic diseases typically have R_0 well above 1; for that situation recall that μ is the mortality rate and γ is the rate of recovery from the disease. To put it another way: $\mu = 1/(\text{mean lifetime})$ and $\gamma = 1/(\text{mean duration of the disease})$ so we typically have $\mu \ll \gamma$. Consequently, in the square root in [6.16] the terms involving μ^2 are dominated by the term involving μ, namely, $-4\gamma\mu(R_0-1) < 0$. Dropping the μ^2 terms we obtain the approximate eigenvalues

$$\lambda_{1,2} \approx -\frac{1}{2}\mu R_0 \pm i\sqrt{\gamma\mu(R_0-1)}. \qquad [6.17]$$

These are complex conjugates, implying that the endemic steady state becomes a spiral when R_0 is large, and the approach to steady state will be oscillatory.

The eigenvalues also give us the (approximate) period of the decaying oscillations. With complex conjugate eigenvalues $\lambda = a \pm ib$, the solutions to the linearized equations are a linear combination of $e^{at}\cos(bt)$ and $e^{at}\sin(bt)$, thus having period $T = 2\pi/b$. For [6.17] we therefore have the period

$$T \approx 2\pi/\sqrt{\gamma\mu(R_0-1)} = 2\pi\frac{1}{\sqrt{\gamma}}\frac{1}{\sqrt{\mu(R_0-1)}}. \qquad [6.18]$$

$1/\gamma$ is the mean duration of the disease. $\mu(R_0-1)$ is the force of infection at the endemic steady state, the probability per unit time of moving from susceptible to infectious. Its inverse is therefore the mean time an individual spends in the susceptible class, which is the mean age at infection in this model since all individuals are born as susceptibles. Calling these δ and α, respectively, we therefore have

$$T \approx 2\pi\sqrt{\alpha\delta}. \qquad [6.19]$$

For comparisons with data we need to estimate the model parameters. Some are easy to come by. N is the total population size; for example, for New York in the prevaccination era we could take 5 million as a representative value. $1/\gamma$ is the mean disease duration. For measles this is estimated at 12 to 14 days, so with time measured in years we could estimate $\gamma \approx 365/13 \approx 28$.

Estimating β is harder. The transmission rate involves the frequency of contacts, and the fraction of contacts that actually lead to disease transmission. Neither of these is easy to observe or estimate directly. In addition, different kinds of contacts (social, in school, within the family, on public transportation, etc.) each occur at different rates and with a different chance per contact of disease being transmitted.

An alternative approach is *calibration*, which means adjusting parameters to make model solutions correspond to the data as well as possible. With calibration, we use the model and data on state variables to infer the value of model parameters. *This assumes that the model is valid*, a dangerous assumption because we are unlikely to have independent tests of the model's validity. Nonetheless calibration is very widely used because it is rare to have a complex biological model in which all parameters can be estimated from direct data on process rates.

Here, we can use calibration to estimate β as follows. Assuming a disease at steady state with constant force of infection, we derived above that $\mu(R_0-1) = 1/\alpha$ where α is the mean age at infection. Rearranging, we get

$$\hat{R}_0 = 1 + \frac{1}{\alpha\mu} = 1 + \frac{\text{mean lifespan}}{\text{mean age at infection}} \qquad [6.20]$$

and then the resulting estimate $\hat{\beta} = \hat{R}_0(\gamma + \mu)/N$. The mean age at infection can be inferred from age-specific case reports, assuming a roughly even age distribution (e.g., Grenfell and Anderson 1985); for measles in England and Wales 1948–1968, this gave $\alpha \approx 5$ years.

Anderson and May (1992) suggested that the oscillations about the endemic steady state in the SIR model with births could account for the observed dynamics in measles and other endemic childhood diseases. Although cycles in the model are damped, Anderson and May (1992) argued that a number of mechanisms would continually perturb the system, leading to sustained oscillations at periods similar to [6.18]. Two suggested mechanisms were finite-population effects—the "coin-tossing" nature of disease transmission, especially when the number of infecteds is low—and seasonal variation in transmission rate due to school vacations. This hypothesis was tested by comparing predicted periods against periodicities observed in the data for a number of diseases (Anderson and May 1992, Table 6.1), and in many cases the fit is good. For example, for measles in developed countries they estimated mean age at infection of 4–5 years, and disease duration of 12 days, giving approximate period $2\pi\sqrt{4.5(12/365)} = 2.4$ years. This compares well to the observed two- and three-year periodicities (Figure 6.5). However, for chickenpox they estimate mean age at infection of 6–8 years, and disease duration 18–23 days, giving predicted periods of 3.4–4.5 years, for which there is no evidence in the data. So the damped-cycles hypothesis is only part of the story.

There is an enormous and still growing literature about the processes underlying dynamic patterns in endemic diseases. For childhood diseases it appears that multiple factors are involved, including those raised by Anderson and May (1992): seasonal variation in transmission rates and demographic stochasticity. Neither of these alone is sufficient. Models without seasonal variation in transmission cannot reproduce the clear annual periodicity observed in virtually all childhood diseases (Schaffer et al. 1990; see Figure 6.6). Deterministic models with seasonal forcing require unrealistically high levels of seasonal variation in order to mimic, through deterministic chaos, the complex multiannual patterns seen in measles (Ellner et al. 1995). But models incorporating both finite-population effects and plausible levels of seasonal variation have been able to account for the main features of the pre-vaccination oscillations in measles (Ellner et al. 1998; Finkenstad and Grenfell 2000; Grenfell et al. 2002), and for effects of vaccination on spatiotemporal patterns in measles and pertussis (Rohani et al. 1999). In addition, variation in birth rates and population size may contribute to changes over longer time scales, such as the transition in the New York measles data from more complex dynamics to a regular biennial oscillation (Ellner et al. 1998; Earn et al. 2000).

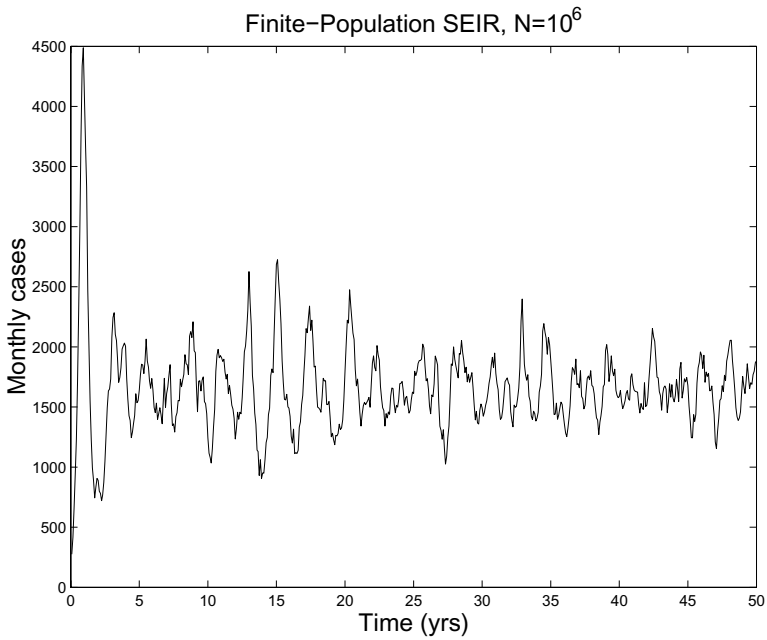

Figure 6.6 Output from a finite-population SEIR model with parameters appropriate for pre-vaccination dynamics of measles in a city of 1 million, but without any seasonal variation in the transmission rate.

6.3.2 Summing Up

We can now return to the biological questions posed at the start of this chapter: what accounts for differing patterns of dynamics in infectious diseases? We have seen that differences in epidemic dynamics emerge from biological differences in the interactions between different pathogens and their host that are reflected in the basic structure of the appropriate model:

- A highly infectious disease with permanent removal or immunity ($S \to I \to R$ model without births, [6.1]), leads to a classic epidemic curve and terminates before all susceptibles are infective.
- A disease where immunity following infection is temporary ($S \rightleftarrows I$ model, [6.6]) leads to a stable endemic state.
- A less infectious disease with permanent removal or immunity ($S \to I \to R$ model with birth and death, [6.9]) can lead to an endemic state with oscillations.

That is not to say that every endemic disease is adequately described by one of these simple models. The essential message is that qualitative properties at the level of individual hosts and pathogens create qualitative differences at the whole-

population level, and the connection between these is made by the dynamic models.

Exercise 6.4. What happens in [6.16] if R_0 becomes really, really large? Is this a realistic possibility—that is, just how large must R_0 be to change our conclusions about the steady state? [Hint: if R_0 is "really, really large" then $(R_0 - 1)/R_0 \approx 1$.]

Exercise 6.5. Starting with the constant-population SIR model with births, equation [6.9], suppose that newborns are vaccinated, with the result that a fraction $p \leq 1$ of all newborn individuals are born as removed rather than susceptible (i.e., we consider here a disease where removed individuals are alive but immune).

(a) Write down the resulting system of differential equations. Note that the population size should still be constant, and that when $p = 0$ your model should reduce to equation [6.9].

(b) Show that as p is increased from 0, the number of infectives at the endemic steady state decreases until it eventually reaches $\bar{I} = 0$ at some value $p < 1$. This is sometimes called *herd immunity*: even though some individuals have not been immunized by vaccination, the disease cannot sustain itself in the population as a whole.

Exercise 6.6. Starting again from [6.9]:

(a) Modify the model to include *vertical transmission*, meaning that offspring of an infected parent have probability v of being born as infected rather than susceptible.

(b) Use linear stability analysis of the trivial steady state ($S = N, I = R = 0$) to study how vertical transmission affects the conditions for persistence of the disease.

Exercise 6.7. Starting once again from [6.9]:

(a) Add a latent phase (E) of individuals who have been infected but are not yet infective—they carry the disease but do not transmit it to others—and write out the resulting system of differential equations. This is called the SEIR model.

(b) Derive the expression for R_0 in this model.

(c) Write a script file to solve this model numerically, for a disease with $R_0 = 15$ (as estimated for measles) in a population of 3,000,000 and with the latent and infectious stages each lasting one week. Do numerical experiments on this model, with seasonal variation in contact rate, to test the claim that the latent proportion $E(t)/(E(t) + I(t))$ remains roughly constant. Describe your

- *methods:* describe the simulations that you conducted;
- *results:* give a verbal summary of the results;
- *conclusions:* Was the claim valid?

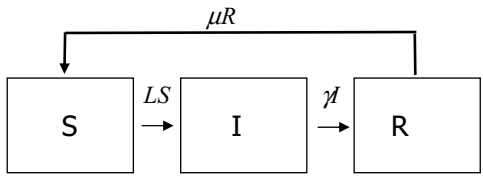

Figure 6.7 Compartment diagram for discrete-event SIR model.

Include a few well-chosen and well-designed graphs to support your claims. [Note: seasonal variation in contact rate is often modeled by $\beta(t) = \bar{\beta}(1 + \phi\cos(2\pi t))$ with time measured in years. Values of ϕ between 0.1 and 0.3 are considered credible for measles, depending on who you ask. Routines for numerically solving differential equations may have trouble for larger values of ϕ, especially if you start with $S(0) \approx N$ so that there is a massive initial outbreak.]

Exercise 6.8. This exercise involves constructing a discrete-event SIR model; an introduction to discrete-event models and how to simulate them is given in the online Computer Lab materials. Figure 6.7 shows the compartment diagram for an SIR model in a population of constant size N. Here L is the force of infection given by $\beta(t)(I + I_0)$, $\beta(t) = \bar{\beta}(1 + \phi\cos(2\pi t))$, where I_0 represents an external pool of infectives (cousins in Connecticut, etc.) whose presence keeps the disease from ever dying out completely. Constant population size is maintained by having each death (from R) balanced by birth of a new susceptible. The corresponding transition matrix for an individual with time step $dt = 1$ day is

$$A = \begin{bmatrix} 1-L & 0 & \mu \\ L & 1-\gamma & 0 \\ 0 & \gamma & 1-\mu \end{bmatrix}. \qquad [6.21]$$

(a) Write a script file to simulate a discrete-event version of the model for 50 years with time step $dt = 1$ day, and produce a plot of the daily number of new cases, using parameter values (for time measured in days) $N = 1,000,000$, $\mu = 0.015/365$, $R_0 = 16$, $\gamma = 1/12$, $I_0 = 1$, $\phi = 0.01$.

(b) Can your model explain the qualitative difference between measles and chickenpox dynamics prior to vaccination—that is, the presence of persistent multiannual periodicities in measles but not in chickenpox? Assume $R_0 = 16$ for measles, 10 for chickenpox, and the disease durations are 12 days for measles, 21 days for chickenpox. For purposes of this exercise accept the current view that $\phi < 0.25$, with smaller values being more plausible. Nothing is really known about I_0. Turn in a write-up explaining how you obtained your answer to this question, with well-chosen graphs or tables to support

your arguments. Some case-report data are available from this book's web page, but remember that only a fraction of cases are reported.

Exercise 6.9. Modify the model (as a second script file) to eliminate the randomness due to finite population size. That is, any Binomial(N, p) random variable is replaced by its expected value Np, making the model a deterministic difference equation with a time-step of 1 day. Can this model explain the difference between measles and chickenpox?

6.4 Gonorrhea Dynamics and Control

In the 1970s Herbert Hethcote and James Yorke used epidemic models to study the control of gonorrhea in the United States. Their work illustrates how qualitative insights derived from simple models can have important practical implications. Bringing models and data into contact also led to an important conceptual advance that now plays an important role in HIV/AIDS research and public health policy, and also has been applied to the spread of computer "viruses." This section is largely based on their monograph (Hethcote and Yorke 1984).

The bacterium causing gonorrhea, *Neisseria gonorrhoeae*, lives only on mucus membranes and dies within seconds outside the human body. In one sexual exposure an infected woman has a 20–30% chance of transmitting the disease, while an infected man has a 50–70% chance. Gonorrhea incidence in the United States tripled between 1965 and 1975, and by the early 1980s there were roughly 1 million cases reported per year, implying roughly 2 million actual cases based on estimated reporting rates. Gonorrhea is a public health concern mainly because of its consequences in women: it is a major cause of pelvic inflammatory disease, infertility, and ectopic pregnancy, and facilitates the transmission of HIV. No vaccine against gonorrhea is available.

6.4.1 A Simple Model and a Paradox

A simple model for gonorrhea can be based on three basic properties. First, infection does not confer immunity. Second, the latent period can be omitted because it is very short (1–2 days) compared to the average duration of infection (about 1 month). Third, because there is only weak seasonal variation in the case reports, it is not necessary to include seasonal variation in the contact rate. The simplest model is therefore an SIS model with constant population size, which reduces to a single equation for the number of infectives,

$$dI/dt = \beta I(N - I) - \gamma I. \qquad [6.22]$$

As noted above, this is a logistic model whose solutions converge to the steady state $I = N(1 - 1/R_0)$ whenever $R_0 > 1$.

The paradox comes from the fact that, as usual, the endemic fraction of susceptibles in model [6.22] is $1/R_0$. The endemic fraction of susceptible therefore provides an estimate of R_0. Hethcote and Yorke (1984) reasoned as follows:

> We estimate that the actual yearly incidence of gonorrhea in the United States is 2.0 million and that the population at risk is approximately 20 million. If the average duration of infection is one month, then the number of cases at any given time is 166,667 which is less [than] 1% of the active population.

So $1 < R_0 < 1.01$, which portrays gonorrhea as a disease on the brink of extinction. That does not square well with its long-term persistence and its threefold increase in incidence within a decade. In addition, a rough estimate of the actual value of R_0 was obtained from the effects of a screening program in 1973–1975 (Yorke, Hethcote, and Nold 1978). It was estimated that the program decreased the average infectious period (and thus R_0) by 10%, and resulted in a 20% reduction in the rate of new case reports. The fraction infectious at any given time was therefore reduced by a fraction $(0.9)(0.8) = 0.72$. Thus, $1 - 1/(0.9R_0) \approx 0.72(1 - 1/R_0)$, giving $R_0 \approx 1.4$. But if R_0 is this large and most of the population is susceptible, the disease should be very rapidly increasing—tripling within a year rather than a decade.

6.4.2 The Core Group

How can we reconcile this contradiction? In order for the disease to be at steady state or slowly growing, the fraction of susceptibles among individuals contacted by an infected individual must be close to $1/R_0$. If we accept that $R_0 \approx 1.4$, then $1/R_0 \approx 0.7$ so the disease incidence must be about 30% *among individuals contacted by an infective*.

Since disease incidence in the general population is much lower than 30%, mixing between susceptibles and infectives cannot be random. Instead, Hethcote and Yorke concluded, there must be a *core group* of individuals, mostly transmitting the disease to each other, in which the disease is at much higher incidence than in the general population.

The importance of the core group had an immediate impact on programs for gonorrhea control (e.g., St. John and Curran 1978, WHO 1978, quoted by Hethcote and Yorke 1984). At that time, the main control measure in the United States was mass screening of women at public health clinics in order to identify asymptomatic carriers, who were considered to be the main reservoir for the disease. The presence of a core group implied that control programs should target the core group rather than the general population. The question, then, is how to do that most effectively.

The simplest model that could be used to examine control strategies is an SIS model that distinguishes between core (group = 1) and noncore (group = 2) sub-

populations:

$$\boxed{S_1} \rightleftarrows \boxed{I_1}$$
$$\boxed{S_2} \rightleftarrows \boxed{I_2}\,. \qquad [6.23]$$

This model ignores a good bit of reality—for example, it does not distinguish men from women, or symptomatic from asymptomatic infectives—and it is simplistic to posit a sharp division between core and noncore. However, it is the natural first step.

Assuming constant size for the core and noncore subpopulations, [6.23] reduces to a pair of differential equations for the number of infecteds in each group:

$$dI_1/dt = (\lambda_{11}I_1 + \lambda_{12}I_2)X_1 - \gamma I_1$$
$$dI_2/dt = (\lambda_{21}I_1 + \lambda_{22}I_2)X_2 - \gamma I_2. \qquad [6.24]$$

Here λ_{ij} is the number of effective (pathogen-transmitting) contacts per unit time of a group-j individual with persons in group i, and $X_i = S_i/N_i$ is the fraction of susceptible individuals in group i. This is Ross's mass action model, where each infected individual has a constant number of contacts, and the rate of new infections is limited by the fraction of contactees already infected.

To use the model we have to specify its parameters, including the as-always unobservable contact rate parameters λ_{ij}. To reduce the number of parameters, Hethcote and Yorke made the so-called *proportionate mixing* assumption, that frequencies of contact between individuals are proportional to their *activity levels* a_i, defined as the average number of effective contacts per unit time for an individual in group i. The fractional activity level of group i is then $b_i = a_iN_i/A$, where $A = a_1N_1 + a_2N_2$. Note that the b's only depend on the relative population sizes and activity levels:

$$b_i = a_iN_i \Big/ \sum_j a_jN_j = 1 \Big/ \sum_j (a_j/a_i)(N_j/N_i). \qquad [6.25]$$

The proportionate mixing assumption is that each individual (of either group) has encounters with core versus noncore individuals in proportions $b_1{:}b_2$.

Proportionate mixing simplifies things considerably—we then have

$$dY_i/dt = a_i\left(\sum_j b_jY_j\right)X_i - \gamma Y_i. \qquad [6.26]$$

That is, the rate at which group i susceptibles contract the disease is given by the product of their activity level a_i—their rate of effective contacts—and the fraction of contacts who are infected, $\sum_j b_jY_j$.

The model is thus specified by the activity levels a_i, the relative sizes of the two groups, and the mean disease duration. Some idea of the core group's size and parameters can also be determined, if the core is identified as individuals

having repeated infections in a relatively short time period. For example, one study cited by Hethcote and Yorke (1984) found that 6.7% of 7347 patients at venereal disease clinics fit this description, and were responsible for over 22% of the cases seen at the clinic.

The importance of the core group results from the fact that the product $b_1 Y_1$ appears in the transmission rate for all groups. If the core is especially active ($a_1 \gg a_2$) and consequently has a much higher disease prevalence ($Y_1 \gg Y_2$), the core group can be the primary source of new cases even if they are a small fraction of the population.

Exercise 6.10. This exercise illustrates the potential importance of a small but active core group. Suppose $N_1 = N_2/50$ and $a_1 = 10a_2$, with the overall contact rate such that the steady-state prevalence in the noncore population is 3%.

(a) By solving [6.26] for the endemic steady state, show that steady-state prevalences $\bar{Y}_i$ scale with activity level according to

$$\frac{\bar{Y}_i}{1-\bar{Y}_i} = a_i \bar{Y}/\gamma$$

where $\bar{Y} = \sum_j b_j \bar{Y}_j$ is the average fraction of infected contacts.

(b) Use the result of (a) to show that $\bar{Y}_1 = 10\bar{Y}_2/(1+9\bar{Y}_2)$.

(c) Show that the fractional activity levels are $b_1 = 1/6, b_2 = 5/6$, and hence $5/6$ of all contacts are with noncore individuals.

(d) Contacts with infecteds in core versus noncore occur in the ratio $(b_1\bar{Y}_1):(b_2\bar{Y}_2)$. Combine the results of (c) and (d) to show that this ratio is > 1; hence most infections are contracted from a core group member.

6.4.3 Implications for Control

The two-group model can now be used to evaluate alternative control measures by adding each to the model, and comparing their effects on disease incidence. For example, a strategy of randomly screening individuals in the at-risk population, and treating those infected, would be modeled as

$$dY_i/dt = a_i \left(\sum_j b_j Y_j \right) X_i - \gamma Y_i - gY_i \qquad [6.27]$$

where g is the screening rate. Hethcote and Yorke (1984) developed similar models for other strategies:

- *Rescreening:* Treated individuals are rescreened a short period after recovery from the disease, and retreated if necessary.

- *Contact tracing:* When a case is treated, try to identify the persons to whom the patient may have given the disease (potential infectees) or from whom they may have gotten it (potential infectors), and treat any who are infected.

Which of these works best? To find out, Hethcote and Yorke (1984) went through the exercise of estimating parameters for these models insofar as possible, and considering a range of possibilities for unknown parameters. They then did the same for a far more complex model with twelve state variables: susceptible, symptomatic infected, and asymptomatic infected among men and women in the core and noncore groups. But those efforts were hardly necessary, because the results can be inferred from how successful each strategy is at finding core group members:

- General random screening finds core versus noncore individuals in proportion to the group size $N_1{:}N_2$, and so does a bad job of treating the core.
- Rescreening finds core versus noncore individuals in proportion to their disease incidence, $I_1{:}I_2 = N_1Y_1{:}N_2Y_2$.
- Potential infectee tracing finds core versus noncore individuals in proportion to their activity levels, $b_1{:}b_2 = N_1a_1{:}N_2a_2$.
- Potential infector tracing finds core versus noncore individuals in proportion to the rate at which they transmit the disease to susceptibles, $b_1Y_1{:}b_2Y_2$.

(The difference between infectee and infector tracing is that potential infectees may or may not have been infected, whereas Hethcote and Yorke assumed that an infected individual will know who gave them the disease.) Because $Y_1 > Y_2$ and $a_1 > a_2$ we see that rescreening and infectee tracing will both outperform general screening, and (by similar comparisons) infector tracing will outperform both of these. Hethcote and Yorke's (1984) simulations confirmed these conclusions: infector tracing was found to be the most effective by far, for all parameter sets considered.

The concept of a core group and the importance of targeting the core for treatment was a major factor in the 1975 revision of U.S. control measures for gonorrhea. The new programs emphasized contact tracing and rescreening of people identified as core group members on the basis of frequent reinfection. The result was an immediate and long-lasting reversal of the increase in gonorrhea incidence (Figure 6.8), indicating that the new measures were substantially more effective.

Later models have relaxed the assumption of a sharp distinction between core and noncore, in two ways. The first is to allow multiple groups. Anderson and May (1992, Chapter 11) review models for HIV with an arbitrary number of groups, where group-j individuals constitute a fraction $P(j)$ of the population and have effective contacts at rate $a_j = aj$. They derived the important formula

$$R_0 = aT(E[j] + \mathrm{Var}(j)/E[j]). \qquad [6.28]$$

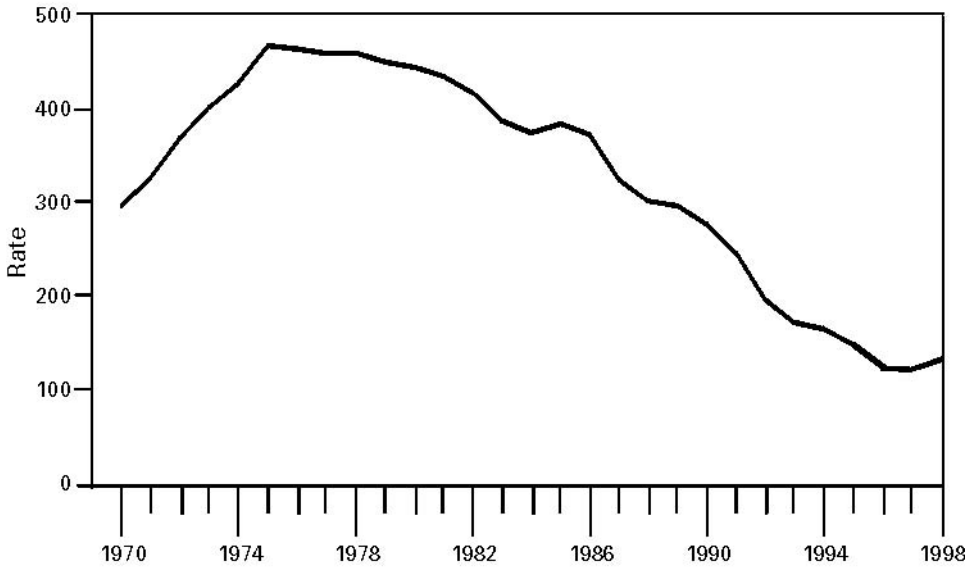

Figure 6.8 Reported gonorrhea rate (cases per 100,000 population) in the United States (from CDC 2000)

Here T is the mean duration of infection, and $E[j]$ and $\mathrm{Var}(j)$ are the mean and variance of the frequency distribution of contact rate: $E[j] = \sum_j jP(j)$, $\mathrm{Var}(j) = \sum_j P(j)\left(j - E[j]\right)^2$. As always the disease persists if $R_0 > 1$, so [6.28] implies that individuals in the upper tail of the activity level distribution have a disproportionate effect on disease persistence.

The second generalization is modeling epidemics on social networks, where each individual is explicitly linked to a finite number of other individuals, and an infected individual has a constant probability per unit time of infecting individuals to whom it is linked. This is a very active research area now; Newman (2003) and Moreno et al. (2002) are good introductions. As in multiple-group models, persistence of the disease can be strongly influenced by the fact that individuals differ greatly in their rate of contacts with others in the population.

Exercise 6.11. We said above that rescreening finds core and noncore individuals in proportion to disease incidence in each group. Why is this? Similarly, give a verbal justification for the expressions given above for the proportions of core versus noncore individuals found by the infector tracing and infectee tracing strategies.

Exercise 6.12. What happens if gonorrhea-infected individuals can't always identify correctly the person who infected them? Propose a model for this situation and discuss its implications.

Exercise 6.13. Consider a power-law distribution for the activity rate in the multigroup HIV model, $P(j) = Cj^{-(2+\gamma)}$, which has been suggested by some recent investigations. Show that for $0 < \gamma \leq 1$, R_0 is infinite. What would this imply about HIV persistence? How does it relate to the idea of targeting the core group for treatment?

Exercise 6.14. Explain why the model in the previous exercise is unrealistic for a finite population, and suggest how it might be modified. Does this change the model's implications for disease control strategies, and if so, how?

6.5 Drug Resistance

Antimicrobial and antiviral drugs have drastically reduced the impact of infectious diseases on humans in developed countries, but their effectiveness is being challenged by the emergence of drug-resistant strains. For example, between 1991 and 1996 the rate of penicillin resistance increased by more than 300%; the rate of cefotaxime resistance in *Streptococcus pneumoniae*, one of the main causes of ear infections in children, increased by more than 1000% (Butler et al. 1996). Within an individual receiving drug treatment, the drug creates an environment where drug-resistant strains are at an advantage. As a result, if treatment fails—for example, if the patient does not comply with the treatment regime or does not complete it—that individual may then be carrying and transmitting drug-resistant strains of the disease. The U.S. Centers for Disease Control and Prevention has identified antibiotic resistance as a significant public health problem and has initiated a National Campaign for Appropriate Antibiotic Use aimed at reducing the spread of antibiotic resistance.

Two current concerns are multidrug resistant (MDR) tuberculosis, and antiretroviral resistant HIV. Tuberculosis (TB) is a major global public health burden, with over 9 million cases per year and $\sim 25\%$ mortality; untreated or drug-resistant cases would have $\sim 50\%$ mortality rate (Dye et al. 2002). TB strains are classified as MDR if they are resistant to the two main first-line drugs used to treat TB, isoniazid and rifampicin. In 2000 only 3.2% of all new TB cases were estimated to be MDR, but MDR prevalence of 10–14% has been estimated for specific regions in Eastern Europe, Asia, and Africa (Dye et al. 2002). The incidence of TB is currently increasing by about 2%/year, and there is concern that this might reflect the emergence of MDR strains that could lead to a significant global rise in TB prevalence.

Combination antiretroviral (ARV) therapies for HIV, involving simultaneous treatment with three or more different drugs, are currently the most effective available treatment. In use since 1996, these have substantially decreased the death rate from AIDS (Blower et al. 2001). However, strains resistant to the three-drug "cocktail" have emerged and have been sexually transmitted. Blower et al.

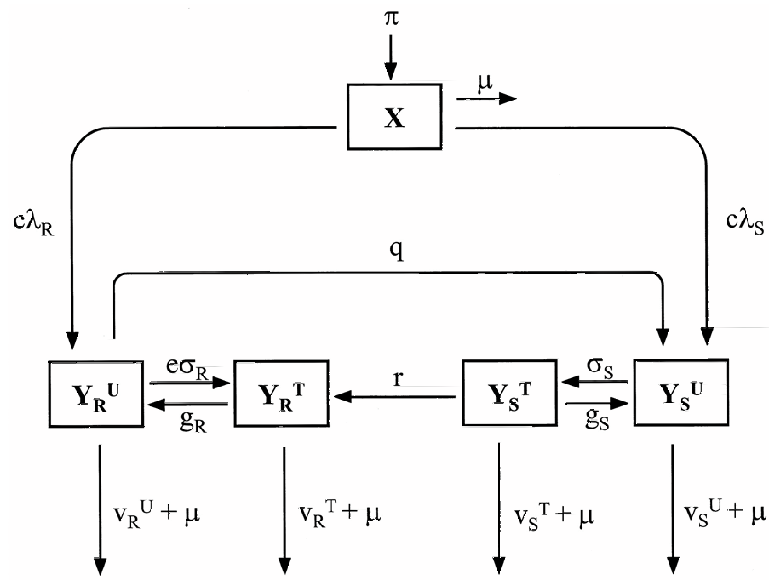

Figure 6.9 Compartment diagram from Blower et al. (2000) for their model for HIV transmission dynamics in the presence of antiretrovivral therapy, with both resistant and non-resistant strains circulating in the population.

(2000, 2001) used a relatively simple model to evaluate the magnitude of the threat posed by ARV-resistant HIV strains, and to evaluate possible responses.

The model (Figure 6.9) describes the transmission dynamics of HIV in the presence of antiretroviral therapy, with resistant and nonresistant strains being transmitted. In structure it is an SIR model with two complications: distinguishing between treated and untreated cases, and between resistant and nonresistant strains of the disease. Dropping the recovered population as usual, the state variables are the numbers of susceptible individuals (X), and four classes of infected individuals (Y) with S and R indicating drug-sensitive versus drug-resistant strains, and T and U indicating treated versus untreated individuals. The λ's are the force of infection, calculated from the number of infected individuals in each category and the infectiousness (β) of each type of infection. Parameters σ and g represent the rates of individuals entering and leaving treatment. Untreated drug-resistant infections revert to drug sensitive at rate q, while treated drug-sensitive infections acquire drug resistance at rate r.

Data-based estimates for several model parameters were available for the San Francisco gay community, while others parameters were less certain and probability distributions were used to represent the relative likelihood of different possible values. For example Blower et al. (2001) allowed the rate r at which ARV drug resistance develops in treated cases to range between 10% and 60% per year. Because little is known about ARV resistant strains, and in particular

about their transmissibility relative to drug-sensitive strains, they allowed the relative transmissibility to vary between 1% and 100% of the transmissibility of the drug-sensitive strain.

Given this wide range of uncertainty, model predictions were generated for a large number of random draws of parameter values according to the distributions representing parameter uncertainty, and then studied statistically as if they were the results of an experiment—which in a sense they are. Blower et al. (2001) were able to compare model predictions with empirical estimates through 1999; data published later allowed comparison out to 2001 (Blower et al. 2003, Figure 6.10).

A surprising prediction from the model is that the transmission of resistance is low, and will remain low at least in the short run: for 2005, it was predicted that most new infections will still be by drug-sensitive strains (median 84.4%, interquartile range 72–94%). Thus, the main source of drug-resistant cases is conversion of drug-sensitive to drug-resistant cases: individuals with a drug-resistant infection are at risk themselves, but do not pose a major threat to the general population. This prediction has some important practical applications. First, it says that combination ARV will remain effective on most new infections, and can continue to be used on newly diagnosed cases. Second, efforts to limit the spread

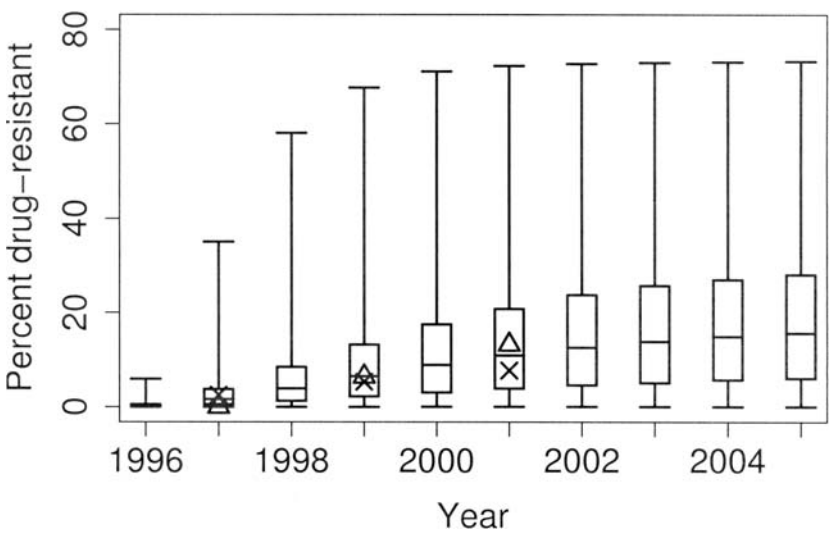

Figure 6.10 Fraction of new HIV infections that are resistant to combination ARV treatment: theoretical predictions versus empirical data for San Francisco (from Blower et al. 2003). Model simulations were run over the time period 1996–2005, with initial conditions corresponding to estimated values for 1996. Boxes enclose the interquartile range (25th to 75th percentiles) of model outcomes and bars show outlier cutoffs; the bars inside the boxes are the median values. Triangles show resistance to non-nucleoside reverse transcriptase inhibitor, and crosses show resistance to protease inhibitor, in a study of 243 newly infected individuals in San Francisco from 1996 to 2001 who had no previous exposure to ARV drugs.

of drug-resistant strains should focus on minimizing the rate of conversion from sensitive to resistant cases—by delaying treatment as long as possible, and trying to enforce strict compliance with the treatment program. Finally, the lack of resources to monitor patient compliance in developing countries implies that drug resistance is likely to be more of a problem than in developed countries.

6.6 Within-Host Dynamics of HIV

Models for infectious diseases at the population level have recently been adapted to model the proliferation of viral diseases within a single host (Nowak and May 2000). In this section we present some relatively simple models that were developed to analyze clinical studies of HIV. Results from these models provided the first inkling of the massive battle waged by the immune system against HIV during the chronic period of infection, and had a major impact on the treatment of HIV infection.

The primary targets of HIV-1 are CD4-positive T-lymphocytes. Infection begins when a viral particle (virion) encounters an activated T-cell, and the viral envelope binds with the CD4 receptor on the cell membrane. The cell membrane and viral envelope fuse, and the viral core enters the cell. The host cell's genetic machinery is commandeered and it begins making multiple copies of the viral RNA. New virions form within the cell and then bud off, carrying along some of the host membrane as a new viral envelope.

Disease progression after HIV infection, if untreated, has three phases. An initial acute phase is marked by high viral loads and flu-like symptoms. The second phase is largely asymptomatic: viral loads fall to a quasi-steady-state and remain there for a period of a few months to a decade. During this time T-cells slowly decline. Finally, there is immune system failure followed by death from opportunistic infections (Figure 6.11).

Because viral and T-cell levels change very slowly, the asymptomatic stage was assumed to be a period when the virus was relatively inactive and nothing much was happening. The development of antiretroviral drugs made it possible to test this assumption. Ho et al. (1995) and Wei et al. (1995) treated chronically infected patients with then newly developed drugs (reverse transcriptase and protease inhibitors) which prevent the virus from infecting additional cells. The surprising result was an extremely rapid exponential decay in viral load (Figure 6.12).

The simplest model to explain these findings posits that viral production is totally shut down by drug treatment. Then

$$dV/dt = -cV \qquad [6.29]$$

where V is the viral load and c the virion clearance rate. The value of c is then the slope of $log\ V$ versus t, and was estimated by fitting a straight line to the

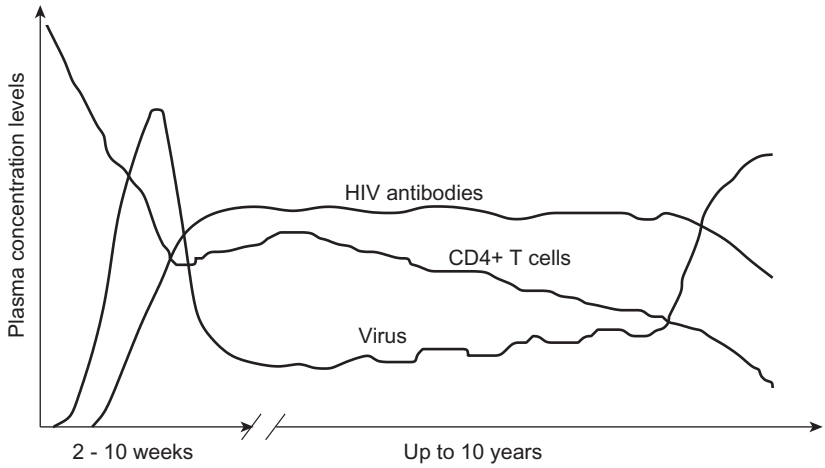

Figure 6.11 Schematic depiction of the typical course of HIV infection in an adult. The early peak in viral load corresponds to the primary infection or "acute" phase. Also shown are T-cell dynamics (from Perelson and Nelson 1999).

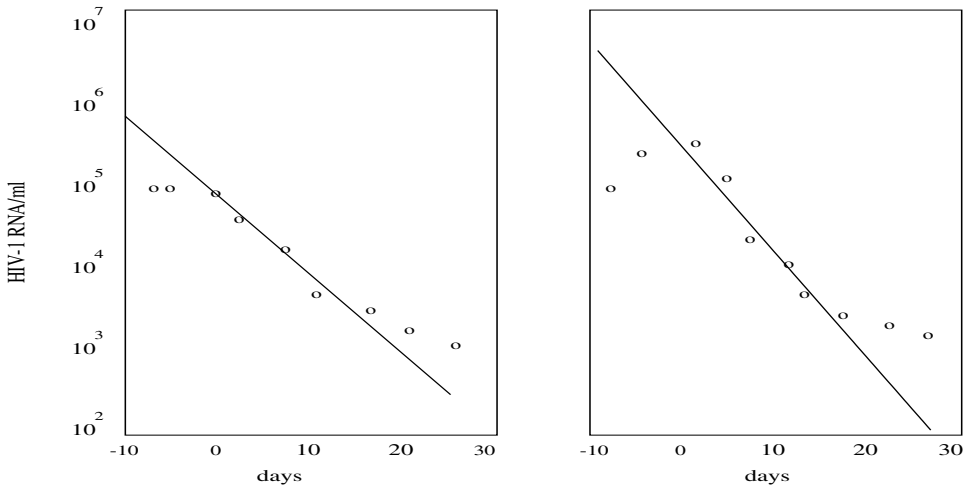

Figure 6.12 Decay of plasma viral load in two patients following treatment with a protease inhibitor. Treatment is initiated at $t = 0$ (from Perelson and Nelson 1999, using data from Ho et al. 1995).

exponential decay phase of the data (Figure 6.12). The viral half-life is then $t_{1/2} = log\ 2/c$. Ho et al. (1995) estimated $t_{1/2} = 2.1 \pm 0.4$ days based on 20 patients, and Wei et al. (1995) estimated $t_{1/2} = 1.8 \pm 0.9$ days based on 22 patients. These estimates revealed that the asymptomatic stage is actually very dynamic, with rapid virus production balanced by rapid clearance.

However, the simple model [6.29] confounds two processes: the clearance of existing free virions, and the clearance of infected cells that are producing new virus. Estimating both clearance rates required a model in which both processes

are explicitly represented (Ho et al. 1995, see also Perelson and Nelson 1999; Nowak and May 2001):

$$\begin{aligned} dT/dt &= \lambda - dT - kVT \\ dT^*/dt &= kVT - \delta T^* \\ dV/dt &= N\delta T^* - cV \end{aligned} \qquad [6.30]$$

where T are uninfected T-cells, T^* are productively infected cells, and V represents free virus. This is essentially an SI model for the spread of infection in a population of T-cells, with infection occurring through contacts between susceptible T-cells and free virions. T-cells are "born" at rate λ, have per capita mortality rate d, and are infected at rate kVT—a "mass action" model for contacts between susceptible cells and virus. Infected cells die or are cleared at a rate δ. Virus V is generated at a rate $N\delta$ per infected T-cell T^* and cleared at a per capita rate c. N is called the "burst size" and represents the total number of free viral particles produced by an infected cell over its lifetime. Since the mean lifetime of an infected cell is $1/\delta$, the burst size N corresponds to an instantaneous virion production rate of $N\delta$.

Perelson et al. (1996) applied this model to clinical data. Five patients received a protease inhibitor, and their HIV-1 RNA concentrations were measured frequently over the next 7 days. Protease inhibitors cause infected cells to produce noninfectious virions. In the presence of a protease inhibitor (assumed for simplicity to be 100% effective), the model becomes

$$\begin{aligned} dT/dt &= \lambda - dT - kV_I T \\ dT^*/dt &= kV_I T - \delta T^* \\ dV_I/dt &= -cV_I \\ dV_{NI}/dt &= N\delta T^* - cV_{NI}. \end{aligned} \qquad [6.31]$$

Here V_I is infectious virus, and V_{NI} is noninfectious virus. Both are cleared at rate c.

Given the short duration of the experiment and assuming a relatively large pool of uninfected cells T, it was assumed that T remained at its steady state value T_0 for the duration of the study. The model then reduces to

$$\begin{aligned} dT^*/dt &= kV_I T_0 - \delta T^* \\ dV_I/dt &= -cV_I \\ dV_{NI}/dt &= N\delta T^* - cV_{NI}. \end{aligned} \qquad [6.32]$$

This can be solved by matrix methods, or sequentially as follows. Prior to treatment all virus is infectious, that is, $V_I(0) = V_0$, therefore

$$V_I(t) = V_0 \exp(-ct).$$

Assuming that the patient was at steady state prior to treatment, the second line of [6.31] implies that

$$kV_0T_0 = \delta T_0^*. \qquad [6.33]$$

Substituting these into the equation for T^* yields

$$dT^*/dt = -\delta T^* + \delta T_0^* \exp(-ct) \qquad [6.34]$$

with initial condition $T^*(0) = T_0^*$. Since this equation is linear it can be solved explicitly (as outlined in exercises below), to obtain

$$T^*(t) = \frac{T_0^*}{(c-\delta)} \cdot [c\exp(-\delta t) - \delta\exp(-ct)].$$

The equation for noninfectious virus can be solved similarly, after substituting in the expression for T^* and imposing the initial condition $V_{NI}(0) = 0$. The resulting expression for the total free virus $V(t) = V_I + V_{NI}$ is then

$$V(t) = V_0\exp(-ct) + \frac{cV_0}{(c-\delta)} \cdot \left\{ \frac{c}{(c-\delta)} \cdot [\exp(-\delta t) - \exp(-ct)] - \delta t\exp(-ct) \right\}. \qquad [6.35]$$

Values of c and δ were estimated for each patient by fitting this equation to the measurements of plasma viral load (Figure 6.13). The clearance rate of free virus c was found to be relatively rapid, with the half-life estimated to be 0.24 ± 0.06 days. Thus, half the virion population is replaced every 6 hours. Death of productively infected cells T^* was found to occur more slowly, with an estimated half-life of 1.55 ± 0.57 days.

During the approximate steady state prior to treatment, virus production must balance the clearance, cV. Using their estimate of c and measured steady-state viral loads V_0, Perelson et al. estimated that 10.3×10^9 free virions were produced per day prior to drug treatment. This was an order of magnitude higher than the original estimates by Ho et al. (1995) and Wei et al. (1995).

The implication for treatment was that HIV has the potential to evolve very rapidly in response to selection imposed by the immune system, or by drug treatment. Combining turnover rate estimates with estimates of genome size and mutation rate, Coffin (1995) concluded that all possible point mutations in HIV arise thousands of times each day in a chronically infected adult. This provided a simple mechanistic explanation for the rapid development of resistance to single-drug treatments. The current practice of administering three-drug cocktails—which would require simultaneous mutations at three different sites to confer drug resistance—arose directly from these conclusions. As noted in the last section, multidrug resistance is developing much more slowly.

Exercise 6.15. Here is one way to find the solution to [6.34]. General theory for first-order linear differential equations tells us to expect a solution of the form $T^*(t) = A\exp(-ct) + B\exp(-\delta t)$ for some constants A and B. Find the solution by

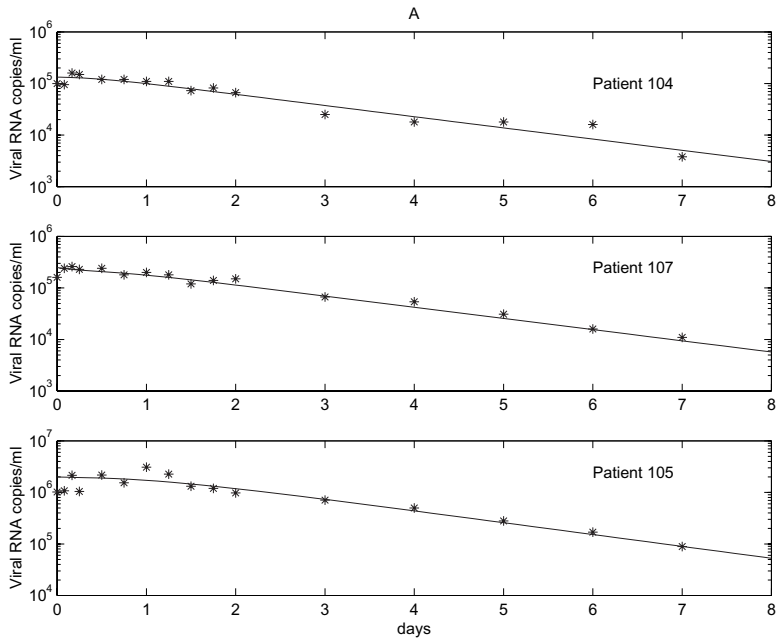

Figure 6.13 Viral load data (symbols) versus model predictions (solid line) during first phase of viral decay after onset of treatment ($t = 0$) (from Perelson and Nelson 1999, after Perelson et al. 1996).

substituting this trial form into [6.34] and finding A, B to satisfy the initial conditions and the differential equation.

Exercise 6.16. Here is another. Equation [6.34] can be rearranged as

$$dT^*/dt + \delta T^* = \delta T_0^* \exp(-ct).$$

(a) Let $x(t) = \exp(\delta t)T^*(t)$ and use the chain rule to show that $\exp(-\delta t)dx/dt = dT^*/dt + \delta T^*$.

(b) Using the result of part (a) show that $x(t)$ satisfies a differential equation of the form $dx/dt = f(t)$, which implies that $x(t) = x(0) + \int_0^t f(s)ds$.

(c) The rest is algebra: plug in $x(0) = T^*(0)$, use [6.33] and simplify, to derive [6.35].

6.7 Conclusions

Infectious disease models provide some of the best examples of the practical value of simple dynamic models. Simple models, deliberately stripped down to bare essentials so that they could be fully understood, led to *qualitative* insights, such as

- endemic steady-states of diseases with permanent immunity will be spirals,
- contact tracing is the best way to target the core group of gonorrhea carriers,
- rapid turnover of HIV-1 allows it to rapidly evolve resistance to any single-drug therapy.

In principle these insights could also be gleaned from more complex and realistic models, that would presumably make the same predictions. But complex models can be as hard to understand as the real-world system that they represent. Simpler models are then an essential starting point for disentangling the complexities of complex models and complex real-world systems.

Infectious diseases also provide examples of situations where complex models are essential because the quantitative details matter—such as predicting just how rapidly drug-resistant HIV will spread, deciding what control measures will be sufficient to halt the spread of foot-and-mouth disease (as described in the Preface), or determining which individuals should be given highest priority for limited vaccine supplies. Our next two chapters are devoted to some of the complexities that we have so far ducked. The next adds *space*, where the variation in state variables across location as well as across time must be taken into accout. After that we look more closely at models that try to represent the complexity of biological systems by modeling at the level of individual agents, and therefore can only be studied by computational methods.

6.8 References

Anderson, R. A., and R. M. May. 1992. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, U.K.

Blower, S. M., H. B. Gershengorn, and R. M. Grant. 2000. A tale of two futures: HIV and antiretroviral therapy in San Francisco. Science 287: 650–654.

Blower, S. M., A. N. Aschenbach, H. B. Gershengorn, and J. O. Kahn. 2001. Predicting the unpredictable: Transmission of drug-resistant HIV. Nature Medicine 7: 1016–1020.

Blower, S. M., A. N. Aschenbach, and J. O. Kahn. 2003. Predicting the transmission of drug-resistant HIV: Comparing theory with data. The Lancet Infectious Diseases 3: 10–11.

Butler, J. C., et al. 1996. The continued emergence of drug-resistant Streptococcus pneumoniae in the United States: An update from the CDC's Pneumococcal Sentinel Surveillance System. Journal of Infectious Diseases 174: 986–993.

CDC (Centers for Disease Control and Prevention). 2000. Gonorrhea—United States, 1998. Morbidity and Mortality Weekly Report 29: 538–542.

J. M. Coffin. 1995. HIV population dynamics in vivo: Implications for genetic variation, pathogenesis, and therapy. Science 267: 483–489.

Campbell, C. L., and L. V. Madden. 1990. *Introduction to Plant Disease Epidemiology*. John Wiley & Sons, New York.

Davidson, D. R. 1995. Influenza Outbreak in Basic Training, Fort Benning, GA. Medical Surveillance Monthly Report 1:7. U.S. Department of Defense, U.S. Army Center for Health Promotion and Preventive Medicine (Provisional), Aberdeen Proving Ground, MD 21010-5422.

Dye, C., B. G. Williams, M. A. Espinal, and M. C. Raviglione. 2002. Erasing the world's slow stain: Strategies to beat multidrug-resistant tuberculosis. Science 295: 2042–2046.

Earn, D.J.D., P. Rohani, B. M. Bolker, and B. T. Grenfell. 2000. A simple model for complex dynamical transitions in epidemics. Science 287: 667–670.

Ellner, S. P., B. A. Bailey, G. V. Bobashev, A. R. Gallant, B. T. Grenfell, and D. W. Nychka. 1998. Noise and nonlinearity in measles epidemics: Combining mechanistic and statistical approaches to population modeling. American Naturalist 151: 425–440.

Finkenstadt, B. F., and B. T. Grenfell. 2000. Time series modelling of childhood diseases: A dynamical systems approach Journal of the Royal Statistical Society Series C—Applied Statistics 49: 187–205.

Grenfell, B. T., and R. M. Anderson. 1985. The estimation of age related rates of infection from case notifications and serological data. Journal of Hygiene 95: 419–436.

Grenfell, B. T., O. N. Bj**o**rnstad, and B. F. Finkenstadt. 2002. Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. Ecological Monographs 72: 185–202.

Hall, A., P. Pomeroy, and J. Harwood. 1992. The descriptive epizootiology of phocine distemper in the UK during 1988/89. Science of the Total Environment 115: 31–44.

Hethcote, H. W., and James A. Yorke. 1984. *Gonorrhea Transmission Dynamics and Control*. Lecture Notes in Biomathematics Vol. 56. Springer-Verlag, New York.[1]

Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373: 123–126.

Hudson, P. J., A. Rizzoli, B. T. Grenfell, H. Heesterbeek, and A. P. Dobson. 2002. *The Ecology of Wildlife Diseases*. Oxford University Press, Oxford.

Kermack, W. O., and W. G. McKendrick. 1927. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London, Series A 115: 700–721.

Kermack, W. O., and W. G. McKendrick. 1932. Contributions to the mathematical theory of epidemics. II. The problem of endemicity. Proceedings of the Royal Society of London, Series A 138: 55–83.

McCallum, H., N. Barlow, and J. Hone. 2002. How should pathogen transmission be modelled? Trends in Ecology and Evolution 16: 295–300.

Moreno, Y., R. Pastor-Satorves, and A. Vespignani. 2002. Epidemic outbreaks in complex heterogeneous networks. European Physics Journal B 26: 521–529.

Newman, M. E. J. 2003. The structure and function of complex networks. SIAM Review 45: 167–256.

Nowak, M. A., and R. M. May. 2000. *Virus Dynamics: Mathematical Principles of Immunology and Virology.* Oxford University Press, Oxford, U.K.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. 1996. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 271: 1582–1586.

Perelson, A. S., and P. W. Nelson. 1999. Mathematical analysis of HIV-1 dynamics in vivo. SIAM Review 41: 3–44.

Rohani, P., D. J. D. Earn, and B. T. Grenfell. 1999. Opposite patterns of synchrony in sympatric disease metapopulations. Science 286: 968–971.

Ross, R. 1916. An application of the theory of probabilities to the study of *a priori* pathometry. Part I. Proceedings of the Royal Society of London, Series A 92: 204–230.

Schaffer, W. M., L. M. Olsen, G. L. Truty, and S. L. Fulmer. 1990. The case for chaos in childhood epidemics. Pages 138–166 in S. Krasner (ed.), *The Ubiquity of Chaos*. American Association for the Advancement of Science, Washington, DC.

Wei, X. P., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw. 1995. Viral dynamics in human-immunodeficiency-virus type-1 infection. Nature 373: 117–122.

York. J. A., H. W. Hethcote, and A. Nold. 1978. Dynamics and control of the transmission of gonorrhea. Sexually Transmitted Diseases 5: 51–56.

[1]This book is out of print but has been placed in the public domain by the authors and is available at http://www.math.uiowa.edu/ftp/hethcote/lnb56.pdf

This page intentionally left blank