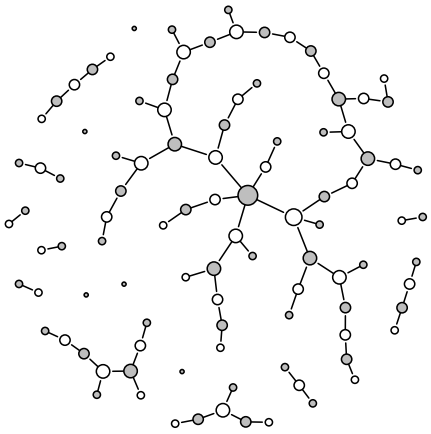
INFECTIOUS DISEASE MODELS AND THEIR USE IN THE SARS-COV-2 PANDEMIC

CHRISTIAN L. ALTHAUS
JULIEN RIOU
EMMA B. HODCROFT
ROLAND REGOES



A three day course at the Swiss Epidemiology Winter School in Wengen, Switzerland

March 31 – April 2, 2022

PD Dr. Christian L. Althaus
Institute of Social and Preventive Medicine (ISPM), University of Bern
https://www.ispm.unibe.ch - christian.althaus@ispm.unibe.ch

Dr. Julien Riou
Institute of Social and Preventive Medicine (ISPM), University of Bern
https://www.ispm.unibe.ch - julien.riou@ispm.unibe.ch

Dr. Emma B. Hodcroft
Institute of Social and Preventive Medicine (ISPM), University of Bern
https://www.ispm.unibe.ch - emma.hodcroft@ispm.unibe.ch

Prof. Dr. Roland R. Regoes
Institute of Integrative Biology, ETH Zurich
https://www.tb.ethz.ch - roland.regoes@env.ethz.ch

Further information about the Swiss Epidemiology Winter School can be found at: http://www.epi-winterschool.org

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

Part I

THEORY

'As a matter of fact all epidemiology, concerned as it is with variation of disease from time to time or from place to place, *must* be considered mathematically (...), if it is to be considered scientifically at all. (...) And the mathematical method of treatment is really nothing but the application of careful reasoning to the problems at hand.'

Ross (1911)

1

INTRODUCTION TO MATHEMATICAL EPIDEMIOLOGY

The history of describing and quantifying infectious diseases by means of mathematics has a long and interesting history. In the late 16^{th} and the early 17^{th} century, when several plague epidemics raged through Europe, John Graunt in London and even earlier Felix Platter in Basel developed the first detailed and methodological accounts of the death toll due to plague. Unsurprisingly, given the enormous death toll (during the peak of plague epidemics up to $4/5$ of the overall mortality was due to plague) the science of demography was at its birth closely intertwined with the infectious diseases. Daniel Bernoulli, from the famous dynasty of mathematicians in Basel, used in 1766 the recently developed differential calculus to study the impact of variolation[1] on human mortality due to small pox.

In the early 1900's several people made key contributions to the modeling of infectious diseases. The British physician William Heaton Hamer in his Milroy Lectures on Epidemic Diseases studied in 1906 the dynamical behavior of measles incidence (Hamer, 1906). In particular he is credited for introducing the concept of mass-action kinetics (see Section 3.1) between infected and susceptible individuals as a feedback in the dynamics of infectious diseases. He argued that the periodic pattern of measles incidence may be due to a feedback between the dynamics of susceptibles and infecteds and need not be explained by a change in virulence of the pathogen over time. Around the same time the British physician Ronald Ross, who was awarded the second Nobel prize in Medicine in 1902 for his discovery of the role of mosquitoes in transmitting malaria, began to apply mathematical approaches to understand malaria transmission. He made many key contributions to epidemiology (Fine, 1975). In particular he was the first to recognize the threshold phenomena for the endemic transmission of infectious diseases and determined a critical density of mosquitoes required for persistence of malaria. This was key to the efforts of control of malaria, because he was able to argue that in order to eradicate malaria it is not necessary to eradicate the mosquito population, which was regarded as impossible, but only bring it below a threshold density.

In 1927 the two British scientists William Kermack and Anderson McKendrick developed the first more generalized models of infectious disease epidemiology and the basic SIR model that will be dis-

1 Variolation is a form of vaccination in which small pox virus from infected individuals was collected and transferred into small wounds of "vaccinees" with very high rates of mortality.

DISEASE	YEARLY MORTALITY
Lower respiratory infections	3.9×10^6
HIV/AIDS	2.8×10^6
Diarrheal diseases	1.8×10^6
Tuberculosis	1.6×10^6
Malaria	1.2×10^6
Measles	0.6×10^6

Table 1: Leading causes of death due to infectious diseases. Data are for 2002 (taken from the WHO World Health Report, 2004).

cussed in Section 3.1 can be found for the first time in their paper (Kermack and McKendrick, 1927). While also recognizing the role of thresholds for the spread of infectious diseases, like Ross they formulated the threshold criterial in terms of a critical host density rather than taking a pathogen centric view. The pathogen centric view of a threshold behavior came later through the development of the concept of the basic reproduction ratio, R_0 (see Heesterbeek (2002) for a history of the concept of R_0). This number describes the average number of secondary infections caused by one primary infected individual placed into a population that is otherwise wholly susceptible, and has become a key concept in theoretical epidemiology and the control of infectious diseases.

1.1 CONTROL OF INFECTIOUS DISEASES

Today, infectious diseases remain a major public health problem worldwide (Table 1). It is therefore important to develop efficient control measures that can contain infections in a variety of ways:

- Vaccination
- Treatment
- Contact tracing (e.g., for sexually transmitted infections (STIs))
- Others such as quarantining (SARS-CoV-2) and culling (animal disease, foot and mouth disease)

It is often hard to anticipate the results of different control scenarios. Empirical studies like experiments or randomized controlled trials (RCTs) are not always feasible and sometimes unethical. This is why mathematical modeling is useful for the study of transmission and for predicting the effects of public health interventions. Different hypothetical control scenarios can be compared with each other in a timely manner - a critical factor in new epidemics. Conversely, new

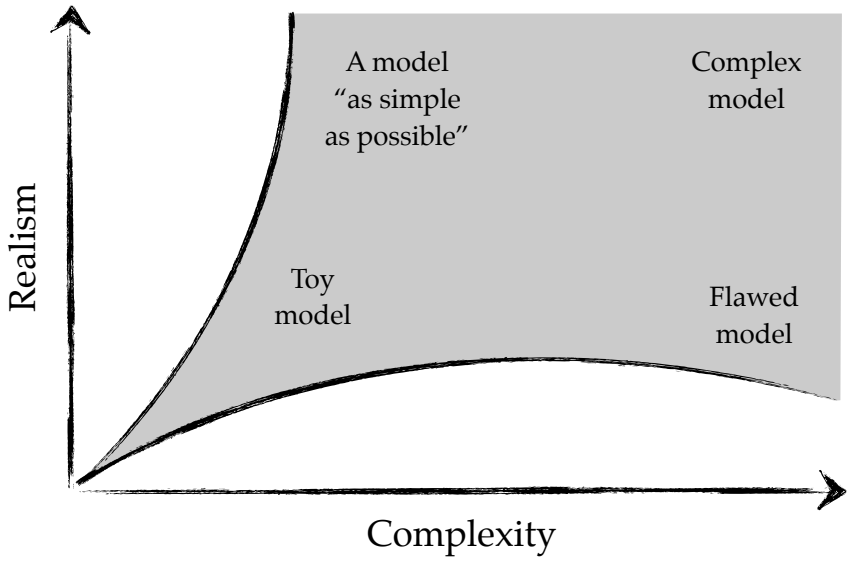


Figure 1: Trade-off between model complexity and realism. Toy models can be realistic enough to describe basic principles of infectious diseases. Complex models offer opportunities to study infections diseases in greater detail, but model development is not useful if it is based on incorrect assumptions and unreliable parameter estimates. A 'good' model is "as simple as possible". It is realistic enough to achieve its purpose, but still tractable.

control efforts may perturb the infectious dynamics of a previously reached endemic steady-state. Mathematical modeling can provide new insights into the dynamics of infectious diseases by analyzing change in the epidemic pattern.

1.2 WHAT MODELS CAN DO (AND CANNOT DO)

Mathematical models can be used to understand and predict the time course of infectious diseases.

In mathematical epidemiology, models have two distinct roles: they help us *understand* and *predict* the time course of infectious diseases. However when we use mathematical models we must recognize the trade-off between realism and complexity (Figure 1). Very simple models, like those introduced in Chapter 3, can be considered as "toy models". They do not incorporate a great number of realism but, but their simplicity makes them useful for understanding the essential properties of an infectious disease. Since the early 1990's, the availability of high performance computing has resulted in a trend towards creating more complex models that incorporate much finer details of human interaction and disease transmission, including spatial structure, air traffic and even climate change. While such complex models provide an intriguing platform for studying the transmission of infectious diseases, it can be extremely difficult to parameterize them. Their simulation is computationally expensive, which often prevents comprehensive analysis. The addition of more 'realism' may result in

unreasonable assumptions and ultimately flawed models that do not capture realism as well as simple toy models.

A 'good' model is *suited to its purpose*, *transparent* and provides enough *flexibility* to be adapted readily to new situations. Using the well-known dictum, such "models should be as simple as possible, but not more so"[2]. For a comprehensive overview of the development and application of mathematical models in infectious disease epidemiology, we recommend the books by Anderson and May (1991); Bjørnstad, Ottar N (2018); Diekmann et al. (2012); Diekmann and Heesterbeek (2000); Keeling and Rohani (2008); Vynnycky and White (2010).

2 This dictum is often attributed to Albert Einstein (e.g., in May (2004)), although he has never said it in this context.

2

BASIC CONCEPTS OF POPULATION DYNAMICS

Mathematical epidemiology describes populations that are infected with a pathogen. Different mathematical frameworks can be used to capture the dynamics of infections over time These can be categorized into approaches that treat time as discrete or continuous. In this course, we will focus on models that describe the dynamics of infectious diseases in continuous time, most commonly by using ordinary differential equations (ODEs).

2.1 ORDINARY DIFFERENTIAL EQUATIONS

An equation that relates the derivatives of a variable (e.g., the change of the size of a population) to the value of the variable itself (the population) as a continuous function of a single parameter (e.g., time) is called an ODE. Let us assume a population that increases or decreases depending on its current value. In mathematical terms this can be expressed as

Ordinary differential equations are the workhorse of mathematical epidemiology.

$$\frac{dN(t)}{dt} = F(N(t)), \tag{1}$$

where F is a function that relates the population's rate of change $dN(t)/dt$ to its size $N(t)$ at the time t.

2.1.1 *Exponential growth and decay*

The simplest model of population growth is obtained if one assumes that the per capita growth rate is constant over time, i.e.,

$$\frac{dN}{dt} = rN, \tag{2}$$

where the growth rate is given by r. The solution to this differential equation is very simple and can be obtained through integration:

$$\int_{N(0)}^{N(t)} \frac{1}{N} dN = \int_0^t r dt, \tag{3}$$

$$\log\left(\frac{N(t)}{N(0)}\right) = rt, \tag{4}$$

$$N(t) = N(0)e^{rt}, \tag{5}$$

where $N(0)$ is the population density at time $t = 0$. Thus, the solution is the well-known exponential function.

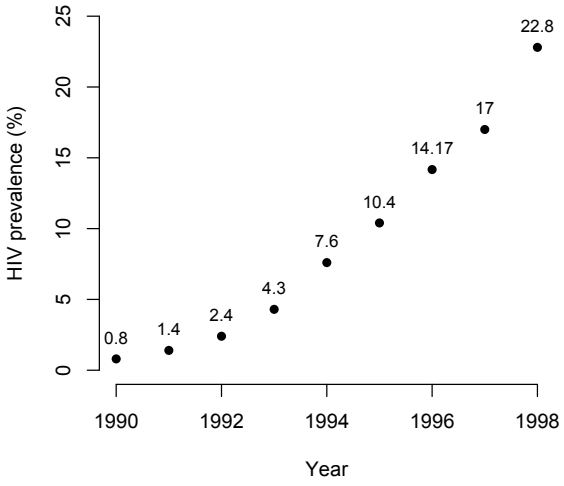


Figure 2: Prevalence of HIV infection in pregnant women in South Africa from 1990 to 1998. It can be seen that in the first few years the prevalence roughly doubles each year. In later years the rate of increase begins to decrease, in part because of increased public awareness of the risks of HIV infection, but also because of the decrease of susceptible individuals. Data from the National Antenatal Sentinel HIV & Syphilis Prevalence Survey in South Africa.

The exponential growth is often a reasonable description for the initial growth of biological populations in the absence of limiting factors. An example of an almost exponential increase of an infection is the early HIV epidemic in South Africa (Figure 2).

Removal or death of individuals can also decrease the size of the population. Mathematically, this can be described by

$$\frac{dN}{dt} = -\delta N, \tag{6}$$

where δ represents the rate at which individuals are removed from the population. A cohort of infected individuals who can clear the infection spontaneously through their immune response can exhibit an exponential decrease in the number of infected individuals (Figure 3).

The interaction between populations is often density dependent.

By combining exponential growth and death (Eqs. 2 and 6), one would expect infections to either grow or die out, depending on whether $(r - \delta)$ is positive or negative. However it is unrealistic to assume a constant per capita growth rate because no population or infection can grow limitlessly. Resource depletion and other factors can result in density-dependent growth rates.

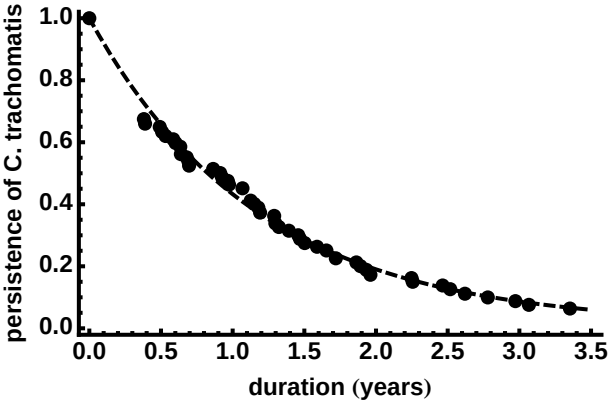


Figure 3: Persistence of chlamydia in a cohort of asymptomatically infected women. Fitting a mathematical model to the data indicates an exponential decrease in the number of infected women. The rate at which women clear the infection was estimated at 0.84 per year, meaning that they are infected for 433 days on average. Figure from Althaus et al. (2010).

2.1.2 *Logistic growth*

The simplest way to introduce a density dependent effect is to assume that the per capita growth rate decreases linearly with population size, i.e., $r(1 - N/K)$, where K is the carrying capacity. The corresponding differential equation, the logistic differential equation, is thus given by

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}). \tag{7}$$

The parameter r now describes maximal per capita growth rate (i.e., the per capita growth rate when the population is vanishingly small). The analytic solution of Equation 7 is given by

$$N(t) = N(0)\frac{K}{N(0) + (K - N(0))e^{-rt}}, \tag{8}$$

where $N(0)$ is again the population size at time $t = 0$. The interpretation of the carrying capacity K becomes clear when the behavior of this equation is inspected for large times. Carrying capacity represents the value that the population attains as time t proceeds to infinity. Figure 4 shows the logistic 'growth' of two populations that reach the same stable equilibrium.

We have so far considered only one population, but density dependent effects come into play if populations interact. This is particularly true for population dynamics of infectious diseases, where the number of infected individuals can grow only as long they come in contact with enough uninfected individuals.

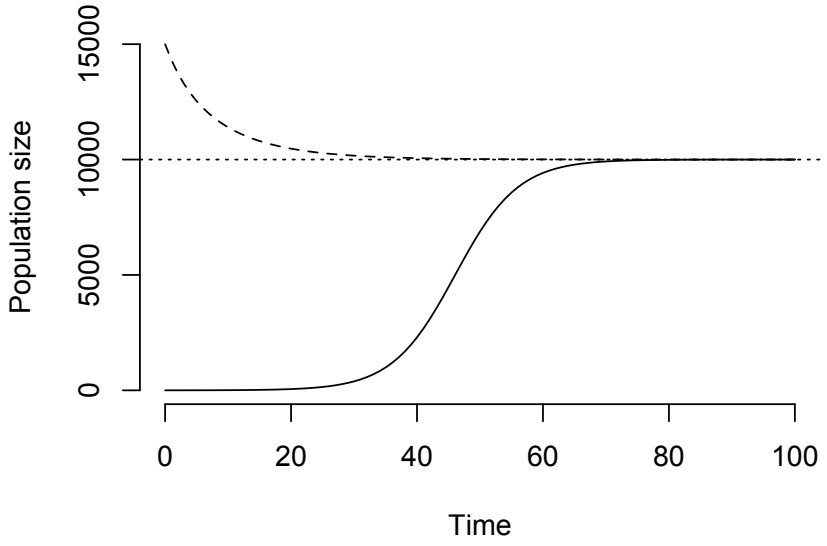


Figure 4: Logistic growth of two populations that reach the same stable equilibrium. Solid line: $r = 0.2$ and $N(0) = 1$. Dashed line: $r = 0.1$ and $N(0) = 1.5 \times 10^4$. Carrying capacity $K = 10^4$ (dotted line).

3 COMPARTMENTAL MODELS

The dynamics of an infection can be followed and the interaction between the different subpopulations can be studied if the population of hosts is divided into different compartments[1].

3.1 SIR MODEL

The very simple SIR model divides the population into three compartments: susceptible hosts S; infected hosts I; and, hosts that recovered from the disease R (Figure 5). Transmission may occur if susceptible hosts come into contact with infected hosts, generating a "flow" from susceptible to infected host populations.

Compartmental models can divide a population into different disease states.

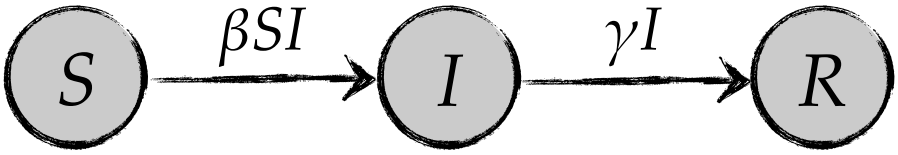


Figure 5: Flow chart of the SIR model. Susceptibles S become infected individuals I at rate βI. Infected individuals I can clear the infection at rate γ to become recovered individuals R.

The flow chart in Figure 5 can be translated easily into a mathematical model. The rate of change of a population is given by the "inflow" into the population minus the "outflow" from that population; it corresponds to the time derivative of the population. The resulting dynamical model is a system of ordinary differential equations (ODEs). Thus, the scheme in Figure 5 translates into the following system of ODEs:

$$\frac{dS}{dt} = -\beta SI, \tag{9}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{10}$$

$$\frac{dR}{dt} = \gamma I, \tag{11}$$

where β denotes the rate at which infectious contacts between susceptible and infected individuals are made. The infection is then cleared

1 Wikipedia provides a detailed entry on compartmental models in epidemiology (http://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology).

at rate γ (see Equation 6). Since the assumption of a constant clearance rate results in exponentially distributed infectious durations, average infectious duration is simply given by $1/\gamma$.

The overall transmission is *density dependent*, i.e., it is proportional to S and I. This follows from the law of *mass action*, which has its origin in chemical reaction kinetics[2]. This assumption is justified if hosts of type S and I are well mixed (homogenous) and the encounters between the two types of host are random: every individual has equal chance of 'meeting' any other individual in the population. Obviously, contact patterns between humans are not fully random but for many applications this proves to be a good first approximation. For many diseases, however, the assumption of random contacts between infected and susceptible individuals is questionable and we will later discuss models that take contact structure more explicitly into account and where transmission is assumed to be *frequency dependent*. Finally, the rate at which susceptible individuals become infected, βI, is called the *force of infection* λ, which can be used to calculate the expected *incidence rate* for a given population: i.e., λS gives the rate of new infected cases in the population of susceptible hosts.

3.2 SEIRS MODEL

While the SIR model is useful for describing a variety of diseases, many infections are characterized by an incubation period and a phase of transient immunity. That is, infected hosts may go through a latent phase during which they are not yet infectious. Also, many diseases do not confer lifelong immunity and recovered individuals may again become susceptible. The SEIRS model structure incorporates these additional complexities (Figure 6). The flow chart can again be translated into a set of ODEs:

$$\frac{dS}{dt} = -\beta SI + \omega R, \tag{12}$$

$$\frac{dE}{dt} = \beta SI - \sigma E, \tag{13}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \tag{14}$$

$$\frac{dR}{dt} = \gamma I - \omega R. \tag{15}$$

Compared to the population dynamics models of Chapter 2, neither the SIR and SEIRS model can be solved analytically. A numerical solution, however, can be easily obtained by numerical integration of the ODEs. The Euler method gives a simple algorithm to integrate ODEs[3] but most mathematical software packages offer more sophis-

2 See http://en.wikipedia.org/wiki/Law_of_mass_action for more information.
3 See http://en.wikipedia.org/wiki/Euler_method for more information.

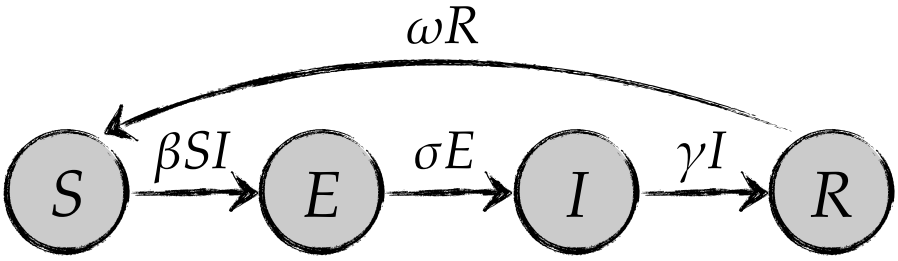


Figure 6: Flow chart of the SEIRS model. Newly infected individuals move through a latent stage E (for exposed) before becoming infectious individuals I at rate σ. Recovered individuals R loose their acquired immunity at rate ω and become susceptible S again.

ticated routines such as the one that we will use in the exercises of Chapter 7 and **??**.

Norovirus infections offer an interesting example of infections with an incubation period and temporary immunity. They are the most common cause of viral gastroenteritis in humans. Outbreaks of norovirus infection often occur in closed or semiclosed communities, such as long-term care facilities, overnight camps, hospitals, prisons, dormitories, and cruise ships, where the infection is transmitted very rapidly either from person to person or through contaminated food. The left panel of Figure 7 shows the dynamics of a norovirus outbreak in a group of 100 individuals (hospital ward, school camp). After the first peak of infection, one can observe a rebound of norovirus at around 50 days. This is only expected to happen if the duration of immunity against norovirus infection is short and no control efforts take place after the first peak.

Since there is uncertainty around the duration of norovirus immunity, one can investigate the long-term equilibrium (steady-state) of the infection as a function of the duration of immunity (Figure 7, right panel). The equilibrium values attained by the system can be determined without having to solve the system of ODEs. The system is in equilibrium if the populations do not change over time, i.e., when the derivatives can be set to zero ($dS/dt = dE/dt = dI/dt = dR/dt = 0$). Setting the right hand side of equations (12)–(15) to zero and assuming a constant total population size of $N = S + E + I + R$, we obtain the solutions for S, E, I and R corresponding to a disease free equilibrium:

$$S_f = N, \qquad E_f = 0, \qquad I_f = 0, \qquad R_f = 0; \tag{16}$$

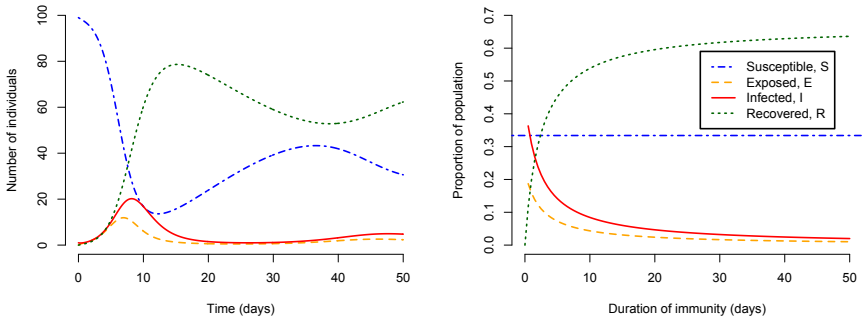


Figure 7: Simulation of a norovirus outbreak. (Left panel) The number of infected individuals reaches a peak around a week after the introduction of the first norovirus case. (Right panel) Steady-state values of S, E, I and R for different durations of immunity. The parameters used in the simulation are: $\beta = 1.9 \times 10^{-2}$ per person per day; $1/\sigma = 0.8$ days; $1/\gamma = 1.6$ days; $1/\omega = 30$ days (panel A); N is set to 100 and 1 in panel A and B, respectively. Parameters were inferred from the data in Vanderpas et al. (2009) using a non-linear least square method (see Section **??**).

and an endemic equilibrium:

$$S_e = \frac{\gamma}{\beta}, \tag{17}$$

$$E_e = \frac{\gamma\omega(\beta N - \gamma)}{\beta(\sigma\omega + \gamma(\sigma + \omega))}, \tag{18}$$

$$I_e = \frac{\sigma\omega(\beta N - \gamma)}{\beta(\sigma\omega + \gamma(\sigma + \omega))}, \tag{19}$$

$$R_e = \frac{\sigma\gamma(\beta N - \gamma)}{\beta(\sigma\omega + \gamma(\sigma + \omega))}. \tag{20}$$

Note, that perhaps somewhat counterintuitively the equilibrium value of susceptible individuals at the endemic steady-state does not depend on ω, i.e., the rate at which immunity is lost. The increased rate of inflow into the susceptible hosts for short durations of immunity is compensated by an increased infection rate of susceptibles, because the density of infected hosts increases. This is related to the paradox of enrichment in ecology[4], where increasing the birth rate of a prey results not in an increase in the prey but the predator population.

We started with the simple SIR-model and added layers of complexity and realism by introducing further host classes (such as recovered or exposed hosts) and processes (such as transient or life-long immunity). These models can be extended similarly, to include other host classes and processes as necessary. Some frequently used examples follow:

4 See http://en.wikipedia.org/wiki/Paradox_of_enrichment for more information.

- Demography, i.e., birth and death of individuals, can be incorporated by adding a death term to each host class and a birth term to the susceptible host class.
- Sexually transmitted infections (STIs) may require separate dynamical equations for the female and male population.
- Age-structure can be used to study the age-related incidence pattern of certain infections.
- Pathogen induced mortality can be included by adding a death term to the infected host class.
- Infection can be divided into symptomatic and asymptomatic host classes with different transmission rates.
- Infectious periods can be modeled as a multi-stage process to account for acute and chronic infection.

3.3 BASIC REPRODUCTION NUMBER R_0

The basic reproduction number is one of the most important concepts of mathematical epidemiology.

Sometimes, it is important to know if an infectious disease can spread and be maintained endemically in a host population. An important measure to understand this is the concept of the basic reproduction number, also called R_0. It is defined as *the expected number of secondary infections arising from a single individual during his or her entire infectious period, in a population of susceptibles*. If the number is larger than 1, the introduction of a single infected case can cause an epidemic that might later result in an endemic disease equilibrium (Figure 8). In a homogeneously mixing population, we can calculate R_0 by the rate at which new cases are produced by an infectious individual[5] multiplied by the average infectious period:

$$R_0 = \text{transmission rate} \times \text{infectious duration} = \beta D = \frac{\beta}{\gamma}. \quad (21)$$

While the intuitive formula above is useful for a variety of diseases, it often fails to be a good estimator for R_0. For example, if the population of hosts is structured by the pattern in which they make contacts (see Section **??**), or their susceptibility to infection, one must calculate R_0 in a more sophisticated way, such as the next generation method introduced by Diekmann et al. (1990). A broader perspective on the basic reproduction number is provided by the excellent review by Heffernan et al. (2005).

The basic reproduction number varies widely across diseases and populations (Table 2) but control efforts have in common that they aim to reduce R_0 below 1 so that the infection cannot sustain itself.

5 We now assume frequency dependent transmission, i.e., the scenario when the total population size is set to $N = 1$.

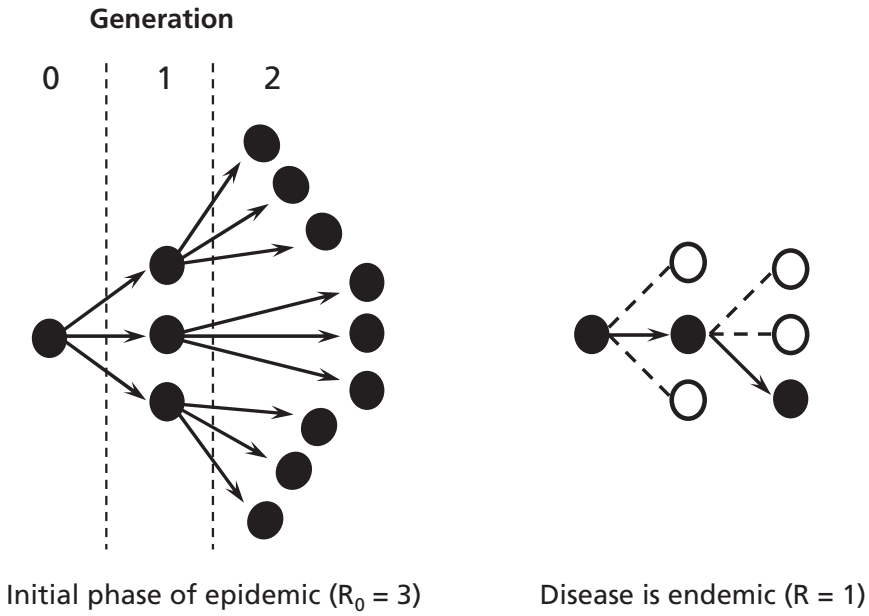


Figure 8: The number of new infections generated when the basic reproductive number (the number of new cases created by a single primary case in a susceptible population) is 3. Cases of disease are represented as dark circles, and immune individuals are represented as open circles. When there is no immunity in the population (left) and the basic reproductive number (R_0) is 3, the initial infectious case generates on average 3 secondary cases, each of which in turn generates 3 additional cases of disease. Once the disease becomes endemic owing to acquired immunity (right), each case generates on average 1 additional case (effective reproductive number $R = 1$). Figure from Pandemic Influenza Outbreak Research Modelling Team (Pan-InfORM) and Fisman (2009).

Treating infected individuals reduces infectious duration D; hygiene measures or quarantining reduce the transmission rate β. The burden of a disease may also be reduced by vaccinating a sufficient proportion of the population. In the exercises in Chapter 7 we will see how such vaccination thresholds can be derived for influenza.

DISEASE	TRANSMISSION	R_0
Measles	Aerosol	12–18
Chickenpox (varicella)	Aerosol	10–12
Mumps	Respiratory droplets	10–12
Rubella	Respiratory droplets	6–7
Pertussis	Respiratory droplets	5.5
SARS-CoV-2 (Delta)	Droplets and aerosol	5.1
Polio	Fecal-oral route	5–7
SARS-CoV-2 (Alpha)	Droplets and aerosol	4–5
Smallpox	Respiratory droplets	3.5–6.0
SARS-CoV-2 (ancestral)	Droplets and aerosol	2.4–3.4
Diphtheria	Saliva	1.7–4.3
HIV/AIDS	Body fluids	2–5
SARS-CoV	Respiratory droplets	2–4
Ebola (2014 outbreak)	Body fluids	1.4–1.8
Influenza (2009 pandemic)	Respiratory droplets	1.3–2.0
Influenza (seasonal)	Respiratory droplets	1.2–1.4
MERS-CoV	Respiratory droplets	0.3–0.8

Table 2: Values of the basic reproduction number R_0 of well-known infectious diseases. Note that the estimated values can strongly depend on the population of interest. Data from http://en.wikipedia.org/wiki/Basic_reproduction_number.

4

STOCHASTIC EFFECTS

Stochastic simulations can account for the expected variability in epidemics.

The infectious disease models in Chapter 3 are deterministic. They are like a clockwork system in which the same initial conditions result in the same epidemic trajectory. Real world epidemics are not like this, and the element of chance plays an important role. For example, if we could rerun an epidemic, the same individuals would not all become infected. Stochastic models deal with the probabilistic element. They help us understand which properties of an epidemic are governed by random events and to what extent. Stochastic models have the following two important properties:

1. Variability between simulations: Multiple simulations result in different outcomes. This is the most important element of stochastic models. Although an infection may possess some general statistical properties (i.e., having a defined value of R_0 Figure 9), we cannot determine the precise epidemic trajectory.

2. Extinction: In many stochastic models, hosts are represented as integer values. Chance fluctuations can thus drive an epidemic to go extinct. This can have important implications for modeling the impact of vaccination strategies or early epidemic outbreaks.

Stochastic models are useful for assessing the variability one expects in an epidemic, and they provide important information about real world behavior. Sometimes it is also important to analyze real world data so we can understand an underlying deterministic process. We will briefly discuss how observational noise can be dealt with in parameter inference. In Section 4.2, we discuss the way the infection process can cause stochastic noise and show how we can replicate it using stochastic computer simulations.

4.1 OBSERVATIONAL NOISE

One way of dealing with noise in infectious disease modeling is to sort out observational noise. One can assume that the underlying epidemic follows a deterministic (predictable) dynamic but that the observed data is prone to error. This type of noise comes from different sources, such as:

1. Misreporting: Several diseases are difficult to diagnose because they share symptoms with other diseases (e.g., influenza), or

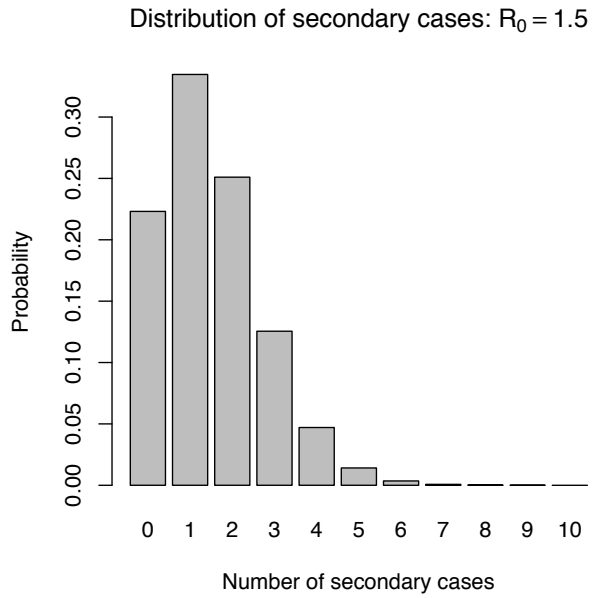


Figure 9: Probability distribution that one infected individual will spread the infection to n other individuals in an entirely susceptible population. The probability distribution shown here is a Poisson distribution with mean $R_0 = 1.5$. A Poisson distribution for the number of secondary cases is obtained if we assume a well-mixed population in which there is a small constant probability of transmission per contact between a susceptible and an infected individual.

because infections remain asymptomatic and are therefore undetected.

2. Uncertainty or error in diagnosis: Laboratory tests can result in false positive (specificity) or false negative (sensitivity) diagnoses. Both errors can be caused by humans (handling of samples) or machines (test kit).

Observational noise does not impact the epidemic itself but stems from reporting data.

3. Small sample sizes: The smaller the testing sample (number of hosts), the larger the expected uncertainty around the proportion of infected individuals.

Observational noise plays an important role in parameter inference. Consider the noise that is expected due to small sample sizes: to obtain estimates of the population prevalence of an infection, individuals from a representative subpopulation must be tested. The second British National Survey of Sexual Attitudes and Lifestyle in Britain (Natsal-2) provides test results for chlamydia, the most common bacterial sexually transmitted infection (STI) among young adults (Fenton et al., 2001; Johnson et al., 2001). Among individuals of 20 years of age, five out of 101 tests are positive for chlamydia infection. This means that the prevalence of chlamydia in the sample population is 4.95%, assuming that the test sensitivity and specificity are 100%.

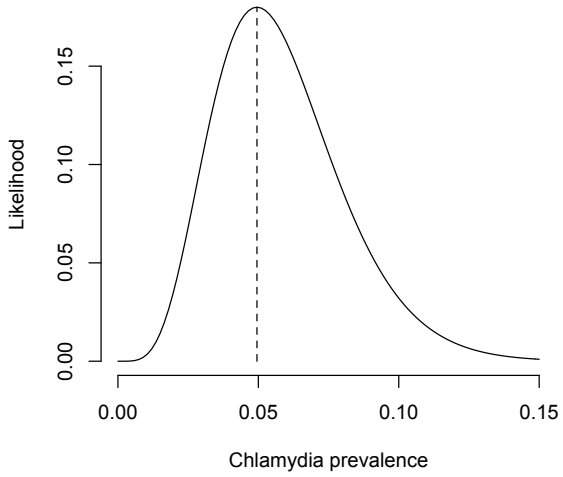


Figure 10: Likelihood of chlamydia prevalence among 20 year olds in Britain. 5 among 101 tests were positive for chlamydia (Fenton et al., 2001). The dashed vertical line denotes the most likely prevalence estimate (4.95%).

But what can we say about the 'true' chlamydia prevalence in the population of 20 year olds in Britain? It could well be that the 5 positive tests in the sample are a result of chance. The true prevalence in the population can be inferred by a likelihood approach. The likelihood that k out of N individuals test positive for chlamydia as a function of the true population prevalence p follows the binomial distribution:

$$\text{Prob}(k|p,N) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{22}$$

Figure 10 shows the likelihood of the prevalence, given the observed test results. The most likely value of the prevalence is indeed 4.95% but it becomes clear that confidence around this estimate is relatively large because noise was generated by the small sample size of the tested population. This kind of likelihood approach is widely used in infectious disease epidemiology (see Section **??**) and in the field of ecology in general (Bolker, 2008).

4.2 PROCESS NOISE

Apart from observational noise, the infection process can cause some randomness in the epidemic dynamics, which has important implications because stochastic effects early in an epidemic may be propagated forward in time. One modeling technique that includes stochasticity considers that the governing parameters of an ordinary differ-

ential equation (ODE) are noisy for two reasons. First, variability in parameters can be caused by an external, unpredictable force; for example, the transmission term β can be modified by certain climatic conditions (viruses often survive better in a cold environment and people might be more likely to use public transport when it rains). Second, parameter fluctuations may stem from the variability of individuals themselves, such as their pattern of contacts. This latter form of variation is expected to average out over a large number of infected individuals.

Inclusion of demographic stochasticity by so-called event-driven approaches is the most important method for including randomness in epidemiological models. These incorporate fluctuations in population processes that arise from the random nature of events at the individual level. We have seen such randomness in the pattern of chlamydia persistence in asymptomatically infected women (Figure 3). Although the infection appears to be cleared at a constant rate (or with the same probability) in the cohort as a whole, the infectious period of each individual woman is a different length. Stochastic computer simulations can take this into account.

4.2.1 *Gillespie algorithm*

Gillespie's algorithm is an exact method to simulate stochastic effects.

In his seminal paper, Gillespie (1977) introduced this algorithm to simulate chemical or biochemical systems of reactions efficiently and accurately. It belongs to the class of Monte Carlo methods and has become popular in the field of infectious disease modeling. It is useful because it represents an exact stochastic representation of a deterministic system that can be formulated by ODEs.

Here, we consider a simple SIS model. This is a variant of the SEIRS model from Section 3.2 without a period of latency and immunity:

$$\frac{dS}{dt} = -\beta SI + \gamma I, \tag{23}$$

$$\frac{dI}{dt} = \beta SI - \gamma I. \tag{24}$$

In the stochastic version of this model, we assume that the state space is integer valued, i.e., that S and I can only take values 0, 1, 2, 3, ... A certain state (S, I) may then change if an individual becomes infected or recovers. All these events occur at the rates that we have in the deterministic SIS model equations (23) and (24). The following table lists all possible events and the rates at which they occur:

$$\begin{pmatrix} S \\ I \end{pmatrix} \xrightarrow{\beta SI} \begin{pmatrix} S-1 \\ I+1 \end{pmatrix}, \tag{25}$$

$$\begin{pmatrix} I \\ S \end{pmatrix} \xrightarrow{\gamma I} \begin{pmatrix} I-1 \\ S+1 \end{pmatrix}. \tag{26}$$

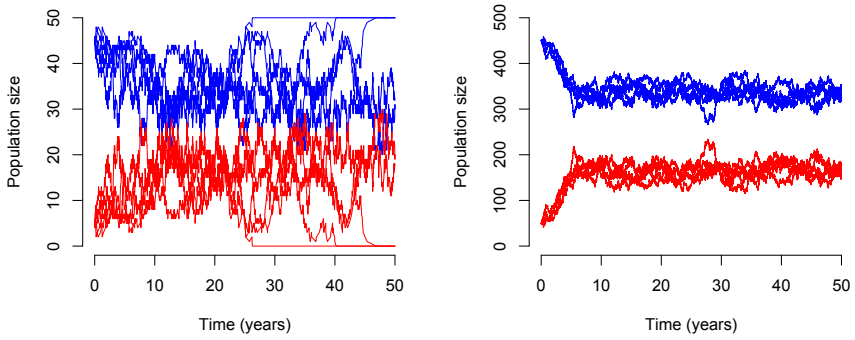


Figure 11: Stochastic simulation runs of the SIS model with a population of $N = 50$ (left panel) and $N = 500$ (right panel). The parameters used in the simulations are: $R_0 = 1.5$; $1/\gamma = 1$ year; $I(0) = N/10$.

How do we simulate this process? The following pseudo code provides an efficient and accurate implementation of Gillespie's direct method (Gillespie, 1977):

1. Initialize the populations in your system ($S = 999$, $I = 1$) and set $t = 0$.
2. For each event ($n = 2$) determine the rate at which it occurs ($a_1 = \beta SI$, $a_2 = \gamma I$).
3. Calculate the rate at which any event occurs ($a_0 = \sum_{\nu=1}^{n} a_\nu$).
4. Generate two random numbers[1] that are uniformly distributed between 0 and 1 (r_1 and r_2).
5. Calculate the time until the next event ($\tau = (1/a_0)\log(1/r_1)$).
6. Choose the next event so that $\sum_{\nu=1}^{\mu-1} a_\nu < r_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu$.
7. Set time to $t = t + \tau$.
8. Adjust the populations according to event μ.
9. Return to step 2.

Figure 11 shows the infection dynamics for the stochastic SIS model with a population size of $N = 50$ and $N = 500$. The infection grows initially and fluctuates around its endemic steady-state at $I_e = (1 - 1/R_0)N$. There is quite a lot of variation between individual simulation runs despite the identical parameters and initial values of the variables across runs. The variation is much larger if the population size is smaller, which is a general result for stochastic processes. In

1 See http://en.wikipedia.org/wiki/Random_number_generation on how random numbers can be computed.

fact, the infection dynamics can come to a halt eventually in the simulations with $N = 50$. This behavior is contrary to that of the deterministic SIR model, in which an infection does not go extinct if $R_0 > 1$ but attains the endemic equilibrium. This phenomenon is also observed empirically and is termed *stochastic fade-out*.

5

INTRODUCTION TO PHYLOGENETIC ANALYSIS[1]

Pathogens spread through rapid replication in one host followed by transmission to another host, and an outbreak can only take off when one infection results in more than one subsequent infection. As the pathogen replicates and spreads, its genome needs to be replicated many times and random mutations – copying mistakes – will accumulate in the genome. Such random mutations can help to track the spread of the pathogen and to learn about its transmission routes and dynamics. Figure 12 shows a transmission tree with a subset of cases that were sampled. In practice, the transmission tree is unknown and typically only rough estimates of the total case counts are available. Genome sequences, however, will allow us to reconstruct the transmission tree. In this example, mutations are indicated on the tree and these mutations allow us to group samples into clusters of closely related viruses that belong to the same transmission chains. Real data will contain many more mutations and many more samples that allow detailed reconstructions.

Genome sequences record pathogen transmission.

5.1 SEQUENCE DATA

Sequencing of pathogens from patient samples and the processing of the raw sequence data won't be discussed here. We start assuming we have genome sequences of the same pathogen sampled from many cases during an outbreak. These will typically come as text files in FASTA format:

```
>Sample_A
ACTGCTCGATCGCTACGCTACGCTA...
>Sample_B
CTGCTCCATCGATCGCTANNNN...
```

Each line that starts with a '>' sign is interpreted as a line that specifies the name of the sample, followed by the sequence. The sequence can be broken over several lines. Sequences obtained in outbreak scenarios will often be incomplete. Unknown bases are typically denoted by 'N', but other characters like '?' or '.' are also common. FASTA files like the above are the point of departure for phylogenetic analyses.

5.2 ALIGNMENT

Partial genome sequences typically don't all start in exactly the same place and mutations not only change the nucleotide content but can

1 This chapter is adapted from https://neherlab.org/201901_krisp_intro.html.

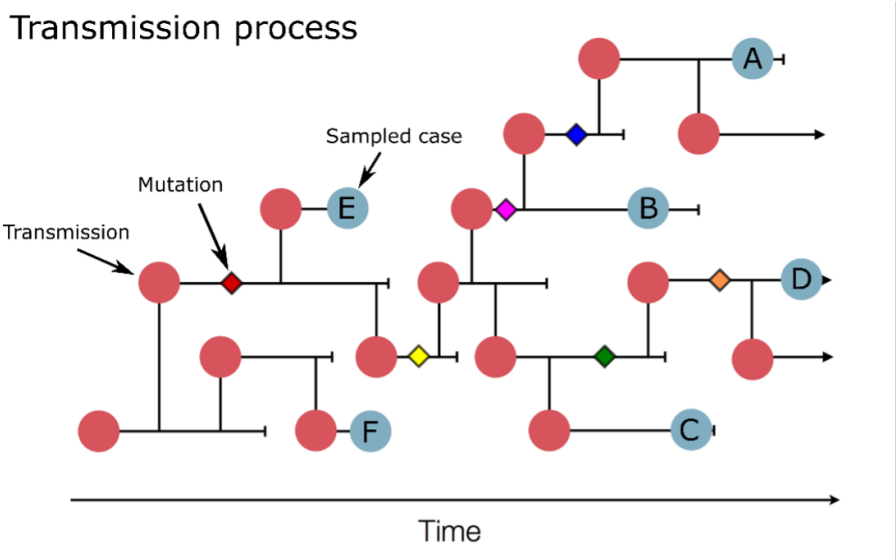


Figure 12: Transmission tree with sampled cases (blue). Mutations are indicated as little diamonds.

also delete or insert a couple of bases. As a consequence, the different sequences are not in register and need to be aligned. Such alignments are a common task in bioinformatics and many robust programs are available to solve this problem. The aim is to insert gaps into the sequences, encoded with character '-', that corresponding characters are in the same column. For the two sequences from above, this would look like this:

```
ACTGCTCGATCGCTACGCTACGCTA...
-CTGCTCCATCGAT-CGCTANNNNA...
```

At most positions, the two sequences now agree. Formally, alignment is trying to insert gaps such that homologous parts (i.e., parts that descent from a common ancestral sequence) of the sequence are in the same column. Common programs to perform this task are:

- MAFFT (https://mafft.cbrc.jp/alignment/software/)
- Muscle (https://www.drive5.com/muscle/)
- ClustalW (http://www.clustal.org)

For SARS-CoV-2, the Nextstrain (https://nextstrain.org) team also developed a specialized aligner called NextAlign[2]. The output of an alignment is best viewed (and edited) in specialized multiple alignment viewers. A very common one is aliview (Figure 13).

2 See https://docs.nextstrain.org/projects/nextclade/en/stable/user/nextalign-cli.html for more information.

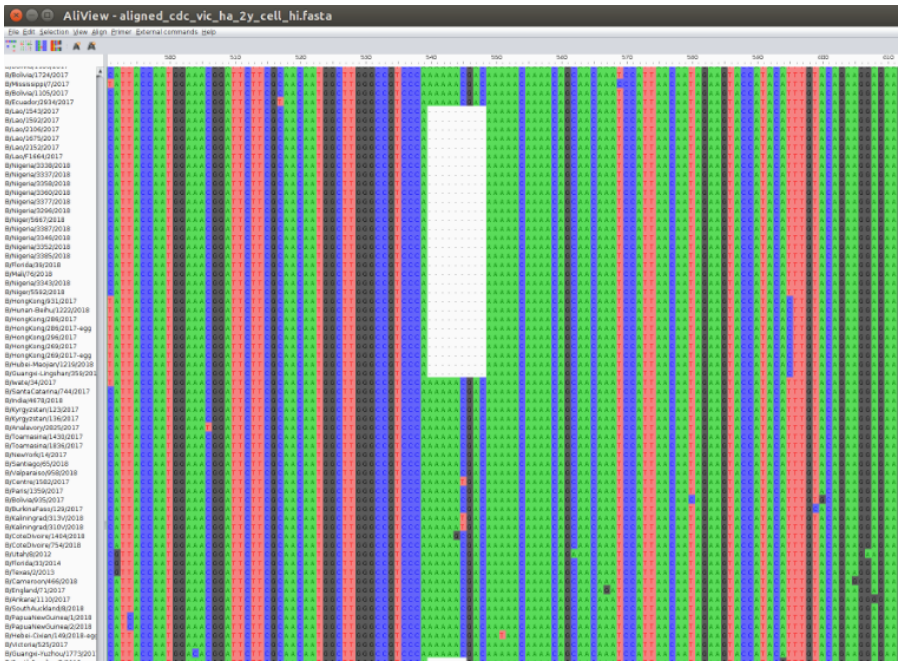


Figure 13: Multiple alignment viewer AliView (http://www.ormbunkar.se/aliview/).

5.3 PHYLOGENETIC TREES

Once sequences are aligned, we can meaningfully look for differences between them and group them into clusters of similar sequences, which in essence boils down to reconstructing their "family tree" or evolutionary history. In Figure 14, it is straightforward to group sequences according to the mutations they carry – in practice finding the tree that explains your data best is a computationally challenging problem. But a large number of highly optimized programs are available to solve this task. Commonly, the following three are used:

- IQ-TREE: Fast and accurate (http://www.iqtree.org)
- RAxML: Current gold standard, but slower than IQ-TREE (https://cme.h-its.org/exelixis/web/software/raxml/index.html)
- FastTree: Fast and rough, but often good enough (depends who you ask) (http://www.microbesonline.org/fasttree/)

The most common output format of these programs is called Newick format. The branch length of trees estimated by the above programs roughly correspond to the fraction of positions in the sequence that changed along the branch (more precisely, the expected number of changes per site). Trees can be graphed in different ways (Figure 15). The most important distinction is whether a tree is rooted or not. In a rooted tree, a particular bifurcation is the ancestor of all sequences

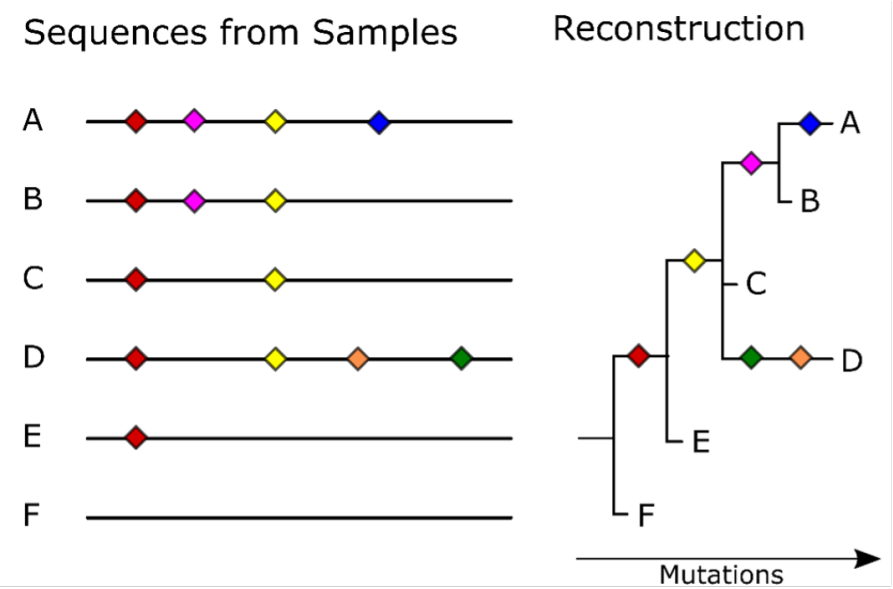


Figure 14: Tree reconstruction. Colored diamonds indicate differences between the sequences that we assume originated in a mutation event in a common ancestor of all sequences that carry the variant. Note that the sequences correspond to the transmission tree in Figure 12.

in the sample. Time scaled phylogenies (where branch length corresponds to time) will always be rooted and the root is by necessity the oldest node. In other applications, it is difficult to single out a root node and a more appropriate layout is and unrooted layout.

5.4 MOLECULAR CLOCK AND TRANSMISSION

The number of changes accumulate in a pathogen genome in a year depends on:

1. the mutation rate
2. the duration of one replication cycle
3. the size of the genome
4. the fraction of mutations that persist and spread

While the first three factors are very natural and intuitive, the last one depends on the biology of the pathogen and the environment it is replicating in. The majority of mutations that change the sequence of important proteins will be detrimental and are quickly removed by natural selection, while other mutations, for example those that help a virus to evade human immunity, will preferentially spread. In addition, a substantial fraction of mutations (mostly mutations at 3rd codon positions) don't affect pathogen replication much and accumulate freely. As a consequence, the exact rate at which changes

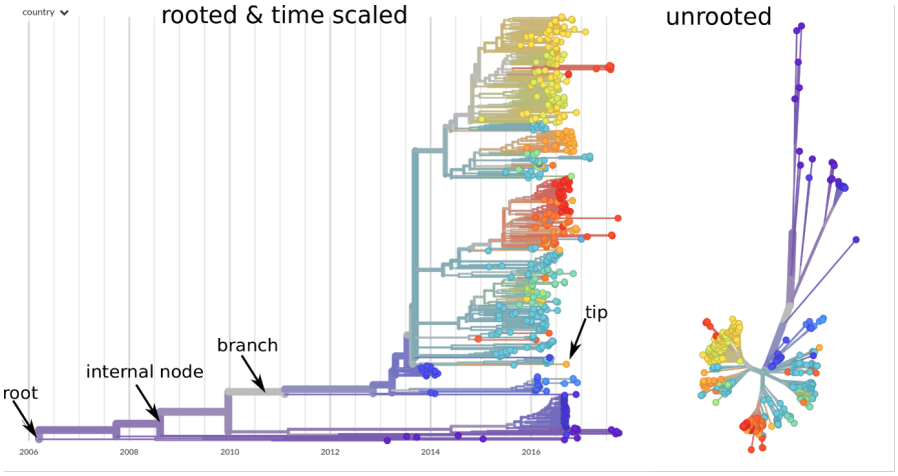


Figure 15: Phylogenetic trees can be rooted (left) or unrooted (right).

accumulate is not known *a priori*. Nevertheless, mutations tend to accumulate at a constant rate and the accumulation can be used as a *molecular clock*.

The molecular clock is a relationship between time and the expected number of changes in the sequence. These substitution rates vary greatly between organisms and within organisms between genomic regions:

- RNA viruses: $0.0001 \ldots 0.01$ substitutions per site and year
- Bacteria: $10^{-8} \ldots 10^{-6}$ substitutions per site and year

Multiplying the rate per site with the size of the genome, we get the expected number of substitutions per year:

- RNA viruses: Genome size $\approx 10^4 \rightarrow 1 - 100$ substitutions year and genome
- Bacteria: Genome size $5 \times 10^6 \rightarrow 0.05 - 5$ substitutions per year and genome

This suggests that whole genome sequencing often has high enough resolution to distinguish pathogens that were sequences a few month apart – sometimes even down to a few weeks.

5.5 TIME RESOLUTION

From basic phylogenetic trees, where branch lengths represent mutational changes, we can infer *time resolved trees* if we have a good selection of samples and know the date these samples were taken. Importantly, how well this will work depends on things like having a wide enough range of samples over time, the mutation rate of our pathogen (and how stable it is), and the presence of things like recombination (this will make clock inference much harder).

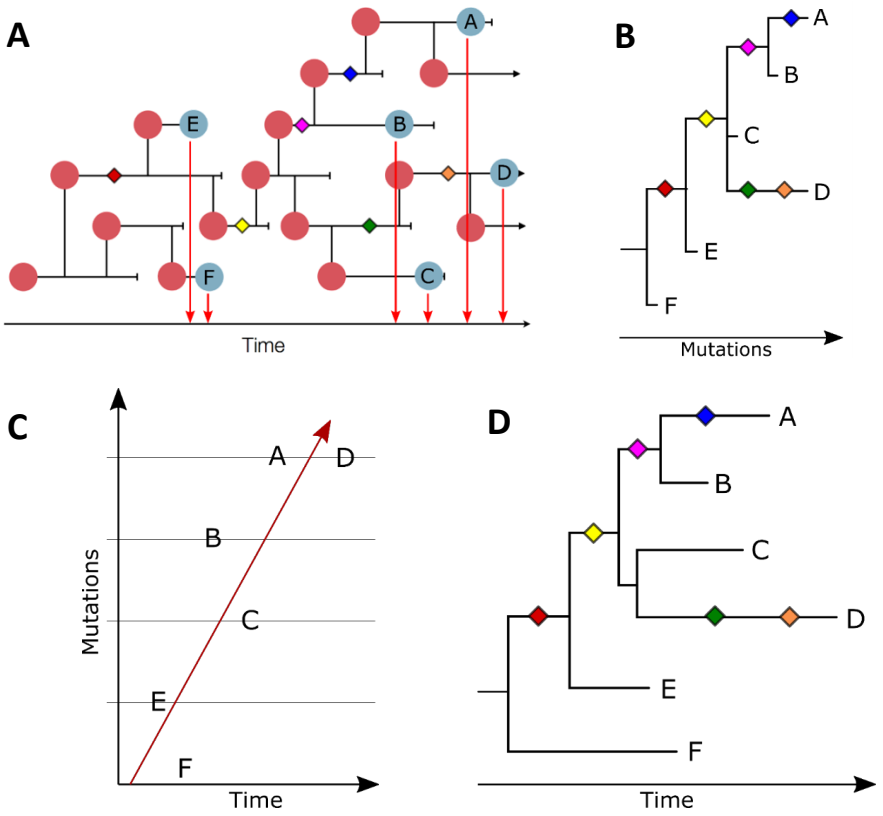


Figure 16: Reconstruction of a time-resolved tree.

Making a time resolved tree essentially infers when hypothetical ancestral sequences (which are *internal nodes* on the tree) existed, based on the mutation rate, and when sampled sequences were taken. In a transmission tree, if the sampled sequences are shown in blue, we can see what time they were taken (Figure 16A). Based on our basic phylogeny (Figure 16B), we can plot these times against the number of mutations from the root we observe in each sample (Figure 16C). This allows us to adjust our mutation-based phylogeny, stretching branch lengths and moving internal nodes, to reflect our inferred clock, and create a time-resolved tree (Figure 16D). Time-resolved trees can be more intuitive, particularly to people less familiar with phylogenetics, as it can be easier to see how a pathogen has spread and changed over time. In addition, it can provide information about the past. However, time-resolved trees are always *inferred*. This means reconstructions will depend on input data, and the confidence intervals of estimates (and the underlying mutational structure) should always be checked – they can be misleading!

5.6 TUTORIAL: NEXTCLADE

Nextclade (`https://clades.nextstrain.org`) is a web tool that provides a great way to get an overview of sequences, including their

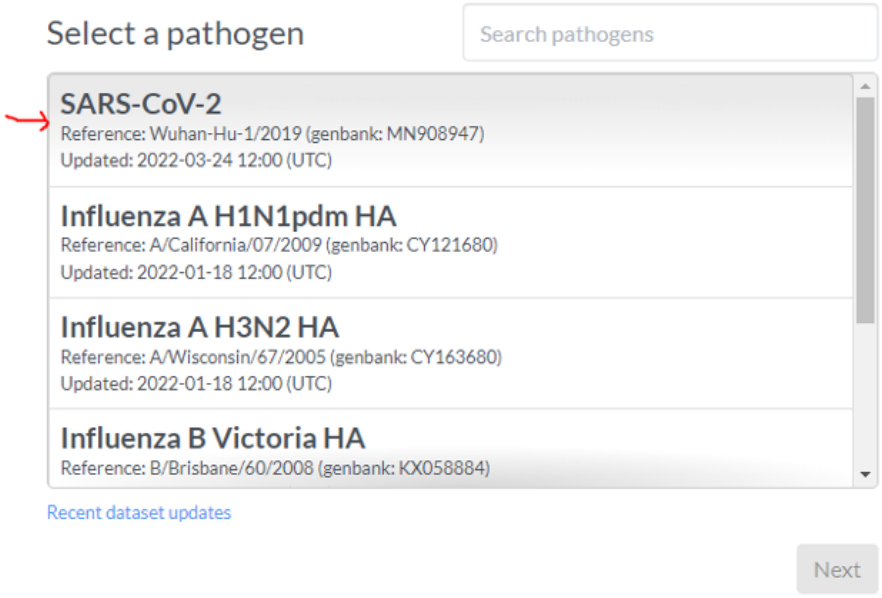


Figure 17: Pathogen selection in Nextclade.

quality ("QC"), what mutations they have, and where they fall in a basic tree. Because it works online, we will use it as a very basic way to do some phylogenetic investigation. However, it is not a replacement for a true phylogenetic analysis. Note that Nextclade is also available as a command-line-interface (CLI) version, for more intensive uses. When you first visit the website, you will need to select the pathogen you are going to analyze. For this tutorial, select SARS-CoV-2 (Figure 17).

You can then drag-and-drop FASTA files onto the website, and press `next` to analyze them. They may take a few minutes to appear (Figure 18). The analysis pipeline comprises the following steps:

1. Alignment: Sequences are aligned to the reference genome using our custom Nextalign alignment algorithm.
2. Translation: Nucleotide sequences are translated into amino acid sequences.
3. Mutation calling: Nucleotide and amino acid changes are identified.
4. PCR primer changes are computed.
5. Phylogenetic placement: Sequences are placed on a reference tree, clades assigned to nearest neighbour, private mutations analyzed.
6. Quality control: Quality control metrics are calculated.

You can mouse over different parts of the interface to learn more about what they show, or the details for a specific sequence, such

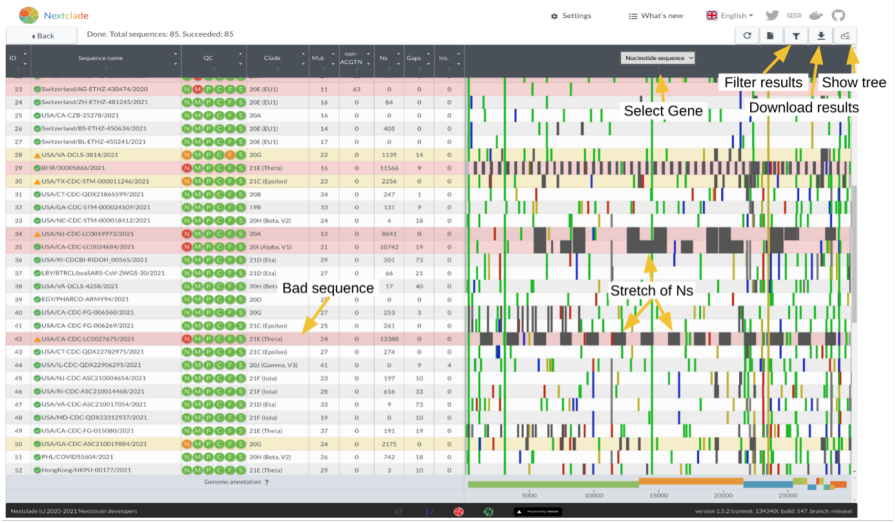


Figure 18: Results overview in Nextclade.

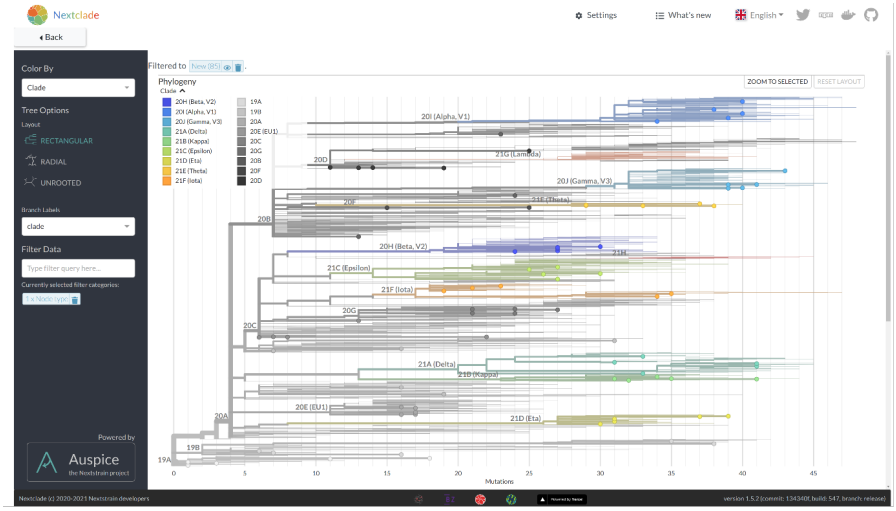


Figure 19: Tree view in Nextclade.

as the mutations it has and their impact on amino-acids. You can also click on the `show tree` button on the top-right to view how the sequences you have analyzed fall into a basic SARS-CoV-2 phylogeny (Figure 19). Note that this view is great for getting an overview, but it is not a replacement for a full phylogenetic analysis.

Part II

EXERCISES

The exercises consist of computer practicals using the programming language R. Example scripts are given to simulate an influenza epidemic and to explore stochastic effects during the early phase of an Ebola outbreak.

6 INTRODUCTION INTO R

R is a powerful language and environment for statistical computing and graphics. It is open source and can be downloaded for Windows, Mac and Linux from the website of The R Project for Statistical Computing (http://www.r-project.org). RStudio (http://www.rstudio.com) offers a convenient integrated development environment that is also freely available.

In Section 6.1, we provide a very brief introduction into some basic commands in R. Additional explanation will be given in the exercises but we also want to refer to the comprehensive information offered on the R website ("An Introduction to R", http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf). You can also download a concise "reference card" that is useful to have around when writing R code[1].

6.1 SOME BASIC COMMANDS

R is an interpreted language typically used through a command line interface. If one types `10^2 + 5.6` at the command prompt and presses enter, the computer calculates the result:

```
> 10^2 + 5.6
[1] 105.6
```

You can also give numbers a name. For example, you can assign the number `3` to the variable `a` by typing

```
> a <- 3
```

The `a` is now called a *scalar* (a single number that is 0-dimensional). To generate a 1-dimensional *vector*, you can use the function `c` to concatenate (paste together) several numbers:

```
> b <- c(a,4,5)
```

If you want to know the third element of `b`, simply type `b[3]`. You can also create 2-dimensional structures, like a table, that are called *matrices*.

There are various built-in functions. For example, if you want to calculate the sum of vector b, you can type

```
> sum(b)
[1] 12
```

1 Download at http://cran.r-project.org/doc/contrib/Short-refcard.pdf

R also offers excellent plotting features for publication quality graphs. By typing

```
> x <- rnorm(1000)
> plot(x)
```

you create a graph with 1000 normally distributed random numbers. Use `help(function)` or `?function` to show some documentation for a given function.

6.2 NECESSARY PACKAGES

Many customized functions for numerical and statistical analysis, plotting, etc. are available through additional packages. For the exercises, you need to download and install the following package for R:

- *deSolve*: Solvers for ordinary differential equations (http://cran.r-project.org/web/packages/deSolve)

The *deSolve* package is part of The Comprehensive R Archive Network (CRAN) and you can download and install it directly by typing `install.packages("deSolve")`.

7 SIMULATING AN INFLUENZA EPIDEMIC

Listing 1 shows the R code for simulating an influenza epidemic in a population of 10,000 individuals. It makes use of the function `ode` from the package *deSolve* for numerically integration the ODEs. Since we want to focus on a single season of an influenza epidemic, we can assume an SIR structure with lifelong immunity. The model is parameterized through R_0 that is then converted into the transmission rate β. The resulting dynamics shown in Figure 20 resembles a typical influenza epidemic during the winter season, peaking around two months after the introduction of the first case.

Listing 1: SIR model of influenza

```
# Library for ordinary differential equation solvers
library(deSolve)

# Definition of the SIR model
SIR <- function(t, x, parms) {
        with(as.list(c(parms, x)), {
                N <- S + I + R
                beta <- R_0*gamma/N
                dS <- - beta*S*I
                dI <- beta*S*I - gamma*I
                dR <- gamma*I
                der <- c(dS, dI, dR)
                list(der)
        })
}

# Set the parameters of the model
parms <- c(R_0 = 1.5, gamma = 1/4)

# Set the initial values
inits <- c(S = 1e4 - 1, I = 1, R = 0)

# Set the time points for evaluation
times <- seq(0, 150, 1)

# Simulate the model using the function ode()
sim <- as.data.frame(ode(inits, times, SIR, parms))

# Plot the results
plot(sim$time, sim$S,
     type = "l", lty = 3, col = "blue", ylim = c(0, sum(inits)),
     xlab = "Time (days)", ylab = "Number of individuals", frame=
         FALSE)
```

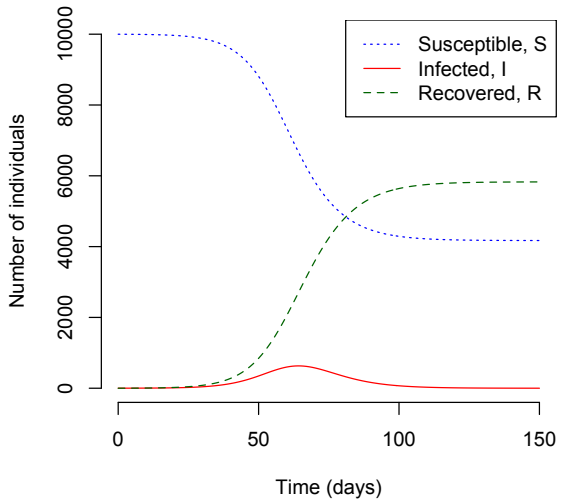


Figure 20: Illustrative time course of an influenza epidemic. The number of infected individuals (red solid line) reaches a peak at about two months after the introduction of a single infected case. $R_0 = 1.5$ and the average infectious duration is 4 days.

```
lines(sim$time, sim$I,
      lty = 1, col = "red")
lines(sim$time, sim$R,
      lty = 2, col = "darkgreen")
legend("topright",
       legend = c("Susceptible, S", "Infected, I", "Recovered, R"
           ),
       col = c("blue", "red", "darkgreen"), lty = c(3,1,2))
```

Exercise:

- Play around with the parameter values by changing their values in `parms`. How does R_0 and γ influence the dynamics? Can you explain this?

7.1 EXPLORING THE IMPACT OF VACCINATION

Vaccination is a widely used intervention against seasonal influenza as it reduces the population of susceptible individuals. If one assumes that vaccination provides 100% sterilizing immunity, all vaccinated individuals can be directly moved into the compartment of recovered individuals R.

Exercises:

- Assume that 20% of the population are vaccinated and set the initial value of R to 20% of the total population. What is the effect on the influenza epidemic?

- Now assume that 50% of the population are vaccinated. What happens? Can you explain it?
- Use pencil and paper to derive an analytic solution of the vaccination threshold, i.e., the level of vaccination needed to prevent an outbreak. Use R to plot this relation, using the function curve().

Advanced exercise:

- Extend the model with temporary immunity so that recovered individuals can become susceptible again. Play around with the parameter values to see how this affects the time course of the epidemics. Can you calculate the expected numbers of susceptible, infected and recovered individuals at the endemic steady-state?

8

STOCHASTIC EFFECTS DURING EBOLA OUTBREAKS

Listing 2 provides an implementation of Gillespie's stochastic simulation algorithm (see Section 4.2.1) for an Ebola transmission model cite (Althaus, 2015). Note that the package *GillespieSSA* (http://cran.r-project.org/web/packages/GillespieSSA) from CRAN provides a much more flexible implementation of Gillespie's stochastic simulation algorithm. However, the provided listing describes all essential steps in detail. Following 100 Ebola outbreaks in absence of any control measures illustrates the variability caused solely by stochastic effects (Figure 21).

Listing 2: Stochastic model of Ebola outbreak

```
# Set random number generator
set.seed(4386)

# Set the parameters of the model
N = 500; seed = 1; R_0 = 1.5; sigma = 1/9; gamma = 1/7; f = 0.7
beta <- R_0*gamma/N

# Set the maximal simulation time
maxtime = 1000

# Initialize plot
plot(NA, NA,
     xlim = c(0, 400), ylim = c(0, N),
     xlab = "Time (days)", ylab = "Cumulative number of cases",
         frame = FALSE)

# Perform 100 simulations
runs <- 100
outbreak <- as.data.frame(matrix(data = NA, nrow = runs, ncol =
    2))
names(outbreak) <- c("duration", "size")
cols <- sample(terrain.colors(runs))
for(i in 1:runs) {
        # Algorithm for Gillespie's direct method
        # Set initial conditions
        S = N - seed; E = 0; I = seed; R = 0; D = 0; C = 0;
            simtime = 0; event = 0
        simulation <- matrix(data = NA, nrow = 3*N, ncol = 7)
        simulation[1, ] <- c(simtime, S, E, I, R, D, C)
        while(simtime < maxtime) {
                # Calculate the event rates
```

```
                rates <- c(beta*S*I, sigma*E, (1-f)*gamma*I, f*
                    gamma*I)
                sum_rates <- sum(rates)
                cum_rates <- cumsum(rates)
                # Interrupt the loop if infection goes extinct
                if(sum_rates == 0) break
                # Time until next event
                tau <- 1/sum_rates*log(1/runif(1))
                simtime <- simtime + tau
                # Choose which event and update populations
                rand <- runif(1, 0, sum_rates)
                if(rand < cum_rates[1]) {
                        S <- S - 1
                        E <- E + 1
                } else {
                        if(rand < cum_rates[2]) {
                                E <- E - 1
                                I <- I + 1
                                C <- C + 1
                        } else {
                                if(rand < cum_rates[3]) {
                                        I <- I - 1
                                        R <- R + 1
                                } else {
                                        I <- I - 1
                                        D <- D + 1
                                }
                        }
                }
                event <- event + 1
                simulation[event + 1, ] <- c(simtime, S, E, I, R,
                     D, C)
        }
        # Plot the cumulative number of cases for each simulation
             with a different color
        lines(simulation[, 1], simulation[, 7], col = cols[i])
        # Store the duration and size of an outbreak
        outbreak[i, ] <- c(simtime, C)
}
```

Exercises:

- Play around with different parameters values and population sizes. If you want to follow the infection over longer time periods, you can increase `maxtime` accordingly. You might also need to change the scale of the x-axis in the function `plot()`.
- We learned that if $R_0 < 1$, an infection cannot grow and will not cause an outbreak in a deterministic model. Try using values of R_0 that are just slightly below 1. Can you explain the observed dynamics?

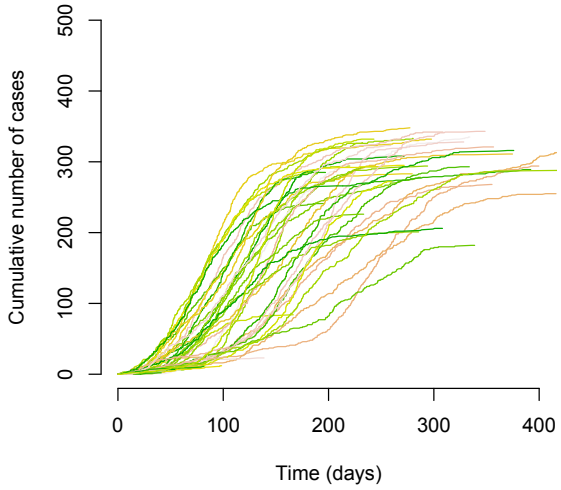


Figure 21: Multiple simulations of an Ebola outbreak. Only around a third of all simulations result in an epidemic outbreak and show substantial variability in the time and the number of infected individuals at the peak of the epidemic. Parameters as in Fig. 20.

- The matrix `outbreak` stores the duration and the final size (the cumulative number of infected cases) of a simulated outbreak in the first and second column, respectively. Use this information to calculate the probability that the infection goes extinct early during the epidemic.
- Explore the variability in the duration and the final size of epidemic outbreaks using the matrix `outbreak`. You can use the function `hist()` to create histograms.

Advanced exercises:

- Investigate how the probability of an outbreak changes as a function of the population size N, the seed size, R_0, the incubation period $1/\sigma$ and the infectious duration $1/\gamma$. You can write your own loop for different parameter values and store the calculated extinction probabilities in an array before you plot them as a function of the parameter.
- A ring vaccination trial of a recombinant vesicular stomatitis virus-Zaire Ebola virus in Guinea indicated that the vaccine might be highly efficacious in preventing Ebola virus disease (Henao-Restrepo et al., 2015). Investigate how the probability of an outbreak changes with increasing proportions of vaccinated individuals.

BIBLIOGRAPHY

Althaus, C. L. (2015). Rapid drop in the reproduction number during the Ebola outbreak in the Democratic Republic of Congo. *PeerJ*, 3:e1418.

Althaus, C. L., Heijne, J. C., Roellin, A., and Low, N. (2010). Transmission dynamics of Chlamydia trachomatis affect the impact of screening programmes. *Epidemics*, 2(3):123 – 131.

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

Bjørnstad, Ottar N (2018). *Epidemics: models and data using R*. Springer.

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ. ISBN: 0691125228.

Diekmann, O., Heesterbeek, H., and Britton, T. (2012). *Mathematical Tools for Understanding Infectious Disease Dynamics:*. Princeton University Press.

Diekmann, O., Heesterbeek, J. A., and Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio Ro in models for infectious diseases in heterogeneous populations. *J Math Biol*, 28(4):365–82.

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

Fenton, K. A., Korovessis, C., Johnson, A. M., McCadden, A., McManus, S., Wellings, K., Mercer, C. H., Carder, C., Copas, A. J., Nanchahal, K., Macdowall, W., Ridgway, G., Field, J., and Erens, B. (2001). Sexual behaviour in Britain: reported sexually transmitted infections and prevalent genital Chlamydia trachomatis infection. *Lancet*, 358(9296):1851–4.

Fine, P. E. (1975). Ross's a priori pathometry - a perspective. *Proc R Soc Med*, 68(9):547–51.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Hamer, W. (1906). *The Milroy lectures on epidemic disease in England: the evidence of variability and of persistency of type*. Milroy lectures. Bedford Press.

Heesterbeek, J. A. P. (2002). A brief history of Ro and a recipe for its calculation. *Acta Biotheor*, 50(3):189–204.

Heffernan, J. M., Smith, R. J., and Wahl, L. M. (2005). Perspectives on the basic reproductive ratio. *J R Soc Interface*, 2(4):281–293.

Henao-Restrepo, A. M., Longini, I. M., Egger, M., Dean, N. E., Edmunds, W. J., Camacho, A., Carroll, M. W., Doumbia, M., Draguez, B., Duraffour, S., Enwere, G., Grais, R., Gunther, S., Hossmann, S., Kondé, M. K., Kone, S., Kuisma, E., Levine, M. M., Mandal, S., Norheim, G., Riveros, X., Soumah, A., Trelle, S., Vicari, A. S., Watson, C. H., Kéïta, S., Kieny, M. P., and Røttingen, J.-A. (2015). Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. *Lancet*, 386(9996):857–66.

Johnson, A. M., Mercer, C. H., Erens, B., Copas, A. J., McManus, S., Wellings, K., Fenton, K. A., Korovessis, C., Macdowall, W., Nanchahal, K., Purdon, S., and Field, J. (2001). Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. *Lancet*, 358(9296):1835–42.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton.

Kermack, W. O. and McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Royal Society of London Proceedings Series A*, 115:700–721.

May, R. M. (2004). Uses and abuses of mathematics in biology. *Science*, 303(5659):790–793.

Pandemic Influenza Outbreak Research Modelling Team (PanInfORM) and Fisman, D. (2009). Modelling an influenza pandemic: A guide for the perplexed. *CMAJ*, 181(3-4):171–3.

Ross, R. (1911). *The Prevention of malaria*. J. Murray, London, 2nd edition.

Vanderpas, J., Louis, J., Reynders, M., Mascart, G., and Vandenberg, O. (2009). Mathematical model for the control of nosocomial norovirus. *J Hosp Infect*, 71(3):214–22.

Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling*. Oxford University Press.