DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

Introduction to Mathematical Modeling of Infectious Diseases

Sourya Shrestha[1] and James O. Lloyd-Smith[2,3]

1. Introduction

Despite major advances in science and public health, infectious diseases continue to cause significant morbidity and mortality in human populations worldwide, with disproportionate impact in developing countries. A recent survey estimated that infectious diseases are responsible for more than half of human deaths in sub-Saharan Africa, and sub-lethal effects of disease impose heavy burdens on quality of life and economic development (Lopez et al., 2002). Impacts of infectious disease extend beyond human populations, exacting tolls on domestic animal, wildlife and plant populations. The combination of complex ecology, rapid evolution in response to changing circumstances, and the on-going emergence of novel pathogens, ensures that infectious diseases will continue to pose serious challenges for the foreseeable future.

Thus there is strong motivation to pursue the scientific study of infectious diseases, and mathematical models make a unique contribution to this endeavor. Infectious diseases can exhibit complex nonlinear dynamics, and mathematical models enable clear and rigorous analysis of the underlying mechanisms. Models provide a crucial link between individual-level 'clinical' knowledge of disease properties and population-scale patterns of incidence and prevalence, and help to establish the relative importance of different processes to focus research and management effort. Models can provide a testing ground to investigate the possible efficacy of alternative disease control policies, in settings where direct experimentation is often ethically or logistically impossible. Finally, the process of formulating and analyzing models has the salutary effect of focusing research questions and forcing clear enunciation of assumptions and hypotheses.

This chapter is a summary of lectures given by JL-S at the DIMACS Advanced Studies Institute at AIMS in 2007. We are grateful to Ottar Bjørnstad and Matt Ferrari for sharing materials that helped with the preparation of these lectures. We are also very grateful to the students and other instructors at the institute, whose questions and comments helped to improve the presentation of this material. JL-S was funded by a Center for Infectious Disease Dynamics Fellowship at Penn State University, and by the RAPIDD program of the Science & Technology Directorate, US Department of Homeland Security, and the Fogarty International Center, National Institutes of Health.

In this article we seek to provide a concise introduction to the concepts and basic methods of infectious disease modelling. We begin by briefly reviewing some key concepts and terminology for the study of infectious disease dynamics, then introduce some classical models and broad decisions faced in model design. We discuss formulation of the transmission term that sits at the heart of infectious disease models, and approaches to incorporating population heterogeneity and structure. We conclude with a survey of approaches to parameter estimation, model fitting, and sensitivity and uncertainty analysis. Our treatment is far from comprehensive and for details we refer the reader to books by Anderson & May (1991), Diekmann & Heesterbeek (2000) or Keeling & Rohani (2007).

2. Basic concepts in infectious disease dynamics

2.1. Pathogens. It is conventional among disease modellers to categorize disease-causing organisms into two broad groups based on their size and life cycle. (i) **Microparasites** are very small organisms that multiply, often to reach very large population sizes, within individual hosts. This category includes viruses (such as influenza or HIV), bacteria (such as *Bordetella pertussis* that causes whooping cough and *Bacillus anthracis* that causes anthrax), protozoa (such as *Trypanosoma* species that cause sleeping sickness and *Plasmodium* species that cause malaria), and pathogenic fungi (such as *Candida* species that cause thrush and *Tinea pedis* that causes Athlete's foot). Microparasitic infections range from acute to chronic, lasting days to years, and cause effects ranging from asymptomatic infection to rapid death. Host individuals that survive infection often have some degree of immunity to re-infection, though the duration of this protection varies greatly among pathogens. Microparasites have a short generation time relative to their host, and undergo many generations in the course of a single infection. Hence microparasites can evolve rapidly, with important applied consequences for development of drug resistance, evasion of host immune responses, and adaptation to changing environmental circumstances.

(ii) **Macroparasites** are larger multicellular organisms, which multiply outside of the host. Macroparasites can be further divided into endoparasites that enter the host's body, such as roundworms and flukes, and ectoparasites that dwell on the surface of the host, such as ticks and mites. Macroparasites typically have longer generation times than microparasites, and their life cycles tend to be complex often comprising stages both within and outside the host and specialized infective stages. Macroparasitic infections are often chronic and sub-lethal, and the host immune response often serves to contain or limit the infection rather than eliminate it.

From a disease modelling standpoint, there are several important distinctions between macroparasites and microparasites. Typically, microparasite models classify hosts according to infection status (e.g. susceptible vs infected), while macroparasite models explicitly quantify the parasite burden within each host. The burden of macroparasites, which accumulates as a result of multiple infection events of the same host individual, is assumed to affect transmission rates from that host and morbidity or mortality of the host. In contrast, parasite burden and multiple infection events are typically assumed to be unimportant for microparasites (i.e. once infected, further exposures do not matter) – though exceptions exist for models addressing multi-strain dynamics or other particular questions.

2.2. Host immune response. Immunology is a vast and fast-moving field, and we certainly do not aim to summarize it here. Instead we present a few deliberately simplified concepts that motivate the formulation of basic disease models. A more in-depth treatment with a focus on dynamical aspects can be found in Perelson & Weisbuch (1997). The host immune response can be broadly divided into two parts. The **innate immune response** is an intrinsic, mostly non-specific response to any foreign intruder, which provides a first line of defense particularly for pathogens that the host has not encountered before. The **acquired** or **adaptive immune response** is a pathogen-specific defense based on recognition of characteristic molecular 'signatures' of pathogens that have been encountered before. The acquired immune response is the basis for the practice of vaccination, which is based on exposing the host to molecules from the parasite to prime the immune system for subsequent exposures.

A greatly simplified timecourse of the immune response to an acute microparasitic infection is as follows. Following infection with a bacterial pathogen, the innate immune response is activated immediately as macrophages recognize molecules in the bacterial wall as foreign particles. Macrophages begin killing the bacteria, and they also produce chemical signals or chemokines that recruit neutrophils and other innate effectors to the site of infection to aid in killing the bacteria. The adaptive part of the immune response begins several days after the beginning of the infection, stimulated by other chemokines and antigens (molecular motifs from the parasite) which stimulate antigen-specific B cells. These stimulated B cells multiply to produce plasma B cells that in turn produce antibodies (large molecules that bind the corresponding antigen with high affinity). These antibodies bind to the antigens on the bacteria, and cytotoxic T cells recognize the bound antibodies and kill the bacteria, eventually eliminating the infection. Apart from producing plasma cells, the stimulated B cells also produce long-lived memory B cells. This helps the immune system to react swiftly and effectively if the host is later exposed to the same antigen, so re-infection by the same bacterial strain is less likely. Depending on the efficacy of the memory response, that host may be said to be 'immune' to that bacterial pathogen. The duration and efficacy of immunity can vary greatly from one pathogen to another, from essentially no lasting immunity in the case of gonorrhea to lifelong protective immunity in the case of measles.

Immunology has mostly been a stand-alone field, but recent advances in immunology as well as epidemiology and disease ecology have generated both the need and the interest to study the immune system in conjunction with epidemiological aspects of disease dynamics. The interaction between pathogens and the host immune system is likely to affect various characteristics of an infection, such as how much pathogen the host will be carrying during the infection or how long the infection will last. These characteristics in turn will influence epidemiological properties of the disease, including transmissibility. These complexities are beyond the scope of this chapter, so we will direct readers to a sampling of recent papers that address the interaction between immunology and epidemiology. (Schmid-Hempel, 2008; Antia et al., 2005; Mideo et al., 2008; King et al., 2009)

2.3. Clinical course of disease. Once a pathogen becomes established in a host, the infection goes through different phases as it runs its course. Understanding the details of the course of infection is crucial when modelling disease spread. One is typically concerned about the time it takes for a newly infected host to become

Table 1. Incubation, latent and infectious periods (in days) for a variety of viral and bacterial infections. Adapted from Anderson and May (1991).

Infectious disease	**Incubation period**	**Latent period**	**Infectious period**
Measles	8 - 13	6 - 9	6 - 7
Mumps	12 - 16	12 - 18	4 - 8
Pertussis	6 - 10	21 - 23	7 - 10
Rubella	14 - 21	7 - 14	11 - 12
Diphtheria	2 - 5	14 - 21	2 - 5
Chicken pox	13 - 17	8 - 12	10 - 11
Hepatitis B	30 - 80	13 - 17	19 - 22
Poliomyelitis	7 - 12	1 - 3	14 - 20
Influenza	1 - 3	1 - 3	2 - 3
Smallpox	10 - 15	8 - 11	2 - 3
Scarlet fever	2 - 3	1 - 2	14 - 21

infectious and begin transmitting, the time it takes for the onset of symptoms, and the duration of infectiousness. These details correspond to basic parameters in epidemic models, and their values will determine basic properties of the resulting epidemic curves, some of which we shall discuss in the later sections. For now we introduce some of the frequently-used terminologies:

Incubation period: the time from infection to the onset of symptoms.

Latent period: the time from infection to the onset of transmissibility. This is also called the pre-patent period for macroparasites.

Infectious period: the time during which individual can transmit disease.

Generation time (or serial interval): the time period between infection of one host and infection of other secondary cases caused by that host.

Table 1 shows the range of values for these periods for some important microparasites. Note that the incubation and latent periods are not always equal, so the onset of transmission can come before or after the onset of symptoms for some pathogens. This has important consequences for the ability to control these disease (Fraser et al., 2004).

2.4. Transmission. Transmission of infection between individuals is the central process of infectious disease dynamics – indeed it is the reason why these disease are 'infectious' – so it is essential to consider transmission carefully in designing an epidemic model. There are many different modes of transmission (see Table 2), and models must account for the relevant modes and how they affect the disease spread. Often one has to confront questions such as:

- What is the host contact structure relevant to this mode of transmission? For example, sexually transmitted diseases (STDs) clearly follow different contact structures from respiratory pathogens.
- Are there important heterogeneities among hosts that will impact transmission?
- What disease control measures are relevant for this mode of transmission?

The answers to these questions gives us a good platform to develop accurate models of disease dynamics. We will return to the topic of different transmission models and their consequences in section 4.1.

2.5. Epidemiological terms and population-level patterns. Infectious disease epidemiology is the study of the spread of disease at the population level. Here are a few basic terms that are used in epidemiological descriptions of infectious diseases:

Incidence: the number of new infections per unit time.

Prevalence: the proportion of population that is infected at a particular time.

Attack rate: the proportion of susceptible individuals in a given setting that become infected during a given interval.

Force of infection: the per capita rate of infection per unit time (i.e. the hazard rate of infection experienced by each susceptible individual).

Seroprevalence: the proportion of population carrying antibodies indicating past exposure to pathogen.

The population-level patterns of disease spread can be broadly categorized as follows (see Fig 1 for schematic depictions):

Endemic infections: infections which do not exhibit wide temporal fluctuations in incidence or prevalence in a defined place. For microparasites, the term 'endemic' can also be used to indicate an infection that can persist locally without the need for reintroduction from outside host communities. Stable endemicity is where the incidence of infection or disease shows no secular trend for increase or decrease.

Simple epidemic: an outbreak of infection entailing a rapid increase in incidence followed by decline and (possibly temporary) disappearance of the pathogen. This epidemic pattern is typical of the microparasitic infections with high transmission rates, short generation times, and long-lasting host immunity. An epidemic usually begins with an exponential rise in the number of cases and a subsequent decline as the population of susceptible hosts is exhausted (or control measures are imposed).

Recurrent epidemics: a population-level pattern in which simple epidemics occur sporadically due to repeated introduction of the pathogen. Following each epidemic there may be a 'refractory period' in which the susceptible population is diminished so that another epidemic cannot occur. After sufficient time the susceptible population will be replenished (by births, immigration, or loss of acquired immunity) and subsequent introduction of the pathogen can lead to another epidemic.

TABLE 2. Different modes of transmission

Modes	**Diseases**
Direct (droplet, aerosol, fomite)	influenza, measles, SARS
Sexual transmission	HIV, gonorrhea, HSV-2, chlamydia
Vector-borne (mosquitoes, tsetse flies, sandflies)	malaria, trypanosomiasis, leishmaniasis
Free-living infectious stages and environmental reservoirs	helminths, anthrax
Waterborne, food-borne, fecal-oral	cholera, polio, Salmonella

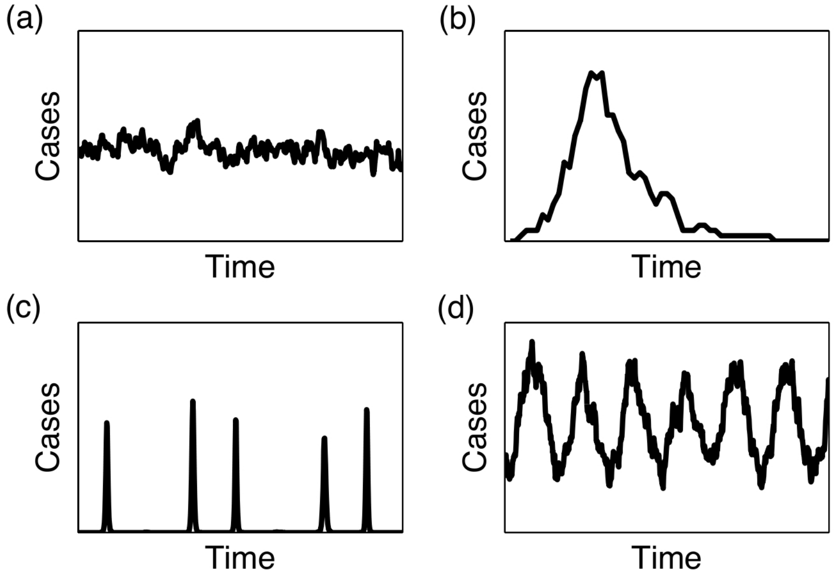


Figure 1. Population level patterns. (a) Endemic infection; (b) Simple epidemic; (c) Recurrent epidemic; and (d) Seasonal endemism.

Seasonal endemism: a situation in which the pathogen circulates persistently within a given population, but incidence rates vary seasonally due to changing contact rates or environmental factors. Such systems may exhibit annual epidemics of varying intensity, or marked multi-annual cycles arising from the interaction of seasonal factors with longer-period cycles of susceptible depletion and replenishment.

3. Introduction to infectious disease modelling

3.1. SIR-type compartment models for microparasite infections. The most common approach to understanding the dynamics of microparasite infections is the family of compartmental models. These models divide the population into compartments, where all individuals in a given compartment share some common attributes — where the choice of these attributes will depend on the questions that you are trying to answer and the complexity of the dynamics that you are modelling. For microparasites, these models are often referred to as SIR-type models, where S, I, and R represent compartments of Susceptible, Infectious, and Removed (or Recovered) populations, respectively. These are represented by state variables $S(t)$, $I(t)$ and $R(t)$ that track how the numbers of individuals in each compartment change with time. The set of equations below describe how the quantities in each of the compartments are changing in a standard SIR model with frequency-dependent transmission.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta\, S\, I}{N} \\
\frac{dI}{dt} &= \frac{\beta\, S\, I}{N} - \gamma\, I \\
\frac{dR}{dt} &= \gamma\, I
\end{aligned} \tag{1}$$

Here, β is the transmission coefficient, which relates the current prevalence of infectiousness to the force of infection, and γ is the per capita rate of recovery for infected individuals. Susceptible individuals become infected upon successful transmission of the pathogen when two individuals from the S and I compartments interact (see section 4.1 for more details). Infectious individuals are removed from the infectious state at the given recovery rate. In this very simple model, recovered individuals remain in the R compartment indefinitely.

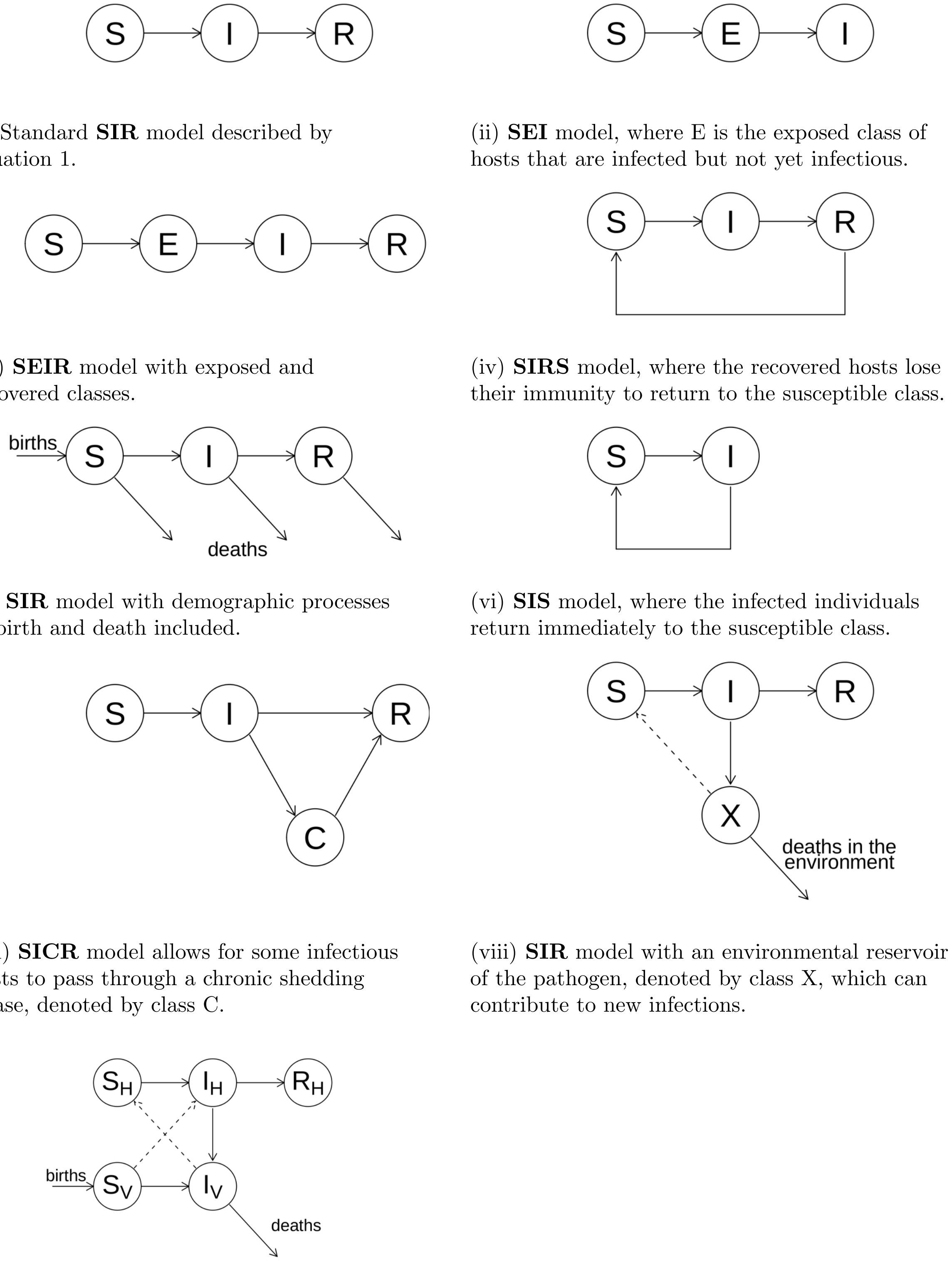
The set of differential equations given by (1) can be solved approximately. Because $S+I+R=N$ is a constant, we can write $\dfrac{dR}{dt} = \gamma(N-S-R)$. Introducing $\hat{\beta} = \frac{\beta}{N}$, we find $\dfrac{dS}{dR} = \dfrac{\frac{dS}{dt}}{\frac{dR}{dt}} = -\dfrac{\hat{\beta}}{\gamma}S$. Now S can be expressed in terms of R as $S = S_0\, e^{-\frac{\hat{\beta}}{\gamma}R}$, where S_0 is the initial condition, i.e. $S(t=0)$. Hence, the differential equation for R can be written in terms of R alone, $\dfrac{dR}{dt} = \gamma(N - S_0\, e^{-\frac{\hat{\beta}}{\gamma}R} - R)$. The approximate solution to this, due to Kermack & McKendrick, is:

$$\begin{aligned}
R &= \frac{\gamma^2}{S_0\, \hat{\beta}^2}\left[\frac{\hat{\beta}}{\gamma}S_0 - 1 + q\tanh\left(\frac{1}{2}q\,\gamma\, t - \phi\right)\right], \\
&\text{where } \phi = \tanh^{-1}\left(\frac{\hat{\beta}}{\gamma}S_0 - 1\right)/q, \\
&\text{and } q = \left[\left(\frac{\hat{\beta}}{\gamma}S_0 - 1\right)^2 + 2\, S_0\, I_0 \frac{\hat{\beta}^2}{\gamma^2}\right]^{1/2}.
\end{aligned} \tag{2}$$

The total size of an epidemic is approximately $\dfrac{2\gamma}{S_0\hat{\beta}}\left(S_0 - \dfrac{\gamma}{\hat{\beta}}\right)$ (Bailey, 1955).

The SIR model can be altered or customized to suit the biological and social details of the disease system being modelled, or the question that one wants answered. A number of these alterations are shown in figure 2, including very common models (e.g. SEIR, SIS) and models that we have designed for particular problems (e.g. SICR).

3.2. Basic reproductive number, R_0. A central concept in the analysis of infectious disease dynamics is the basic reproductive number, or R_0. R_0 is defined as the expected number of new infections caused by a typical infectious individual in a wholly susceptible population. It is a threshold parameter that determines whether a disease starting from a typical infectious individual can cause an epidemic. In most cases, the same threshold also determines whether a disease can persist at the population scale (i.e. is there an ‘endemic equilibrium’ for the model or is



(i) Standard **SIR** model described by equation 1.

(ii) **SEI** model, where E is the exposed class of hosts that are infected but not yet infectious.

(iii) **SEIR** model with exposed and recovered classes.

(iv) **SIRS** model, where the recovered hosts lose their immunity to return to the susceptible class.

(v) **SIR** model with demographic processes of birth and death included.

(vi) **SIS** model, where the infected individuals return immediately to the susceptible class.

(vii) **SICR** model allows for some infectious hosts to pass through a chronic shedding phase, denoted by class C.

(viii) **SIR** model with an environmental reservoir of the pathogen, denoted by class X, which can contribute to new infections.

(ix) For vector-borne diseases, one tracks epidemic dynamics in both host and vector populations, where transmission occurs via vector bites.

Figure 2. Various adaptations of the SIR-type family of compartmental models.

the 'disease-free equilibrium' globally asymptotically stable?). If $R_0 \leq 1$, then on average an infectious individual will spread the disease to less than one individual, and the disease will not be able to spread and cause a major epidemic (though note that it can 'stutter along' for some time before it goes extinct). By contrast, if $R_0 > 1$, then a typical infectious individual will spread the disease to more than one other individual on average, and hence the disease can take off and result in an epidemic (though note that $R_0 > 1$ does not guarantee that an epidemic will occur, due to stochastic effects that we discuss below). If an epidemic does occur, then the value of R_0 gives important information about the initial growth rate of the epidemic, the prevalence at the peak of an epidemic, and the proportion of the population ultimately infected by a simple epidemic. In an endemic setting, R_0 gives insight into the endemic equilibrium prevalence and the mean age of infection.

For a simple SIR model, R_0 can be calculated in a fairly intuitive manner: it is the product of the per capita rate of infecting others, and the mean duration of infection. Under frequency-dependent transmission, the per capita rate of infecting others is $\beta S/N$ (which reduces to β if the population is wholly susceptible), and the duration of the infection is simply the reciprocal of the recovery rate, so $R_0 = \beta/\gamma$.

Generally, R_0 is taken in the context of a wholly susceptible population. To talk about a population that is not wholly susceptible, we define $R_{\text{effective}}$, which is the expected number of cases caused by a typical infectious individual in a population that is not wholly susceptible. In the simplest case, $R_{\text{effective}} = R_0 \times S/N$. This formulation will help us calculate the endemic level of disease in the population, since at equilibrium $R_{\text{effective}} = 1$ (by definition, because at equilibrium each infected case replaces itself), so the endemic equilibrium level of disease in the population is $1/R_0$. This simple formulation of $R_{\text{effective}}$ also allows us to understand the concept of **herd immunity**. If there is a sufficiently low proportion of the population that is susceptible to the disease, then $R_{\text{effective}}$ will be below 1 and the infection will be unable to circulate persistently in the population. The critical proportion of the population that needs to be immune is determined simply: since $R_{\text{effective}} = R_0 \times S/N$, then for $R_{\text{effective}} < 1$ we need $S/N < 1/R_0$. Hence, we need a proportion $1 - 1/R_0$ of the population to be immune, and the remainder of the population may be individually susceptible but is protected by herd immunity. This sets a target level for vaccination campaigns to stop disease circulation. $R_{\text{effective}}$ is sometimes also used to denote the reproductive number when other control measures have been put in place (e.g. drug treatment or case isolation). In these cases, one can use R_0 and $R_{\text{effective}}$ to determine the threshold coverage (proportion of the population that is subject to the control measure) and efficacy (proportional reduction in transmission) needed to stop disease circulation.

3.3. Models for macroparasite infection. As noted above, the basic model for macroparasite infection differs fundamentally from SIR-type models for microparasites, because for macroparasites the intensity of the infection needs to be taken into consideration. The number of disease-causing macroparasites within a host affects the way in which the disease gets transmitted. Given below is a basic model for a directly-transmitted macroparasite (Anderson & May, 1991).

$$\begin{aligned}\frac{dM}{dt} &= d_1\beta\, L(t-\tau_1) - (\mu+\mu_1)\, M \\ \frac{dL}{dt} &= s\, d_2\, \lambda\, N\, M(t-\tau_2) - \mu_2\, L - \beta\, N\, L\end{aligned} \tag{3}$$

where $N(t)$ represents the size of the host population, $M(t)$ represents the mean number of sexually mature worms in the host population, and $L(t)$ represents the number of infective larvae in the habitat. The parameter β is the infection rate; λ is the per capita fecundity rate of the parasites; μ, μ_1 and μ_2 are the death rates of the hosts, adult worms within the hosts, and larvae in the environment, respectively; d_1 and d_2 are the proportions of ingested larvae that survive to adulthood, and eggs shed that survive to become infective larvae, respectively; τ_1 is the time delay for parasite maturation to the reproductive stage; τ_2 is the time delay for maturation from egg to infective larva; and s is the proportion of offspring that are female. We will not go into greater depth on macroparasite modelling, but note that this model is a truly simplified depiction of macroparasite dynamics, and that even simple models typically account for complexities such as the density dependence of parasite fecundity and mortality, and the well-known pattern of parasite aggregation within host individuals.

3.4. R_0 and $R_{\textbf{effective}}$ for macroparasites. For macroparasites, R_0 is the average number of female offspring (or just offspring in the case of hermaphroditic species) produced throughout the lifetime of a mature female parasite, which themselves achieve reproductive maturity, in the absence of density-dependent constraints on the parasite establishment, survival or reproduction. $R_{\text{effective}}$ is the average number of female offspring produced in a host population within which density dependent constraints limit parasite population growth.

Note the distinction in usage of $R_{\text{effective}}$ across classes of parasites. For microparasites, $R_{\text{effective}}$ is the reproductive number in the presence of competition for hosts at the population scale (i.e. due to herd immunity), whereas for macroparasites, $R_{\text{effective}}$ is the reproductive number in the presence of competition at the within-host scale (i.e. due to density dependence within the host). For both cases, $R_{\text{effective}} = 1$ under the conditions of stable endemic infection.

3.5. Modelling decisions. Even after choosing an appropriate compartmental structure, there are many more decisions faced in the course of designing a model for a particular system and problem. Different modelling approaches are suited to different kinds of analysis, and important assumptions are often implicit in model designs. Here we review a few key dimensions of model design.

3.5.1. *Deterministic versus stochastic models.* One basic assumption made in compartmental models based on ordinary differential equations is that the course of the dynamics is **deterministic**. That is, given a model structure, parameter values and initial conditions, the model output is absolutely determined with no variation. This deterministic framework does not capture the effects of chance, or of variability in the parameters and processes involved in the model. Sometimes this is an acceptable approximation, and deterministic models do have the advantage that they are well studied and often easier to analyze. The results, when attained are definite and easy to intrepret.

In certain situations, however, it is important to consider the influence of chance events by using a **stochastic** model which incorporates variability in dynamic outcomes by modelling some processes in a probabilistic manner. One such situation is when some crucial processes in the model involve small populations or rare events. In these situations, substantial variations in population-level outcome can arise because individual-level events are discrete and uncertain. For example, an infection may have $R_0 = 1.5$ but each infected host will transmit the infection to an integer number of other hosts (0,1,2,3,...). If there is only one infected host in a population, then it matters a great deal whether he/she transmits to 0 or 1 or 2 other hosts, because the infection is very vulnerable to going extinct in that host population. (If there are 1000 infected hosts, then this individual-level uncertainty is less important because the population average outcome will dominate.) This uncertainty arising from the discreteness of individual-level events is called 'demographic stochasticity'. A second source of variation in dynamic outcomes is 'environmental stochasticity', which arises from fluctuations in the environment. Note that this includes factors in the natural environment, such as the weather, but also any other factors or processes that are not explicitly included in the model.

There are many different approaches to stochastic modelling. **Monte Carlo simulations** are models that use computers' pseudo-random number generators to determine the occurrence of events according to appropriate probability distributions. **Branching processes** and **birth-death processes** are simple analytic approaches to modelling invasion in relatively large susceptible populations. These models allow for flexibility in the distribution of secondary infections, and can be solved explicitly for quantities such as the probability of extinction, but they do not account for depletion of the susceptible population. **Chain binomials** or **Reed-Frost models** are used to model stochastic epidemics in finite populations. For each generation of transmission, these models enable us to calculate the number of newly infected individuals as a binomial random draw from the remaining susceptibles. **Diffusion processes** can be used to model infectious disease spread in large populations. Here the number of infectious individuals is modeled as a random walk around the equilibrium value, giving rise to a so-called 'quasi-stationary distribution' which is the long-term steady state distribution of values, conditioned on non-extinction.

3.5.2. *Continuous-time versus discrete-time models.* A fundamental decision in formulating a dynamic model is how to treat time. The models we have described using differential equations are continuous-time models, i.e. the time variable can take any real value. The key advantage of continuous-time models is that they are well-suited for mathematical analysis. They also allow for full flexibility in quantities like residence times and other durations. However, there are several factors that may favour the use of discrete-time models, in which the time variable changes by fixed increments. Data are typically reported in discrete time intervals, and it can be convenient to design the model to match the data. Discrete-time models can also match natural time scales of the system, such as the generation time or length of a season. Some people find discrete-time models more intuitive, and simulation code is easy to write as a basic 'for' loop. It can also be argued that for most epidemic models, especially the ones that are more complicated, it is not possible to obtain analytical results for any models; then continuous-time models have to be solved numerically, which involves discretizing time anyways.

While most discrete-time models can be compared to continuous-time models in the limit of small time steps, there are properties of discrete-time models that differ fundamentally from continuous-time models. The possibility of chaotic dynamics is one property of some discrete-time models that does not translate into continuous-time models. Whether chaotic dynamics are biologically meaningful is a topic of open debate. As ever, whichever formulation is used, one should be judicious in interpreting model results and always wary of artefacts (i.e. model behaviors that arise from quirks of the model design rather than from the basic biological processes under study). One common guideline to guard against artefactual results from discrete-time models is to repeat your model simulations with time-steps that are twice as long and half as long. If most of your work uses a time-step of 1 day, then check whether the model gives wildly different results using time-steps of 2 days or half a day. If it does, then examine the results closely to ensure that your choice of time-step – which is to some degree arbitrary – is not unduly influencing your results.

3.5.3. *Residence time distributions.* A related topic concerns the probability distributions of so-called 'residence times' that are implied by the model. The residence time refers to the duration that individuals in the model spend in a given compartment. In continuous-time models, a constant per capita rate of leaving a compartment gives rise to an exponentially distributed residence time, e.g. in equation 1, the constant recovery rate implies that the duration that individuals remain infectious is exponentially distributed. Analogously, in discrete-time models with constant probabilities per time step of leaving a compartment, the residence time is geometrically distributed.

For example, the constant recovery rate in equation 1 implies that the distribution of infectious periods in the population is exponential (and hence the most common infectious period duration approaches zero!), whereas data from real infections invariably show that such infectious period distributions are peaked at non-zero values. Numerous studies over the years have shown that modelling the incubation and infectious period distributions inaccurately can introduce undesirable biases into model results (Keeling & Grenfell, 1998; Lloyd, 2001; Wearing et al., 2005). Fortunately there is a useful trick to adapt the ordinary differential equation framework to reflect more realistic residence time distributions. This is known informally as the 'box-car' model. In this box-car model, a particular compartment is broken into a chain of sub-compartments, e.g. the E compartment could be broken into $E_1, E_2, \ldots, E_n$, as shown in figure 3. This changes the distribution of residence times from exponential in the usual framework to the gamma distribution, which can adopt a range of shapes depending on the number of sub-compartments included in the box-car model (Wearing et al., 2005).

3.5.4. *Continuous versus discrete state variables.* State variables are continuous in the models described by differential equations, in the sense that the state variables (such as the number of infected individuals) can take any positive real value. While these models have the advantage of being tractable mathematically, their precise biological interpretation is often vague. Sometimes the state variables are defined as densities, so non-integer values are acceptable. In other cases the state variables are meant to describe numbers of individuals, with the hope that the numbers in most compartments are sufficiently high that 'fractional individuals' do not exert an important influence on dynamics. This assumption can lead to serious

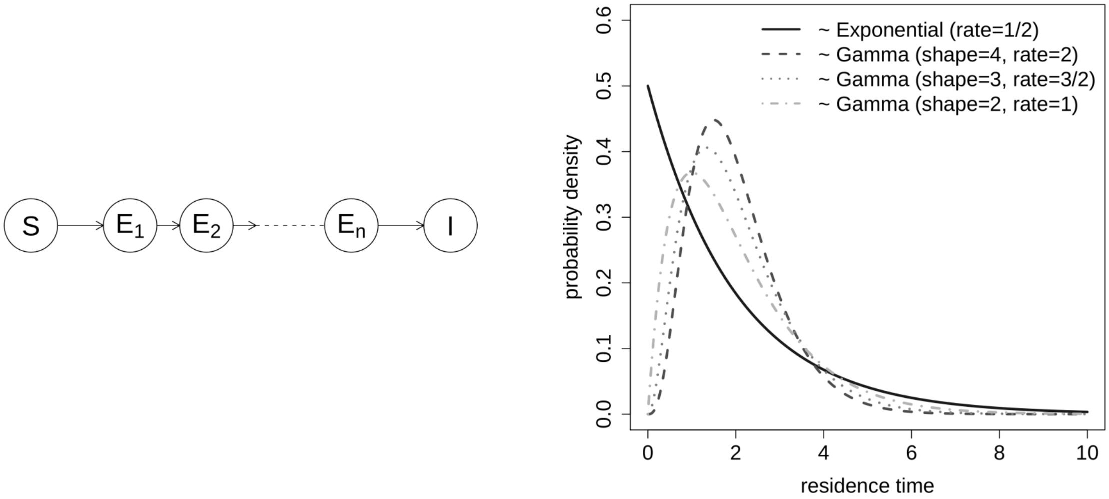


FIGURE 3. In a 'box-car' model, individuals move through a chain of sub-compartments to generate a more realistic distribution of residence times, e.g. for the Exposed compartment. If the rate of moving between sub-compartments is constant, then the total residence time in the Exposed compartment is gamma distributed. The gamma distribution can assume a range of forms depending its 'shape' parameter, which corresponds to the number of sub-compartments in the box-car model. By varying the 'rate' parameter for movement between sub-compartments, one can ensure that all of the distributions have the same mean (in the example shown, this is 2 time units). Note that the exponential distribution is actually a special case of the gamma distribution – in the example shown it is gamma (shape=1, rate=1/2).

problems, however, as famously exemplified by the "atto-fox" problem identified by Mollison (1991), wherein infinitesimal fractions of infected foxes allowed rabies to persist between epidemics in a differential equation model. It is thus crucial to recognize situations when key state variables may approach very low values during your system's dynamics, and to design your model accordingly. One convenient work-around is to impose a so-called quasi-extinction threshold, which asserts that a continuous-valued state variable in a deterministic model goes extinct if its value passes below some arbitrary lower bound. Discrete state variables, which count individuals in integer units, arise naturally in many stochastic models. These models are sometimes less amenable to mathematical analysis, but they are also safe from yielding artefactual results.

3.5.5. *Models for population structure and heterogeneities.* Simple compartmental models make two other important assumptions. These models assume that each individual in a compartment is identical to any other (i.e. that the populations are homogeneous), and that all individuals in the modeled population mix randomly

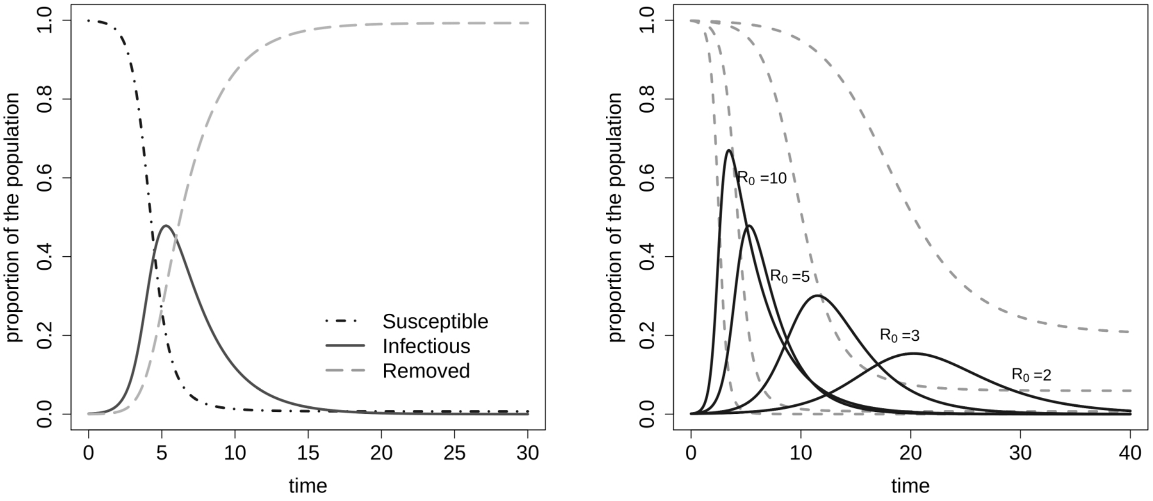


Figure 4. Epidemic curves for SIR model with frequency-dependent transmission. (Left) Typical population level epidemic pattern. The parameter values used here are $\beta = 2$ and $\gamma = 0.4$, which gives $R_0 = 5$. (Right) Epidemic curves for different values of R_0.

and completely. It is increasingly recognized that for almost all real-world populations, neither of these assumptions is valid, so they need to be considered carefully when formulating a disease model. 'Population structure' refers to non-random patterns of mixing in a population, due to factors such as spatial separation or social contact patterns. 'Population heterogeneity' refers to situations where basic properties such as infectiousness or contact rate vary among individuals in a population. There are numerous frameworks for modeling population structure and heterogeneity with varying degrees of complexity; these will be discussed in section 5.

3.6. Basic analysis of SIR-type compartmental model. The SIR family of models has been studied in great detail, and many analytic results can be derived from simple ordinary differential equation formulations. Here we summarize a few expressions that can be derived from the basic models, to give a sense of what is possible. For details we refer the reader to the classic treatment of these analyses given by Anderson & May (1991). For the simple deterministic SIR model without host demographics, diseases with $R_0 > 1$ will initiate an epidemic, which eventually burns itself out because the pool of susceptible individuals is depleted. Figure 4 shows graphs of epidemic curves generated by the SIR model with frequency-dependent transmission described by equation 1. In this case, the **exponential growth rate** at the outset of an epidemic is given by $r = (R_0 - 1)/D$, where D is the mean duration of infectiousness. The fraction of hosts infected when the epidemic reaches its peak, or **peak prevalence**, is given by $I_{\max} = 1 - (1 + \ln R_0)/R_0$.

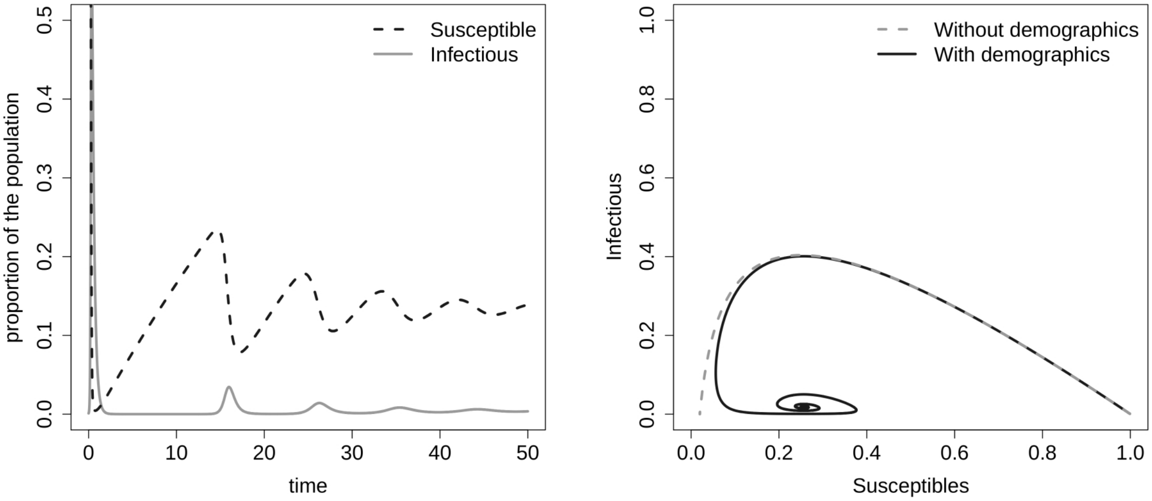


Figure 5. (Left) Typical population level epidemic pattern for SIR model with demographics. Introduction of demographics allows persistence and introduces periodicity due to repeated cyclic epidemics. (Right) S-I phase plane with phase diagram for both SIR models, with and without demographics.

The **final proportion susceptible** in the host population, after the epidemic has run its course and burned out, is given by $f = \exp(-R_0(1-f)) \approx \exp(R_0)$.

Introduction of host demographics changes the model as shown in equation 4, and can fundamentally change the dynamics exhibited by the system.

$$\begin{aligned} \frac{dS}{dt} &= \mu\, N - \frac{\beta\, S\, I}{N} - \mu\, S \\ \frac{dI}{dt} &= \frac{\beta\, S\, I}{N} - \gamma\, I - \mu\, I \\ \frac{dR}{dt} &= \gamma\, I - \mu\, R \end{aligned} \tag{4}$$

Here we have introduced host births and deaths, which both occur at per capita rate μ so the population size remains constant. Depending on parameter values, this model can now exhibit endemic persistence, and can give rise to periodicity in incidence that was not present in the simpler model without demographics (Figure 5). The period of the cycles can be calculated roughly, $T \approx 2\,\pi\,(A\,D)^{1/2}$, where A is the mean age of infection, and D is the mean disease generation interval Anderson & May (1991). The reproductive number, $R_0 = \beta/(\gamma+\mu)$. Intuitively, this is because the average duration an individual host will remain infected will be shortened to $1/(\gamma+\mu)$ from $1/\gamma$, which was the case in the simpler model described by (1). One can find a more thorough and general mathematical treatment in van den Driesche & Watmough (2002).

3.7. Roles and types of data. Mathematical modelling of infectious disease dynamics is a scientific endeavor aimed at improving our understanding of real biological systems, so it is desirable to relate our models to reality by incorporating data whenever possible. Broadly speaking, there are two primary uses of data: to construct and parameterize models, and to validate model output.

When considering what data are needed to construct a model, one can delineate three rough categories. The first category includes data describing the progression of infection at the scale of individual hosts, including parameters describing clinical progression of the infection such as the infectious period, the latency period and the duration and efficacy of immunity once it is attained. For some pathogens, these data are readily available from the medical or veterinary literature, although fine points (such as how infectiousness varies through time) may be unknown. For many pathogens, particularly those circulating in wildlife hosts or newly emerged in human populations, even these basic data aren't available.

The second category of data entails information about the host population in the absence of disease, such as the size and structure of the host population, basic demographic characteristics such as rates of birth, death, and migration, and rates of contacts between people in a community and between communities. These types of data are very useful in designing epidemic models, particularly with regard to population structure and contact patterns.

The third category of data describes epidemiological patterns of disease circulation in the host population. These data can take the form of time series of estimated incidence or prevalence, or point estimates of quantities such as prevalence or seroprevalence. The data may be structured by host age or spatial location. Occasionally the data will be very detailed, such as contact tracing reports of transmission chains, or household studies that track generations of transmission among household members.

Epidemiological data can be used either to construct or to validate a model. If there are significant unknowns remaining after other classes of data have been used, then fitting the model to epidemiological data is an excellent way to estimate values for remaining free parameters. The most fundamental quantity is the transmissibility of the pathogen (i.e. R_0), which is rarely directly observable but can sometimes be estimated indirectly from other quantities (see section 7.1). Other aspects of transmissibility such as its seasonality or density dependence can also be estimated. If it is possible to construct and parameterize the model without using all the data that are available, then there is an opportunity to validate the model by making appropriate predictions and comparing them to data that were not used in model construction. If the model predictions are relatively close to these 'out-of-sample' data, then there are grounds for much greater confidence that the model is an accurate depiction of the system dynamics. This is particularly true (though rare in practice) if the model is able to predict responses to perturbations that were not included in the training data.

It is important to note that individual, population and epidemiological data are widely available, in the scientific literature and elsewhere. The websites for The World Health Organization (WHO), and The Center for Disease Control (CDC) are excellent places to search for epidemiological data. Two recent books full of data on important global health problems can be downloaded for free from the following websites: `http://www.dcp2.org/pubs/GBD` and `http://www.dcp2.org/pubs/DCP`.

4. Incidence function and population threshold

4.1. Incidence rate. The incidence rate is the rate at which new infections arise in a host population. Since the production of new infectious individuals drives the dynamics of infectious disease spread, the formulation of the incidence term is of central importance to epidemic modelling. Formulation of the incidence rate will reflect the biological and social processes inherent in disease transmission, as well as the assumptions made by the modeler. It is typically formulated as a product of (i) the number of susceptible individuals in the population, S, and (ii) the force of infection experienced by each susceptible, λ. The force of infection, in turn, constitutes of a number of other factors, depending on the transmission mechanism of the disease and the structure of the host population. In the most basic terms, the force of infection can be expressed as $\lambda = c(N)\, p\, I/N$, where $c(N)$ is the per capita rate of contacting other hosts (with an arbitrary dependence on the population size, N); p is the probability of transmission given that contact with an infectious host has occurred; and I/N is the probability that a randomly chosen host is infectious. Hence, the incidence rate can be expressed as:

$$f(S,I) = c(N)\, p\, \frac{S\, I}{N} \tag{5}$$

The contact rate $c(N)$ is commonly modeled in two different ways. Under **density-dependent transmission**, the contact rate is assumed to depend linearly on the host population size, i.e. $c(N) = k\, N$. Then the incidence rate is $f(S,I) = k\, N\, p\, S\, I/N = \beta_{\mathrm{DD}}\, S\, I$, where $\beta_{\mathrm{DD}} = k\, p$. In contrast, under **frequency-dependent transmission** the contact rate is assumed to be constant with respect to the host population, i.e. $c(N) = c_0$. In this case, the incidence rate is $f(S,I) = c_0\, p\, S\, I/N = \beta_{\mathrm{FD}}\, S\, I/N$, where $\beta_{\mathrm{FD}} = c_0\, p$. The frequency-dependent model is sometimes also called the standard incidence. The density-dependent model used to be called "mass action", in reference to the model from chemistry which was the original inspiration for early disease models, but since the mid 1990s the terminology has become confused with some authors calling the frequency-dependent model "mass action" (McCallum et al., 2001).

There is an ongoing debate about which transmission model is a more realistic representation of disease spread in human and natural populations. Classically, it was assumed that the transmission rate increases with population size, because contacts increase with crowding. For that reason, the density-dependent model was dominant until the 1980s. However, it has since been argued that many modes of contact (e.g. sexual contact) are governed more by social norms than by density, favoring use of frequency-dependent transmission (Hethcote & Van Ark, 1987). In recent years, evidence has been accumulating from many human and animal studies that even for infections transmitted by casual contact there is little evidence for clear density dependence in transmission rates (Bjørnstad et al., 2002; Begon et al., 1999).

Putting aside the debate, it is always worthwhile to think carefully about the natural history of the disease-host interaction — with particular attention to the mechanism of transmission — in formulating the incidence term. As a general rule frequency-dependent transmission is more appropriate in a large well-mixed population, whereas density-dependent transmission probably applies in smaller populations where individuals' contact budgets may not be saturated. A more realistic general model may be one which links these two regimes via a simple saturating function (Antonovics et al., 1995). It is important to realize the shortcomings and limitations of all these simple transmission models, and to think about the underlying population structure and mechanisms of mixing at the temporal and spatial scales which apply for the pathogen being considered. In many circumstances an explicit population structure is appropriate to capture the relevant mechanisms (see section 5). We refer the reader to McCallum et al. (2001) for further discussion regarding the choice of the transmission model and related issues.

4.2. Population thresholds in epidemic dynamics.

4.2.1. *Population threshold for invasion.* The basic reproduction number, R_0, has been the central concept in epidemic dynamics since the early 1980s, thanks largely to the work by Anderson and May (Heesterbeek, 2002). Long before this, people studying epidemic dynamics have focused on population thresholds. In their classic study, Kermack & McKendrick (1927) focused on the **population threshold for invasion**, which is defined as the host population size below which the parasite cannot invade. This is readily derived from the density-dependent transmission model, in which $R_0 = \beta\, N\, D$ where β is the rate of transmission, N is the size of the host population, and, D is the average duration of an infection. Because R_0 is positively related to N, an increase in the host population size increases R_0 and reduces the probability of extinction of a pathogen in a host population. In particular, since a pathogen can only successfully invade when $R_0 > 1$, this implies that there is a threshold population size N_T, such that the pathogen has a positive probability of successful invasion only if the host population $N > N_T$. This N_T can be thought of as the population threshold for invasion. Note that this phenomenon is not observed in the model with frequency-dependent transmission, where $R_0 = \beta\, D$ and since R_0 is not a function of N there is no population threshold for invasion.

Apart from looking at the population threshold in a fully susceptible host population, one can also examine a similar threshold for a population that is only partially susceptible. As introduced in section 3.2, $R_{\text{effective}} = R_0\, S/N$, where S and N are the susceptible and total host populations, respectively. An epidemic is possible in a population if $R_{\text{effective}} > 1$. This translates into a threshold condition on the fraction of the population that is susceptible, i.e., the pathogen can invade only if $S/N > 1/R_0$. This susceptibility threshold for disease invasion is applicable to either density-dependent or frequency-dependent transmission models. Conversely, the pathogen fails to survive in a host population when $S/N \leq 1/R_0$, and this is the basis for herd immunity as described above.

Despite its conceptual simplicity, real-world evidence for invasion thresholds is hard to find, for two major reasons (Lloyd-Smith et al., 2005a). First, while we can gather various kinds of epidemiological data for a successful disease invasions, it is very difficult to gather data for unsuccessful pathogen invasion. Failure of a disease invasion implies that the disease did not spread, and such an event usually does not get noticed. Thus there is an inherent bias in the observation process. Second, even with perfect observation, demographic stochasticity leads to substantial variation in outbreak sizes which can obscure the distinction between populations below and above the invasion threshold. Consider the stochastic simulation results shown in figure 6, which show the distribution of outbreak sizes following introduction of a single infected individual into populations of various sizes, with various values of

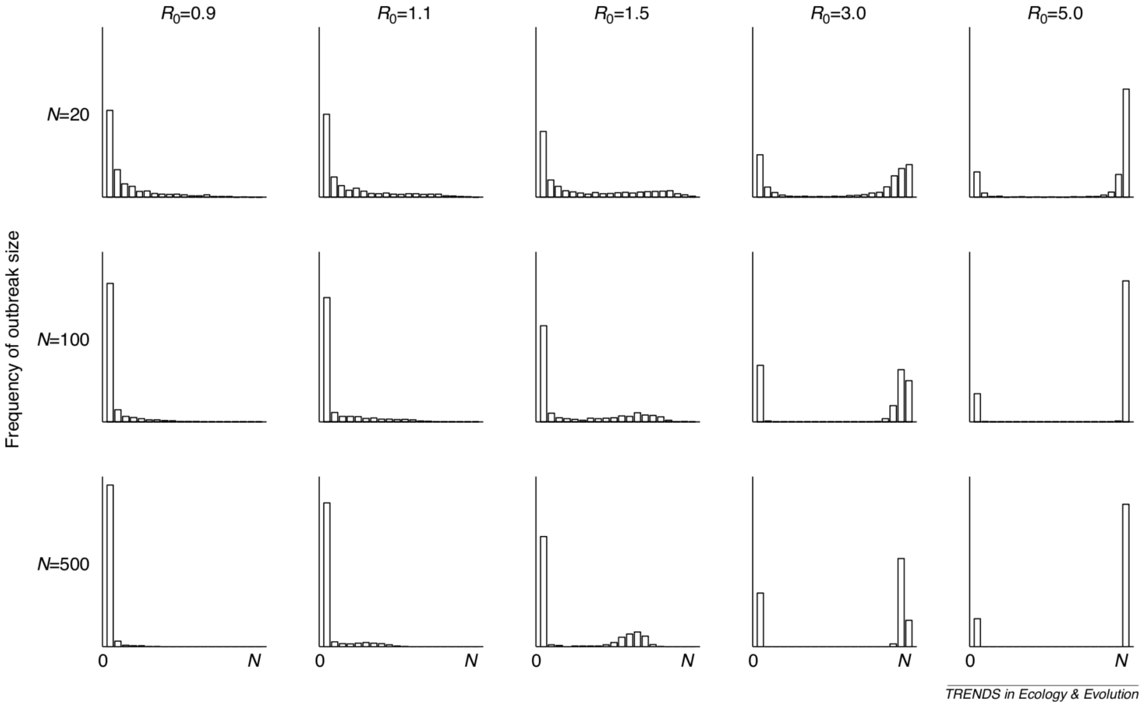


Figure 6. Histograms showing the distribution of total outbreak size for simulated stochastic outbreaks in populations of three sizes under five values of R_0. Figure reproduced from Lloyd-Smith et al (2005).

R_0. Higher population sizes and higher values of R_0 lead to relatively clear bimodal distributions, where success versus failure of invasion can be clearly delineated. In contrast, small N or $R_0 \approx 1$ lead to distributions with density throughout the range of possible outbreak sizes. For example, compare the outbreak size distributions for populations of 100 hosts, and R_0 values of 0.9 and 1.1. Here the two values of R_0 are on the two sides of the threshold of $R_0 = 1$, so in principle outbreaks are possible for one and not for the other. Yet limited chains of transmission can still occur for $R_0 < 1$, while epidemics can still die out by chance when $R_0 > 1$, and the resulting distributions of outbreak size look very similar. Considering these intrinsic challenges in light of the substantial logistical problems in completely observing the outcome of many replicate disease introductions, we can see why clear empirical demonstrations of threshold population sizes for disease invasion are so rare.

4.2.2. *Population threshold for persistence.* A disease that successfully invades a host population is not necessarily guaranteed to persist long-term. There are two broad mechanisms whereby an established disease can fail to persist, or fade out. **Endemic fadeout** can occur when a disease is prevalent in the host population at a relatively low endemic level, but then the transmission chain is broken (i.e. the number of infected hosts goes to zero) due to a stochastic fluctuation. **Epidemic fadeout** occurs following a major epidemic, when the pool of susceptible individuals is depleted and the pathogen runs out of new individuals to infect.

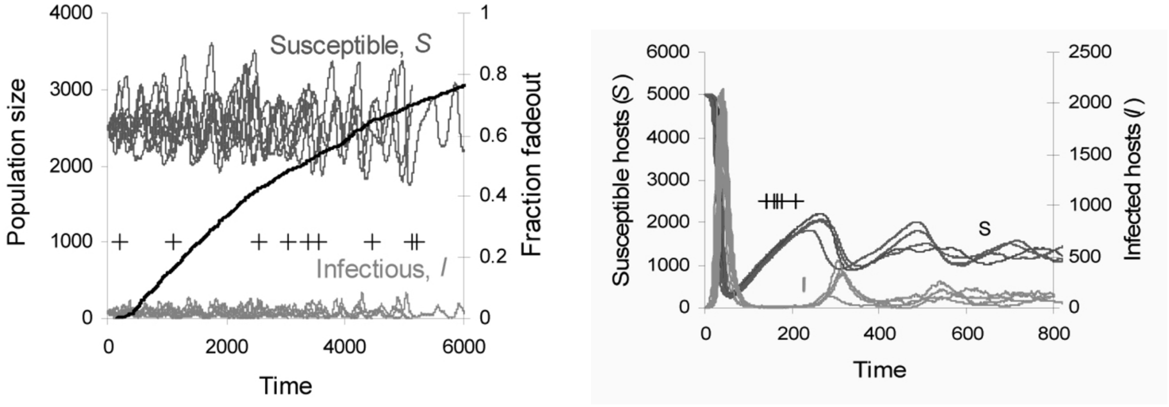


Figure 7. Endemic (left) and epidemic (right) fadeouts for a stochastic SIR model with frequency-dependent transmission and $R_0 = 4$. The graphs plot the number of susceptibles (in dark grey) and number of infecteds (in light grey) in 10 simulations. + signs show times when the disease fades out in particular simulations. The bold black line in the left graph indicates the fraction of 1000 simulations that had faded out as a function of time. Figure reproduced from Lloyd-Smith et al. (2005a).

Epidemic fadeout typically occurs for highly infectious pathogens which exhibit strongly overcompensatory dynamics with very dramatic outbreaks and hence deep post-epidemic troughs in the susceptible population. Fig 7 illustrates endemic and epidemic fadeouts for a stochastic SIR model.

Measles is the classic example of a pathogen for which epidemic fadeouts are relevant, although the phenomenon certainly applies to other acute immunizing infections. In his seminal work, Bartlett (1957) observed that measles frequently disappeared from smaller towns in Britain, but circulated persistently in larger cities (note that this work was conducted before the measles vaccine was available). By analyzing the likelihood of fadeouts as a function of the population of the towns or cities, Bartlett concluded that there was a **critical community size** required for long-term persistence of the virus in a given population. This phenomenon has since been corroborated and analyzed in great detail (Anderson & May, 1991; Keeling & Grenfell, 1997; Bjørnstad et al., 2002; Grenfell et al., 2002). It is important to note, though, that the critical community size is not simply a fixed population size, but arises from a complex interaction between population size, demographic rates, and epidemiological properties of the disease (Lloyd-Smith et al., 2005a). Endemic fadeouts have proved more difficult to identify in empirical data, but Nåsell (2005) has developed important theory for defining and estimating such a threshold.

5. Population structure, heterogeneity and mixing

The simple models we have discussed so far have assumed that each individual host in a chosen compartment is exactly the same as all the others, neglecting variation in host genetics and condition, patterns of interaction with other individuals, and spatial location among many other factors. These assumptions are bound to have significant bearing on the results we obtain from the model. In this section, we will attempt to lay out several ways in which we can relax these assumptions, and discuss the consequences of such extensions. We can roughly group these methods into two very broad categories—methods that focus on population heterogeneity and those that focus on population structure. **Heterogeneity** refers to differences among individuals or groups in a given population, for instance due to different occupations or levels of nutrition. By **population structure** we mean the deviations in the mixing patterns of the population from the standard random mixing assumption in the SIR model. These deviations could be due to spatial proximity (you may be likely to interact with another host living nearby, but much less likely with someone in a different continent) or social factors (you are more likely to mingle co-workers than with strangers, regardless of where you live). This nomenclature is somewhat arbitrary and can be somewhat confusing — for instance models that include heterogeneity in host age are called 'age-structured', while models where parameters vary through space may be called 'spatially heterogeneous'.

5.1. Models for population structure. The most simple models for population structure, implicit in differential equations of the simple SIR-type models, are **random mixing** or **mean-field** models. In these models, every individual in the population is assumed to have equal probability of contacting any other individual. As we have discovered, they are mathematically tractable in the simple mass-action formulations analogous to mixing of small particles in chemistry models. The number of contacts that result in transmission simplifies to the product of the number of susceptible and number of infected individuals, and hence the incidence rate, $f(S, I) = \beta\, S\, I$. Sometimes, this form is modified to $\beta\, S^a\, I^b$ by including exponents as a phenomenological representation of non-random mixing. The dynamical properties of the resulting system is also more rich (Lui et al., 1987), and one needs to be cognizant of unintended consequences while adopting this technique. See Koelle et al. (2005) for an example of application of this technique to model cholera dynamics.

Multi-group, **multi-patch** or **metapopulation** models incorporate differences in host mixing behavior in a more mechanistic way, by dividing the host population into multiple groups based on the spatial location (or some other attribute that is expected to differentiate host mixing, such as social groups). The transmission function resulting from this mixing can be modeled in a couple of different ways. The most common approach is to construct a matrix that specifies the rate of transmission between individuals of any two chosen groups — often called the 'who acquires infection from whom' or WAIFW matrix. For a multi-group model with n groups, one constructs a $n \times n$ matrix whose component β_{ij} gives the transmission rate from an individual in group i to group j. β_{ii} is taken to be the transmission rate between two individual hosts within group i. Hence, the force of infection on a susceptible host in group j is:

$$\sum_i \beta_{ij}\, \frac{I_i}{N_i}.$$

See Anderson & May (1991) or Keeling & Grenfell (1997) for examples of the construction and use of WAIFW matrices. Here the authors subdivide the host population into groups based on their age — age being an important criterion for mixing in relation to measles epidemic.

An alternative approach to modelling transmission in multi-group models is to assume that transmission occurs only within groups (often using a random mixing assumption within the group), but to explicitly model movements of individual hosts between groups. The advantage of this approach is that the stochasticity involved in between-group movements can be captured, which is important if the groups are small or movement events are rare (see section 5.3.1).

Metapopulation models are particularly useful to study spatial patterns. To explore spatial dynamics of raccoon rabies epidemic in Connecticut, for example, Smith et al. (2002) divide the host population into 169 spatial groups corresponding to townships. By constructing a transmission matrix, the authors model both transmission between adjacent townships and longer-distance translocations of infected hosts, and as a result are able to quantify the roles of human habitats and rivers in the spatial spread of the disease.

For some disease systems (or research questions), it is important to consider the fine structure in interactions among hosts, or to capture the reality of persistent relationships between host individuals. As an example, for STDs, the number and identity of sexual partners for each infected host — which tend to vary substantially from one individual to another — will play a fundamental role in the spread of the disease. **Network models**, based on a branch of mathematics called graph theory, are well-suited to address problems where the details or 'memory' in the host contact structure are important. A network, or a 'graph', is a collection of nodes and edges, where nodes represent individual hosts and edges between two nodes represent contact between those hosts. When appropriate, an edge can have a weight and a direction to represent the intensity of the contact and the direction of the contact. This information is captured in an adjacency matrix, A, where values of a component, A_{ij}, denote the strength of contact between hosts i and j. By tracing the edges starting from a node, referred to as the index case, one can find which other hosts can be infected in the network. This type of contact tracing can be useful for assessing the risk of and to an individual host and for designing control strategies. There are several basic mathematical tools that have been developed to study different properties of these networks. These include the degree distribution, which the distribution of number of edges per node, and the clustering coefficient, which for a given node is the number of neighboring nodes that also share edges between themselves. Finally, it is important to realize that, despite their complexity, many network models make the very strong assumption that the contact structure is static, which can have the effect of artificially limited disease spread. There is increasing interest in dynamics network models which allow for changes in the network structure over time. We refer readers to Newman (2002), Keeling & Eames (2005) and Bansal et al. (2007) for reviews of network models in epidemiology.

Individual-based models or **microsimulation models** go even further in capturing the differences between individuals. Every individual in the model carries its own attributes, such as age, sex, location, contact behavior, or any other property that is deemed important. This extremely flexible framework allows one to represent arbitrarily complex systems and ask detailed questions. But this additional complexity comes at a high price. The dynamics of the model are simulated stochastically, often requiring major computational facilities. Results from model simulations can be difficult to reproduce and understand, and lack the clear cause-and-effect relationships of simpler models. Also, for each parameter that the modeler incorporates, he/she will then need to figure out what values they take, and how it interacts with others in the model. Often there are not sufficiently detailed data to estimate all these parameters, and sensitivity analyses can be burdensome. STDSIM is a famous example of this type of model used to study transmission and control of STDs including HIV in East Africa.

5.2. Models for heterogeneity. Incorporation of population heterogeneity into epidemic models has been a major topic of research for decades, motivated by both the clear influence of heterogeneity on patterns of disease spread and the potential to take advantage of population differences to design targeted control measures. The topic is treated in depth elsewhere (Anderson & May, 1991; Keeling & Rohani, 2007; Lloyd-Smith et al., 2006; Diekmann & Heesterbeek, 2000; Becker & Marschner, 1990); here we will briefly summarize a few key concepts.

5.2.1. *Group-level heterogeneity.* The most common approach to modelling heterogeneity is to divide the population into several sub-groups, which have some differences in parameter values but are themselves homogeneous. The model can then be analyzed using well-developed methods related to the metapopulation approaches described above. Again we can generalize R_0 for a multi-group population by defining R_{ij} as the expected number of new infections caused in group j due to the infected individuals in group i. Under the assumption that group membership is static, $R_{ij} = D_i \, \beta_{ij}$ where D_i is the expected infectious period, spent entirely in group i, and β_{ij} is the transmission rate from group i to group j. The value of R_0

If we assume that hosts can move from one group to another but transmission can occur only locally, then we can define D_{ij} to be the expected time spent in group j, while still infectious, by an individual infected in group i, and β_j to be the transmission rate within group j, so that $R_{ij} = D_{ij} \, \beta_j$. This helps us construct a $n \times n$ matrix of reproductive numbers for an n-group model. If the movement rules are Markovian, then the process can be defined by an absorbing Markov chain, with overall transition matrix:

$$\begin{bmatrix} P_{n\times n} & m \\ 0 & 1 \end{bmatrix},$$

where p_{ij} is the per-timestep probability that an individual moves from group i to group j, and m_j is the probability that an infected individual will recover or die while in group j. The expected residence times D_{ij} are then given by the fundamental matrix $\mathbf{D} = (\mathbf{I} - \mathbf{P})^{\mathbf{-1}}$, and so $R_{ij} = D_{ij} \, \beta_j$. These arguments are generalized, and further conclusions are derived about the invasion dynamics of a disease in a heterogeneous population with movement among groups, in Schreiber & Lloyd-Smith (in press).

5.2.2. *Individual heterogeneity and superspreaders.* Individuals differ from one another in many ways, but from the perspective of a disease modeler, the most important heterogeneities are those that affect disease transmission. There are many accounts, for many diseases, of 'superspreader' individuals who are responsible for much more transmission than is typical for the disease (Lloyd-Smith et al., 2005b). For sexually-transmitted and vector-borne infections, this phenomenon has been attributed to the long-recognized fact that contact rates (i.e. contacts with sexual partners or bites by vectors) differ greatly across individuals. Indeed, Woolhouse et al. (1997) analyzed contact rate data for a suite of sexually-transmitted

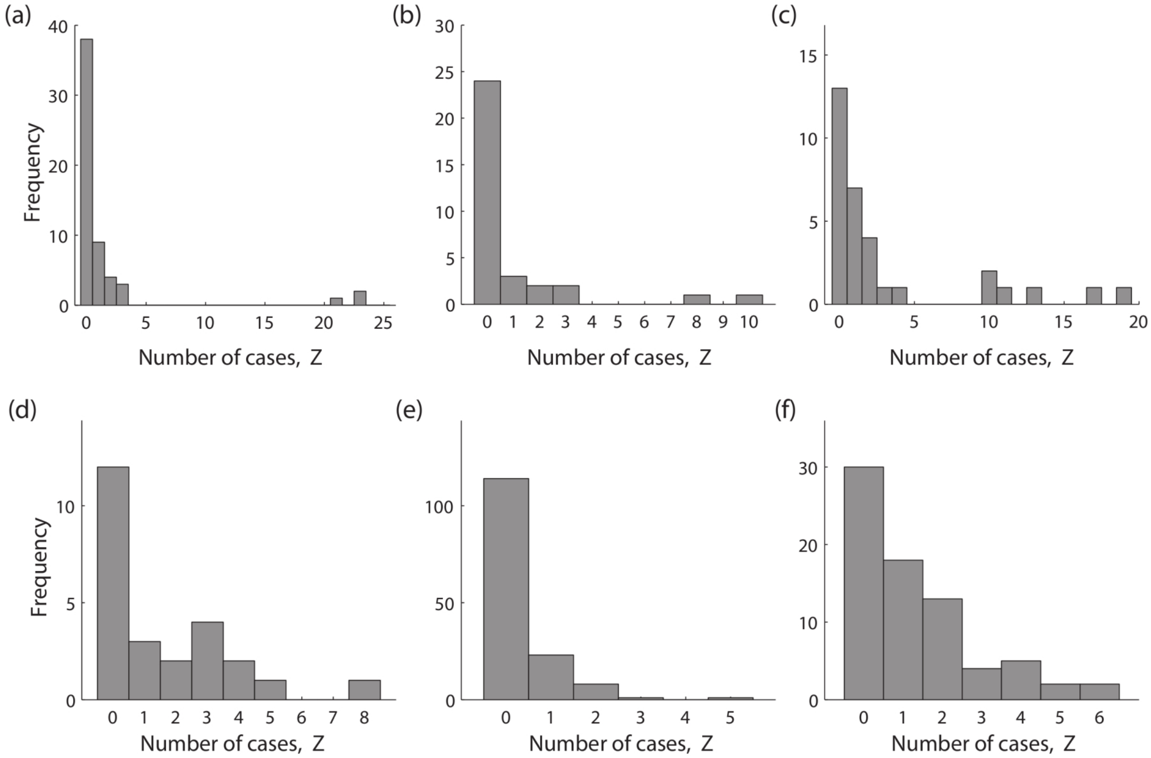


Figure 8. Data from contact tracing studies highlight the substantial individual-level heterogeneity in infectiousness. The variability in the number of secondary cases caused by a single infected individual is shown for various infectious diseases: (a) SARS in Singapore; (b) SARS in Beijing; (c) Smallpox; (d) Variola minor; (e) Monkeypox; and (f) Pneumonic plague.

and vector-borne infections, and proposed that they are described by a 20/80 rule which states that 20% of hosts are responsible for 80% of the net transmission potential. This approach cannot be applied for many directly-transmitted infections, however, because it is not possible to quantify contact rates (e.g. for influenza or SARS or any other infections transmitted in casual contact, contacts are not discrete and identifiable events). Furthermore, for all infections, many other factors contribute to a host's infectiousness – including properties of the host, pathogen and environment – so we expect the degree of infectiousness to be distributed continuously in a population, and not identifiable *a priori*. For all of these reasons, the conventional group-based approach to modelling population heterogeneity is not well-suited to the challenge of capturing individual variation in infectiousness.

One approach to this challenge was presented by Lloyd-Smith et al. (2005b), who proposed an extension of the idea of the basic reproductive number, R_0, to a population distribution. They defined the individual reproductive number, ν, as the expected number of infections caused by a particular infectious individual in a susceptible population. This quantity is allowed to vary among individuals in the population according to any continuous distribution with a mean of R_0. The actual number of cases caused by a particular individual, Z, is determined by this expectation filtered by demographic stochasticity in the transmission process, and under simple assumptions is given by $Z \sim \text{Poisson}(\nu)$. Lloyd-Smith et al. propose a flexible model wherein ν follows a gamma distribution, so that Z follows a negative binomial distribution. By comparing this model with data from real outbreaks on the distribution of the number of secondary cases resulting from single infected individuals (Figure 8), they were able to estimate the distribution of ν for several infections. The 20/80 rule did not hold as a generality, but for all diseases there was a substantial degree of individual heterogeneity, and it was shown that the standard models assuming homogeneous infectiousness are statistically inconsistent with many of the available data sets. Their analysis went on to analyze the stochastic dynamics of disease outbreaks using the branching process framework introduced in section 3.5.1, showing that increased heterogeneity leads to a higher probability that a disease outbreak will go extinct, but that those outbreaks that do succeed will grow faster in their early stages.

A second approach to modelling superspreading events was suggested by James et al. (2006). In this model infections occur in two ways: ‘normal’ infections can be caused by all individuals according to a branching process with Z following a Poisson distribution, and superspreading events occur as a Poisson process with rate ρ. Each superspreading event is assumed to cause a number of infections that is Poisson distributed with mean λ. The authors show that this model can fit some data sets about as well as the model proposed by Lloyd-Smith et al. (2005b). Closer examination of epidemiological reports indicates that reality probably falls somewhere between the two models, with heterogeneity in infectiousness arising sometimes from individual differences in transmission rate, and sometimes from events that occur unpredictably.

5.3. The importance of mechanism. In formulating an epidemic model, it is essential to think carefully about the mechanisms of contact underlying transmission for the system at hand. Sometimes these considerations can lead to greater complexity in your model, but for particular scenarios it is often possible to reduce this complexity by working in various limits. As a rule, one needs to consider the timescales of the relevant processes, such as mixing, recovery and transmission, to determine what limits are applicable. In situations where these timescales have similar magnitudes, it is often important to use mechanistic models to capture the dynamics accurately.

5.3.1. *Contact between groups.* For populations structured into groups, transmission occurs within groups and large-scale spread arises when infected individuals move between groups. As described in section 5.1, this situation can be modelled by explicitly tracking the movement of individuals or by introducing a ‘between-group’ transmission rate that implicitly describes both movements and transmission in other groups. Keeling & Rohani (2002) show that, when movements between groups are rapid, these two approaches give similar results. Subsequent work by Cross et al. (2005), however, shows that this similarity breaks down as the movements between groups become less frequent. In particular, Cross et al. examine how the timescales of host movement and disease recovery interact to determine invasion success in structured populations. They show that it is essential to consider the discrete and stochastic nature of movement for systems where group sizes are small and between-group moves occur on the same (or slower) timescales as disease recovery. That is, if the expected number of between-group moves by a host during its infectious period is 1 or lower, then a widespread disease outbreak is unlikely even if the local $R_0 \gg 1$ (Figure 9). Thus we expect acute and chronic diseases to exhibit different invasion dynamics in a structured population, even if they have the same R_0.

So we see that R_0 can fail to properly predict the invasion threshold for a structured population. For a metapopulation, it is then appropriate to develop a similar threshold quantity. Ball et al. (1997) introduce a quantity R_*, which is defined to be the expected number of groups infected by the first infected group, and present elegant analyses of this quantity in systems with large numbers of groups and without tracking host movements. Analytical expressions for R_0 or R_* are hard to find for systems with mechanistic movement, finite group sizes and finite numbers of groups, so Cross et al. (2005) tested the applicability of the R_* concept to their model by computing 'empirical' values, $\hat{R}_0$ and $\hat{R}_*$, from their simulation results. Figure 10 shows how the two quantities compare in their ability to predict invasion in a structured population. A system can have $\hat{R}_0 \gg 1$ and not experience a major outbreak, because movements between groups are too infrequent, but if $\hat{R}_*$ is greater than about 2 then a major outbreak is very likely. Clearly R_* is a better predictor of invasion.

We have summarized one study of the complexities arising in group-structured populations, but there is an extensive literature addressing these issues. We refer the reader to papers by Ball et al. (1997) and Ball & Neal (2002) for further details on mixing at different levels, and to Cross et al. (2004, 2005, 2007) for further analysis on disease invasion in structured populations.

5.3.2. *Pair formation and STD transmission.* Sexually-transmitted diseases are often modeled using ordinary differential equations with a frequency-dependent incidence rate. Following the presentation introduced in section 4.1, the incidence rate, $f(S, I)$ is given by the following function:

$$f(S, I) = c_{\mathrm{FD}}\, p_{\mathrm{FD}} \left(\frac{S}{N}\right) I, \tag{6}$$

where c_{FD} is the rate of acquiring new partners, p_{FD} is the probability of transmission in S-I partnership, S/N is the probability that a partner is susceptible, and I is the abundance of infected hosts. But the derivation of the incidence rate given earlier was based upon random mixing in a population, whereas most sexual transmission takes place within monogamous pairs of varying duration. Lloyd-Smith et al. (2004) examined the underlying dynamics of pair formation and dissolution to understand how this basic mechanism related to the common frequency-dependent formulation of incidence for these diseases.

Consider a population of host individuals that can be either susceptible, X_{S} or infected, X_{I}. These individuals form pairs $P_{\mathrm{SS}}, P_{\mathrm{SI}}$,and P_{II} where the subscripts indicate the infection status of the individuals in the pair. Pairs form at per capita rate k and dissolve at per-pair rate l. For generality, and to incorporate the observation that infected individuals often exhibit different contact behavior Lloyd-Smith et al. (2004), we let k_S and k_I be the pairing rates for susceptible and infected hosts, and l_{SS}, l_{SI}, l_{IS},and l_{II} be the dissolution rates for pairs with different compositions. We introduce m_{SS}, m_{SI}, m_{IS}, and m_{SS} as the mixing rates between individuals of different categories.

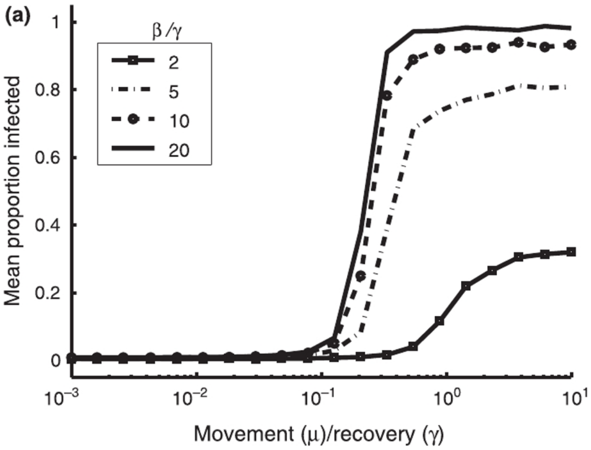


FIGURE 9. The interaction of movement rate (μ) and recovery rate (γ) determines the mean proportion of a structured population that becomes infected in an outbreak. The reproductive number within a group is given by β/γ. These results are the average of 1000 stochastic simulations of disease invasions initiated by a single infected individual introduced into a toroidal 11×11 array of groups, where each group began with 10 hosts. Figure reproduced from Cross et al. (2005).

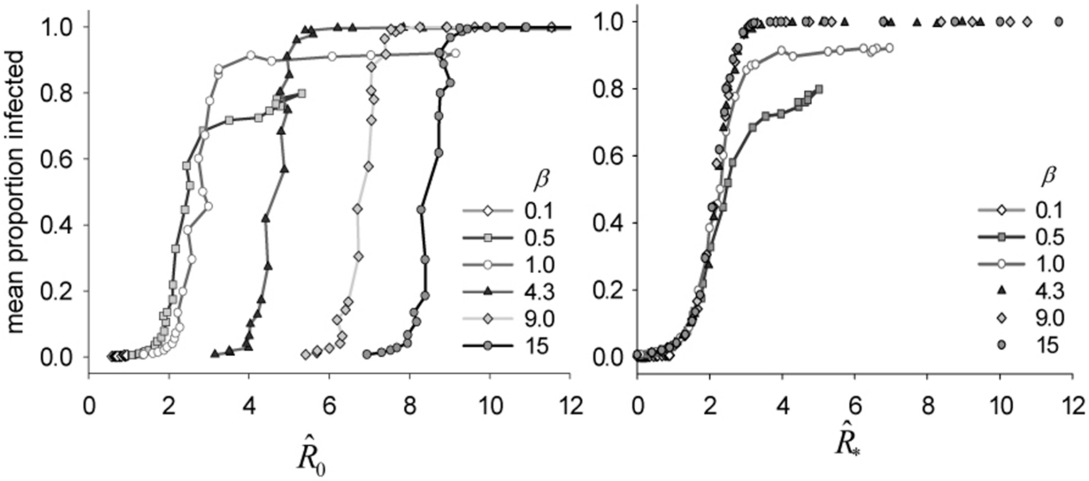


FIGURE 10. Comparing the ability of $\hat{R}_0$ and $\hat{R}_*$ to predict invasion success in a structured population. Details as in Figure 9, with $\gamma = 0.1$. Figure reproduced from Cross et al. (2005).

Now we introduce a strong assumption that pair formation and dissolution dynamics are fast compared to the disease dynamics (i.e. transmission and recovery). Under this assumption we can separate the timescales of pairing and disease, and assume that the pairing dynamics are at a quasi-steady-state relative to the disease dynamics. This means that disease status is constant on the timescales relevant to pairing dynamics, so the dynamics of the pairing system are described by:

$$\begin{aligned}
\frac{dX_{\mathrm{S}}}{dt} &= -k_{\mathrm{S}}\,X_{\mathrm{S}} + 2l_{\mathrm{SS}}\,P_{\mathrm{SS}} + l_{\mathrm{SI}}\,P_{\mathrm{SI}}\\
\frac{dX_{\mathrm{I}}}{dt} &= -k_{\mathrm{I}}\,X_{\mathrm{I}} + 2l_{\mathrm{II}}\,P_{\mathrm{II}} + l_{\mathrm{SI}}\,P_{\mathrm{SI}}\\
\frac{dP_{\mathrm{SS}}}{dt} &= \frac{1}{2}k_{\mathrm{S}}\,m_{\mathrm{SS}}\,X_{\mathrm{S}} - l_{\mathrm{SS}}\,P_{\mathrm{SS}}\\
\frac{dP_{\mathrm{SI}}}{dt} &= \frac{1}{2}k_{\mathrm{S}}\,m_{\mathrm{SI}}\,X_{\mathrm{S}} + \frac{1}{2}k_{\mathrm{I}}\,m_{\mathrm{IS}}\,X_{\mathrm{I}} - l_{\mathrm{SI}}\,P_{\mathrm{SI}}\\
\frac{dP_{\mathrm{II}}}{dt} &= \frac{1}{2}k_{\mathrm{I}}\,m_{\mathrm{II}}\,X_{\mathrm{I}} - l_{\mathrm{II}}\,P_{\mathrm{II}}
\end{aligned} \tag{7}$$

The disease dynamics take place on a much slower timescale. Transmission can only occur in discordant S-I pairs, so we are interested in the quasi-steady-state abundance P^*_{SI}. If transmission occurs within such pairs at rate β_{pair}, then the incidence rate for the whole population is $\beta_{\mathrm{pair}}\,P^*_{\mathrm{SI}}$. The disease dynamics are then captured by the following equations:

$$\begin{aligned}
\frac{dS}{dt} &= \lambda - \beta\mathrm{pair}\,P^*_{\mathrm{SI}} + \sigma\,I - \mu\,S\\
\frac{dI}{dt} &= \beta\mathrm{pair}\,P^*_{\mathrm{SI}} - (\sigma+\mu)\,I
\end{aligned} \tag{8}$$

where λ is the recruitment rate of new susceptibles, σ is the rate of recovery to the susceptible class, and μ is the rate of leaving the population (by death, emigration, or entering a permanent relationship). Then to understand the incidence term we must solve the system for P^*_{SI}. Lloyd-Smith et al. (2004) show that it is straightforward to calculate P^*_{SI} under the assumption that partners are chosen at random (although the choice is weighted by the abundance and pairing rates of unpaired S and I individuals), by solving for the steady state of equation 7.

In the simplest case where infection status does not affect pairing behavior, we can take a single pairing rate k and a single break-up rate l and the incidence rate is

$$f(S,I) = \beta_{\mathrm{pair}}\,P^*_{\mathrm{SI}} = \beta_{\mathrm{pair}}\left(\frac{k}{k+l}\right)\frac{SI}{N}. \tag{9}$$

Noting the similarity to equation 6, we recall that c_{FD} is the per capita rate of acquiring new partners, so

$$c_{\mathrm{FD}} = \frac{1}{1/l + 1/k},$$

and p_{FD} is the probability of transmission in S-I partnership, so

$$p_{\mathrm{FD}} = 1 - \exp(-\beta_{\mathrm{pair}}/l) \approx \beta\mathrm{pair}/l, \qquad \text{since}\quad \beta_{\mathrm{pair}} \ll 1.$$

Thus we find that

$$c_{\text{FD}}\, p_{\text{FD}} \left(\frac{S}{I}\right) I \approx \beta_{\text{pair}} \left(\frac{k}{k+l}\right) \frac{SI}{N}.$$

i.e., we have derived the frequency-dependent incidence rate from a mechanistic pair-formation model. This means that this standard formulation of STD incidence can represent pair-based transmission, but several strong assumptions were required – most notably the timescale approximation that pairing dynamics are much faster than disease dynamics. (Conversely, we know that STD dynamics are driven by pair-based transmission, which means that models using frequency-dependent incidence are implicitly making this timescale approximation.)

To assess the validity of this approximation, Lloyd-Smith et al. (2004) compared epidemics simulated using frequency-dependent transmission with those simulated using the full pair-formation/epidemic system, for parameters corresponding to different types of STD. For transient and highly transmissible diseases such as gonorrhea, frequency-dependent incidence is a good depiction of pair-based transmission only for populations that change sexual partners very frequently (i.e. mean partnership duration 1 day). For chronic and less-transmissible diseases such as HIV, frequency dependence is a reasonable model for populations with mean partnership duration of several months. For populations outside these ranges, mechanistic models that track pairing dynamics should be used. Note also that generalized forms of frequency-dependent incidence can be derived for situations where pairing behavior depends on infection status.

6. Disease Control

6.1. Approaches to disease control. The primary goal of disease control measures is to reduce morbidity and mortality due to disease. Sometimes the control measures are focused on protecting vulnerable populations, such as elderly people for influenza, or endangered populations of wildlife. Most often, however, the aim is to reduce the disease burden in the whole population, by reducing transmission of the infection.

In section 3.2, we established that the effective reproductive rate for transmission within a population can be expressed as $R_{\text{effective}} = c\,p\,D(S/N)$, where, c is the contact rate, p is the probability of transmission given contact, D is the duration of infection, and S/N is the proportion of the population that is susceptible. Examination of this expression allows us to classify control measures according to which component of $R_{\text{effective}}$ they address. The contact rate, c, can be reduced by measures such as quarantine, case isolation, promotion of abstinence, and reduction of mass gatherings (e.g. in theaters or schools). In animal populations, there is also the option of culling hosts which reduces the population density and hence the contact rate. The probability of transmission, p, can be reduced by promoting the use of condoms and safe sex practices, increasing barrier precautions such as masks, gloves, gowns etc., by male circumcision (which now is known to reduce female to male transmission of HIV (Williams et al., 2006)), and by prophylactic treatment. The duration of infectiousness, D, can be reduced by treatment, by contact tracing and case isolation, by improved diagnostics (which enable faster identification of cases) and again by culling of infected hosts. Finally, one can reduce the proportion of susceptibles, S/N in the population by carrying out different types of vaccination

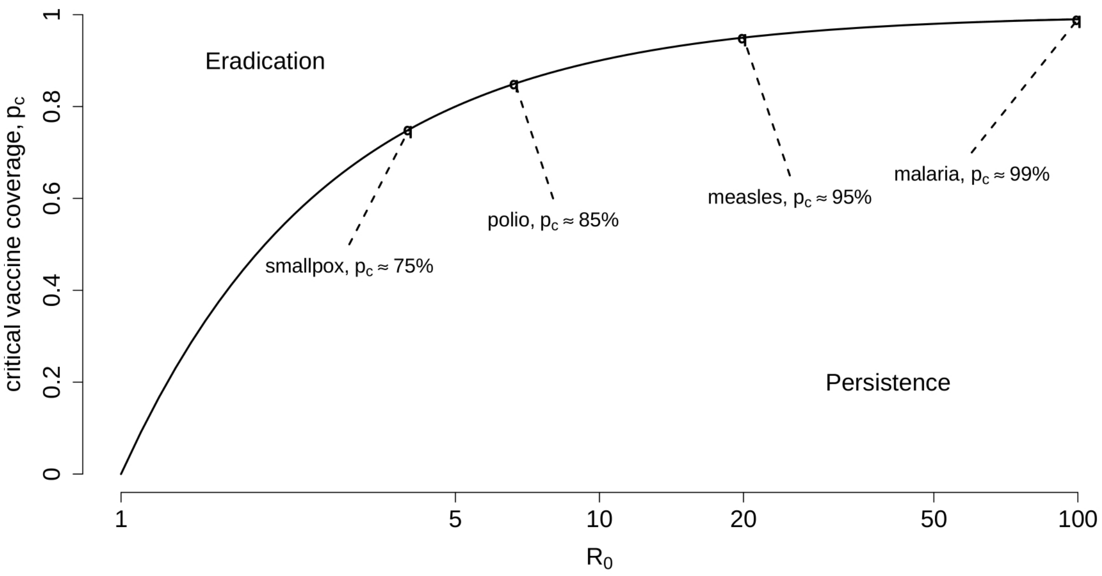


FIGURE 11. Critical vaccination coverage for various diseases, which is determined by R_0.

programs. Mass vaccination is the simplest program, and involves trying to vaccinate the entire population experiencing or threatened by an infectious disease. Ring vaccination involves vaccinating individuals epidemiologically linked to suspected cases, e.g. by tracing contacts of the cases and vaccinating them. Spatially targeted vaccination targets neighborhoods of affected cities and is particularly effective at containing transmission localized to specific geographical area. Other sub-groups in a population can also be targeted if they are thought to be epidemiologically important.

Another aim of some control measures is to reduce transmission between groups (i.e. to reduce R_* rather than R_0). Such measures include movement restrictions and ring vaccination for human hosts, and fencing and ring culling for animals. For vector-borne diseases, further measures are available in addition to the approaches listed above. Bednets and insect repellents are used to reduce biting rates. Many measures aim to reduce the vector population, including the application of larvicides or insecticides, removal of standing water, and use of biological control methods such as larval predators or fungal pathogens.

6.2. Basic theory of disease control. The fundamental question that arises in disease control is "what amount of control is needed to stop sustained transmission of the pathogen?" The question is posed most simply in terms of vaccination, i.e. "what fraction of the population do we need to vaccinate?", but can apply equally to other modes of control. The question is readily answered with reference to the R_0 and $R_{\text{effective}}$ concepts introduced above. Since $R_{\text{effective}} = R_0 \times S/N$, we need $S/N < 1/R_0$ to achieve $R_{\text{effective}} < 1$. Therefore the critical vaccination coverage to eradicate a disease is $p_c = 1 - 1/R_0$. Figure 11 shows this critical vaccination coverage, p_c for various infectious diseases. Note that the disease can be eradicated without vaccinating the entire population, due to herd immunity (Anderson & May, 1985b,a). However, as R_0 increases, eradication by vaccination becomes very challenging due to logistical problems in achieving high coverage levels. It is also important to remember that this calculation makes numerous strong assumptions, including mass untargeted vaccination in a randomly mixing homogeneous population, and that vaccination occurs at birth and is 100% effective. The critical vaccination coverage should thus be treated as a rough guideline, and more detailed mechanistic models are needed to calculate more refined targets.

Based on the considerations of section 5, it is clear that heterogeneity or population structure will influence the efficacy of control measures. For instance, in a classic study May & Anderson (1984) considered vaccination strategies in a patchy population, with host patches of different sizes (e.g. a city and surrounding villages). They assumed that transmission was density dependent so it varied strongly among patches. They showed that, if vaccination was applied uniformly (i.e. the same fraction vaccinated on each patch), spatial heterogeneity increased the critical vaccination coverage relative to the expectation for a homogeneous and well-mixed population. However, if information about the heterogeneity is used to design a targeted campaign, then the disease can be eradicated with a lower overall coverage. Under the assumption of density dependent transmission, the optimal strategy is the one that leaves the same number of individuals susceptible on each patch. In an important follow-up study, Hethcote & Van Ark (1987) showed that weaker density dependence in transmission greatly reduced the magnitude of these effects, but the qualitative result still held.

Eames & Keeling (2003) studied the efficacy of contact tracing in a network epidemic model, considering the critical 'efficiency' of contact tracing (i.e. the proportion of contacts of an infected individual that need to be traced and treated) required to stop transmission. They found that for a model with non-random mixing the critical tracing efficiency was $\approx 1 - 1/R_0$, as predicted for models based on random mixing, provided the network had a negligible degree of clustering. For clustered networks, they found that a lower tracing efficiency was sufficient to contain an epidemic.

Other studies have emphasized the importance of individual heterogeneity for control efforts. Woolhouse et al. (1997) highlighted the marked heterogeneity in contact rates for STDs and vector-borne diseases, and showed how control measures targeting the hosts with highest contact rates could disproportionately help to reduce transmission. Lloyd-Smith et al. (2005b) broadened this point to directly-transmitted diseases, showing that the existence of individual variation in infectiousness – and superspreaders – presented an opportunity for potent targeted control, but also emphasized the challenges in identifying highly-infectious individuals *a priori*. This study also introduced the distinction between two idealized classes of control measures: **population-wide** control measures that are partially effective, such as wearing masks or reducing time in public areas, and **individual-specific** measures that are wholly effective, such as case isolation. (Note that these two extremes are also relevant to contemporary issues in vaccination, since there is interest in the impacts of partially-effective vaccines for diseases like HIV or malaria.

Such vaccines can act by protecting all individuals partially, or by protecting a proportion of individuals fully.) For a given reduction in R_0, individual-specific control measures are likely to stop a disease outbreak (because they lead to increased heterogeneity among individuals; see section 5.2.2).

6.3. Success stories. A number of major success stories demonstrate the potential efficacy of disease control measures. The greatest example is the global eradication of smallpox, via a worldwide effort led by the World Health Organization, which culminated with the last naturally occurring case in Somalia in 1977. The epidemiology, control and eradication of smallpox is described in a fascinating book (available online) written by the leaders of the effort (Fenner et al., 1988). Smallpox was a terrible disease, and has been estimated to have killed 300-500 million humans during the 20th century alone. However, it had certain characteristics that made it vulnerable to eradication. It was not too highly transmissible ($R_0 \approx 3.5 - 6$), requiring a critical vaccine coverage of about 70-85% (Gani & Leach, 2001). It had a relatively long incubation period ($\approx$ 21 days) giving sufficient time for control if an outbreak is detected in its early stages (Ferguson et al., 2003). Smallpox had conspicuous symptoms and few subclinical cases, leading to easy case finding. Crucially, smallpox was a disease of humans only, so there was no animal reservoir from which the infection could re-emerge (although there is concern that monkeypox, which is a zoonotic infection, may fill the niche left vacant by smallpox eradication). Despite these beneficial characteristics, however, the eradication effort struggled in its final phases to stop transmission in the last hot-spots of smallpox transmission. Interestingly, the success of local eradication efforts depended on both vaccination coverage and population density – i.e. factors beyond the predictions of the basic theory of control – and involved intensive contact tracing and ring vaccination. In recent years there has been renewed interest in the dynamics of smallpox transmission, due to concern about its possible use as a bioterror agent (Ferguson et al., 2003).

Another significant success story is the containment of SARS in 2002-2003 (Anderson et al., 2004). SARS is the best case study of an emerging viral pathogen, which had potential for pandemic spread ($R_0 \approx 3$) but was arrested by global efforts. This outbreak has been extensively modelled (Bauch et al., 2005), with particular focus on quarantine and case isolation measures (since there were no drugs or vaccines available to fight this new disease). There was also considerable focus on hospital-based control measures, since $18 - 63\%$ of SARS cases were reported in health care workers (Lloyd-Smith et al., 2003).

A more controversial example is the containment of the foot and mouth disease epizootic in Britain in 2001. The effort was successful in eliminating the infection, but the massive targeted culling program which led to the success has remained controversial. This outbreak is an important example of epidemic models playing a clear role in guiding policy, with two early modelling studies being quite influential (Ferguson et al., 2001; Keeling et al., 2001). Further studies have weighed prophylactic and reactive vaccination strategies, as well as the impact of heterogeneities in the landscape (Keeling et al., 2003; Tildesley et al., 2006).

6.4. Challenges. Disease control efforts are also confronted by many challenges, with causes ranging from fundamental biology to human behaviour. An important example of a biological challenge is drug resistance. Rapid evolutionary rates of pathogens, coupled with strong selection pressure imposed by drug treatments, often result in evolution of drug-resistant pathogen strains. Drug resistance poses grave public health threats worldwide, from methicillin-resistant *Staphylococcus aureus* (MRSA) to chloroquine-resistant malaria. Anti-retroviral resistance is a major concern in the effort to treat HIV and AIDS victims all over the world. Epidemic modellers can play a useful role in integrating information from clinical, experimental and epidemiological data to understand the spread of drug-resistant pathogens. It is important to recognize, though, that some crucial aspects of the basic biology of drug resistance are complex or incompletely understood, such as the relationship between drug dose regimens and evolutionary selection on the pathogen strains (Lipsitch & Samore, 2002) and the relative transmissibility of resistant strains. Models should be designed to reflect (or better still, address) these uncertainties.

Human behaviour can pose a challenge to disease control through vaccine scares, when the belief spreads in a population that a particular vaccine has more adverse effects than benefits. It is true that vaccination typically carries a slight risk, but these risks are much less than those arising from infection with the pathogen. Vaccine scares can arise from an overestimation of the risk of vaccines, or an underestimation of the risk of infection (e.g. for diseases like measles, which have have been eliminated from developed countries for long enough that people may have forgotten that infection can cause serious side effects or death). It is also possible that vaccine refusal arises from selfish behaviour, since if enough of a population is vaccinated then an unvaccinated individual can get all the benefits of herd immunity while avoiding the small risk of vaccination. The individual benefits from refusing vaccine, but the costs of lower coverage are shared among the group – this is a classic 'tragedy of the commons' situation, and accordingly there have been efforts to apply game theory to vaccination behaviour (Bauch & Earn, 2004). Drops in vaccine coverage resulting from vaccine scares have had severe public health consequences. Vaccine scares in the late 1970s for whooping cough in the UK led to slump in immunization, resulting in several major epidemics affecting millions of children (Rohani et al., 1999). The decline in MMR (measles-mumps-rubella) vaccine coverage through the late 1990s and early 2000s in Britain led to increased outbreak size (Jansen et al., 2003). A vaccine scare in Nigeria in 2004 led to increased circulation of polio virus, which was then re-seeded in several other countries by infected travellers, dealing a serious blow to the current campaign to eradicate polio worldwide (Minor, 2004).

7. Parameter estimation, model fitting, and sensitivity analysis

7.1. Estimating R_0. Given its central role in infectious disease dynamics, R_0 is the one parameter we would most like to estimate for any disease. We have already seen that in its simplest form, $R_0 = \beta/\gamma = c\,p\,D$, where, c is the contact rate, p is the probability of transmission given contact, D is the duration of infectiousness, β is the transmission rate, and γ is the recovery rate. It is tempting to think that if we can estimate these parameters at the individual level, then we would be able to estimate R_0. Unfortunately, estimation of these parameters at the individual level is often difficult. The contact rate is particularly challenging, because for many infectious diseases 'contacts' are not defined precisely. This ambiguity also complicates estimation of p. Furthermore, estimates based on R_0 expressions are highly model-dependent, so any assumptions made in model construction will influence the resulting estimate. Finally, it is important to note that in general parameters, c, p and D are not independent of each other. So, the expectation, $\mathrm{E}(c\,p\,D) \neq \mathrm{E}(c)\,\mathrm{E}(p)\,\mathrm{E}(D)$, and any correlations among the constituent parameters would need to be accounted for.

Fortunately, there are many techniques available to estimate R_0 from epidemic data (Dietz, 1993; Anderson & May, 1991). Simple analysis of the SIR model yields two useful approaches. First, during the initial phase of an epidemic the growth rate in the number of cases is roughly exponential. If the exponential growth rate is r, then $R_0 = 1 + rD$. Equivalently, if t_d is the doubling time of the number of infected individuals, then

$$R_0 = 1 + \frac{D \ln 2}{t_d}.$$

On the other hand, if s_0 and s_∞ are the susceptible proportions before the epidemic and after it runs to completion, respectively, then

$$R_0 = \frac{\ln(s_0) - \ln(s_\infty)}{(s_0 - s_\infty)}.$$

All of these estimates are based on simple differential equation models, and hence assume exponentially distributed infectious periods (see section 3.5.3). Wallinga & Lipsitch (2007) provide a general analysis of how the distribution of the serial interval influences the relationship between r and R_0. They find that $R_0 = \frac{1}{M(-r)}$, where $M(z)$ is the moment generating function for the distribution of the serial interval. This allows one to calculate R_0 from r for any arbitrary distribution of the serial interval. Furthermore, they prove that the upper bound on R_0 is $R_0 = e^{rT}$, where T is the mean serial interval.

If the case data are collected in discrete intervals, estimation from continuous-time models is moe difficult. Ferrari et al. (2005) derive an approach based on chain binomial models that provides a maximum-likelihood estimator for R_0 and the associated uncertainty. Note that, like the previous approach based on s_∞, this approach requires that the epidemic runs to its natural completion. The reproductive number can also be estimated based on data from the self-limiting outbreaks that occur when $R_0 < 1$. Branching process models allow analysis of the distribution of outbreak sizes to make inference about the effective reproductive number (Farrington & Whitaker, 2003).

It is also possible to estimate R_0 from epidemiological data in endemic settings. Anderson & May (1991) present simple expressions based on the mean age at first infection, A, and the mean lifespan in a population L. The precise results depend on the age-dependent rate of mortality in the population, $\mu(a)$, where they treat two idealized cases. In "Type I" mortality, all individuals live to age L, so that:

$$\begin{aligned} \mu(a) &= 0 && \text{for} \quad a < L \\ \mu(a) &= \infty && \text{for} \quad a > L. \end{aligned}$$

In "Type II" mortality, individuals are subject to a constant death rate:

$$\mu(a) = \mu = \text{constant} = 1/L.$$

For a population with Type I mortality, $R_0 \approx L/A$. For Type II mortality, the relationship is exact, $R_0 = L/A$. These simple estimates are based on a number of strong assumptions, including random mixing, the absence of heterogeneities or age dependence in the force of infection, and a constant population size. For more advanced treatments, see later chapters of Anderson & May (1991), or Dietz (1993).

R_0 can also be estimated from data on the seroprevalence as a function of age ("age-seroprevalence curves"), by calculating the age-dependent force of infection and making assumptions on the WAIFW matrix for the age-structured population. We refer the readers to Farrington et al. (2001) for a comprehensive review in this topic.

7.2. Parameter estimation. In this and the next section, we give a very cursory overview of various approaches to the estimation of parameter values and their associated uncertainties. Our aim is to introduce the reader to the general principles, and perhaps provide enough basic concepts and vocabulary to enable independent forays into the literature to search for greater detail. In a few instances we have followed the presentation of material by Bolker (2008), and we encourage readers to refer to this excellent book for a more in-depth treatment of these issues (and tips on how to implement estimation algorithms in R).

7.2.1. *Method of moments.* Many problems in parameter estimation can be reduced to the problem of estimating the parameters of a relevant probability distribution to describe the data. For most standard probability distributions, the parameters can be expressed in terms of the moments of the distribution, such as the mean and the variance. The most simple method of estimating your parameter is using these moments directly. For example, consider an exponential distribution whose probability density function, $f(x) = \lambda \exp(-\lambda x)$. If μ is the mean of the distribution, then the exponential rate parameter can be expressed as $\lambda = 1/\mu$. One can calculate the sample moments from the data, and substitute them in the expression, i.e. in our example, $\lambda \approx 1/\bar{\mu}$, where $\bar{\mu}$ is the estimate for μ derived from the data. This approach provides a quick and simple way of getting rough a estimate for your parameter, but the estimates are sometimes biased.

7.2.2. *Maximum likelihood.* For a more formal approach to fitting single distributions, one can use the method of maximum likelihood. Maximum likelihood estimates have certain desirable statistical properties when the data points are independent and identically distributed ("iid") and as the number of data points gets large (the "asymptotic limit"). In particular, maximum likelihood estimates are asymptotically unbiased, asymptotically efficient (i.e. no other asymptotically unbiased estimator has lower uncertainty), and asymptotically normal (i.e. the sampling distribution of the maximum likelihood estimate is normal). These properties, in addition to other useful properties of maximum likelihood estimates (see section 7.5), contribute to the great popularity of this approach for relatively simple estimation problems.

The **likelihood** is the probability of observing the data given the model (which includes the parameter values for the model). If Y is the data set, such that $Y = \{Y_1, Y_2, \ldots, Y_n\}$ and θ is the set of model parameters, then the likelihood is $\mathcal{L} = \Pr(Y|\theta)$, where the model (e.g. the probability distribution we are fitting) will determine the form the probability. The basic idea of the maximum likelihood estimation method is to find the parameter set that maximizes the likelihood of observing the data that you have.

As a simple example, consider the binomial distribution. The binomial distribution describes the number of successes out of N independent trials, if each trial has a probability p of success. In a disease context, this could be the number of individuals infected in a day, out of total susceptible population of N. If the per capita probability of infection is p, then the probability that k out of N are infected is

$$\Pr(k|p, N) = \binom{N}{k} p^k (1-p)^{N-k}$$

Now imagine that we have n independent observations of this process, each with the same initial number of susceptible individuals, N. If the number of individuals infected on the i^{th} observation is k_i, then the likelihood is

$$\mathcal{L} = \prod_{i=1}^{n} \binom{N}{k_i} p^{k_i} (1-p)^{N-k_i}$$

Note that we have applied the principle that the joint probability of several independent events is the product of the probabilities of the single events.

It is conventional to work with the log-likelihood, $L = \log(\mathcal{L})$ for two reasons: (1) it turns the product into a sum, and (2) the probabilities are often very small numbers (usually $\ll 1$), and working with a product of small numbers causes numerical problems in computation. The log-likelihood for the binomial example is:

$$L = \sum_{i=1}^{n} \left(\log \binom{N}{k_i} + k_i \log p + (N - k_i) \log(1-p) \right)$$

For our example, the parameter N is known and we want to find the parameter p that maximizes L. For this simple example, it can actually be calculated analytically (see Bolker (2008)), and yields an intuitive answer:

$$\hat{p} = \frac{\sum_{i=1}^{n} k_i}{nN}$$

Other common probability distributions also have maximum likelihood estimators for their parameters that have simple, intuitive forms. For instance, the maximum likelihood estimators for the means of the Poisson, normal, exponential, gamma, and the negative binomial distributions are all equal to the mean of the data.

For most problems, though, this optimization cannot be solved analytically, so we optimize numerically. It is conventional to do this by minimizing the negative log-likelihood (NLL). For more complex distributions, this can be a multidimensional optimization problem as we may be trying to estimate values for several parameters at once. The optimization can be handled using basic commands in various scientific computing environments (e.g. `fminsearch` and related commands in Matlab, and `optim` in R).

In addition to finding the the optimal parameter value(s) via the maximum likelihood estimate, it is very useful to examine the values of the likelihood for other nearby values of the parameter(s). For a univariate estimation problem this is called the likelihood curve, and gives information about the uncertainty in your estimate (see section 7.4) as well as an indication of any pathological behaviour of your model. For a multivariate problem, this will be a likelihood surface, and it gives information about correlations between parameter estimates as well as their uncertainty.

7.2.3. *Bayesian approaches.* The maximum likelihood approach is an example of frequentist statistics. In frequentist statistics, parameters are assumed to have fixed values that we are trying to estimate as precisely as possible. In Bayesian statistics, in contrast, parameters are treated as random variables, with probabilities assigned to particular values of a parameter to reflect the degree of evidence for that value. Bayesian estimation of distribution parameters is also based on the likelihood, but there are two major differences from MLE (Bolker, 2008):

1. The likelihood is combined with a **prior probability distribution**, which represents information from other sources regarding the values of the parameters. These elements are combined to yield a **posterior probability distribution**, which represents our best estimate of the probability that the parameter takes certain values.
2. The Bayesian parameter estimates are usually given as the mean of the posterior distribution rather than its mode (as in maximum likelihood estimation), because the mean encapsulates more information about the shape of the distribution.

In the setting of parameter estimation, if we have a data set Y and model parameters θ, then Bayes' theorem states that the posterior distribution on θ is

$$\Pr(\theta|Y) = \frac{\Pr(Y|\theta)\Pr(\theta)}{\Pr(Y)},$$

where $\Pr(Y|\theta)$ is the likelihood (as in the previous section), $\Pr(\theta)$ is the prior distribution, which we define, and $\Pr(Y)$ is the probability of observing the data.

The choice of the prior distribution is somewhat subjective, and there are diverse opinions as to how it should be done. The prior distribution be viewed as a useful tool to incorporate background knowledge regarding the parameter, in which case an 'informative prior' can be chosen that places higher density on parameter values that reflect your knowledge. The prior distribution can also be viewed as a necessary evil required to use the tools of Bayesian inference, in which case a 'flat prior' can be chosen that does not include much information about particular values (e.g. a uniform distribution over a plausible range of values). In this case a prior can also be chosen to simplify the computation.

Another challenge in applying Bayesian methods is how to handle the $\Pr(Y)$ term. For very simple situations, the probability of observing a particular data set can be calculated formally, but in most real-world situations this term is somewhat mysterious. There are two important facts to remember about the $\Pr(Y)$ term. The first is that it's a constant for any given data set. This fact is very useful for some numerical techniques which work with the ratio of posterior probabilities, so the $\Pr(Y)$ terms factor out. The second fact is that the posterior probability distribution must be normalized, so it can be re-written:

$$\Pr(\theta|Y) = \frac{\Pr(Y|\theta)\Pr(\theta)}{\int \Pr(Y|\theta)\Pr(\theta)\, d\theta}.$$

For simple problems this integral can be calculated numerically, but for higher dimensional problems we need other tricks. The important insight to derive from this expression is that the posterior distribution is proportional to the product of the likelihood and the prior. It follows that the more 'informative' the prior, the more it will influence the shape of the posterior distribution.

This technique can readily be generalized to estimation of multiple parameters. For example, for a negative binomial distribution with parameters m and k, we can estimate the joint posterior distribution:

$$\Pr(m,k|Y) = \frac{\Pr(Y|m,k)\Pr(m,k)}{\int\int \Pr(Y|m,k)\Pr(m,k)\,dm\,dk}.$$

Then to learn about parameters individually, you can examine the marginal posterior distributions, e.g.

$$\Pr(k|Y) = \int \Pr(m,k|Y)\,m\,dm$$

or take the mean values, e.g.

$$\bar{k} = \int \Pr(k|Y)\,k\,dk.$$

7.3. Fitting more complex models. The approaches described above are readily applied to the problem of fitting a probability distribution to a set of iid data points, but often we will want to estimate parameters from more complex models — sometimes even from our whole dynamic model. The challenge here is to define the likelihood for the model. Because the likelihood is based on probabilities, it requires that we think about the stochastic components of the processes that generated the data — including both the underlying mechanisms and the observation process. This is not a simple problem, in general, but there are two main approaches:

(1) Consider whether the basic mechanisms of the model corresponds to a clearly defined stochastic process.
(2) Do a rough fit of the model to the data, and examine the residuals to look for systematic patterns that corresponds to basic distributions.

We can apply some rules of thumb to choosing a likelihood model, drawn from the excellent discussion of this problem by Hilborn & Mangel (1997). When the quantities in the data are proportions, consider a binomial distribution; if the quantities are rare events, consider a Poisson distribution, or a negative binomial if there seems to be over-dispersion; if the quantities are sums of many contributions, consider a normal distribution, and if they are products of multiplicative probabilities, consider a log-normal distribution. Always examine the residuals from any model you fit to the data, and be alert for patterns that indicate that the likelihood model is inappropriate. Also remember that these are not definitive rules but rather guidelines.

In epidemic models, one example of a clear stochastic mechanism enabling use of a simple likelihood is the situation where the number of susceptible and infectious individuals are known at each point in time. In this case, the likelihood describing the number of new infections in a given interval is binomial. Let $S(t)$ be the number of susceptible individuals, and $\lambda = \beta\, I(t)/N(t)$ be the force of infection. Then the probability that each susceptible becomes infected in time Δt is $p(t) = 1 - \exp(-\lambda(t)\Delta t)$, and the number of new infections generated in $(t,\, t+\Delta t) \sim \text{Binomial}\,(S(t),\, p(t))$. A more sophisticated example of a mechanistic approach to defining the likelihood for an epidemic model is presented by Eichner & Dietz (2003).

If the model is too complex to infer the appropriate likelihood function, one can resort a simpler method. Many studies are published based on simpler fitting procedures, most frequently the method of least squares, or its close relative the χ^2 goodness-of-fit. This approach is based on minimizing the statistic

$$G = \sum_i (O_i - E_i)^2/E_i,$$

where O_i is the observed value at point i from the the real data, and E_i is the expected value at point i from the model output. While this approach does not have the theoretical foundation or powerful ancillary tools associated with the maximum likelihood or Bayesian approaches (though note that least-squares fitting is closely related to maximum likelihood for a normally distributed likelihood (Hilborn & Mangel, 1997)), it will typically yield a decent fit to the data.

7.4. Estimating uncertainties. In any estimation problem, it is essential to obtain some measure of uncertainty that describes the precision and reliability of the point estimate. In frequentist approaches this is often presented as a **confidence interval**, which is defined as an estimated range of values that will contain the unknown parameter with a given probability. In Bayesian approaches, the uncertainty is often presented as a **credible interval**, defined as the region in the center of the posterior distribution containing a given proportion of the density.

7.4.1. *Likelihood profiles.* Likelihood curves and surfaces map out not only the best fit but also the "badness-of-fit" of different parameter values to the data. We can analyze the curvature of these surfaces — or actually, of surfaces of the negative log-likelihood (NLL) — to find confidence intervals for our parameter estimates. For a one-parameter problem, the confidence interval can be calculated directly from the likelihood curve. For higher-dimensional problems, it is often necessary to reduce the dimensionality to focus on the contribution of one parameter to the overall likelihood of the model. The correct approach to this problem is via "likelihood profiles". Likelihood profiles are generated by choosing a range of values for the focal parameter, and for each value maximizing the likelihood with respect to all other parameters. The multiple parameter problem has now been reduced to a one-dimensional profile likelihood curve, which can then be analyzed by the same method that is used for a one-parameter problem to estimate a confidence interval for that parameter value.

The difference in NLL values between the maximum likelihood estimate and other points on a likelihood profile is asymptotically χ^2-distributed with one degree of freedom. To find the $(1-\alpha)\%$ confidence limits on our estimate, we find the parameter values corresponding to NLL values of $\mathbf{NNL}_{\mathrm{MLE}} + \chi^2_1(1-\alpha)/2$, where $\chi^2_1(x)$ is the x^{th} quantile of the cumulative distribution function of the χ^2 distribution with one degree of freedom. For a 95% confidence interval, the increment is $\chi^2_1(0.95)/2$ or 1.92 log-likelihood units. These methods can be generalized to higher dimensions, e.g. to calculate joint confidence regions for two parameters by generating the profile likelihood surface across a fixed grid of values for the parameters. See Bolker (2008) for more details.

7.4.2. *Quadratic approximations.* The likelihood profile approach works very well when you have a small number of parameters, but becomes computationally impractical for models with many parameters, since for an n-parameter model, you have to optimize $n-1$ parameters for each point on your likelihood profile. Luckily, classical likelihood theory tells us that we can learn about the variance of our estimate by considering the second derivative of the likelihood curve — essentially by using a quadratic approximation to the region around the minimum. Here we follow the excellent presentation by Bolker (2008) to summarize this approach.

As mentioned above, maximum likelihood estimates have the property of asymptotically normality. For large enough samples, the sampling distribution for the parameter estimate is approximately normal with standard deviation $\left(\frac{d^2L}{dp^2}\right)^{-1/2}$, where recall that $L = \log \mathcal{L}$. The width of the interval that gives a $(1-\alpha)$ confidence interval is then $N(\alpha)\left(\frac{d^2L}{dp^2}\right)^{-1/2}$, where $N(\alpha)$ is the appropriate quantile from the standard normal distribution. The second derivative with respect to parameter p at a given value p_0 can be computed numerically using the basic formula:

$$\left.\frac{d^2f}{dp^2}\right|_{p=p_0} \approx \frac{f(p_0 + 2\triangle p) - 2f(p_0 + \triangle p) + f(p_0)}{(\triangle p)^2}.$$

For models with several parameters, the same idea applies, but we need to work with the matrix of the second derivatives (the Hessian). For example, if we are estimating the negative binomial parameters m and k:

$$\begin{pmatrix} \frac{\partial^2 L}{\partial m^2} & \frac{\partial^2 L}{\partial m \partial k} \\ \frac{\partial^2 L}{\partial m \partial k} & \frac{\partial^2 L}{\partial k^2} \end{pmatrix}$$

If we evaluate the Hessian at the maximum likelihood estimate and invert it, we obtain the variance-covariance matrix for the parameters:

$$V = \begin{pmatrix} \sigma_m^2 & \sigma_{mk} \\ \sigma_{mk} & \sigma_k^2 \end{pmatrix}$$

Here the diagonal terms can be used to estimate confidence intervals on single parameters, as above, and the off-diagonal terms reflect any correlation between the parameter estimates.

7.4.3. *Bootstrapping.* Bootstrapping is a completely different approach to estimating uncertainties. It is completely non-parametric, meaning that it does not depend on any assumptions about the probability distributions that underlie your data and instead relies on heavy computation to examine the statistical properties of the data set directly. The basic idea is to simulate new data sets by randomly re-sampling with replacement from the observed data. One can then calculate the parameter of interest for each of these simulated data sets, and the distribution of these estimates gives an indication of the uncertainty in your true estimate.

7.5. Model selection. To this point we have focused entirely on how to fit parameters of a single model to the observed data. It is often informative to ask whether the model we are using is the best one, or more generally, to select the best model from a set of candidate models using some objective criterion. The field of model selection and multi-model inference addresses the challenge of comparing models and weighting their outputs in the context of data. It is treated comprehensively by Burnham & Anderson (2002). There are many approaches to model selection, but all have two properties in common: (i) models that fit the data better are preferred; and (ii) parsimonious models are preferred (i.e. there is a penalty for having more parameters).

A classical approach to model selection is the likelihood ratio test. It provides a pair-wise comparison between two models when one model is nested within the other. Model A is said to be nested within model B if it corresponds to some special case of model B where one or more parameters have particular values. For example, $f(x) = ax^2 + c$ is nested in $g(x) = ax^2 + bx + c$ for the value of $b = 0$. An epidemiological example might be whether an additional parameter is justified to describe the possible effect of male circumcision on male-to-female transmission of HIV. For any such pair of models for which likelihood function can be defined, the likelihood ratio test test computes a statistic that compares the log-likelihoods calculated from the two models, and determines whether the additional complexity is justified by the data.

The Akaike information criterion (AIC) provides a more flexible framework for model selection, that does not require models to be nested and can compare many models at once. It is important to note that the AIC can be applied only to models which have been fit by the method of maximum likelihood. The AIC statistic takes the value $\text{AIC} = -2L + 2K$, where, L is the log-likelihood of the maximum likelihood estimate and K is the number of free parameters in the model. For small sample sizes, i.e. when the number of data points N is such that $N/K < 40$, a corrected AIC should be used. Here,

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2K(K+1)}{N-K-1}.$$

In fact the AIC_c converges to the AIC statistic as the sample size grows, so it is always safer to use the AIC_c.

Many models can be compared by simply comparing their AIC values, and the model with the lowest AIC value is generally preferred. Only the relative values of the AIC matter, so they are often reduced to differences from the lowest value obtained, AIC_min:

$$\Delta\text{AIC}_i = \text{AIC}_i - \text{AIC}_\text{min}.$$

As a rule of thumb, models with AIC < 2 units apart have roughly equivalent support, models with AIC between 4 and 7 units apart are clearly distinguishable, and models with AIC > 10 units apart are definitely different.

The ΔAIC values can also be used to calculate the Akaike weight associated with each model i,

$$w_i = \frac{e^{\Delta\text{AIC}_i/2}}{\sum_j e^{\Delta\text{AIC}_j/2}}.$$

These weights are often interpreted as the probability that a given model is the best of the candidate models considered, but this is not formally accurate (Burnham & Anderson, 2002). The weights can be used for model averaging, i.e. to generate an “average” output from several models that is weighted by the support for each model from the data.

7.6. Sensitivity and Uncertainty Analysis. After constructing a model, choosing parameter values, and generating some simulated epidemics, it is important to keep a critical eye on your model output via sensitivity and uncertainty analyses. Model outputs contain uncertainties from two main sources: parameter values and model structure. Parameter values are inevitably uncertain, either because of statistical uncertainties in the estimation process, or because there are no data from which to make an estimate so you have hypothesized a plausible range of values. Model structure is also a source of uncertainty, because there are many possible models that could be built to describe a given system, and all of them make simplifying assumptions while trying to capture the important mechanisms.

Uncertainty analysis aims to assess the variability in model outputs that arises from uncertainty in model inputs. This will determine how much confidence one should place in any quantitative projections that are generated by the model. **Sensitivity analysis** aims to determine which parameters or changes in model structure are most important in determining the model output. It quantifies the influence of each parameter or modelled process on the particular outputs you have obtained. By revealing the relative importance of different mechanisms in the model, sensitivity analysis also tells us which components of the system are good targets for possible disease control measures, or where further efforts to collect data and information should be focused.

A formal definition of **sensitivity** of an outcome λ to the parameter value θ is $S = \frac{\partial \lambda}{\partial \theta}$. By holding all other parameters constant, the partial derivative gives us exactly the rate of change of the outcome, λ, with respect to change in the chosen parameter value θ. However different parameters often have different units or scales of measurement, making sensitivity values difficult to compare. To address this we define **elasticity**, which is the proportional response of an outcome to a proportional perturbation to a parameter. The elasticity of the outcome λ to the value of parameter θ is $E = \frac{\theta}{\lambda}\frac{\partial \lambda}{\partial \theta} = \frac{\partial \log \lambda}{\partial \log \theta}$.

For complex models these quantities typically cannot be calculated analytically, but there are various methods to investigate the influence of parameter values on the model outputs. At the *ad hoc* extreme, it is fairly common to explore a model's range of behaviours by examining different sets of parameters chosen to represent a range of 'scenarios'. This is certainly preferable to no exploration of parameter space, but is not a formal sensitivity analysis. More formally, it is common to examine the sensitivity or elasticity of an output to one or two parameters by running simulations for a range of values of the parameters of interest, while all other parameters are held constant. However, in nonlinear models there can be complex interactions among parameter values, so it is desirable to consider the influence of each parameter in the context of all plausible values of the other parameters. For a model with more than a few parameters, it becomes computationally intensive to investigate all parameters in a full-factorial manner, so more efficient approaches to sampling parameter space have been devised. A popular approach for epidemic models is **Latin Hypercube Sampling**, which is introduced clearly in an influential paper by Blower & Dowlatabadi (1994). Results from this approach are often presented in terms of partial rank correlation coefficients, which describe the influence of each parameter on a given model output in a non-parametric manner.

Structural sensitivity analysis, on the other hand, describes how changes in the design of a model influence its output. There are many subjective decisions, and many assumptions, involved in constructing a model — but too few modelling studies take the time to test explicitly how aspects of their model structure affect their findings. This is probably because it is not a well-defined problem (because there are an infinite number of possible model structures), and there are not established methods for these analyses. Testing for structural sensitivity is an important endeavor, though, because we do not want our assumptions to bias the output of our modelling studies. As a modeller, it is often clear which structural assumptions are most tenuous or have greatest potential to influence the results, and it is good practice to test whether your conclusions are robust to these. The ultimate structural sensitivity analysis occurs when several independent groups of researchers work on the same problem, as has been the case for some recent outbreaks such as H5N1 influenza, SARS, and foot and mouth disease in Britain.

References

R. M. Anderson, C. Fraser, A. C. Ghani, C. A. Donnelly, S. Riley, N. M. Ferguson, G. M. Leung, T. H. Lam, & A. J. Hedley (2004). 'Epidemiology, transmission dynamics and control of SARS: The 2002-2003 epidemic.' *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **359**:1091–1105.

R. M. Anderson & R. M. May (1985a). 'Herd immunity to helminth infection and implications for parasite control.' *Nature* **315**:493–496.

R. M. Anderson & R. M. May (1985b). 'Vaccination and herd immunity to infectious diseases.' *Nature* **318**:323–329.

R. M. Anderson & R. M. C. May (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

R. Antia, V. V. Ganusov, & R. Ahmed (2005). 'The role of models in understanding CD8+ T-cell memory.' *Nat. Rev. Immunol.* **5**(2):101–111.

J. Antonovics, Y. Iwasa, & M. P. Hassell (1995). 'A generalized model of parasitoid, veneral, and vector-based transmission processes.' *American Naturalist* **145**:661–675.

N. T. J. Bailey (1955). 'Some problems in the statistical analysis of epidemic data.' *Journal of the Royal Statistical Society. Series B* **17**:35–68.

F. Ball, D. Mollison, & G. Scalia-Tomba (1997). 'Epidemics with two levels of mixing.' *Ann. Appl. Probab.* **7**:46–89.

F. Ball & P. Neal (2002). 'A general model for stochastic SIR epidemics with two levels of mixing.' *Mathematical Biosciences* **180**:73–102.

S. Bansal, B. T. Grenfell, & L. A. Meyers (2007). 'When individual behaviour matters: homogeneous and network models in epidemiology.' *Journal of the Royal Society Interface* **4**:879–891.

M. S. Bartlett (1957). 'Measles periodicity and community size.' *Journal of the Royal Statistical Society. Series A(General)* **120**:48–70.

C. T. Bauch & D. J. D. Earn (2004). 'Vaccination and the theory of games.' *Proceedings of the National Academy of Sciences of the United States of America* **101**:13391–13394.

C. T. Bauch, J. O. Lloyd-Smith, M. Coffee, & A. P. Galvani (2005). 'Dynamically modeling SARS and respiratory EIDs: past, present, future.' *Epidemiology* **16**:791–801.

N. Becker & I. Marschner (1990). 'The effect of heterogeneity on the spread of disease.' In P. Picard, J. P. Gabriel, & C. Lefevre (eds.), *Stochastic Processes in Epidemic Theory*, vol. 86 of *Lecture Notes in Biomathematics*, pp. 90–103. Springer-Verlag, New York.

M. Begon, S. M. Hazel, D. Baxby, K. Brown, R. Cavanagh, J. Chantrey, T. Jones, & M. Bennett (1999). 'Transmission dynamics of a zoonotic pathogen within and between wildlife host species.' *Proceedings of the Royal Society of London Series B-Biological Sciences* **266**:1939—1945.

O. N. Bjørnstad, B. F. Finkendtädt, & B. T. Grenfell (2002). 'Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model.' *Ecological Monographs* **72**:169–184.

S. M. Blower & H. Dowlatabadi (1994). 'Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example.' *International Statistical Review* **62**:229–243.

B. M. Bolker (2008). *Ecological models and data in R.* Princeton University Press.

K. P. Burnham & D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

P. C. Cross, P. L. F. Johnson, J. O. Lloyd-Smith, & W. M. Getz (2007). 'Utility of R_0 as a predictor of disease invasion in structured populations.' *Journal of the Royal Society Interface* **4**:315–324.

P. C. Cross, J. O. Lloyd-Smith, J. Bowers, C. Hay, M. Hofmeyr, & W. M. Getz (2004). 'Integrating association data and disease dynamics in a social ungulate: bovine tuberculosis in african buffalo in kruger national park.' *Annales Zoologici Fennici* **41**:879–892.

P. C. Cross, J. O. Lloyd-Smith, P. L. F. Johnson, & W. M. Getz (2005). 'Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations.' *Ecology Letters* **8**(6):587–595.

O. Diekmann & J. A. P. Heesterbeek (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation.* Wiley, Chichester.

K. Dietz (1993). 'The estimation of the basic reproduction number for infectious diseases.' *Statistical methods in medical research* **2**:23–41.

K. T. D. Eames & M. J. Keeling (2003). 'Contact tracing and disease control.' *Proceedings of the Royal Society, London, Series B* **270**:2565–2571.

M. Eichner & K. Dietz (2003). 'Transmission potential of Smallpox: Estimates based on detailed data from an outbreak.' *American Journal of Epidemiology* **158**:110–117.

C. P. Farrington, M. N. Kanaan, & N. J. Gay (2001). 'Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data.' *Journal of Royal Statistical Society* **50**:251–292.

C. P. Farrington & H. J. Whitaker (2003). 'Estimation of effective reproduction numbers for infectious diseases using serological survey data.' *Biostatistics* **4**:621–632.

F. Fenner, D. Henderson, I. Arita, Z. Jezek, & I. Ladnyi (1988). *Smallpox and Its Eradication.* World Health Organization, Geneva.

N. M. Ferguson, C. A. Donnelly, & R. Anderson (2001). 'The Foot-and-Mouth epidemic in Great Britian: Pattern of spread and impact of intervensions.' *Science* **292**:1155—1160.

N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, & S. Leach (2003). 'Planning for Smallpox outbreaks.' *Nature* **425**:681–685.

M. J. Ferrari, O. N. Bjørnstad, & A. P. Dobson (2005). 'Estimation and inference of R_0 of an infectious pathogen by removal method.' *Mathematical Biosciences* **198**:14–26.

C. Fraser, S. Riley, R. M. Anderson, & N. M. Ferguson (2004). 'Factors that make an infectious disease outbreak controlable.' *Proceedings of the National Academy of Sciences of the United States of America* **101**:6146—6151.

R. Gani & S. Leach (2001). 'Transmission potential of smallpox in contemporary populations.' *Nature* **414**:748–751.

B. T. Grenfell, O. N. Bjørnstad, & B. F. Finkendtädt (2002). 'Dynamics of measles epidemics. ii. scaling noise, determinism and predictability with the times series SIR model.' *Ecological Monographs* **72**:185—2002.

J. A. P. Heesterbeek (2002). 'A brief history of R_0 and a recipe for its calculation.' *Acta Biotheoretica* **50**:189–204.

H. W. Hethcote & J. W. Van Ark (1987). 'Epidemiological models for heterogenous populations: Proportionate mixing, parameter estimation, and immunization programs.' *Mathematical Biosciences* **84**:85–118.

R. Hilborn & M. Mangel (1997). *The ecological detective: confronting models with data*. Princeton University Press.

A. James, J. W. Pitchford, & M. J. Plank (2006). 'An event-based model of superspreading in epidemics.' *Proceedings of the Royal Society of London Series B-Biological Sciences* **274**:741–747.

V. A. A. Jansen, N. Stollenwerk, H. J. Jensen, M. E. Ramsay, W. J. Edmunds, & C. J. Rhodes (2003). 'Measles outbreaks in a population with declining vaccine uptake.' *Science* **301**.

M. J. Keeling & K. T. D. Eames (2005). 'Networks and epidemic models.' *J. R. Soc. Interface* **2**(4):295–307.

M. J. Keeling & B. T. Grenfell (1997). 'Disease extinction and community size: Modeling the persistence of measles.' *Science* **275**:65–67.

M. J. Keeling & B. T. Grenfell (1998). 'Effect of variability in infection period on the persistence and spatial spread of infectious diseases.' *Mathematical Biosciences* **147**:207–226.

M. J. Keeling & P. Rohani (2002). 'Estimating spatial coupling in epidemiological systems: a mechanistic approach.' *Ecology Letters* **5**:20–29.

M. J. Keeling & P. Rohani (2007). *Modeling infectious diseases in humans and animals*. Princeton University Press.

M. J. Keeling, M. E. J. Woolhouse, R. M. May, & B. T. Grenfell (2003). 'Modelling vaccination strategies against foot-and-mouth disease.' *Nature* **421**:136–142.

M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, & B. T. Grenfell (2001). 'Dynamics of the 2001 UK Foot and Mouth epidemic: Stochastic dispersal in a heterogeneous landscape.' *Science* **294**:813—817.

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

A. A. King, S. Shrestha, E. T. Harvell, & O. N. Bjørnstad (2009). 'Evolution of acute infections and the invasion-persistence trade-off.' *The American Naturalist* **173**:446—455.

K. Koelle, X. Rodo, M. Pascual, M. Yunus, & G. Mostafa (2005). 'Refractory periods and climate forcing in cholera dynamics.' *Nature* **436**(7051):696–700.

M. Lipsitch & M. H. Samore (2002). 'Antimicrobial use and anitmicrobial resistance: a population perspective.' *Trends in Microbiology* **8**:347–354.

A. L. Lloyd (2001). 'Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods.' *Proceedings of the Royal Society of London Series B-Biological Sciences* **268**:985–993.

J. O. Lloyd-Smith, P. C. Cross, C. J. Briggs, M. Daugherty, W. M. Getz, J. Latto, M. S. Sanchez, A. B. Smith, & A. Swei (2005a). 'Should we expect population thresholds for wildlife disease?' *Trends in Ecology and Evolution* **20**(9):511–519.

J. O. Lloyd-Smith, A. P. Galvani, & W. M. Getz (2003). 'Curtailing transmission of severe acute respiratory syndrome within a community and its hospital.' *Proceedings of the Royal Society, London, Series B* **270**:1979–1989.

J. O. Lloyd-Smith, W. M. Getz, & H. V. Westerhoff (2004). 'Frequency-dependent incidence in models of sexually transmitted diseases: portrayal of pair-based transmission and effects of illness on contact behaviour.' *Proceedings of the Royal Society of London, Series B* **271**:625–635.

J. O. Lloyd-Smith, S. J. Schreiber, & W. M. Getz (2006). 'Moving beyond averages: individual-level variation in disease transmission.' In A. B. Gumel, C. Castillo-Chavez, R. E. Mickes, & D. P. Clemence (eds.), *Mathematical studies of human disease dynamics: emerging paradigms and challenges*, vol. 410 of *AMS Contemporary Mathematics*, pp. 235–258. American Mathematical Society.

J. O. Lloyd-Smith, S. J. Schreiber, P. E. Jopp, & W. M. Getz (2005b). 'Superspreading and the effect of individual variation on disease emergence.' *Nature* **438**:355–359.

A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, & C. J. L. Murray (eds.) (2002). *Global Burden of Disease and Risk Factors.* Oxford University Press, Oxford.

W. M. Lui, H. W. Hethcote, & S. A. Levin (1987). 'Dynamical behaviour of epidemiological models with nonlinear incidence rates.' *Journal of Mathematical Biology* **25**:359—380.

R. M. May & R. M. Anderson (1984). 'Spatial heterogeneity and the design of immunization programs.' *Mathematical Biosciences* **72**:83–111.

H. McCallum, N. Barlow, & J. Hone (2001). 'How should pathogen transmission be modelled?' *Trends in Ecology and Evolution* **16**:295–300.

N. Mideo, S. Alizon, & T. Day (2008). 'Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases.' *Trends Ecol. Evol.* .

P. D. Minor (2004). 'Polio eradication, cessation of vaccination and re-emergence of disease.' *Nature Reviews Microbiology* **2**:473–482.

D. Mollison (1991). 'Dependence of epidemic and population velocities on basic parameters.' *Mathematical Biosciences* **107**:255–287.

I. Nåsell (2005). 'A new look at the critical community size for childhood infections.' *Theor. Popul. Biol.* **67**(3):203–216.

M. E. J. Newman (2002). 'Spread of epidemic disease on networks.' *Physical Review E* **66**.

A. S. Perelson & G. Weisbuch (1997). 'Immunology for physicists.' *Reviews in Modern Physics* **69**:1219–1268.

P. Rohani, D. J. D. Earn, & B. T. Grenfell (1999). 'Oppopsite patterns of synchrony in sympatric disease metapopulations.' *Science* **286**:968–971.

P. Schmid-Hempel (2008). 'Parasite immune evasion: a momentous molecular war.' *Trends Ecol. Evol.* **23**:318—326.

S. J. Schreiber & J. O. Lloyd-Smith (in press). 'Invasion success and extinction risk in spatially heterogeneous environments.' *The American Naturalist* .

D. L. Smith, B. Lucey, L. A. Waller, J. E. Childs, & R. L. A. (2002). 'Predicting the spatial dynamics of rabies epidemics on heterogenous landscapes.' *Proceedings of* *the National Academy of Sciences of the United States of America* **99**:3668–3672.

M. J. Tildesley, N. J. Savill, D. J. Shaw, R. Deardon, S. P. Brooks, M. E. J. Woolhouse, B. T. Grenfell, & M. J. Keeling (2006). 'Modelling vaccination strategies against foot-and-mouth disease.' *Nature* **440**:83–86.

P. van den Driesche & J. Watmough (2002). 'Reproduction number and subthreshold endemic equilibria for compartmental models of disease transmission.' *Mathematical Biosciences* **180**:29–48.

J. Wallinga & M. Lipsitch (2007). 'How generation intervals shape the relationship between growth rates and reproductive numbers.' *Proceedings of the Royal Society, Series B* **274**:599–604.

H. J. Wearing, P. Rohani, & M. J. Keeling (2005). 'Appropriate models for the management of infectious diseases.' *PLoS Medicine* **2**:e174.

B. G. Williams, J. O. Lloyd-Smith, E. Gouws, C. Hankins, W. M. Getz, J. Hargrove, I. de Zoysa, C. Dye, & B. Auvert (2006). 'The potential impact of male circumcision on HIV in sub-Saharan Africa.' *PLoS Medicine* **7**:e262.

M. E. J. Woolhouse, C. Dye, J. F. Etard, T. Smith, J. D. Charlwood, G. P. Garnett, P. Hagan, J. L. K. Hii, P. D. Ndhlovu, R. J. Quinnell, C. H. Watts, S. K. Chandiwana, & R. M. Anderson (1997). 'Heterogeneities in the transmission of infectious agents: Implications for the design of control programs.' *Proceedings of the National Academy of Sciences of the United States of America* **94**:338–342.

[1] Applied and Interdisciplinary Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109-1043

[2] Department of Ecology & Evolutionary Biology, University of California, Los Angeles, 90095,

[3] Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA