

Block 3:

Additional methods and dynamics

**(stochastic models,
health economics,
phylogenetics)**

Lecture notes

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 17: Stochastic modelling and its applications Lecture

Objectives

All the models covered in the course until now have been deterministic, and have aimed to describe what may happen on average in a population. The objectives of this session are to:

- Introduce you to a number of different models that incorporate chance in whether events happen and can therefore be used to predict the variation in outcomes that may occur (even if all the parameters remain unchanged). These models are known as “stochastic” models.
- Summarise some of the advantages and disadvantages of stochastic models.

Some examples of when stochastic models might be used

We will start with a few examples, so that you get some appreciation of the role that chance may play and why stochastic models may be useful in epidemiology.

Estimating the likely size of an outbreak

Suppose that one measles case is introduced into a population of 10 susceptibles. What is the most likely size of the subsequent outbreak?

The following diagram summarizes predictions from a deterministic model, similar to one which you set up early in the course. This model predicts that, on average, after 10 days, there should be 3- 4 infectious cases in the population, and that by the 25th day, only 1 individual should still be infectious. In practice, predictions from this model are not very meaningful in this situation since the small numbers of susceptibles in the population means that chance may greatly influence how many cases occur throughout the outbreak, or indeed whether an outbreak occurs at all.

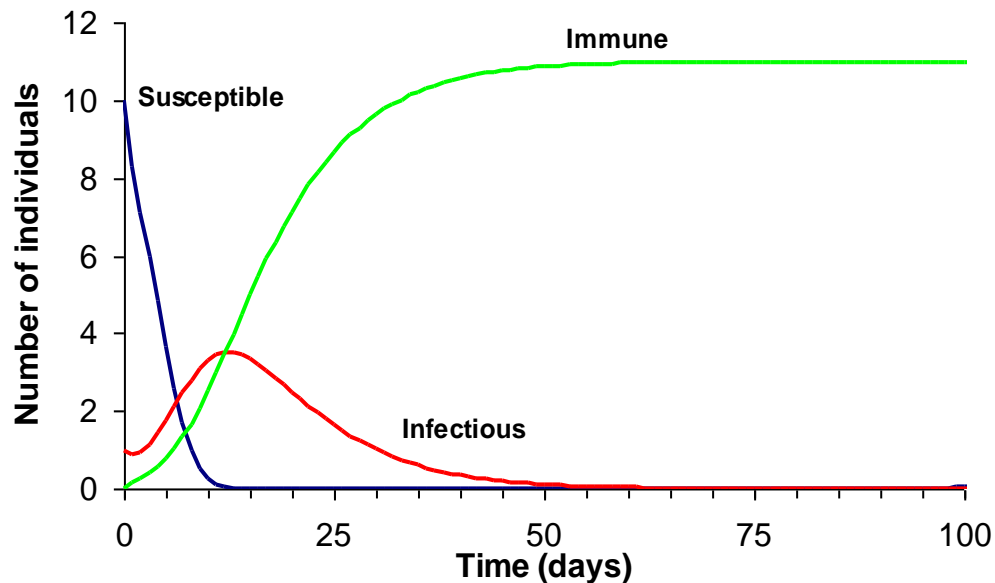


Figure 1: Predictions of the number of susceptibles, infectious and immune individuals following the introduction of 1 infectious case into a population comprising 10 susceptible individuals, assuming that the basic reproduction number is 13.

The chance of an outbreak (of a given size) and the basic reproduction number (R_0)

You should recall that with deterministic models major epidemics always occur when R_0 is greater than one (when an infectious individual is introduced in a fully susceptible population), and never when it is less than one. This is not very realistic. In the real world, if there is one initial infectious individual he or she may, by chance, infect no-one even when R_0 is greater than one. In this case, no epidemic would occur even though there was the potential for one. In a stochastic model, even when R_0 is greater than one, sometimes the epidemic will come to an end by chance after only a small number of secondary cases, or indeed none (when an epidemic comes to an end because there are no more infectious individuals we say it has *faded out*). However, provided R_0 is greater than one, there is also a chance that a major epidemic occurs and very large numbers are infected. If we could do the real experiment of introducing a case into a susceptible population and seeing if an epidemic occurs then we would probably want to repeat this experiment a number of times to get an estimate of the average behaviour, as we would expect that chance may well play a role. It is the same with a computer experiment. If we use a stochastic model that explicitly takes account of the role of chance (examples of which will follow), then we would probably want to repeat the experiment (simulations) a number of times to get an idea of the average behaviour, and the variation in the behaviour of the system that might be expected. We will do exactly this in the practical sessions.

The important thing to note is that we may well get different outcomes even with the same initial conditions and parameter values, and just because R_0 is bigger than one and a case is introduced into a susceptible population does not guarantee that an epidemic will occur. SARS provides a good example of this. R_0 for SARS was about 3. Roughly the same number of SARS cases were imported into British Columbia during the outbreak as were into Ontario. Yet Toronto had a large outbreak and BC had none. Was this due to the diligence of the BC public health authorities, or just chance?

Similarly, when R_0 is less than one (and greater than zero) there is still a chance that each infectious individual will infect others; quite long chains of transmission are still possible, and increasingly likely as R_0 gets closer to one. However, because each person infects, on average, less than one person, the self-sustaining chain reaction needed for a major epidemic cannot occur. Epidemic fadeout will occur before very large numbers are infected.

The relationship between R_0 and the distribution of the total number of cases infected in an outbreak has some important epidemiological applications. De Serres *et al.* (2000) showed how the size of measles outbreaks resulting from imported cases can be used to assess the status of disease elimination. The same idea has been used more recently to assess the risk of a major epidemic from an emerging pathogen (Ferguson *et al.*, 2004). We will return to this idea in the practical sessions.

Extinction of epidemics and persistence of infections

Look at Figure 1 again. At the end of the epidemic the deterministic model predicts that there is only a tiny fraction of an infectious person left. This cannot be, as people come in discrete units. So if you wanted to know when the epidemic will be over, you have to make some arbitrary assumptions e.g. less than 1 case, or less than 0.5 cases (so when you round to a whole number there is 0 cases) and so on. Either way, your estimate of the length of the epidemic will be somewhat arbitrary. Stochastic models deal in discrete units, and can therefore tell you when the epidemic is expected to finish, and the variation in this.

This also has implications for the persistence of an endemic disease. Before vaccination against measles (for instance), it was observed that measles persisted (was always present), in large cities but faded out in smaller populations (i.e. there were periods when measles was not around). For measles to take off again in these smaller populations it had to be re-introduced at some point after the number of susceptible had exceeded a critical density. Clearly a deterministic model cannot properly capture these patterns as the infection always persists in a deterministic framework, and the re-introduction of a case is an inherently stochastic process. This can also have very real public health consequences, particularly when looking at elimination and eradication and under these circumstances stochastic models may well be preferred. For instance Eichner *et al.* (1994 and 1996) used stochastic models to assess whether polio can be eradicated. This allowed them, for instance, to quantify how certain you would be that polio had been eliminated, given no cases had been observed for x amount of time (most polio infections are subclinical).

Clearly there are many more instances where chance may have a major role in governing the epidemiological patterns that emerge, and where the variation in outcomes may be important, and so stochastic models may be preferred. We will now turn to how to implement stochastic epidemic models.

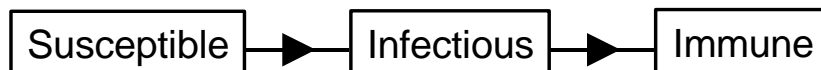
Methods to incorporate chance into epidemic models

Returning to our simple SIR example above, there are two possible approaches for analysing this question:

1. Set up a model which tracks the disease dynamics for *each individual* in the population and allow chance to determine whether or not he/she becomes infected and develops disease over time.
2. Set up a model which tracks populations of susceptible, infectious and recovered individuals and allows chance to determine the *number of secondary cases* which result from the infectious cases at time t .

Within each of these possible approaches (i.e. *individual-based* and *population-based* models), there are further modelling possibilities, depending on whether you model time in discrete units or whether you model time as a continuous process. In the very first practical session of the course you modelled time in discrete units (i.e. t , $t+\Delta t$, and so on) and your model was a series of difference equations. After which you modelled time as a continuous process (where the time step (Δt) tends to zero), and your model was a set of differential equations. Analysts face the same choice with stochastic models. In this lecture and accompanying practicals we will show you two ways of setting up discrete time stochastic models (i.e. an individual-based model and a population-based model), and in the appendix to these lecture notes are details of a 'continuous time' stochastic model.

The notes will use the example of the classic SIR model, i.e.



In the first part of the lecture, we will assume that all infected individuals become infectious after a fixed time, and take a time step of 1 serial interval (the average time between ...). Later we relax this and reformulate the model with different time step sizes.

Method 1 – An individual-based stochastic model, the Reed-Frost Model

The “mathematical” approach for allowing chance to determine whether or not events occur (e.g. an individual is infected) is to draw a number at random and to specify the range in which it should lie for the event occurs. If the random number falls outside that range, then the events does not occur. For infection, this range is based on the expected probability of infection at that time point.

For example, if the risk of getting infected is 20%, then in a model which simulates the dynamics for each individual in the population, it would be sensible to draw a random number between 0 and 1 and to specify that it has to lie between 0 and 0.2 for an individual to become infected and between 0.2 and 1 for the individual to remain susceptible.

Thus if the random number drawn was 0.85, then as 0.85 is larger than 0.2, the individual in question doesn't become infected. If the random number had been 0.12, then because 0.12 is less than 0.2, that individual would have become infected in the model.

To calculate the outbreak size for the simple problem above, it would be necessary to keep on drawing random numbers for each susceptible individual, update the number of cases and repeat until there are no further cases and transmission ceases. The following illustrates the succession of steps which would have to be followed (e.g. in a spreadsheet) to answer this problem:

Step 1: Calculate the risk λ_t that a susceptible individual becomes infected in the next time interval

As the population in the model is small, a susceptible individual may contact more than one infectious case. The risk that a susceptible individual is infected (and becomes infectious) between time t and $t+1$ is given by the Reed-Frost model formula $1 - (1 - p)^{I_t}$, where p is the probability that 2 specific individuals come into effective contact and I_t is the number of infectious individuals at time t . The derivation of this expression is described later in this lecture. An effective contact is defined as one sufficient to lead to transmission from an infectious to a susceptible individual.

Step 2: Draw a random number between 0 and 1 for each of the susceptible individuals.

Step 3: If the random number drawn for any individual is less than λ_t , then that individual becomes infected and hence a case by time $t+1$; otherwise, that individual remains susceptible.

Step 4: Count up the number of cases at time $t+1$ (I_{t+1}), assuming that all those who were cases at time t are now immune.

Step 5: If $I_{t+1}=0$, transmission ceases and the size of the outbreak is given by the sum of the number of cases at time $t=1, 2, 3, 4, \dots, t$; otherwise return to step 1.

An illustration of method 1

Assuming for now that p (the risk of 2 specific individuals coming into effective contact per unit time) is 0.15, the risk, λ_0 , that a susceptible individual becomes infected between time $t=0$ and $t=1$ is $1 - (1 - 0.15)^1 = 0.15$. The following table shows the number of cases at time $t=1$, after drawing a random number for each of the susceptible cases to determine whether or not he/she becomes infected:

Individual number	Random number	Status by $t=1$
1	0.764571	Sus
2	0.067925	Case
3	0.304373	Sus
4	0.462942	Sus
5	0.762053	Sus
6	0.331372	Sus
7	0.61417	Sus
8	0.975166	Sus
9	0.312151	Sus
10	0.850425	Sus

In this instance, only one random number, that drawn for individual 2, was less than 0.15, and so in the model, only this individual becomes infected and is infectious by time $t=1$ (ie $I_1=1$).

Returning to step 1, and substituting for $I_1=1$ into the formula gives the risk that a susceptible individual becomes infected between times $t=1$ and $t=2$ of $\lambda_1=1-(1-0.15)^1 = 0.15$. Repeating

step 2 (drawing a random number for each of the susceptible individuals) and step 3 (translating the random numbers into whether or not the individual becomes infected) gives the following table for the number of cases at time $t=2$:

Individual number	Random number	Status by $t=2$
1	0.239757	
2	-	
3	0.863884	
4	0.412843	
5	0.737687	
6	0.039088	
7	0.094879	
8	0.020703	
9	0.535499	
10	0.347521	

Exercise: Fill in the disease status for each individual in the table. What would have happened to individual 2 during this time step?

In this instance, three random numbers were less than 0.15, and so these three are infected and develop disease by time $t=2$ and so $I_2=3$.

Returning to step 1 and substituting for $I_2=3$ into the Reed-Frost formula gives the risk, λ_2 , that a susceptible individual becomes infected between times $t=2$ and $t=3$ of $1-(1-0.15)^3 = 0.386$.

Repeating the same process at time $t=3$ leads to the following table:

Individual number	Random number	Status by $t=3$
1	0.215361	Case
2	-	Imm
3	0.270405	Case
4	0.862182	Sus
5	0.696761	Sus
6	-	Imm
7	-	Imm
8	-	Imm
9	0.098544	Case
10	0.012308	Case

In this instance, $I_3=4$ and $\lambda_3=1-(1-0.15)^4 = 0.478$.

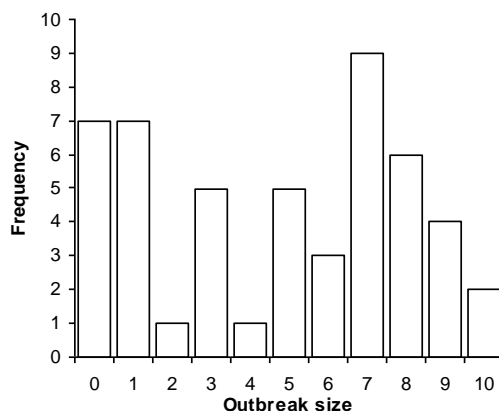
Repeating the same process to identify the status of individuals by time $t=4$ leads to the following table:

Individual number	Random number	Status by t=4
1	-	
2	-	
3	-	
4	0.751125	
5	0.602339	
6	-	
7	-	
8	-	
9	-	
10	-	

Exercise: Fill in the disease status for each individual. What happens to transmission at this time step?

In this instance, none of the random numbers drawn for the susceptible individuals was smaller than λ_4 and so none were infected and therefore $I_4=0$. Thus transmission ceases at time $t=4$, and so the total size of the outbreak in this simulation is given by $I_1 + I_2 + I_3 = 1+3+4 = 8$.

Because this method relies on numbers drawn at random, re-running the steps above another time may lead to a different outbreak size. This is illustrated in the following figure, showing the frequency distribution of outbreak sizes resulting from repeating the algorithm 50 times (eg in a spreadsheet):



This illustrates that with a stochastic model, the results from a single simulation are difficult to interpret. Because of the broad variation in the outbreak sizes, predictions from a deterministic model would not be very meaningful for this problem.

The derivation of the Reed-Frost formula

The expression $1 - (1 - p)^{I_t}$ for the risk that a susceptible individual is infected (and becomes infectious) between time t and $t+1$ (the Reed-Frost model formula) can be derived by applying the following logic:

1. In order to become infected (and infectious), an individual must (obviously) come into contact with **at least one** infectious case. The probability of this occurring is equivalent to $1 - \{\text{the probability that an individual avoids contact with all } I_t \text{ cases}\}$.
2. The probability that an individual avoids contact with all I_t cases is given by the expression $(1-p)^{I_t}$. This follows from the fact that if p is the probability that 2 specific individuals come into effective contact between time t and $t+1$, then $(1-p)$ is the probability that an individual avoids contact with 1 case, $(1-p) \times (1-p)$ is the probability that an individual avoids contact with 2 cases, $(1-p) \times (1-p) \times (1-p)$ is the probability that that individual avoids contact with 3 cases, $(1-p)^{I_t}$ is the probability that an individual avoids contact with each of the I_t cases.
3. Combining the logic from steps 1 and 2 gives the Reed-Frost formula.

Calculating “p” – the probability that 2 specific individuals come into effective contact between time t and $t+1$

The probability “p” for the probability that 2 specific individuals come into effective contact between time t and $t+1$ is given by the expression:

$$p = R_0 \times \{\text{size of time step}\} / \{\text{infectious period} \times \text{total population size}\}$$

p is thus very closely related to the parameter β which is used in models implemented using differential equations: p reflects the probability of an effective contact in a given **discrete time step (ie between t and $t+1$)** whereas β reflects the probability of an effective contact **per unit time**. However, if the model deals with a large population size, then according to the above formula, p is very small and it can be shown (see optional reading below) that the risk $1 - (1-p)^{I_t}$ that a susceptible individual is infected between time t and $t+1$ is approximately equal to pI_t

Combining this expression with the fact that, in the continuous time model, the rate at which individuals are infected is given by βI_t gives the result that

$$p \approx \beta$$

when the population being considered is large (typically >1000 individuals)

(Optional reading) Derivation of the result that $1 - (1-p)^{I_t} \approx pI_t$

From your past mathematical training, you may recall that the expression $(1-p)^{I_t}$ can be expanded to give the expression:

$$1 - pI_t + \frac{I_t(I_t-1)p^2}{2} - \frac{I_t(I_t-1)(I_t-2)}{3!}p^3 + \dots$$

When p is very small, terms involving p^2 , p^3 and higher order terms are negligible in comparison with the first two terms. In this situation

$$(1-p)^{I_t} \approx 1 - pI_t$$

and as a result $1 - (1-p)^{I_t} \approx 1 - (1 - pI_t) = pI_t$

Method 2 – A population-based stochastic model

Instead of keeping track of the status of each individual as we did in method 1, we can keep track of the *total* number of susceptibles and cases at each time step (in a similar way to the population-based models that you have been using up to now in the course). As before we assume that the time step is the same as the serial interval. Random numbers are used to select the *total number of susceptibles infected in each generation* from an appropriate *distribution*.

In the case of the population-based equivalent of the Reed-Frost individual-based model, the appropriate distribution for the number of susceptibles infected in the next generation is a *binomial* distribution. This can be explained as follows.

For each susceptible in the population, there is a probability, λ , that they will be infected in the next generation, and a corresponding probability $(1-\lambda)$ that they will not be infected. An experiment (like a coin toss) whose outcome is random and in which there are two possible outcomes is sometimes called a Bernoulli trial. This is the situation here for each susceptible person. The binomial distribution gives the probability of k successes of n independent Bernoulli trials, so can give us the number of individuals who would be infected ('successes') in each generation (the total number, n , of Bernoulli trials).

To calculate the binomial probability (i.e. the probability that k susceptibles will be infected), assume:

- S = number of independent Bernoulli trials (= number of susceptibles)
 - We can call these Bernoulli trials Z_1, Z_2, \dots, Z_n
- λ = the probability of the event occurring (which is the same in each trial). Thus,
 - $\text{Prob}[Z_1=1] = \lambda$,
 - $\text{Prob}[Z_2=1] = \lambda, \dots$
 - $\text{Prob}[Z_n=1] = \lambda$

Therefore the probability of observing exactly k infections (successes) is

$$\lambda^k (1-\lambda)^{S-k}$$

Or, rewriting it a bit more formally, the probability that exactly k out of S_t susceptibles at time t are infected and develop disease by time $t+1$ is given by the following equation:

$$P(I_{t+1} = k) = \binom{S_t}{k} \lambda_t^k (1 - \lambda_t)^{S_t - k}$$

where λ_t is defined as above (the risk that a susceptible individual is infected by at least one case between time t and $t+1$). This formula is the standard Binomial expression for the probability of k successes out of S_t trials, which can be derived by combining the facts that:

1. For k individuals to become infected, $S_t - k$ individuals need to avoid becoming infected.

2. There are $\binom{S_t}{k}$ ways of choosing k out of S_t individuals to become infected
3. The probability that k individuals are infected is λ_t^k and the probability that $S_t - k$ individuals avoid becoming infected is $(1 - \lambda_t)^{S_t - k}$.

In practice many computer programmes will draw a binomial random number. For instance, the code in Berkley Madonna is:

Binomial(p,n),

where p is the probability of success (in our case λ) and n is the number of trials (in our case the number of susceptibles at time t). So if you are using Berkley Madonna, the steps you would have to follow are simply:

Step 1: Calculate the risk that a susceptible individual becomes infected in the next time interval using (for instance) the Reed-Frost formula of $\lambda_t = 1 - (1 - p)^{I_t}$

Step 2: Draw a number from the binomial distribution with probability of success = λ , and number of trials = S_t

Step 3: Move this number of susceptibles into the infected class.

The Berkley Madonna Code for the population-based version of the Reed-Frost models is given in Figure 2.

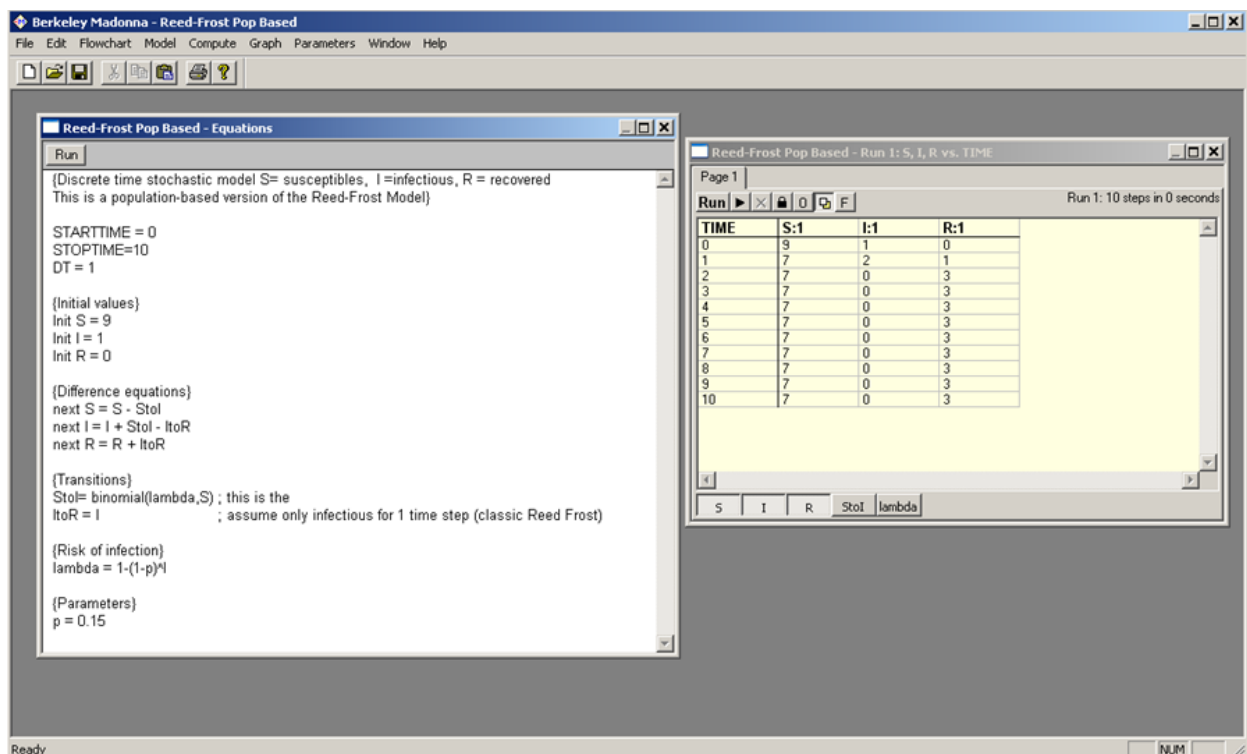
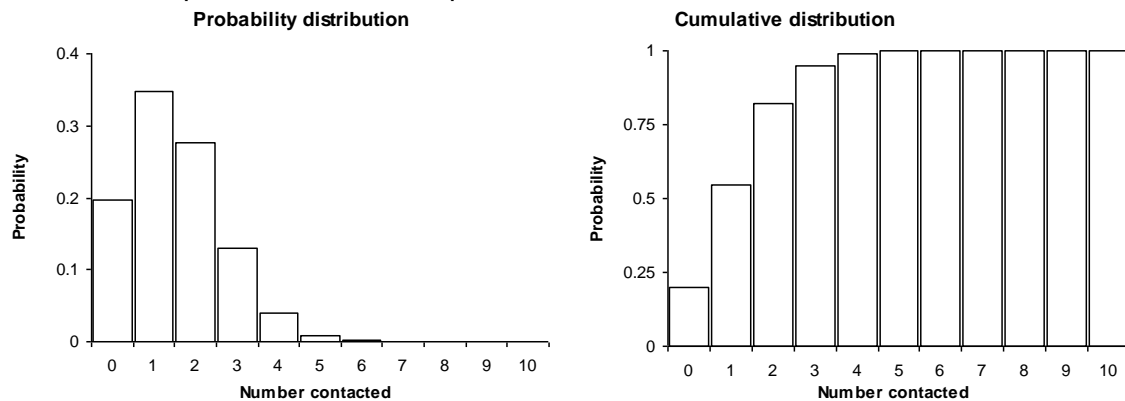


Figure 2: BM code for discrete time, population-based stochastic model (Reed-Frost)

The mechanics of this are given below, and if you were doing this in Excel you would have to follow these steps.

The expression for exactly k out of S_t susceptibles being infected leads to the following distribution and cumulative probability distributions for the number of cases at time $t=1$. In this example we assume, as for the first model, that the probability of an effective contact between two specific individuals is $p=0.15$, $C_0=1$, $S_0=10$.



Then draw a random number between 0 and 1, and see what number of infections this corresponds to. For instance if the random number drawn was 0.10, then the number of contacts which corresponds to a cumulative distribution of 0.10 is 0, so in this simulation no susceptibles would be contacted and transmission ceases.

As for method 1, the total outbreak size would be derived by repeating the process of drawing random numbers and relating their size to the number of susceptibles contacted until there are no further cases in the population and transmission ceases. The following shows the succession of steps which would need to be implemented (eg in a spreadsheet) to derive the outbreak size:

Step 1: Calculate the risk that a susceptible individual becomes infected in the next time interval using the Reed-Frost formula of $\lambda_t = 1 - (1 - p)^{I_t}$

Step 2: Draw a random number between 0 and 1.

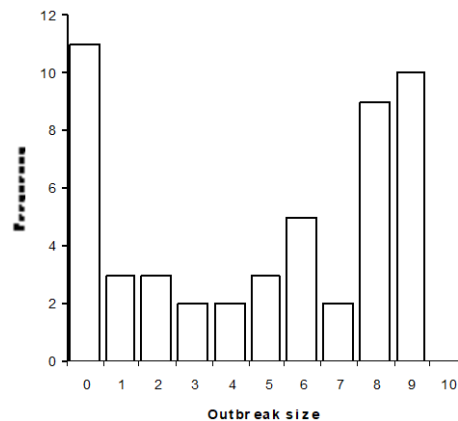
Step 3: Find the value for k for which the cumulative probability distribution (based on S_t and λ_t) of the number of cases, leads to the value of the random number in step 2. I_{t+1} is then given by k and S_{t+1} is given by $S_t - k$.

Step 4: If $I_{t+1}=0$, transmission ceases and the size of the outbreak is given by the sum of the number of cases at time $t=1, 2, 3, 4, \dots, t$; otherwise return to step 1.

An illustration of method 2

The table below shows the result of one simulation run of this model, using the same values for p that were used in the last example (note that there is only one case at time $t=0$ ($I_0=1$), which resulted in an outbreak with 8 cases. This highlights the fact that as the number of susceptibles and infectious individuals changes, the distribution of the number of individuals contacted by each case (and by implication, number of cases observed in the next generation) changes.

The following shows the distribution of outbreak sizes predicted by implementing the above algorithm 50 times:



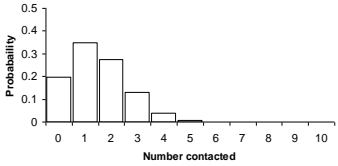
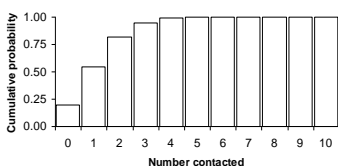
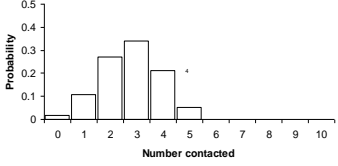
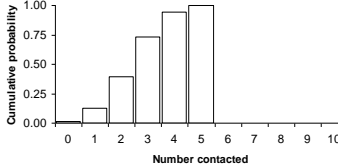
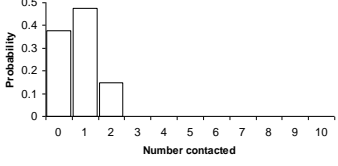
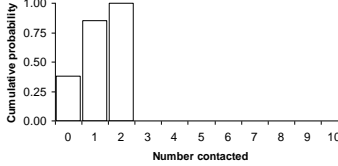
Time	λ_t	Susceptibles at time t (S_t)	Distribution of the number of cases in the next time step		random number	I_{t+1}
			Probability distribution	Cumulative probability distribution		
0	0.15	10			0.991479	5
1	0.556295	5			0.531408	3
2	0.385875	2			0.168033	0

Table: Summary of the results of one simulation run using method 2 of the effect of the introduction of one infectious case into a totally susceptible population comprising 10 individuals.

Note that the shape of this distribution differs from that predicted for method 1. This is due to the relatively small number of simulations on which it is based – for practical purposes, predictions from stochastic models such as these would be based on hundreds or thousands of simulations. The results from stochastic models would typically be summarized in an analogous way to those from any epidemiological study eg showing the average outcome and the 95% range in which the outcomes occurred.

Extensions to methods 1 and 2

Methods 1 and 2 which were described in the last lecture are simplistic, in that only one kind of transition is considered, namely susceptible individuals becoming infected and then infectious. They do not explicitly describe cases recovering from disease and becoming immune and transitions (susceptibles coming into contact and becoming cases) are assumed to occur after fixed time steps, implicitly assumed to be one serial interval. As such, the model is a stochastic realization of the following difference equations

$$S_{t+1} = S_t - m(S_t)$$

$$I_{t+1} = m(S_t)$$

where $m(S_t)$ is the number of susceptible individuals who are infected and become cases between time t and $t+1$, and since transitions occur after 1 serial interval, I_t is both the prevalence and incidence of infectious cases.

Both methods can be adapted to deal with other transitions eg to allow chance to determine the number of infectious cases who recover and become immune, or to allow time steps to be of less than one serial interval (e.g. of size Δt). For example, to incorporate both these features, it would be necessary to set up a model which solves the following equations:

$$S_{t+\Delta t} = S_t - m(S_t)$$

$$I_{t+\Delta t} = I_t + m(S_t) - r(I_t)$$

$$R_{t+\Delta t} = R_t + r(I_t)$$

where I_t is the number of infectious cases at time t ,

R_t is the number of immune individuals at time t

$m(S_t)$ is the number of susceptible individuals who are infected and become cases between time t and $t+ \Delta t$

$r(I_t)$ is the number of infectious cases who recover between time t and $t+\Delta t$.

The number of susceptible individuals who are infected (and become infectious) in each time interval is determined in the same way as described using method 2. That is, get your programme to do this if it is capable of doing so (e.g. Berkely Madonna) or programme it yourself using the above algorithm (draw a random number and identify the value for the number of newly infected individuals for which the cumulative probability distribution of the number of cases would have to be to result in the given random number).

The number of infectious individuals who recover between time t and $t+\Delta t$ is calculated using an analogous method. That is determined by the standard Binomial expression for the probability of k successes out of a given number of trials:

$$P(r(I_t) = k) = \binom{I_t}{k} r_t^k (1 - r_t)^{I_t - k}$$

where r_t is defined as the risk that an infectious individual recovers between time t and $t+\Delta t$. For most practical purposes, small time steps are used in the calculations, and r_t is approximated by the rate at which individuals recover from infectious disease, multiplied by the size of the time step Δt .

The steps in the calculations can be summarized as follows:

Step 1:

a) Calculate the risk that a susceptible individual becomes infected in the next time interval using the Reed-Frost formula of $\lambda_t = 1 - (1 - p)^{I_t}$ or βI_t if the population size is large.

b) Calculate the average proportion of cases who should recover during the next time interval using the expression $r_t = r \Delta t$ where Δt is the size of the time interval and r is the rate at which cases recover (given by $1/\text{duration of infectiousness}$).

Step 2: If your programme is capable of drawing random numbers from a binomial distribution, do this for both the infection step, and recovery step, then go to Step 3

Or

Step 2a) Draw two random numbers between 0 and 1, n_1 and n_2 .

Step 2b):

i) Find the value for k for which the cumulative probability distribution (based on S_t and λ_t) of the number of cases, leads to the value of the random number n_1 in step 2a. $m(S_t)$ is then given by k .

ii) Find the value for k for which the cumulative probability distribution (based on I_t and r_t) of the number of infectious cases who recover, leads to the value of the random number n_2 in step 2. $r(I_t)$ is then given by k .

Step 3: Update the equations:

$$S_{t+\Delta t} = S_t - m(S_t)$$

$$I_{t+\Delta t} = I_t + m(S_t) - r(I_t)$$

$$R_{t+\Delta t} = R_t + r(I_t)$$

based on the results in Step 2.

Step 4: If $I_{t+\Delta t} = 0$, transmission ceases; otherwise return to step 1.

Advantages and disadvantages of stochastic models

The above examples are obviously very simplistic, but illustrate two of the basic methods used in stochastic models – most stochastic models in the literature use one or several of these principles.

Stochastic models (often referred to as Monte Carlo simulation models, given their reliance on chance for different outcomes) have several advantages over deterministic models, namely :

- They can be much more realistic than deterministic models. For example, epidemic fadeout cannot be adequately modelled with deterministic models, but can be with stochastic models. Deterministic models can be considered to be approximations to the average behaviour of stochastic models applicable in large populations when fade-out has not occurred.
- They take account of chance variation.

Their main disadvantages are:

- many simulations need to be run in order to obtain useful predictions.
- Individual-based stochastic models can be laborious to set up, difficult to check and slow to run

With the development of faster computers, these disadvantages have become less of a problem in recent years, and stochastic models have become more common. Indeed, individual-based stochastic models (which tend to be the slowest sort to implement) have become much more common in recent years (see for instance Ferguson et al. 2006 and Germann et al. 2006). Many of these individual-based simulation models (such as the two just mentioned) are extensions of the individual-based Reed-Frost model. In the following practical we will see how it is possible to extend the Reed-Frost model to build up complicated models incorporating heterogeneous mixing.

Further reading

Vynnycky E and White RG (2010) An introduction to infectious disease modelling. Oxford University Press. Chapter 6.

References

De Serres, G., Gay, N.J. & Farrington, C.P. Epidemiology of transmissible diseases after elimination. *Am J Epidemiol* 2000. 151, 1039-48; discussion 1049-52.

Ferguson NM, Fraser C, Donnelly CA, Ghani AC, Anderson RM. Public health risk from the avian H5N1 influenza epidemic. *Science*. 2004 304:968-9.

Eichner M, Dietz K Eradication of polio: when can one be sure that polio virus transmission has been terminated *Am J Epidemiol*, 1996; 143:816-22

Eichner M, Haderer KP, Dietz K Stochastic models for the eradication of poliomyelitis: minimum population size for polio virus persistence. in Model for Infectious Diseases. Their structure and relation to data eds: V Isham, G Medley, Cambridge University Press, 1996

Abbey H An examination of the Reed-Frost theory of epidemics *Human Biology*, 1952;24:201-233

Appendix

Method 3 – Stochastic implementations of differential equations

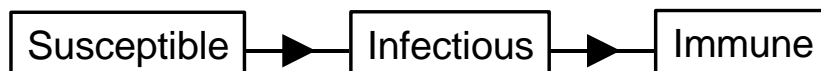
Both methods 1 and 2 assume that all transitions occur after the same fixed time step Δt , and are known as *discrete time* stochastic models. They are analogous to deterministic difference equations. A third method uses chance to determine when the next event occurs (i.e. the size of Δt) and, if more than one kind of transition (e.g. susceptibles becoming infected or infectious cases recovering) is considered, the type of transition which occurs. The size of the time steps can be small or large, but because it is variable models that use it are known as *continuous time* stochastic models. They can be considered to be a stochastic implementation of differential equations.

The steps in the calculations are as follows:

Step 1: Calculate the total rate (M_t) at which individuals can change their current status (i.e. become infected, recover from infectious disease, die etc).

M_t is known as a *hazard rate*, as it gives the hazard or chance of an event occurring over a small time interval. Mathematically, the chance of any event leading to a change of status in a small interval of time δt is approximately given by $M_t \delta t$. The approximation can be made as accurate as we like by choosing a small enough δt .

In the following model, only two kinds of transitions are described explicitly, i.e. susceptible individuals are infected and become infectious, and infectious individuals recover and become immune



It would be written using the following differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta S(t)I(t) \\ \frac{dI}{dt} &= \beta S(t)I(t) - rI(t) \\ \frac{dR}{dt} &= rI(t)\end{aligned}$$

where β is the probability of an effective contact between 2 specific individuals per unit time, and r is the rate at which individuals recover from infectious disease. In this example, the total rate at which individuals could change their current status is given by the sum of the rates at which individuals become infected and infectious i.e.

$$M_t = \beta S(t) I(t) + rI(t)$$

Step 2: Draw a uniform random number, n_1 , between 0 and 1 and calculate the time after which the next transition occurs, given by:

$$T = -\ln(n_1)/M_t$$

The reasoning behind this is given below. If you're not interested in the derivation you can skip the next two paragraphs and jump to step 3.

The above expression for T is derived from the fact that when events occur at random with a constant hazard rate, the time until the event occurs follows the negative exponential distribution. Given that the hazard rate is M_t , it can be shown (using elementary calculus) that the probability there has been no event by time T is $e^{-M_t T}$.

If we perform millions of simulations, each time choosing the time of the next event, T , to equal $-\ln(n_1)/M_t$ (where n_1 is drawn from a uniform distribution between 0 and 1) then the probability that there has been no event by time T is just the proportion of the simulations where $-\ln(n_1)/M_t$ is greater than T . $-\ln(n_1)/M_t$ is equal to T when $n_1 = e^{-M_t T}$ (to see this just rearrange $T = -\ln(n_1)/M_t$) and therefore greater than T whenever n_1 is less than $e^{-M_t T}$ (since the smaller n_1 is the larger $-\ln(n_1)$ will be). From the definition of the uniform distribution, the probability that n_1 is less than $e^{-M_t T}$ is exactly $e^{-M_t T}$, hence the probability there has been no event by time T is also $e^{-M_t T}$, as required. This shows that setting $T = -\ln(n_1)/M_t$ will correctly sample the time to the next event from a negative exponential distribution.

Step 3: Calculate the probability that each type of transition (e.g. susceptible -> infected or infectious -> immune) will occur. Use this to calculate the range in which a number drawn at random must lie for a given transition to occur.

In the above example, the probability that one susceptible individual becomes infected (and then infectious) during the next time step is given by $\beta S(t)I(t)/M_t$; the probability that one infectious individual recovers during the next time step is $r(t)I(t)/M_t$.

Thus if the random number drawn is in the range $0 - \beta S(t)I(t)/M_t$, then one susceptible individual is infected (and becomes infectious); otherwise one infectious case recovers and becomes immune.

Step 4: Draw a uniform random number n_2 to determine the transition event which occurs next.

Step 5: Use the result from step 4 to update the number of susceptible, infectious and immune individuals present in the population and return to step 1.

An illustration of Method 3

Table A1 and Figure A1 summarize the results of simulating the introduction of one infectious case into a totally susceptible population consisting of 10 individuals, using a value for β of 0.169/day (which is equivalent to assuming that the R_0 is 13) and an average duration of infectiousness of 7 days (which is equivalent to a recovery rate of 0.143/day). The calculations during the first few time steps are summarized below.

First time step

Step 1 – calculate M_0 , the total hazard rate for individuals to change their current status

At the start of the simulations, there are 10 susceptible individuals, 1 infectious individual and 0 immune individuals.

The rate at which individuals who can become newly infected (and in this model, infectious) during the next time step is given by $\beta S(0)I(0) = 0.169 \times 10 \times 1 = 1.69$.

The rate at which individuals can recover and become immune during the next time step is given by $rI(0) = 0.143 \times 1 = 0.143$.

Thus M_0 is given by $1.69 + 0.143 = 1.83$.

Step 2: draw a uniform random number n_1 to determine the time T at which the next transition occurs.

In this instance, the random number $n_1 = 0.55$ was drawn and the equation $T = -\ln(n_1)/M_0$ gives the result that the next transition occurs at time $T = 0.326$ days.

Step 3: Calculate the probability that each type of transition occurs next and specify the range in which a number drawn at random will have to lie for the next transition to be a given type.

The probability that one susceptible individual becomes infected is given by $\beta S(0)I(0)/M_0 = 1.69/1.83 = 0.922$.

The probability that one infectious individual recovers and becomes immune is given by $rI(0)/M_0 = 0.143/1.83 = 0.078$.

It would therefore be sensible to specify that if the random number lies in the range 0-0.922, then the next event will be an infection event, and otherwise, 1 infectious individual will recover and become immune.

Step 4: Draw a uniform random number n_2 to determine the type of transition event which occurs next

In this example, the random number drawn is 0.56. As this lies between 0 and 0.922, the next event will be the infection of 1 susceptible individual.

Step 5: Update the number of individuals in each category

At time $t=0.325$, there are 9 susceptible individuals, 1 infectious individual and 0 immune individuals.

Return to step 1 etc

Table A1: Summary of the expected number of individuals observed in different compartments after the introduction of 1 infectious individual into a population consisting of 10 susceptible individuals, as predicted using method 3, assuming that $\beta=0.169/\text{day}$ and the recovery rate is $0.143/\text{day}$

Time	Susceptible s	Infectious	Immune	M	random number	T	probability that the next event is:		random number	next event
							infection event	recovery event		
0.000	10	1	0	1.83	0.55	0.326	0.922	0.078	0.56	S->I
0.326	9	2	0	3.32	0.44	0.248	0.914	0.086	0.95	I->R
0.575	9	1	1	1.66	0.31	0.714	0.914	0.086	0.12	S->I
1.289	8	2	1	2.99	0.65	0.147	0.904	0.096	0.16	S->I
1.436	7	3	1	3.97	0.17	0.439	0.892	0.108	0.99	I->R
1.875	7	2	2	2.65	0.85	0.059	0.892	0.108	0.44	S->I
1.935	6	3	2	3.47	0.88	0.038	0.876	0.124	0.15	S->I
1.973	5	4	2	3.95	0.22	0.385	0.855	0.145	0.95	I->R
2.358	5	3	3	2.96	0.53	0.211	0.855	0.145	0.60	S->I
2.569	4	4	3	3.27	0.21	0.484	0.825	0.175	0.00	S->I
3.053	3	5	3	3.25	0.48	0.224	0.780	0.220	0.56	S->I
3.278	2	6	3	2.88	0.11	0.766	0.703	0.297	0.53	S->I
4.044	1	7	3	2.18	0.23	0.670	0.542	0.458	0.85	I->R
4.714	1	6	4	1.87	0.23	0.791	0.542	0.458	0.51	S->I
5.505	0	7	4	1.00	0.97	0.033	0.000	1.000	0.12	I->R
5.538	0	6	5	0.86	0.33	1.303	0.000	1.000	0.36	I->R
6.841	0	5	6	0.71	0.33	1.568	0.000	1.000	0.81	I->R
8.409	0	4	7	0.57	0.80	0.387	0.000	1.000	0.56	I->R
8.796	0	3	8	0.43	0.94	0.138	0.000	1.000	0.03	I->R
8.934	0	2	9	0.29	0.52	2.279	0.000	1.000	0.15	I->R
11.213	0	1	10	0.14	0.47	5.296	0.000	1.000	0.78	I->R
16.509	0	0	11	-	-	-	-	-	-	-

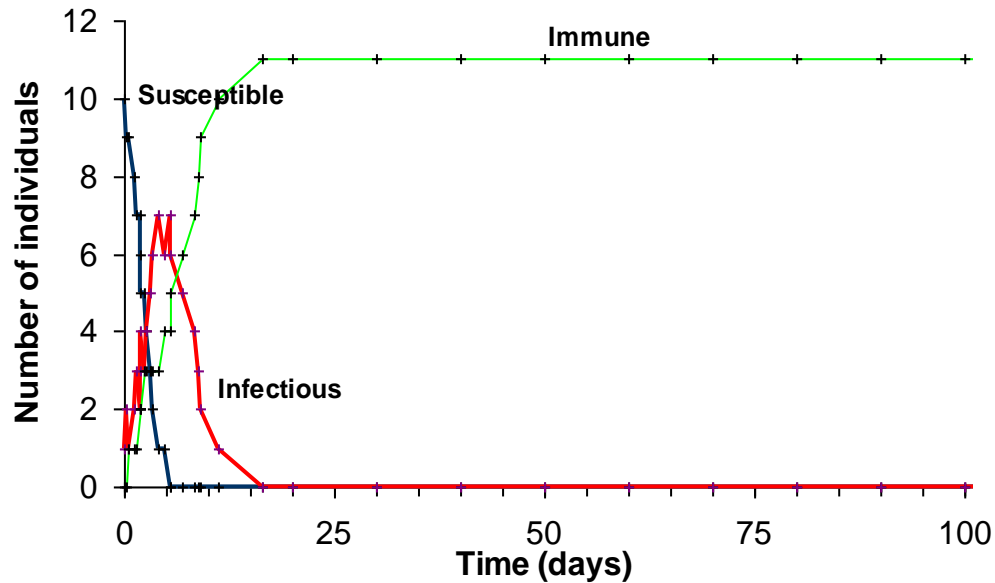


Figure A1: Summary of the expected number of individuals observed in different compartments after the introduction of 1 infectious individual into a population consisting of 10 susceptible individuals, as predicted using method 3, assuming that $\beta=0.169/\text{day}$ and the recovery rate is $0.143/\text{day}$

Method 3 can be implemented in a spreadsheet, though this would only be practical for describing the transmission dynamics of an infection in small populations, and typically, it is implemented using a programming language.

Advantages and disadvantages of method 3, as compared with method 2

In an analogous way that differential equations are more accurate than difference equations, the continuous time formulation (method 3) is more accurate than the discrete time formulation (method 2). Indeed, method 2 can be made to provide a very good approximation of method 3 by taking a smaller and smaller fixed time step (when the time step is infinitely small, the methods are in fact identical!). The disadvantage of the continuous time model is that if events are quite common, then a continuous time model may be much slower to run than a discrete time model.

Introduction to Infectious Disease Modelling and its Applications - 2018

Session 18: Economic evaluation of infectious disease interventions

Lecture

Objectives

By the end of this lecture you should be able to:

- Understand the motivation for conducting economic evaluations of infectious disease interventions
- Understand the advantages and limitations of different methods of economic evaluations of infectious disease interventions.

Fundamental concepts

Why should we bother about economics when we evaluate the impact of an infectious disease intervention like vaccination against an infectious pathogen or screening and treating infectious people? To appreciate the role of economic evaluation, first of all we need to understand some basic concepts in economics.

The first fundamental concept is *scarcity*. Scarcity is called the fundamental economic problem. A resource is said to be scarce if the supply of the resource is limited and is sufficient to supply all human demand for it. This is a familiar concept in health care. Resources that are needed to produce health care such as money, staff time and hospital beds are scarce.

The second concept is *opportunity cost*. The opportunity cost of a choice is a consequence of scarcity. When a resource is scarce, then using it for something involves giving up other possible uses of it. For instance, if we decide to vaccinate children during an influenza pandemic, then we are giving up other uses for the money used to buy vaccines, the healthcare staff needed to deliver the vaccines, and the parents' and children's time needed to come to the vaccination centres. We could have used the same resources to conduct more contact tracing and isolate the infected people for instance. We could even have used it in a different area – to build a new school or a new airport. The opportunity cost of a choice is the value of the best alternative use of the resources consumed by the choice. We are constantly making choices that involve opportunity costs as individuals, organisations and societies. Economics helps us to make the trade-offs (i.e. the opportunity costs) of each choice explicit.

The third concept is *rationing* or *priority setting*. When resources are scarce, not everyone who wants them will be able to have them. Hence we need to have some way to allocate resources to people who want them – this is called our system of rationing. For instance, not everyone who wants health care will necessarily receive it immediately. There are a number of ways we can ration (allocate) health care.

- We can have *no explicit system* at all – in which case the allocation of health care is still limited by supply, but the people who want health care do not know by what criteria it is allocated.

- We could ration by *price* – this is what happens in a completely free market in which providers are willing to supply health care to consumers who are willing to adequately compensate (pay) the providers. This is often how many other goods are allocated in many countries. However, a completely unregulated market for health care is usually seen as both ethically undesirable and inefficient because of special features of health and health care that we don't have time to go into today.
- We could ration by *waiting list* – we could supply health care to everyone who wants it in the order at which they demand it. When we use up our resources (staff, money, hospital beds etc.) then anyone else who wants health care has to wait until new resources become available. But in theory anyone with the ability and willingness to wait will eventually receive health care.
- We could ration according to *need*. This is what health economic evaluation tries to achieve. It means that we allocate health care based on an assessment of people's needs. For instance, we could allocate our supply of health care so as to maximise the health of the population. Or we could use a different principle, such as wanting to ensure that everyone's health is as equal as possible.

A fourth concept is that of *externalities*. An externality is a cost or a benefit borne by someone other than the person producing or consuming an economic good. This is especially important when we look at the economics of infectious disease interventions. If someone gets vaccinated, there is a positive externality – other people who did not consume vaccination (i.e. get vaccinated) may still be protected because someone else cannot be infected and go on to transmit the infection to them. We've seen this effect before - epidemiologists call it herd protection. Similarly there are negative externalities – for instance, someone consuming antibiotics may contribute to antibiotic resistance in the community and hence impose a cost on everyone else.

Economic evaluations

Contrary to what some people think, the aim of health economics isn't to save money *per se*, but to make the best use of scarce resources in order to maximise health in line with the values of society. One tool for assisting such decisions is a health economic evaluation.

A (full) economic evaluation compares the incremental costs and consequences of an intervention to a comparator. Let's suppose we want to evaluate the introduction of vaccination against influenza. Our comparator is that we continue with our current standard of care without vaccinating – so people who get influenza will go for treatment. Hence there are costs involved i.e. the cost of treating people with influenza, and there are negative health consequences involved, i.e. the discomfort (and potentially death) of people who get influenza. Now when we introduce influenza vaccination, fewer people have the 'flu. Hence the cost of treating people with influenza decreases, as does the number of people who are sick (and potentially die) of influenza. However, a new cost is imposed – the cost of actually paying to buy the vaccine and deliver it to people getting vaccinated.

There are several ways of comparing the two options (i.e. introducing vaccination compared to continuing with present care).

1. The simplest is a *cost-minimisation analysis* in which we simply compare the net costs of both options and ignore the health consequences. For instance we might want to know if it costs more to introduce influenza vaccination or to treat the people who may otherwise have had the 'flu. This analysis is usually not terribly helpful on its own, since we often want to introduce a health intervention that saves lives and makes people healthier even if it doesn't save us money. But there are some interventions that may be introduced because they are cost saving without worsening anyone's health, such as replacing an expensive branded drug with an equally efficacious generic competitor.
2. The second is a *cost-effectiveness analysis*. In this analysis, we work out the *incremental cost-effectiveness ratio* (or ICER), which is the incremental cost of an option compared to its comparator (i.e. the difference in net costs of the two options), divided by the incremental health benefit of the option compared to its comparator (i.e. the difference in net health benefits of the two options). Thus we can work out the incremental cost per unit of health benefit. Improved health can be measured in lots of different ways – for example, we can use episodes of disease averted, hospitalisations averted, deaths prevented or years of life gained as units of health. Hence we have an equation like the following:

Incremental cost- effectiveness ratio (ICER)

$$= \frac{\text{Incremental cost of intervention compared to comparator}}{\text{Incremental health benefit of intervention compared to comparator}}$$

$$= \frac{\text{Cost of intervention} + \text{Treatment costs of intervention} - \text{Treatment costs of comparator}}{\text{Health burden with intervention} - \text{Health burden with comparator}}$$

3. A *cost-utility analysis* is a special case of a cost-effectiveness analysis. With this analysis, we measure health using a generic measure of health utility such as a quality-adjusted life year (QALY) or disability-adjusted life year (DALY). We'll get on to that in a moment.
4. Finally, we have a *cost-benefit analysis*. In a cost-benefit analysis, we convert all our benefits (both health and non-health) into monetary terms, based on the value of those benefits to individuals. Hence both sides of the ratio (costs and benefits) are expressed in the same units (money), and so we can give the outcome of the analysis in terms of a dimensionless benefit-to-cost ratio. For instance, 15:1 is a very beneficial intervention, while 1:2 is an intervention that costs more than it returns in benefits.

Cost-effectiveness analysis

Let's look at how cost-effectiveness analysis (including cost-utility analysis) can be used a bit more closely, since this is the most common type of economic evaluation performed in the health care sector.

First, it is important when performing an economic evaluation that we calculate costs and benefits *incrementally*, i.e. we work out the difference with the closest comparator rather than compared to doing nothing. For instance, suppose we had a drug with three possible dosage regimens (see Figure 1). The third dosage regimen, 80 mg/d, costs \$25,800 per life-year gained in the patients that receive it, compared to receiving no drug at all. However, compared to the next available option, a 40 mg/d regimen, it costs a lot more: \$84,300 per life-year gained. That's because most of the benefits of the higher dosage can be obtained at a lower

dosage, at a much reduced cost. And it is the incremental cost-effectiveness ratio (i.e. the ratio when comparing 80 mg/d to 40 mg/d) that we are interested in.

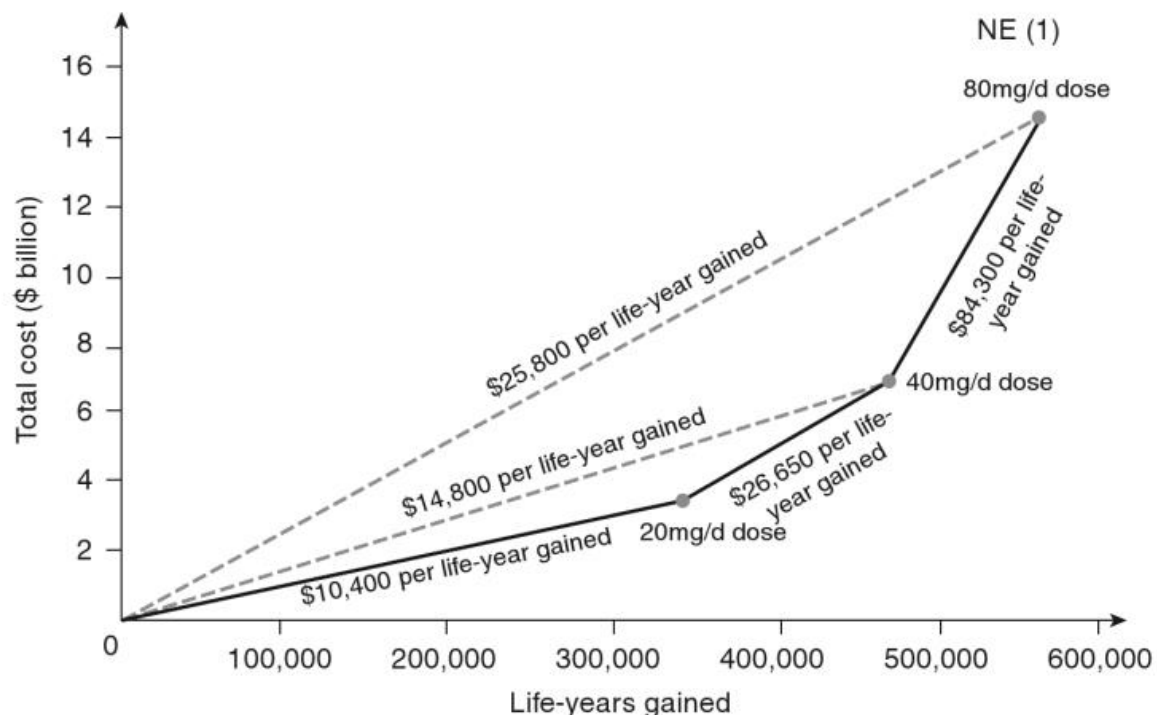


Figure 1. Costs and life-years gained for a drug with three possible dosage regimens. From Gray *et al.* Applied Methods of Cost-effectiveness Analysis in Healthcare, p. 15.

The ICER is then compared to a *willingness to pay threshold*, which is the maximum amount that would normally be spent to gain one health unit (eg. one QALY) before an intervention stops being cost-effective. The idea of the threshold is as follows (the example comes from Claxton *et al.* BMJ 2008; 336:551). Suppose the threshold is £20,000 per QALY gained (the threshold often used by the UK's National Institute for Health and Care Excellence or NICE for health technologies). If an intervention is funded which costs (for example) £60,000 and gains 2 QALYs each time it is used, then it will displace existing interventions costing £60,000 and hence generating at least 3 QALYs ($60,000/20,000 = 3$). Hence there will be a net loss of 1 QALY for each use of this new intervention, making it bad value for money. On the other hand if the intervention only costs £20,000 but still gains 2 QALYs each time it is used, then it will displace existing interventions costing £20,000 and hence generating at least 1 QALY ($20,000/20,000 = 1$). Hence it is probably good value for money.

The willingness to pay threshold varies across countries; Table 1 shows the willingness to pay threshold for a QALY or DALY across several countries.

Country	Decision making body	Outcome measure	Willingness to pay for a QALY or DALY
Australia	Pharmaceutical Benefits Advisory Committee (PBAC)	QALY	A\$50,000 ¹
Ireland	Department of Health & Health Service Executive	QALY	€45,000 ²
Netherlands	Health Care Insurance Board (CVZ)	QALY	€20,000-€80,000 ³
UK	National Institute for Health and Care Excellence (NICE)	QALY	£20,000 - £30,000 ⁴
USA	Preventive Services Task Force (PSTF)	QALY	\$50,000 - \$100,000 ⁵
Thailand	Health Intervention and Technology Assessment Program	QALY	Bt160,000 ⁶
Global	Commission on Macroeconomics and Health	DALY	1-3 x GDP per capita ⁷

Table 1. Willingness to pay thresholds for cost-effectiveness analyses in different countries/organisations. References: (1) Lopert et al. Commonwealth Fund Pub 2009; 1297(60):1; Harris et al. Med Decis Making 2008; 28:713; (2) IPHA DOE & HSC. 2012. Framework Agreement; (3) Boersma et al. Value in Health 2010; 13:853; (4) Guide to the methods of technology appraisal. NICE, 2013; (5) Ubel et al. Arch Intern Med 2003; 163:1637-1641; (6) Teerawattananon et al. Z. Evidenz, Fortbildung und Qualität im Gesundheitswesen 2014; 108:397–404; (7) Newall et al. Pharmacoeconomics 2014; 32:525.

It gets more complicated if either (or both) the incremental costs or incremental health benefits of the intervention (compared to a comparator) are negative. We can think of a *cost-effectiveness plane* with the incremental costs as the y-axis and incremental benefits as the x-axis (see Figure 2). In the top-left quadrant of the plane, incremental costs are positive but incremental consequences are negative. Hence the intervention is always a “bad buy” regardless of the threshold – the intervention is then said to *be dominated* by its comparator. In the bottom-left quadrant, incremental costs are negative but incremental consequences are positive. Hence the intervention now becomes a guaranteed “good buy” regardless of the threshold – the intervention is then said to *dominate* its comparator. In the top-right and bottom-right quadrants, the costs and consequences have the same sign, so the intervention neither dominates nor is dominated by its comparator. Instead, we have to look at the ICER to see if it is above or below the threshold. Note that in the top-right corner incremental costs and consequences are positive, so we are having to pay in order to improve health. In the bottom-right corner incremental costs and consequences are negative, so we are losing health in return for saving money.

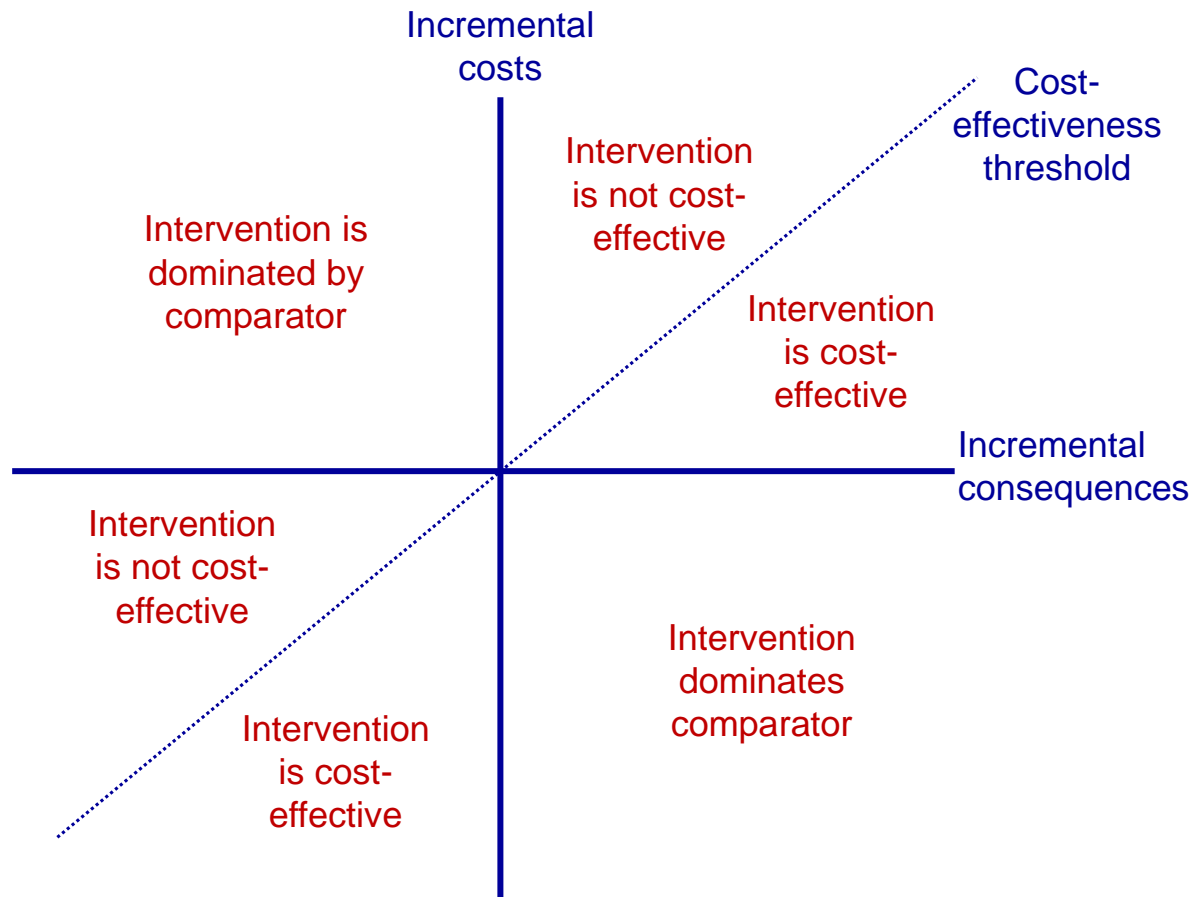


Figure 3. Cost-effectiveness acceptability plane.

The ICER is rarely known precisely. The uncertainty around the value of the ICER can be expressed in terms of a probability distribution. The parameters of the distribution can be calculated from a model using techniques such as *probabilistic sensitivity analysis*. When this happens, instead of being able to say for certain that the intervention is cost-effective or not cost-effective, we say that there is a probability that the intervention will be cost-effective. This can be illustrated using a graph called a *cost-effectiveness acceptability curve* (see Figure 4).

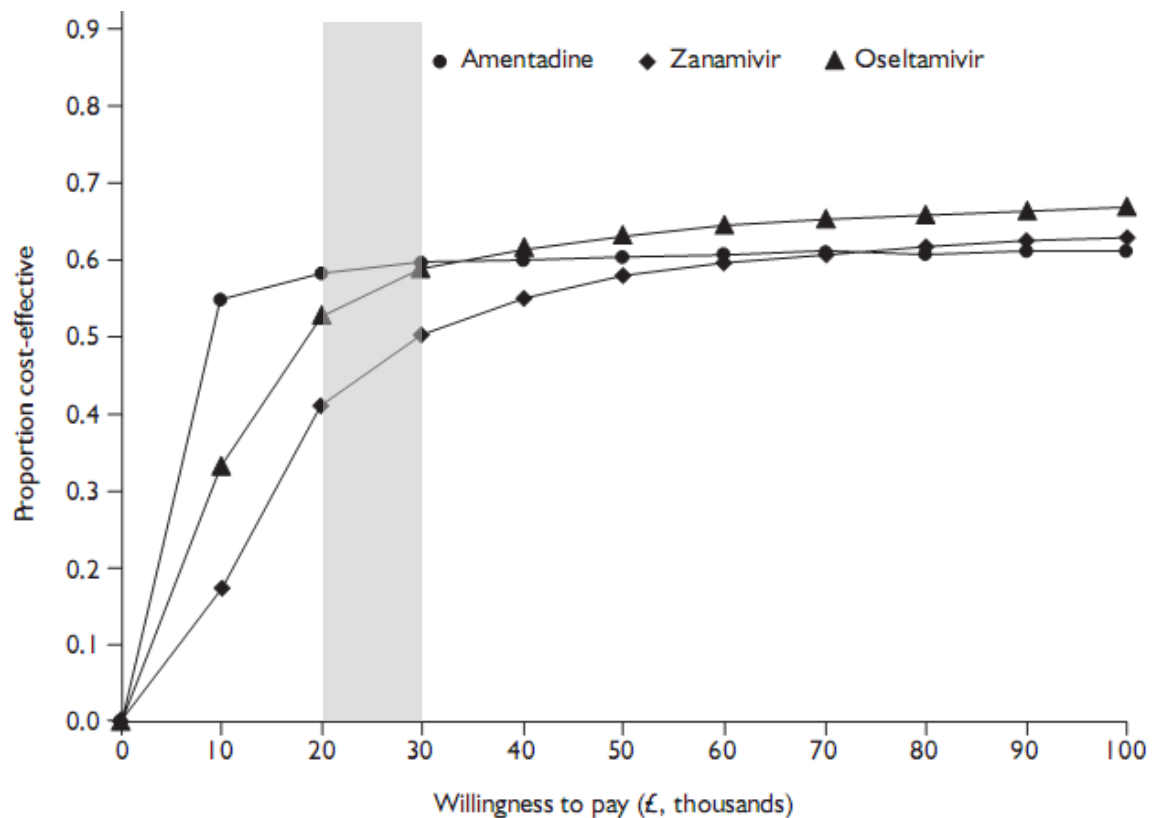


Figure 4. Cost-effectiveness acceptability curves for three kinds of antivirals against influenza. The shaded region shows the willingness to pay threshold of £20,000 - £30,000 per QALY gained. Adapted from Turner *et al.* Health Technol Assess 2003; 7(35).

Measuring health utilities: QALYs and DALYs

What units should we use to measure health consequences, the denominator of the ICER? We could use a disease-specific measure such as the number of cases of influenza or diarrhoea. But this limits comparability between different diseases. Not all diseases have the same impact on health; we are likely to put more value in preventing a case of Ebola than preventing a mild case of pharyngitis. We could use being alive or dead as the measure of health, but this would undervalue preventing diseases which do not normally cause mortality but which have chronic consequences (eg. HIV in the antiretroviral treatment era) or cause long-term sequelae (eg. polio).

An alternative is to use a *generic measure of health utility* such as a QALY or a DALY. These measures combine both the mortality and morbidity consequences of disease into a single measure that allows comparisons between diseases. To calculate QALYs, every health state is given a quality of life value (or weight) between 0 (dead) and 1 (perfect health). (Values below 0, representing health states considered worse than dead, are sometimes admitted.) QALYs are then calculated as the product of the quality of life value and the duration of time spent in that state. Let's look at Figure 5 for an example of how QALYs are calculated.

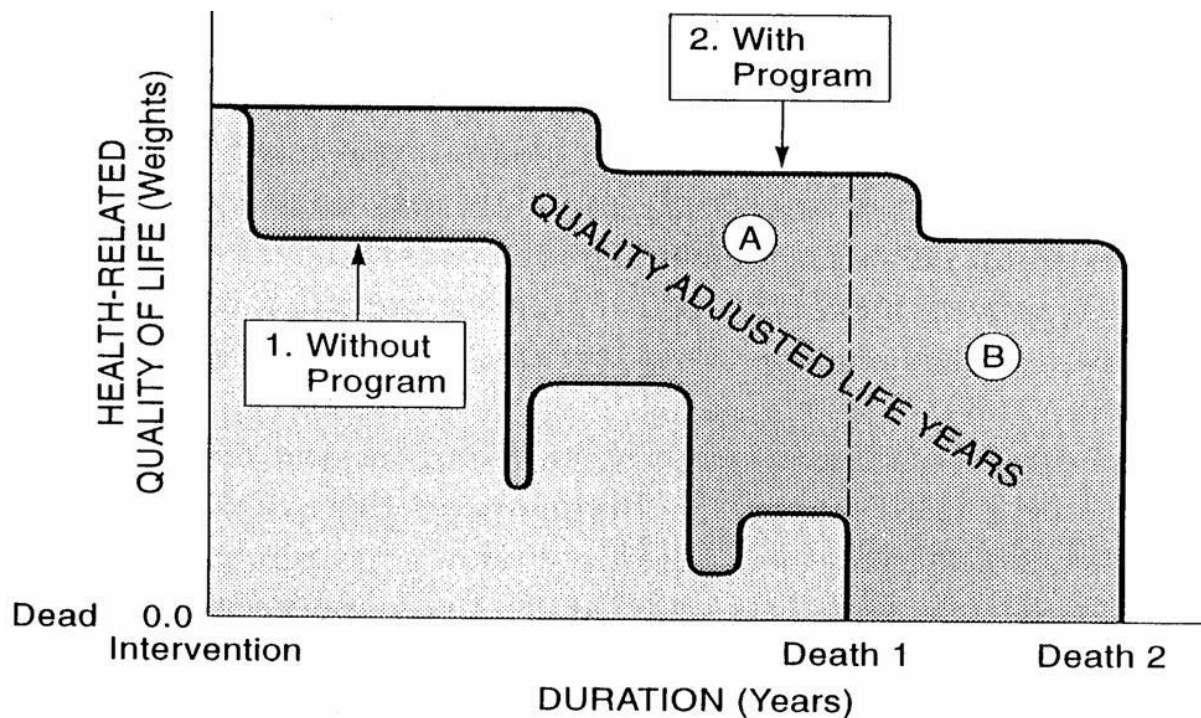


Figure 5. Quality of life weights for someone with a disease that is ultimately faithful, with and without a treatment programme. From Gold et al. *Cost-effectiveness in health and medicine*. OUP, 1996.

In Figure 5, we see the quality of life weights for someone with a chronic disease that eventually leads to her death. Without any treatment programme, her health-related quality of life gradually deteriorates until she dies at time Death 1. However, with a treatment programme, her health-related quality of life and she dies later at time Death 2. The incremental number of QALYs gained by the treatment is the area between the “Without program” and “With program” lines on the graph, i.e. the sum of region A (representing QALYs gained due to quality of life improvements) and region B (representing QALYs gained due to life extension).

So how do we actually estimate the quality of life weight to give someone in a particular state? There are several ways to do this but the simplest is usually to use a generic instrument for measuring quality of life. The most common instrument for measuring quality of life weights used in calculating QALYs is the EuroQoL group’s EQ-5D-5L instrument. This contains five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). Each of these dimensions can occupy one of five different levels ranging from “no problems” to “extreme problems”. (An earlier version of the instrument called the EQ-5D-3L only had three levels per dimension.) So the EQ-5D can be administered to the respondent, who chooses one level per dimension based on her current health state. The five health states are then converted into a quality of life weight using a conversion algorithm called a *tariff*. The values in the tariff are determined by population-wide surveys, but we won’t go into the methodology here – the book by Drummond et al. in the further reading section has details of this if you’re interested.

So for example, if you wanted to estimate how many QALYs are lost from an episode of influenza, you could administer the EQ-5D to patients with influenza from the time of symptom onset to the time they recover. This would give you an estimate of their quality of life at different time points, which you could use to plot a trajectory of their quality of life over their illness. The area under that trajectory would then be the QALYs associated with the episode of illness.

So we've looked at QALYs. What about DALYs? DALYs are a very similar instrument but there are a couple of key differences:

1. **Provenance.** QALYs were developed by American health service researchers in the 1970s, and are now used by health economists worldwide. DALYs were developed by investigators on the Global Burden of Disease programme of work that was initiated in the 1990s to measure the impact of various diseases around the world. Originally this resided in the World Bank. It then migrated to the World Health Organization and now finally resides at the Institute for Health Metrics and Evaluation, part of the University of Washington in the USA.
2. **Geographical span.** QALYs are mainly used in high income countries, although they are becoming increasingly popular in middle income countries. DALYs are used in low income countries and some middle income countries. Organisations such as the Bill and Melinda Gates Foundation, and the World Health Organization, also support the use of DALYs.
3. **Direction of scale.** Both measures estimate both the morbidity and mortality impact of disease, although morbidity is viewed in terms of a quality of life decrement with QALYs and a disability in DALYs. They differ in terms of direction. QALYs measure quality of life on a scale that ranges upwards from 0 (dead) to 1 (perfect health). DALYs measure disability, i.e. *reduction* in health, so they range downward from 1 (dead) to 0 (no disability).
4. **Valuation method for weights.** QALY weights for different diseases are usually estimated using instruments such as the EQ-5D. The results are then converted into weights using a tariff that is based on valuations of descriptions of health states by the general population. This can be very country-specific, so ideally requires data collection to derive new tariffs in each country. Since this process can be resource-intensive, not all countries have their own EQ-5D tariff. DALY weights are estimated based on universal set of standard values for diagnostic groups (eg. diarrheal diseases), that are rated by respondents around the world. This does not require additional data collection to derive new weights, but may not reflect values of particular countries. Also, the number of disease categories available to convert into weights is limited.

Discounting

One other issue to take into account when calculating the cost and benefit implications of an intervention in a cost-effectiveness analysis (or indeed any economic evaluation) is *discounting*. This is an issue that often confuses people.

To get an idea of the need for discounting, let's do a thought experiment. Imagine we live in a world with no inflation. Yesterday, I borrowed £1000 from you. Today, I offer you a choice. I can either return the £1000 to you today, or I can return it to you in 10 years' time. Which of the two would you prefer? It is extremely likely that you'd say today. But why's that? Perhaps your reasoning goes something like this. If I get the money today, I have more options. I can spend it today, or I can keep it and still spend it in 10 years' time if I don't need it now, so I haven't lost anything. If I'm a savvy investor, I can put it in the stock market and perhaps have more than £1,000 in 10 years' time.

Economists call this change in the value of money over time “discounting”. Discounting reflects our time preference... we would rather have good things (money, opportunities to consume) now but postpone bad things (loss of money, loss of health) to the future. In a cost-effectiveness analysis, we adjust for this by decreasing the value of both costs and health effects that occur in the future according to the following formula:

$$Y_n = \frac{X_n}{(1 + r)^n}$$

In the above formula X_n is the undiscounted value of the costs incurred (or health benefits received) in year n , Y_n is the discounted value of those costs or health benefits this year (called its *present value*) and r is the discount rate.

Discounting is imposed on all economic appraisals conducted by governments and private companies, not just in the health sector. The discount rate differs between countries based on financial calculations that each country does. For instance, if you want to know how the discount rate in the UK is calculated you can look up HM Treasury’s “Green Book” on appraisal and evaluation in central government which is on the internet. In some countries, the discount rate also differs between costs and health benefits. Whether costs and health benefits should have the same or different discount rates is an area of active debate among economists, but we won’t go into that here. Table 2 shows the discount rates used in a couple of countries.

Country	Discount rate: costs	Discount rate: effects
Mexico	5%	5%
Netherlands	4%	1.5%
Thailand	3%	3%
South Africa	5%	5%
UK	3.5%	3.5%
USA	3%	3%

Table 2. Discount rate for costs and health effects in different countries.

Discount rates tend to matter a lot for preventive interventions, such as vaccinating against an infectious disease. Let’s take vaccination against human papillomavirus (HPV) as an example. HPV is a sexually transmitted infection that can cause cervical cancer. But the cancer will usually occur many decades after the initial infection. So if we started vaccinating women against HPV today, we would only see reductions in cervical cancer in several decades time. Figure 6 shows what the stream of costs and health effects looks like with and without discounting. As you can see, discounting has a much bigger impact on health effects (cancers prevented) than on costs, since the effects only start to occur many years in the future. Hence discounting will make HPV vaccination look less cost-effectiveness; the higher the discount rate the less cost-effective it will look.

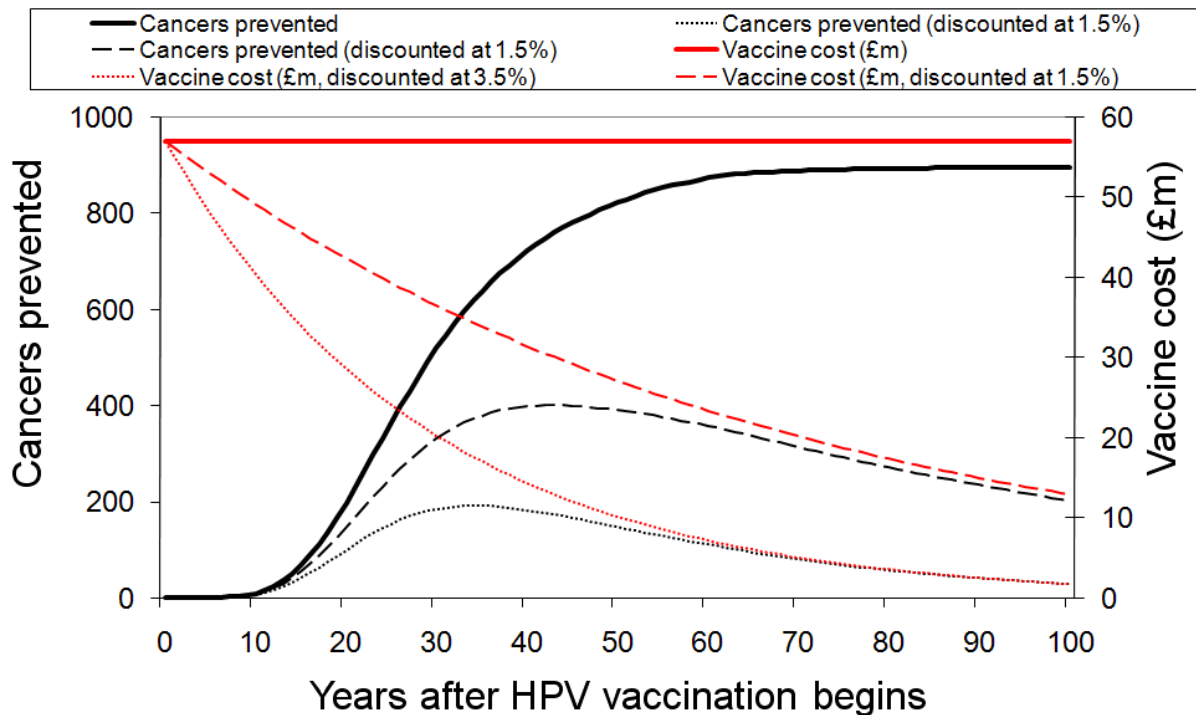


Figure 6. Vaccine costs and health effects (cancers prevented) by HPV vaccination, with and without discounting.

Economic evaluation of infectious disease interventions

So far almost everything we've looked at applies equally to infectious and non-infectious diseases. You can work out the cost-effectiveness of a new breast cancer drug, show uncertainty in your results and discount the outcomes in the same way that you'd do for a vaccine against influenza. So are there any differences in the way economic evaluations of infectious disease interventions are conducted?

In principle we should apply the same rules regardless of the kind of intervention or disease we're looking at. However, the implications of those rules may be different for infectious diseases, because they have some special features that affect their economic impact. We need to be aware of those special features when we're doing our calculations.

1. **Pathogen-host interactions.** If you've studied epidemiology, you'll know that understanding an infectious disease involves consideration of all three interacting parts of the "disease triangle": the pathogen, the host and the environment. Infectious disease interventions can have complex effects on both the pathogen and the host (and sometimes on the environment as well) that need to be captured in an economic evaluation. For instance, vaccinating against the malaria parasite may slow down the development of host immunity against malaria. Hence when the vaccine wanes, vaccinated individuals may be at greater risk of malaria than they would otherwise have been. This could reduce the cost-effectiveness of malaria vaccination and so host immunity needs to be captured in an economic model.
2. **Ecological externalities.** As mentioned before, infectious disease interventions have strong externalities – effects on people not producing or consuming the intervention. For instance, vaccinating someone against an infectious disease has large positive

externalities in the form of herd protection for non-vaccinated individuals. This generally improves the cost-effectiveness of vaccination. But there may also be negative externalities. For instance, *Streptococcus pneumoniae* is a bacteria with more than 90 serotypes, and only a handful of them are protected against by vaccination. Vaccinating people against *S. pneumoniae* can open an ecological niche that allows non-vaccine serotypes to proliferate in the population. This tends to reduce the cost-effectiveness of pneumococcal vaccination.

3. **Time horizons.** The time horizon is the length of time over which costs and benefits are tracked in an economic evaluation. The effects of most health interventions only last a short time, or at most until the end of the lifetime of the person receiving the intervention. However, infectious disease interventions can have lasting effects on entire populations long after the lifetime of the people receiving the intervention. For instance, eradicating a disease such as smallpox still benefits the world many decades after smallpox vaccination has ended. Hence economic evaluations of infectious disease interventions may require much longer time horizons than is usual in the field. This means that the effect of discounting is greatly magnified.
4. **Economic scope.** The economic evaluations we have looked at so far only consider *microeconomic* effects: economic effects on the people actually receiving the intervention, or at most on the people they infect or don't infect and the healthcare services that are used by sick patients. However, many epidemic and pandemic diseases can have *macroeconomic* effects far beyond the people who actually get sick. For instance, an outbreak of cholera or Middle East respiratory syndrome (MERS) can have a big impact on the wider economy of a nation by causing tourists to avoid visiting the country. An influenza pandemic can have an even more devastating blow, by stopping people from going to work or shopping.

Because of the special features of infectious disease interventions, economic models of these interventions may need to have special features. In the last part of this lecture, we'll look at two special types of models: *transmission dynamic models* and *macroeconomic models*.

Transmission dynamic models

Most models used for health economic evaluations are *static*. They only look at changing the risk of disease in people who actually receive the intervention. This is perfectly reasonable for non-infectious interventions. Someone receiving chemotherapy for cancer isn't going to change the risk of cancer progression for another patient in the next ward. However, for many infectious disease interventions this assumption breaks down because of the ecological effects we've discussed: herd protection, serotype replacement and antimicrobial resistance, to name just some. These effects normally require *transmission dynamic* models to capture their effects: models that take into account the change in risk of disease on other people besides those receiving the intervention.

Figure 7 shows the four most common types of models used in economic evaluations. Two of them are static: decision trees and Markov models. Another two are dynamic: compartmental dynamic models and individual-based models. You've already seen compartmental dynamic models and individual-based models from other parts of this course. We aren't going to spend too much time on decision trees and Markov models in this course because they are covered quite well in introductory health economics modelling courses (if you want to know more a

good book to refer to is Briggs et al. *Decision Modelling for Health Economic Evaluation*. OUP, 2006). But let's look at bit more closely at Markov models and see how they differ from compartmental dynamic models.

A Markov model involves flows between different compartments, just like a transmission dynamic model. The difference is that in a Markov model, all the rate of flows just depend on the source compartment. Hence the rate of flow between the susceptible and infected compartment in Figure 7 simply depends on how many susceptible people there are. In other words, the force of infection is constant. Conversely, in a transmission dynamic model, the rate of flow between the susceptible and infected compartment depends on both the number of susceptible people and the number of infected people. In other words, the force of infection depends on the number of people infected at that point of time.

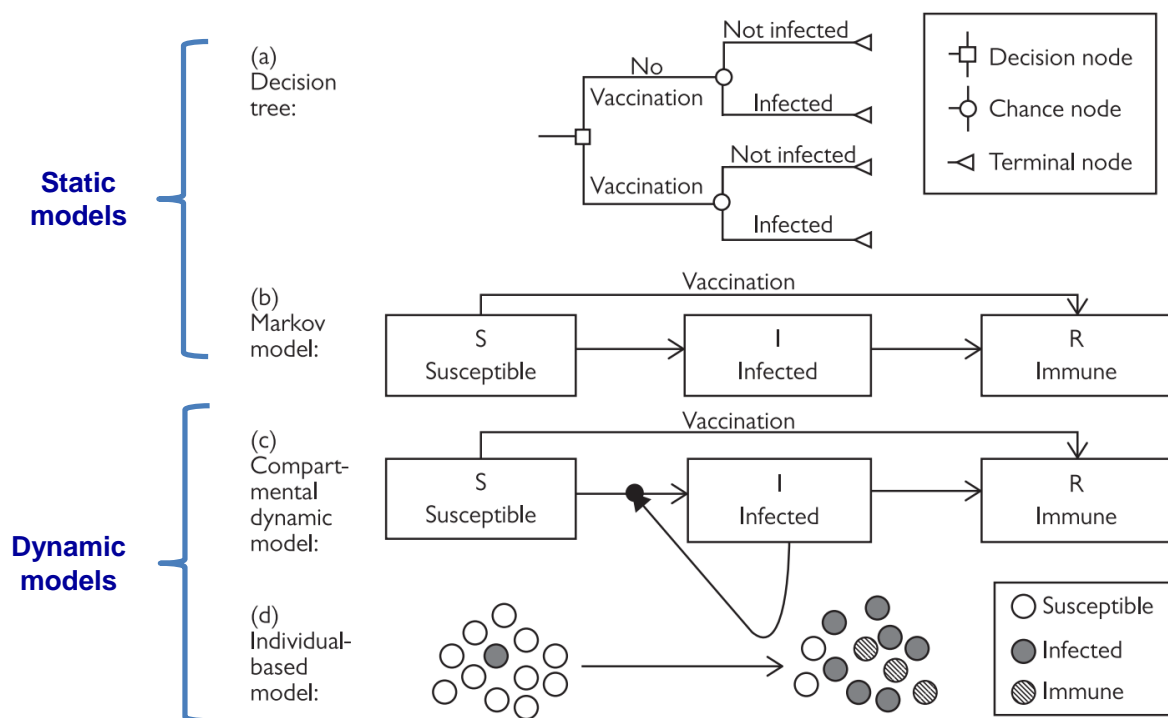


Figure 7. Types of models used in economic evaluations. From: Jit and White. Chapter 17: Economic analysis of interventions against infectious diseases. In: Oxford Specialist Handbook of Infectious Disease Epidemiology. OUP: Oxford, 2015.

Transmission dynamic models tend to take more effort to set up compared to static models – you have to estimate the probability of transmission between infected and susceptible individuals, which may depend on their age, location, disease severity etc. So when should a dynamic model be used and when is it sufficient to use a static model? Figure 8 gives a simple algorithm that can be used. In general, if a static model shows that an intervention is cost-effective, a dynamic model won't change the conclusion – it will suggest that it is even more cost-effective. So the static model can be used as a conservative approximation (i.e. an underestimate). However, if a static model shows that an intervention is not cost-effective, then we aren't sure whether a dynamic model will change the conclusion, so the static model is not helpful. There are some exceptions where we need dynamic models because the dynamic effects may make the intervention less cost-effective. This includes cases where we may see serotype replacement, increases in antimicrobial resistance, shifts in severity

because severity is age-dependent, or where we are comparing two interventions with each other. In general, you need a good grasp of the epidemiology of the organism and the effects of the intervention in order to be able to tell whether a static model is a reasonable conservative approximation. If in doubt, a dynamic model is probably the safer choice.

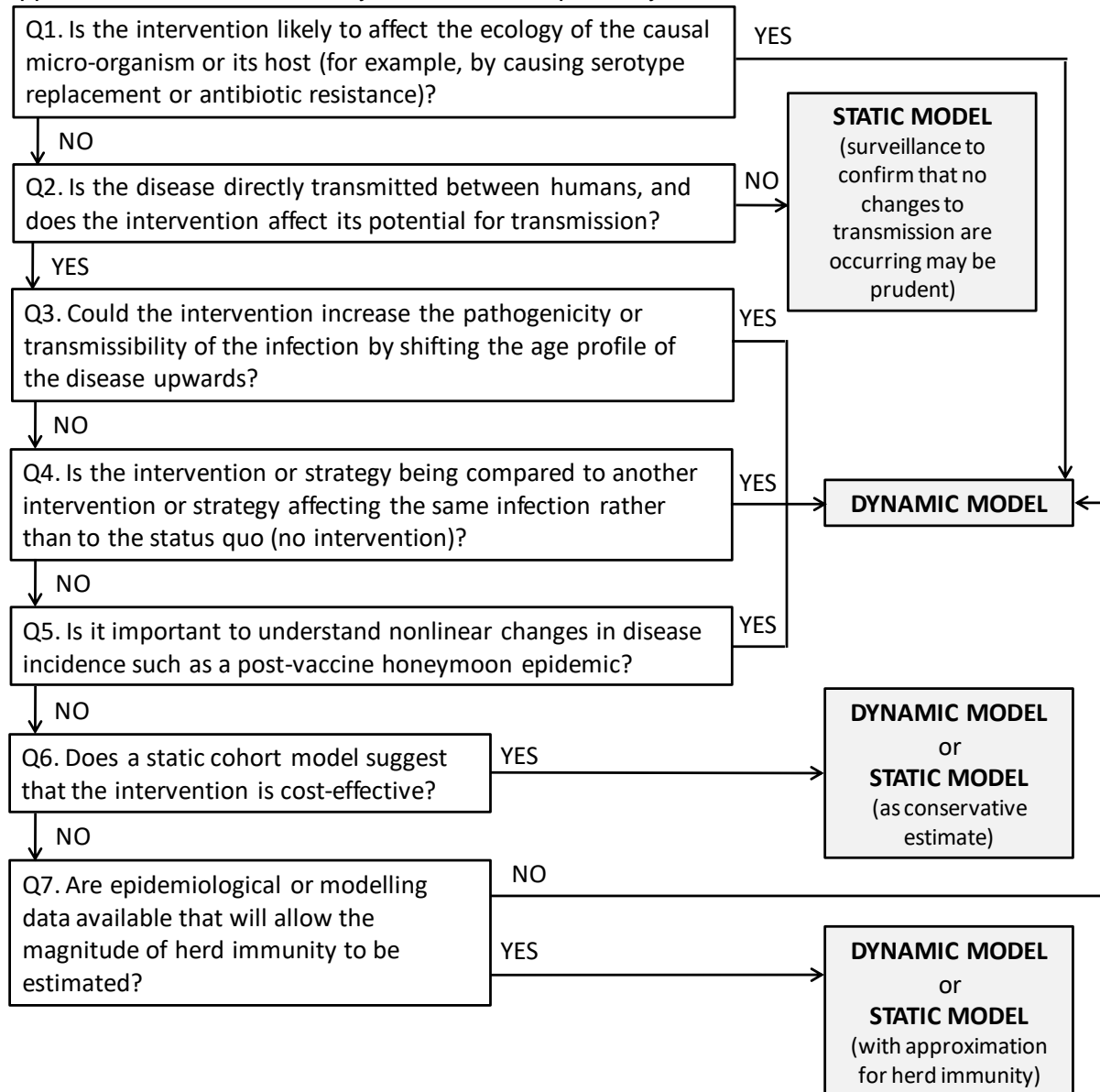


Figure 8. Flow diagram showing when a static model can be used and when a dynamic model is needed. From Jit and Brisson. *Pharmacoeconomics* 2011; 29:371.

Interestingly, most economic models of infectious disease interventions in the literature have been static. For instance a recent review of reviews found that only 8 out of 84 economic evaluations of vaccination published between 1993 to 2014 were fully dynamic (Jit and Mibe. *Vaccine* 2015; 33:378). In some cases this may have been justified, since a simpler and less data-demanding static model may have been sufficient to show that a vaccination programme is cost-effective. However, in other cases, where a static model shows that vaccination is not cost-effective, or where the externalities of vaccination could be negative, this may lead to unjustified conclusions. An example is pneumococcal conjugate vaccination which can lead to serotype replacement; indeed a recent review (Wu et al. *Vaccine*. 2015 Mar 30;33(14):1633-58) suggested that ecological externalities can have a large impact on the cost-effectiveness of different variants of this vaccine. The predominance of static models in the literature may

reflect lack of training for many health economists in the use of transmission dynamic models, although this is gradually changing.

Macroeconomic models

As mentioned before, all the models we have considered so far are microeconomic models. They consider the impact of sick people on the economy of their households, the healthcare services they use for treatment and perhaps the firms they work in. The assumption we have to make is that any changes to the economy of these households, healthcare services and firms will not have any ripple effects on the wider economy. This is called a *partial equilibrium* assumption, and is usually a reasonable assumption.

However, it becomes less reasonable when an infectious disease outbreak causes large changes in the behaviour of consumers, major workplace absenteeism and the capacity of healthcare service providers. This may happen in a severe epidemic of diseases such as cholera or dengue which could cause tourists and other visitors to avoid a country. It may also happen during a pandemic or spread of an emerging infection, as we saw during outbreaks of influenza, severe acute respiratory illness, Ebola and Zika. In these situations, the outbreak is likely to have large ripple effects on the wider economy. We need a different kind of model to capture them – a *general equilibrium* model.

In this kind of model, the entire economy is simulated, as well as the relationship between different sectors like the retail, transport and healthcare sectors. Initially the economy is assumed to be in equilibrium. When a large exogenous shock such as a pandemic occurs, labour supply and consumption in different sectors are affected. The new production levels in the economy after the shock are simulated and this allows an estimation of the economic impact of the outbreak on the wider economy.

Concluding messages

1. Health economic evaluation is used to ensure that resource allocation decisions are made on explicit, evidence-based and needs-based criteria.
2. Economic evaluation of vaccination is a specialised field which requires analysts familiar with both health economics as well as the special epidemiological features of vaccine-preventable diseases.
3. It is important to understand the assumptions and limitations behind different types of economic models as these can have a large impact on model results.

Further reading

If you want a comprehensive overview of the techniques of health economic evaluation then the “go to” textbook on this is *Methods for the economic evaluation of health care programmes* by Drummond et al. (OUP, 2015) which is now in its fourth edition.

A shorter and more infection-focused introduction can be found in chapter 17 (“Economic analysis of interventions against infectious diseases”) of the *Oxford Specialist Handbook of Infectious Disease Epidemiology* (OUP, 2015).

For more focused discussions of economic evaluations of infectious diseases in particular the following articles might be useful:

- Beutels P, Scuffham PA, MacIntyre CR. Funding of drugs: do vaccines warrant a different approach? *Lancet Infect Dis* 2008; 8:727.
- Jit M, Brisson M. Modelling the Epidemiology of Infectious Diseases for Decision Analysis: A Primer. *Pharmacoeconomics* 2011; 29:371.

More information about the importance of dynamic models in economic evaluation of many infectious disease interventions can be found in this article:

- Brisson M, Edmunds WJ. Economic evaluation of vaccination programs: the impact of herd immunity. *Med Decis Making* 2003 Jan-Feb;23(1):76-82.

More information about generalised equilibrium models can be found in these articles:

- Beutels P, Edmunds WJ, Smith RD. Partially wrong? Partial equilibrium and the economic analysis of public health emergencies of international concern. *Health Econ.* 2008; 17(11):1317-22.
- Keogh-Brown MR, Wren-Lewis S, Edmunds WJ, Beutels P, Smith RD. The possible macroeconomic impact on the UK of an influenza pandemic. *Health Econ* 2010; 19(11):1345-60.

Introduction to Infectious Disease Modelling and its Applications - 2018

Session 19: Fitting models to data II: numerical optimisation and sensitivity analysis

Lecture

Overview and Objectives

This is the second lecture in a two-part series on fitting models to data. In this first lecture, we looked at how we measure the goodness of fit of a model to data. We'll briefly recap the key principles we learnt in this lecture. After that, we'll look at numerical algorithms used to achieve the best fits, and discuss methods for undertaking sensitivity analysis.

By the end of this lecture you should be able to:

1. Explain the purpose and some shortcomings of numerical optimisation algorithms, using gradient descent as an example.
2. Explain the need for sensitivity analysis to explore changes in results when input parameters are varied.
3. Conduct one-way sensitivity analysis using Berkeley Madonna.
4. Explain the purpose of multi-way sensitivity analysis and the principles behind different methods of doing this (grid search, random sampling, Latin hypercube sampling).
5. Use and interpret histograms and tornado graphs to show the results of sensitivity analyses.

Review of previous lecture

If you remember, during the previous lecture on model fitting, we looked at a general model with a vector (set) of m input parameters $\mathbf{x} = (x_1, \dots, x_m)$. We also had n data points (observations) which we call O_1, \dots, O_n , which we compared to corresponding model outputs (at the same time points for instance) that depend on the value of the input vector \mathbf{x} . We call these model outputs $E_1(\mathbf{x}), \dots, E_n(\mathbf{x})$, where $E_1(\mathbf{x})$ is the model output at time point t_1 when the input parameters take the value \mathbf{x} and so forth.

Hence we want to compare O_1 to $E_1(\mathbf{x})$, O_2 to $E_2(\mathbf{x})$ and so on, such that O_i is as close as possible to $E_i(\mathbf{x})$. In other words, we need a goodness of fit function $g(E_1(\mathbf{x}), \dots, E_n(\mathbf{x}), O_1, \dots, O_n)$ which takes as its inputs both the model predictions as well as the observations, and returns a single number that signifies how close $E_1(\mathbf{x}), \dots, E_n(\mathbf{x})$ is to O_1, \dots, O_n .

We looked at a couple of ways of writing down a goodness of fit function, based on techniques such as minimizing the sum of squared residuals and maximizing the model likelihood. Regardless of the technique, we want a function such that the smaller (or larger, depending on the function) it is, the better the model fit to data. For instance, we might choose a function that takes the value 0 when $E_1(\mathbf{x}) = O_1$, $E_2(\mathbf{x}) = O_2$ and so forth, and get bigger the further away we get from this perfect fit.

That leads us to our second question. Once we have a goodness of fit metric, how do we find the parameters that “best fit”, i.e. those that make the metric take its most favourable (best fitting) values? In order to answer that question, we need a fitting algorithm (that is, a series of instructions that a computer can follow in order to end up at a best fitting value for a set of model parameters). Finding robust and efficient algorithms is called numerical optimisation, and a very large body of mathematical theory exists about it in the field of numerical analysis. We are only going to cover the very bare essentials of this field!

Fitting algorithms

Consider a function $f(x)$ of a single variable x . The gradient, or steepness of the slope, of $f(x)$ at a given point x is written as $Df(x)$ or df/dx . When $Df(x) = 0$, we know that the function is absolutely flat. So we know that when $f(x)$ is minimized, $Df(x) = 0$, as Figure 1 shows.

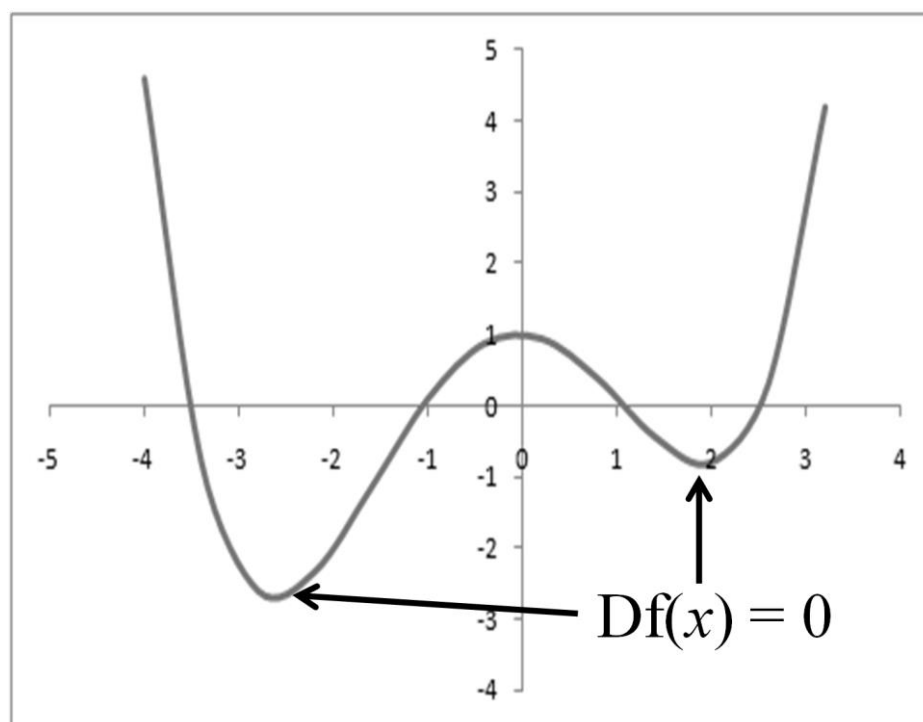


Figure 1. The gradient $Df(x)$ of a function $f(x)$ is 0 when the function is minimised.

In practice, $f(x)$ may be a very complicated expression. For instance, $f(x)$ may actually be an SIR model of an infectious disease at a time point $t=x$. In such a situation, it is extremely difficult to write down $f(x)$ in an explicit expression, let alone its derivative. However, we can estimate $Df(x)$ numerically for a given $x=x_0$. This idea gives rise to one of the simplest numerical optimisation algorithms, called gradient descent.

Gradient descent

Gradient descent is one of the simplest numerical optimisation algorithms for finding the point where a given function $f(x)$ is minimised. It involves the following steps, shown in Figure 2.

1. Choose a starting point x_0 .
2. Search in the direction that f is decreasing most rapidly (the downhill gradient $-Df(x_0)$).

3. Move in that direction a certain distance δx , where δx is a small number that you choose.
4. Get to a new point $x_1 = x_0 - \delta x Df(x_0)$.
5. Repeat this process until $Df(x)$ is sufficiently small, meaning that you are very close to a minimum point.

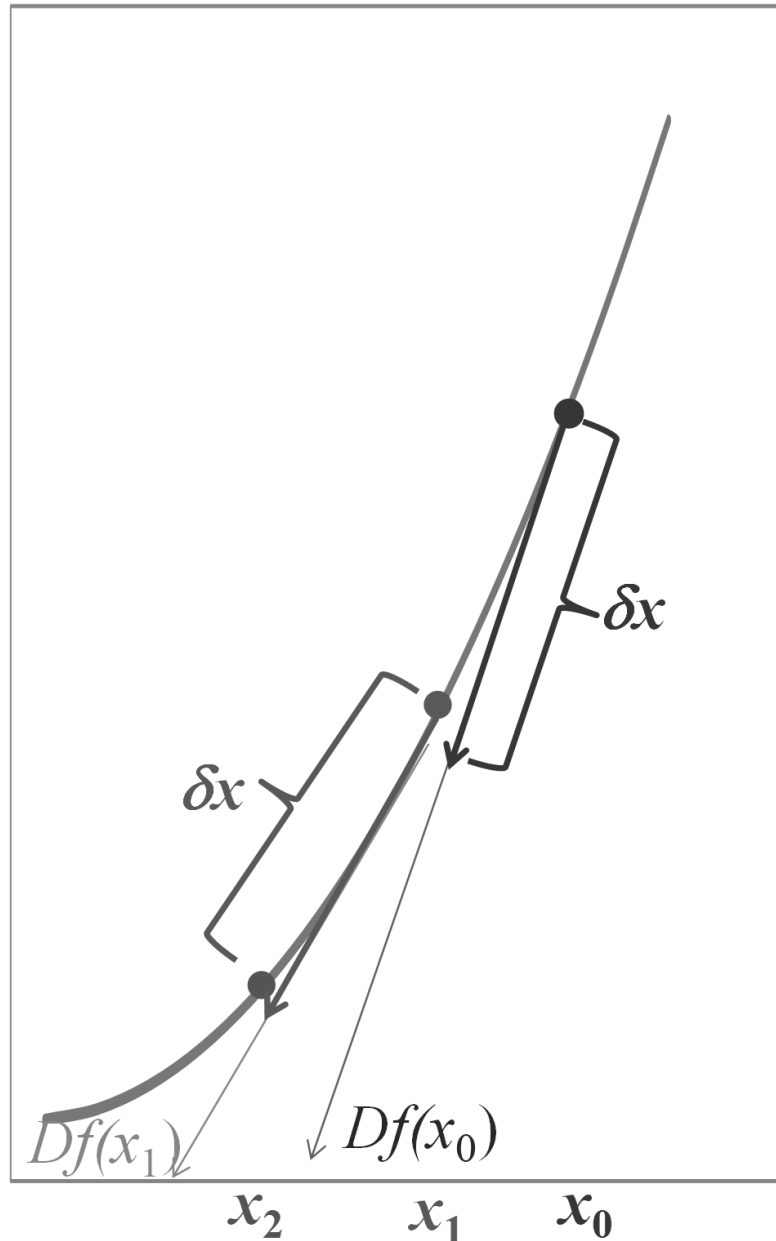


Figure 2. The gradient descent algorithm, moving from a starting point x_0 to x_1 and then x_2 which are increasingly closer to the minimum of $f(x)$.

You may be thinking that this looks fine in the case where $f(x)$ is a very simple function of just one variable. What if you get more complicated functions where \mathbf{x} is a vector of several variables? Well, let's look at the case where $f(\mathbf{x})$ is a function of two variables $\mathbf{x}=(x_1, x_2)$, we look for \mathbf{x} such that $\nabla f(\mathbf{x}) = 0$. $\nabla f(\mathbf{x})$, pronounced 'del $f(\mathbf{x})$ ', is a generalisation of $Df(\mathbf{x})$ when \mathbf{x} is a vector rather than a scalar (a point in more than one dimension). $\nabla f(\mathbf{x})$ points us in the direction of the downward slope, but this time $f(\mathbf{x})$ is more like a hill where you have to choose your direction in both the north-south as well as the east-west axis. Figure 3 shows one way of visualising this process

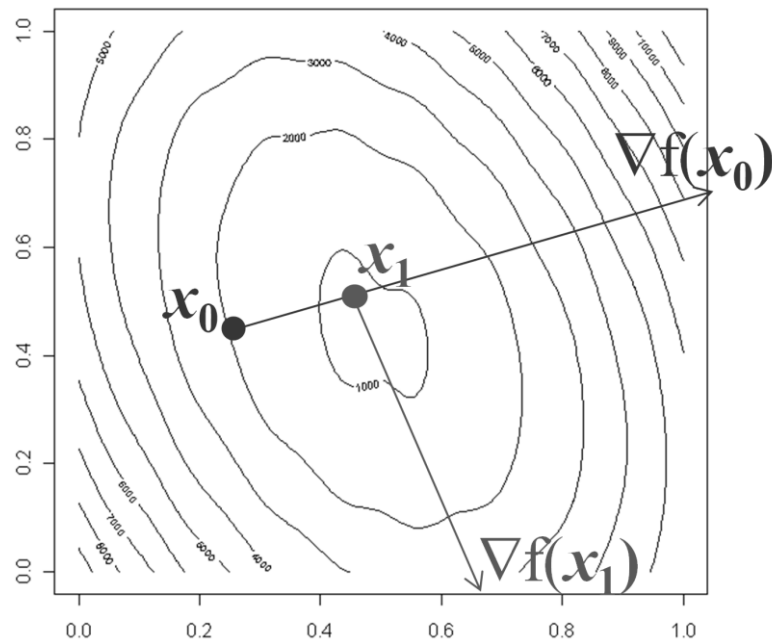


Figure 3. The gradient descent algorithm when $f(\mathbf{x})$ is a function of two variables $\mathbf{x}=(x_1, x_2)$.

In fact we can generalise this to $f(\mathbf{x})$ being a function of n variables $\mathbf{x}=(x_1, \dots, x_n)$. This becomes quite hard to visualise on a graph, but computationally it is no different.

Gradient descent runs into problems when you have multiple local minima, that is points which are lower than any of the neighbouring points but not necessarily lower than any other point on the curve. For instance, in Figure 4, when you start at point $\mathbf{x}=\mathbf{x}_0$ and use gradient descent, you will end up close to a local minimum around $\mathbf{x}=2$. However, there is another minimum at around $\mathbf{x}=-2.5$ which is actually lower, and which there is no way of reaching from $\mathbf{x}=\mathbf{x}_0$. There's an example of this in the CJD practical next week.

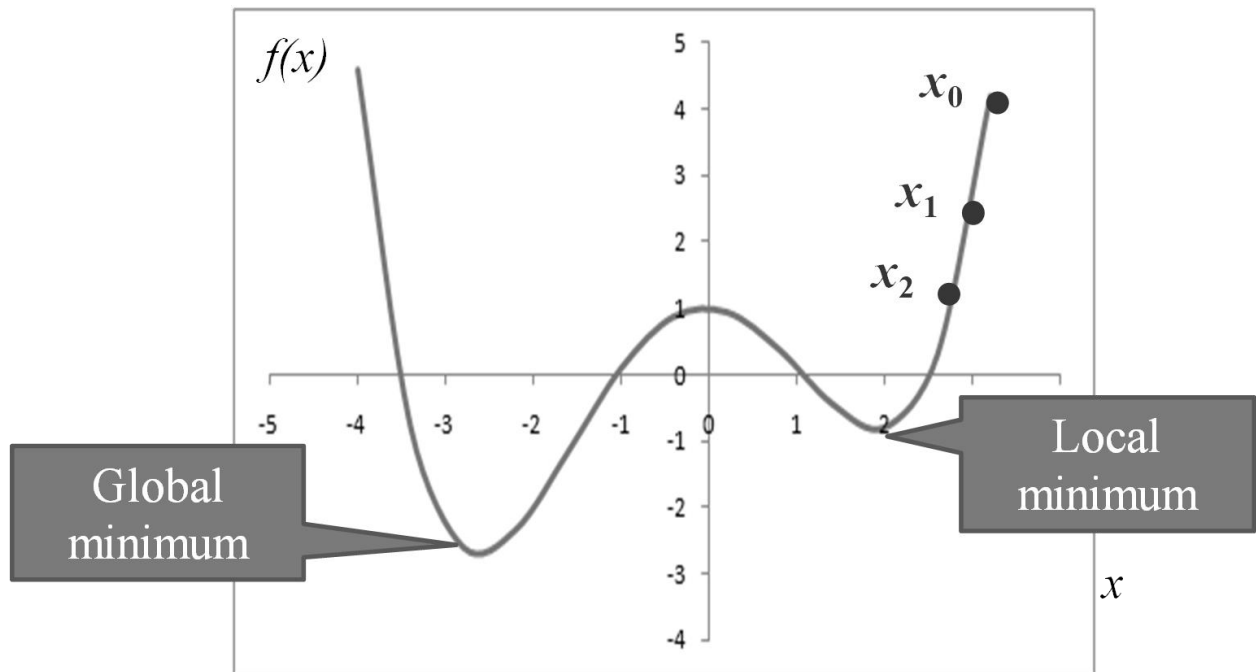


Figure 4. Problems with gradient descent when there are multiple local minima. Starting the algorithm at $x=x_0$ will bring you to a local minimum rather than a global minimum.

How do we deal with this problem? There actually isn't a foolproof algorithm for avoiding this problem. It's usually a good idea to try to start reasonably close to the global minimum if you have an idea where it is. For instance if you are able to see what the function looks like you can choose a point near the minimum. Alternatively, if you have some insight into the model, you may be able to pick a region that is close to what is likely to be the minimum for biological and physical reasons.

Otherwise, you could pick multiple starting points and run gradient descent at each of them, to see if you end up with a number of local minima. You can then check to see which one gives the lowest value of $f(x)$. Also you can make sure that they are physically and biologically plausible. For example, a negative value of R_0 , or one that is in the hundreds for an infection such as influenza, is probably implausible.

There are also probabilistic algorithms such as simulated annealing that try to avoid getting "trapped" in a local minimum by occasionally taking steps in the upward rather than downward direction. These tend to be slower than deterministic algorithms but can be better at avoiding this problem.

In practice, gradient descent is not used for problems of significant complexity because it is inefficient. Instead, we tend to use algorithms with names like Levenberg-Marquardt (used by Berkeley Madonna, MATLAB and Mathematica), generalised reduced gradient (used by Excel Solver), Nelder-Mead (also known as downhill simplex). Technical details of these algorithms aren't really necessary, but they all suffer from similar problems as gradient descent (though to a lesser extent). This is partly why Excel Solver and Berkeley Madonna have problems with fitting (you need to start them off at the right place).

Sensitivity analysis

We've now looked at the issue of fitting models to data, both in terms of finding a goodness of fit metric and also algorithms for reaching a point where the goodness of fit metric is as big (or small) as possible. In this second section, we're going to look at the issue of **sensitivity analysis**.

Sensitivity analysis explores the change in results when input parameters are varied. (To be precise, *parametric* sensitivity analysis explores this. There are other forms of sensitivity analysis which explore changes in results when other aspects of the model besides the input parameters are varied. We'll consider them briefly at the end of these notes.)

Input parameters to a model are inherently uncertain. It's impossible to precisely measure or know the probability of transmission of infection between two individuals, or the duration that someone will be infected or will have natural immunity. To what extent does uncertainty in these parameters affect results?

Two things are important: First, we need to know the magnitude of uncertainty around each parameter. Second, we need to know how important each parameter is in influencing the results. Particularly in non-linear models (such as models of infection transmission), some parameters are more influential than others. Hence a highly influential and highly uncertain parameter can cause the results to be highly uncertain.

One-way sensitivity analysis

One way to explore the influence of each parameter on the results is to use one-way sensitivity analysis. This involves changing the value of individual parameters one at a time, while keeping the remaining parameters fixed, and seeing what effect this has on outcomes of interest. Berkeley Madonna has several options under the "Parameters" menu that can do this. Sliders and the parameter window allow you to vary each parameter and see the effect on results, while batch runs allow you to generate many different scenarios.

The problem is that varying parameters one at a time while holding others at base case values does not give a complete description of the sensitivity of the model to each parameter. For instance, let's think of an SIR model - a model of an infectious disease like measles or pertussis which conveys permanent natural immunity upon recovery. After a certain time, the proportion of the population that is infected (i.e. the prevalence of infection) will reach an endemic equilibrium in which the rate at which new individuals become infected is equal to the rate at which infection is cleared. Suppose we now introduce vaccination in our model at $t=500$ (see Figure 5). Vaccination will cause the infectious equilibrium to decrease, i.e. a smaller proportion of the population will be infected at any given time. The higher our vaccination coverage, the lower the infectious equilibrium, until it reaches zero when coverage hits the herd immunity threshold.

Now let's consider the difference between 20% and 50% vaccine coverage. Clearly at 50% coverage, the infectious equilibrium will be lower than at 20% coverage. But how much lower? Well that depends on the probability of transmission. If the transmission probability is low, then increasing coverage will have a large effect on the infectious equilibrium, because there is less transmission to block (look at the left graph of Figure 5 for this). But if the transmission probability is high, then increasing coverage will have a smaller effect, because there is so much transmission going on that just blocking some transmission events won't be enough to stop most people from getting infected (right graph of Figure 5).

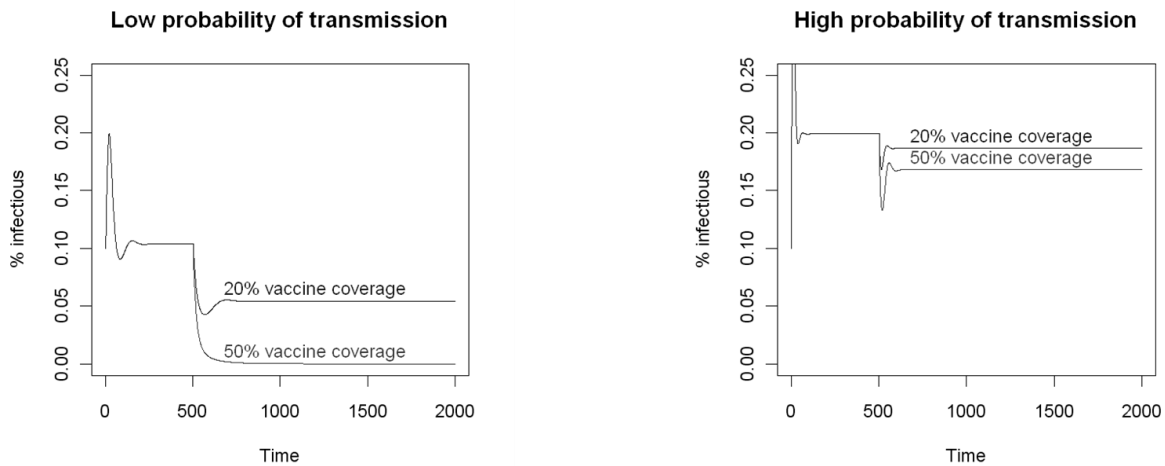


Figure 5. Effect of increasing vaccine coverage from 20% to 50% in an SIR model with low (left) or high (right) transmission probability.

Hence we see that we can't determine how sensitive the model is to vaccine coverage without knowing what is happening to other parameters, such as the probability of transmission. That's the weakness of one-way sensitivity analysis – many models become more sensitive to certain parameters when other parameters take particular values. Hence we have to evaluate sensitivity to all the parameters at the same time, not just one at a time.

One-way sensitivity analysis

So how do we evaluate the sensitivity of a model to more than one parameter at the same time, i.e. how do we do **multi-way sensitivity analysis**? The most obvious method is to do a grid search – to systematically search the joint parameter space of all relevant parameters and evaluate the result at each combination of parameters.

For instance, suppose we had a model with two parameters, A and B . We could evaluate each parameter at three values (say a_1 , a_2 and a_3 for parameter A , and b_1 , b_2 and b_3 for parameter B). If we want to evaluate the model at every combination of these three values for each parameter, we will need to sample $3^2 = 9$ parameter sets (Figure 6). Clearly you can see that if number of parameters or number of values for each parameter is large then grid search becomes unfeasible. With 5 parameters and 10 values each, we need $10^5 = 100,000$ sets.

		A		
		a_1	a_2	a_3
B	b_1	X	X	X
	b_2	X	X	X
	b_3	X	X	X

Figure 6. Grid search of a model with two parameters A and B taking three values each.

Instead of systematic exploration of the parameter space, we can use a technique called **Monte Carlo sampling**, named the city of Monte Carlo which is famous for its casinos. Nick Metropolis, one of the inventors of this method, thought of the name because his co-inventor Stan Ulam “had an uncle who would borrow money from a relative because he just had to go to Monte Carlo” (Metropolis N, The beginning of the Monte Carlo method, *Los Alamos Science Special Issue* 1987). Monte Carlo sampling works like this:

1. Pick a value for each parameter we are uncertain about from some distribution. (For example, we may pick A and B uniformly from the range 0.01 – 0.5).
2. Evaluate the outcome measure by solving the model for that set of parameters.
3. Repeat this process many times (e.g. 100,000).

This form of sensitivity analysis is called **probabilistic sensitivity analysis**. It avoids having to sample every single combination of values for each parameter, while still ensuring a roughly even distribution of samples across the parameter space (when the process of sampling is repeated enough times).

One advantage of probabilistic sensitivity analysis is that it allows us to choose appropriate probability distributions for each parameter, which can reflect our belief about how likely it is that a parameter will take certain values. For instance, if we think that a parameter is equally likely to take any value between 0 and 1, then we can sample its values from a uniform distribution on $[0,1]$. Or if we think that it is more likely to be closer to 0.5, we could instead sample from a triangular distribution on $[0,1]$ with the mode at 0.5.

We could also construct distributions from corresponding outcomes in epidemiological studies, or syntheses (eg. meta-analyses) of several studies. In this case we are likely to use the sampling distribution of the mean of the appropriate outcome eg. normal or lognormal. If there are no data available, there are methods that allow us to elicit appropriate distributions from an expert or panel of experts, although expert elicitation should always be a last resort.

It is important to remember that the probability distribution around a parameter should represent the **uncertainty** around that parameter rather than the **variability**. Uncertainty **represents** our (lack of) knowledge about a quantity, which can be reduced by further study (eg. taking a larger sample). Variability on the other hand represents heterogeneity between individuals in a population, which is inherent in the population and will not be reduced by further study. For instance, suppose we wanted to estimate the mean height of everyone in this class. If we measured everyone's height precisely and took the mean, there would be no uncertainty in our estimate of the mean (since we were able to include every single member of the population), but a lot of variability. On the other hand, if we only sampled two individuals from this class and they both had a height of 1.7 metres, there would be no variability in our estimate but a great deal of uncertainty.

Latin hypercube sampling

If we sample completely at random from the probability distribution of each parameter, the method is inefficient as it doesn't ensure full coverage of parameter space. Instead we may get clusters of points (i.e. over-sample certain parts of the parameter space and under-sample others), simply because of the luck of the draw (see Figure 7).

A more efficient way to sample is to use **Latin hypercube sampling**. This uses the following method:

1. Divide the probability distribution of each parameter into N equal probability sections.
2. Sample each parameter from one of its available sections (without replacement).
3. Repeat until you have built up N equal parameter sets that together encompass all sections of all parameters.

The values chosen are more evenly distributed than in a simple random Monte Carlo sample.

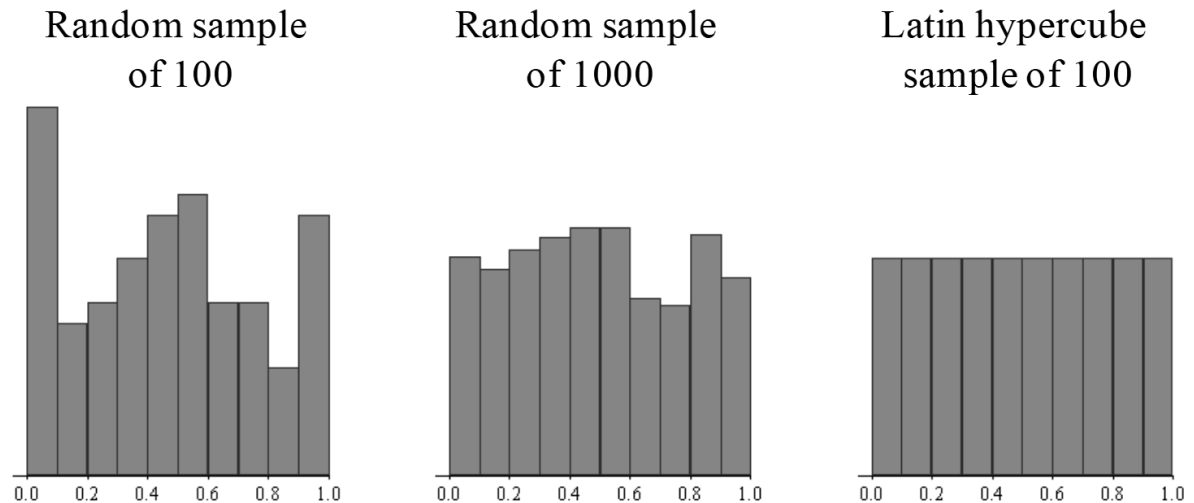


Figure 7. Typical distribution of values when sampling from a parameter using either completely random sampling or Latin hypercube sampling.

Representing results

How do we display the results of our uncertainty analysis? The simplest way is to use a tornado graph (see Figure 8). A tornado graph is a stacked bar plot in which each bar shows how the outcome parameter (eg. prevalence of infection at equilibrium) varies as each input variable is varied. Hence longer bars represent variables that contribute more to the uncertainty in the outcome.

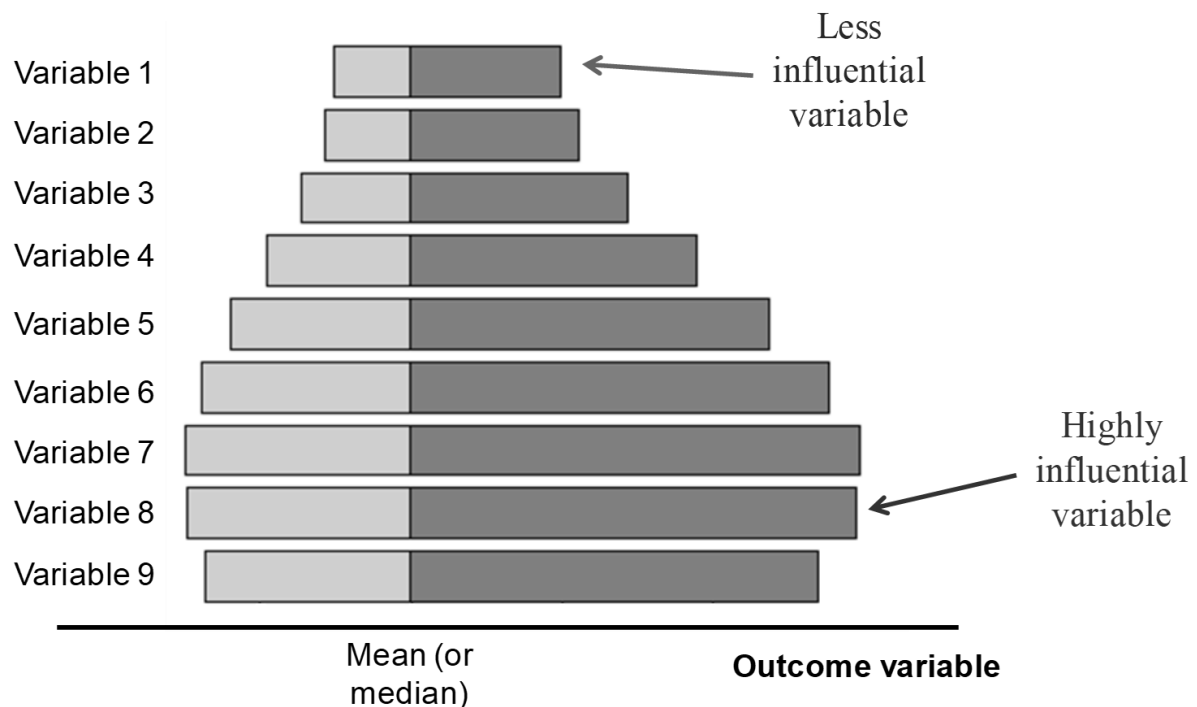


Figure 8. Tornado plot showing how an outcome variable varies as different input variables are varied.

So how do you construct a tornado plot? If you're doing one-way sensitivity analysis, this is easy. Just vary each input parameter within a certain range, and record how the output variable changes. Now adjust the width of the corresponding bar in the tornado plot based on the range in the outcome.

If you're doing multi-way (eg. probabilistic) sensitivity analysis this becomes a bit more complicated. Remember that in multi-way sensitivity analysis, the influence a parameter has on the outcome depends not only on how it is varied, but also on how all the other parameters are varied. To capture the joint effect of all the parameters varying at the same time, we need to use a statistical model. So once you have your set of jointly sampled input parameters and corresponding outcomes, construct a statistical (eg. linear) model of the association between the outcome variable and each of the input parameters, parameterised using the sampled parameter sets. Now vary each input parameter within its uncertainty range in the model and see how the outcome changes. This gives you the marginal effect of changing that one parameter (the effect of changing that parameter regardless of what values the other parameters take).

For example, let's consider our SIR model again (Figure 9). Remember that this has three input parameters: β (probability of transmission), r (rate of clearing infection) and w (rate of natural immunity waning) that are uncertain. Let's start by doing some one-way sensitivity analysis: let's vary β (probability of transmission) while keeping r and w fixed, and then see what the (pre-vaccination) infectious equilibrium will be. We find that as we increase β , the infectious equilibrium increases (Figure 10). This makes sense; if there is more transmission going on, then a larger proportion of the population is going to be infected at any one time.

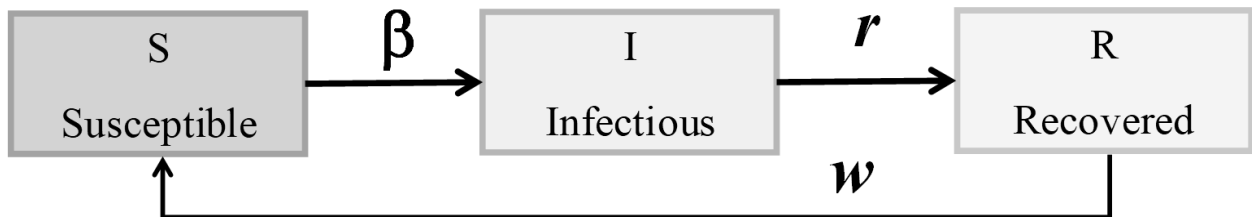


Figure 9. SIR model with three input parameters: β (probability of transmission), r (rate of clearing infection) and w (rate of natural immunity waning).

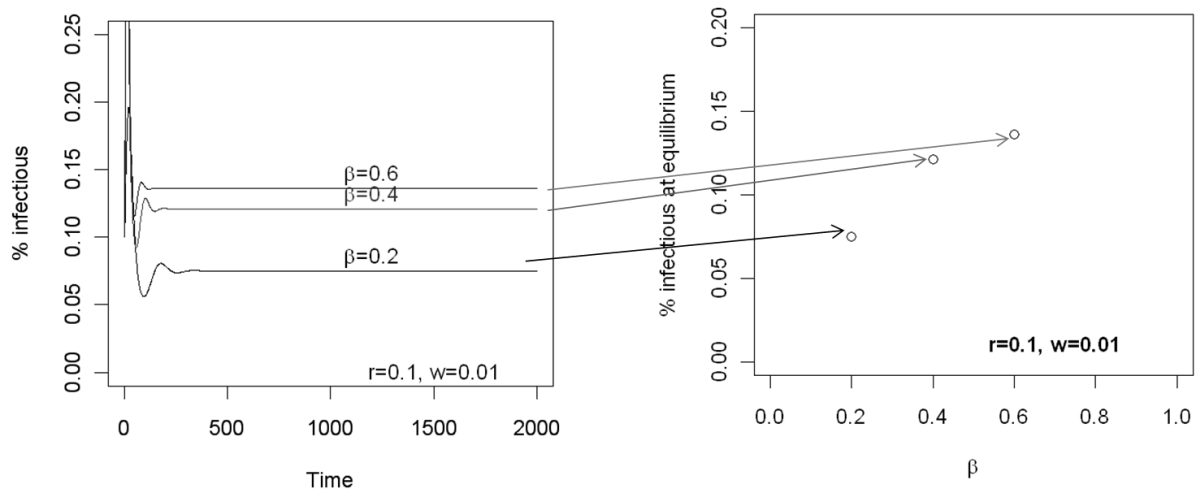


Figure 10. Effect of varying β (probability of transmission) on the infectious equilibrium in the SIR model in Figure 9.

Now let's vary r (the rate at which infection wanes), while continuing to vary β . The results are in Figure 11. Clearly the effect that β has on the infectious equilibrium changes for different values of r . When we take a small value of r , β has a larger effect. This is as expected - a short duration of infection (higher rate at which infection wanes) dampens the importance of the probability of transmission.

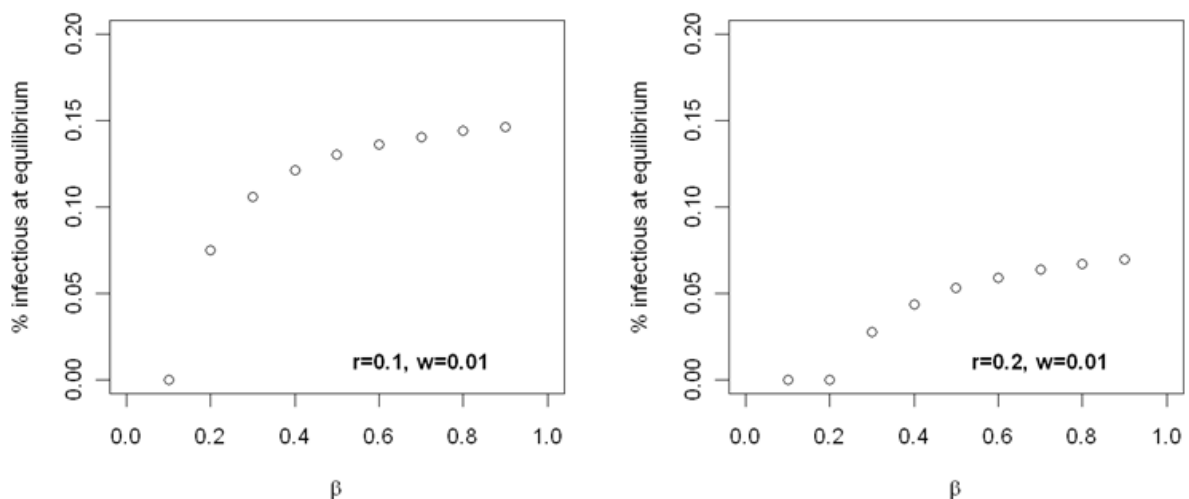


Figure 10. Effect of varying β (probability of transmission) on the infectious equilibrium in the SIR model in Figure 9, for two values of r (rate of waning of infection).

We can take this even further, and investigate the joint effect of varying all three parameters at the same time. To this, let's sample β , r , w from uniform distributions. Let's have β sampled from a uniform distribution on $[0.1, 0.9]$, r from $[0.01, 0.09]$, and w from $[0.01, 0.09]$. When we run 10,000 random samples through the SIR model, the distribution of outcomes looks like Figure 11.

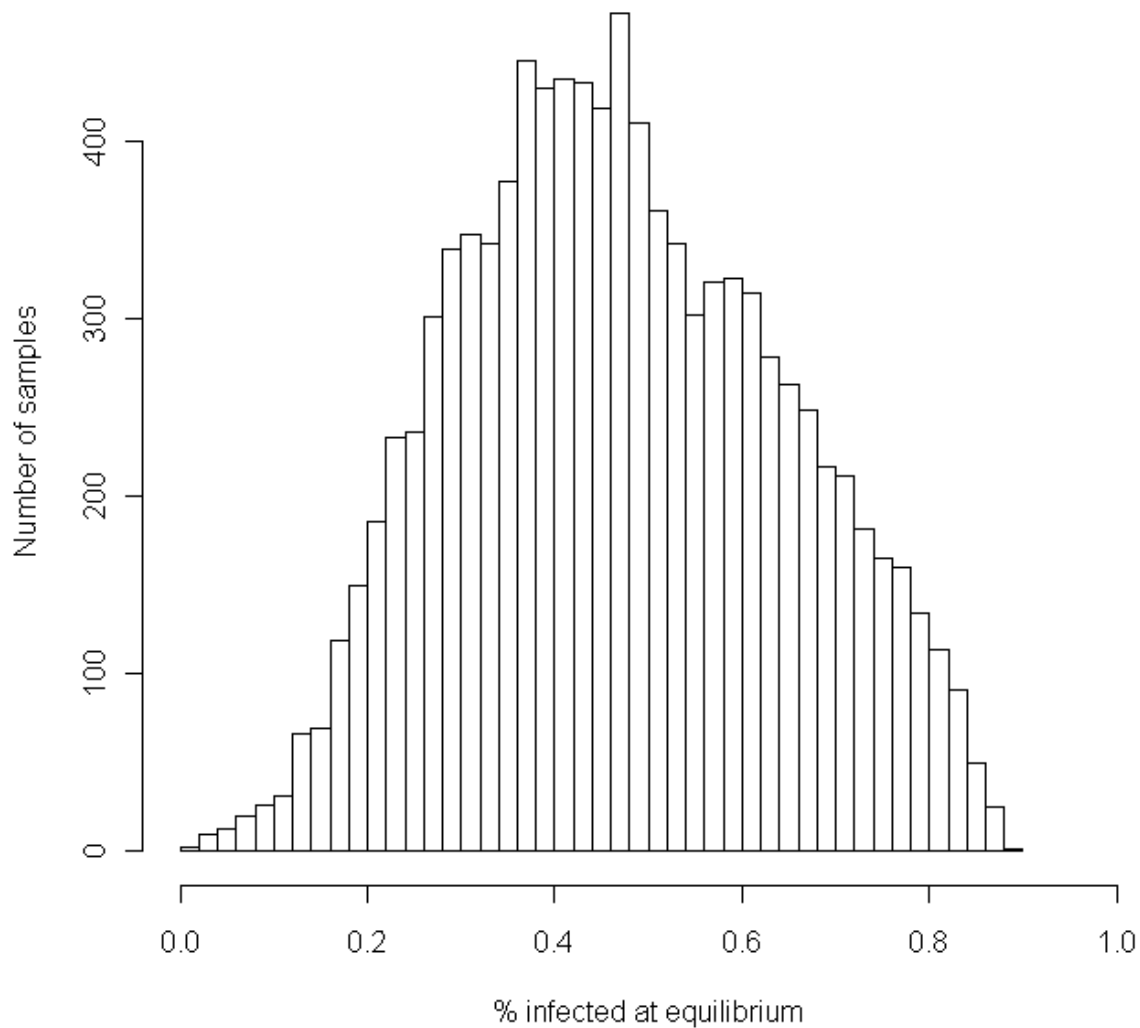


Figure 11. Histogram showing the distribution of outcome values across 10,000 samples of input parameters in the SIR model in Figure 9.

We can now take all the 10,000 combinations of input parameters and outcomes and fit a linear model through them:

$$y_i \sim c_1\beta_i + c_2r_i + c_3w_i + \varepsilon_i (i=1, \dots, 1000)$$

In the above regression equation, the dependent variables b_i , r_i and w_i are the 10,000 sampled values of the probability of transmission, rate of infection waning and rate of natural immunity waning respectively. The independent variable y_i is the infectious equilibrium. Notice that the model has no interaction terms between variables. This isn't realistic, but is helpful for the purposes of sensitivity analysis. We aren't trying to recreate the SIR model (it isn't even linear to start with), we are trying to represent it with a simpler model to capture the marginal effect of each input parameter on the outcome.

So let's start by trying to find the marginal effect of β . To do this, use the linear regression model, and assume that r and w are held constant and take their mean values. This is okay

because there are no interactions between variables in our linear model. It is obviously an approximation, but it allows us to get a better handle on the influence of β than varying β in the full model. When we run the regression we may end up with results such as the ones in Table 1 below:

Coefficient	Estimate	Standard Error	t-value	P-value
Intercept	0.486526	0.001859	261.7	<0.001
c_1	0.208670	0.002075	100.6	<0.001
c_2	-5.940816	0.020759	-286.2	<0.001
c_3	3.485241	0.020783	167.7	<0.001

Table 1. Regression coefficients for the linear model relating infectious equilibrium with input variables.

These describe the linear marginal relationship between each input variable and the outcomes. We could plot the straight line relationship between β (say) and the infectious equilibrium, together with the position of the $(\beta, \text{infectious equilibrium})$ pairs in our 10,000 samples. If we do that we will get a graph like Figure 12.

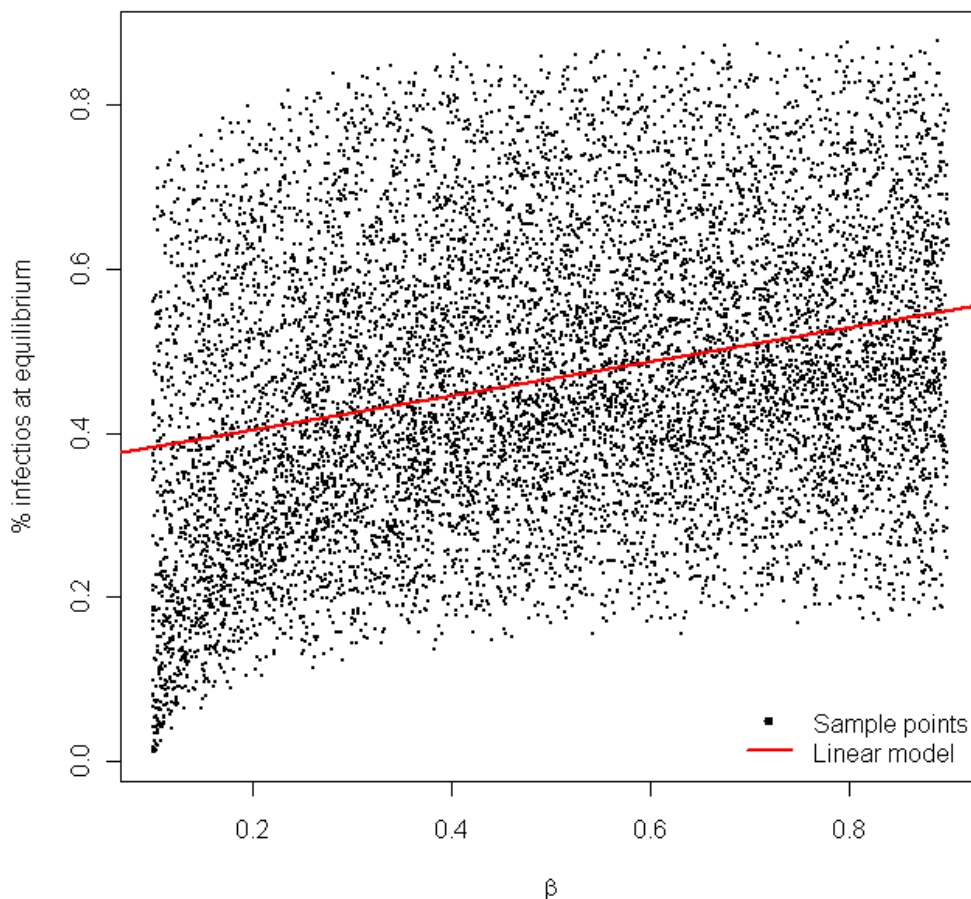


Figure 12. Values of β (probability of transmission) and the infectious equilibrium in the 10,000 samples, together with the best fitting relationship from a multivariable linear regression.

Similarly we can plot the relationship between r (rate of infection waning) and the infectious equilibrium (Figure 13). This time, the relationship is negative. This is as we might expect – when the duration of infectiousness is short (rate of losing infectiousness is high), the endemic equilibrium level of infectiousness is lower.

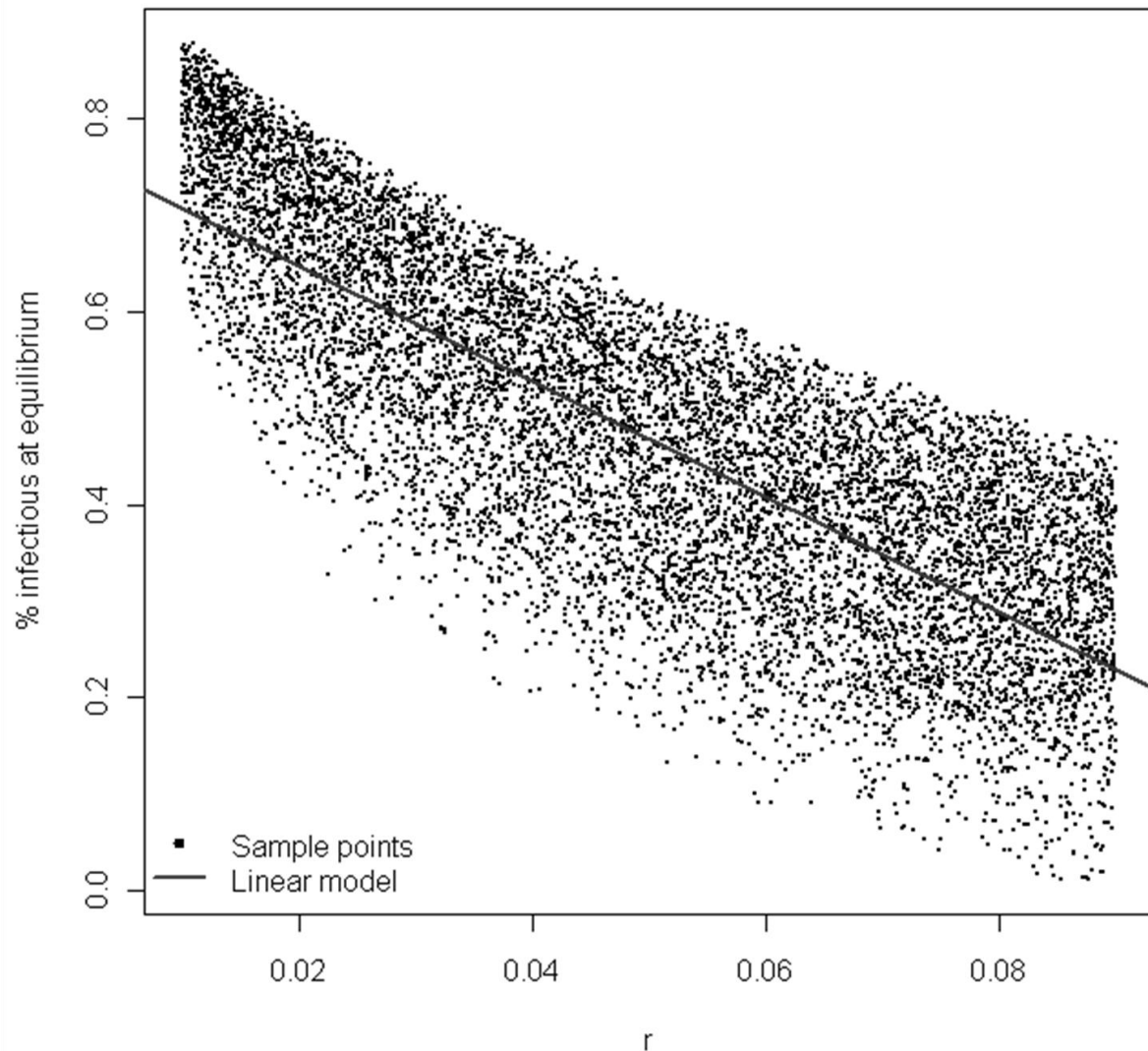


Figure 13. Values of r (rate of infection waning) and the infectious equilibrium in the 10,000 samples, together with the best fitting relationship from a multivariable linear regression.

We can make probability statements about outcomes from the results of a probabilistic sensitivity analysis. For instance, suppose for 2.5% of simulations the outcome Y is below Y_1 and for 2.5% it is above Y_2 . Then (Y_1, Y_2) is a 95% uncertainty interval for Y (see Figure 14). This depends on the probability distributions chosen for the parameters.

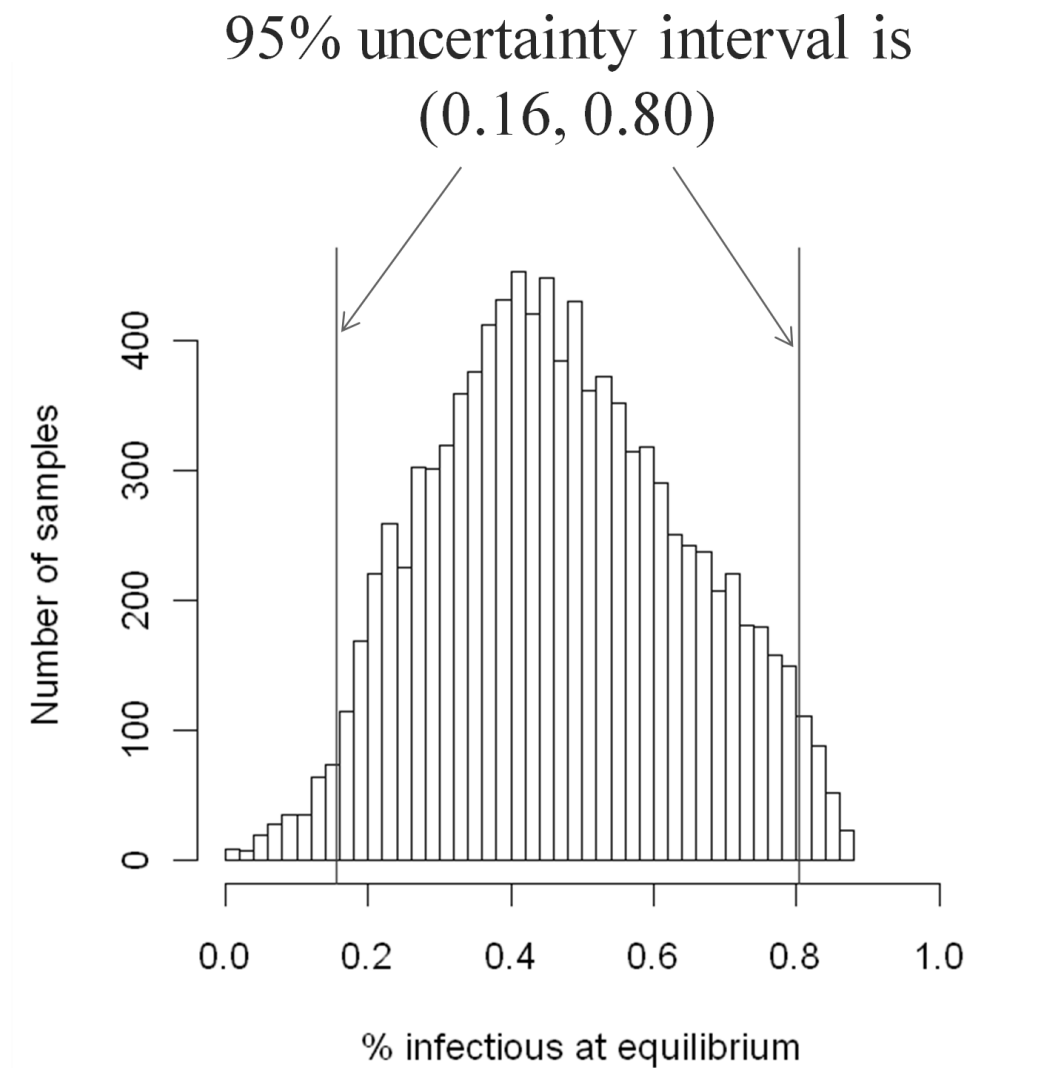


Figure 14. Distribution of the outcome (infectious equilibrium) together with its 95% uncertainty interval.

Other sources of uncertainty

So far, we've only considered sensitivity of results due to input parameters. However, you may also need to consider how uncertainty in the outcome of a model can stem from other sources of uncertainty in the model, such as:

- **Model structure.** For example, you could build SIS, SIR and SIRS models and see how outcomes change across different structural choices.
- **Type of model.** You may need to consider how outcomes may change if for example you used a stochastic model instead of a deterministic model.
- **Initial conditions.** This should always be checked, but is unlikely to be important for deterministic models that are run to endemic equilibrium. It is more important for stochastic low-prevalence models and models of an unfolding epidemic.

We don't have time to go into these topics in any more detail, but the following paper has more information if you are interested:

Bilcke J *et al.* Accounting for Methodological, Structural, and Parameter Uncertainty in Decision-Analytic Models: A Practical Guide. *Med Decis Making* 2011; 31(4): 675-92.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 21: An introduction to Phylodynamics

Lecture

Objectives

By the end of this lecture, you should:

- Understand how transmission patterns are inferred from pathogen gene sequences
- Know how to read and interpret a phylogenetic tree
- Understand how the timing, spatial dispersal and population dynamics of an epidemic are modelled from genetic information

Introduction – The Molecular Epidemiology of Infectious Diseases

Most epidemiological studies rely on information collected from patients or clinicians. An alternative approach consists in extracting information from nucleotide or amino acid sequences of pathogens sampled from an infected population. Since the diversity and rapid evolution of pathogens is measurable in near real time, it is possible to use genetic markers, such as particular mutations, to establish epidemiological linkage between sampled individuals.

This approach, termed **molecular epidemiology**, involves delineating the phylogenetic relationship between pathogen through gene sequence similarity, and using these relationships to study transmission pathways. Molecular epidemiology studies are frequently conducted in order to:

- Shed light on the origins of an epidemic (e.g. HIV¹, SARS², P. falciparum³)
- Monitor the spread of specific genetic markers over time (e.g. drug resistance⁴)
- Characterise the patterns and correlates of a pathogen's transmission⁵
- Assess the reliability of patient-derived information⁶

The starting point of all molecular epidemiological studies is a **phylogenetic tree**.

1 - Phylogenetic Inference

The fast evolution of pathogens produces genetic changes that are observable on a human time scale. The accumulation of these genetic 'footprints' occurs at a near constant pace (see Section 3), which means that the degree of relatedness of two or more pathogenic isolates can be inferred from their **genetic distance**.

By comparing the gene sequences of a set of infectious organisms, one can therefore reconstruct their evolutionary histories and establish their relative relatedness. A phylogenetic tree, or **phylogeny**, is the graphic representation of these evolutionary relationships (like a genealogy or a pedigree).

Phylogenetic trees

A phylogenetic tree is made of **branches** and **nodes** (**Figure 1**). The units one is comparing (here, molecular sequences) are called **taxa** (plural for taxon). Taxa are located at the tip of the tree, or external branches. Related taxa are linked by a node, which corresponds to their most recent common ancestor. The length of a branch represents the genetic distance between two nodes or between a node and a taxon, i.e. the number of mutations accumulated since they evolved away one from another. Two or more sequences descending from a node form a **clade** (or cluster). The length of horizontal branches is usually expressed as the number of nucleotide substitutions per site. Branches may be labelled with a numerical value indicating their reliability. These **branch support** values are derived from statistical tests of confidence, such as bootstrap resampling, and are usually expressed as percentage confidence. A tree sometimes has a **root**. It corresponds to the position of the most recent common ancestor of all the taxa.

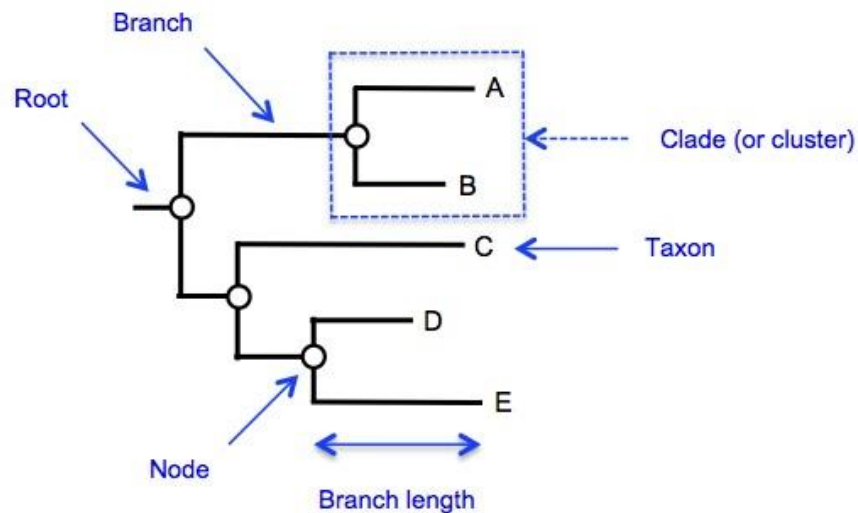


Figure 1. Components of a phylogenetic tree

To infer the phylogenetic relationship of a set of molecular sequences, these need to be 'aligned'. That is, nucleotide positions need to be arranged in columns to ensure that we compare **homologous** positions one to another. In evolutionary biology, homology means similarity due to descent from a common ancestor. Two genes are homologous if they descend from an ancestral gene. Likewise, two nucleotides in different sequences are homologous if they correspond to the same nucleotide position in the ancestral gene.

Many methods for reconstructing phylogenies have been developed. The most popular ones broadly fall into two categories: (i) **distance-based methods** and (ii) **character-based methods**.

(i) Distance based methods involve the calculation of a matrix of pairwise genetic distances. The matrix is then used to build the tree in a step-wise manner, grouping sequences (or groups of sequences) in order of increasing dissimilarity. A good example of a distance-based method is **neighbor-joining**, developed by Satou and Nei in 1987⁷.

(ii) Character-based methods infer trees directly from the genetic patterns seen in the sequence alignment. Given a set of sequences, (almost) all possible trees are considered. The idea is then to identify the tree (or trees) that maximises the probability of observing the sequences in the alignment, given a model of evolution (see below). This search for the 'best' tree can be done using a likelihood (e.g. maximum likelihood method⁸) or Bayesian Markov chain Monte Carlo⁹ statistical framework.

Models of evolution

Both method categories require an explicit **model of nucleotide substitution**. This is a probabilistic model that describes the relative rates of substitution between discrete nucleotide characters (i.e. A, C, G or T). These models are used during the calculation of the likelihood of a tree in order to estimate the difference between observed and expected evolutionary distances. This is necessary because multiple substitutions may occur at the same nucleotide position, resulting in an underestimation of pairwise genetic distances if not corrected for. Many models of nucleotide substitutions have been developed, amongst which the Jukes & Cantor, Kimura 2 parameters, Hasegawa-Kishino-Yano (HKY) or General Time Reversible models¹⁰.

Substitution models make various assumptions about the biological processes underlying molecular evolution, with varying accuracy and complexity. It is customary to test a set of models against a given sequence alignment and select the model that fit the data best prior to the phylogenetic reconstruction. Model selection is traditionally preformed through the implementation of statistical tests such as the likelihood ratio test, Akaike Information Criterion (AIC) or Bayes factor calculation^{10,11}, depending on the framework used.

2 - Phylogenetic Inference of Pathogen Transmission

Phylogenetics is a powerful and well-established approach to trace transmission events in an infected population. Pathogens acquire significant amounts of genetic diversity during replication, and a fraction of the genetic changes acquired in one host will be fixed and passed on to the next host. Thus, the level of genetic similarity between pathogens correlated with the likelihood of a common source of infection. One can therefore use these genetic markers to trace a lineage in a larger population (represented by a phylogenetic tree), hence establishing **epidemiological linkage** between the patients represented by pathogen sequences in the tree.

The identification of discrete transmission chains within a pathogen phylogeny involves the identification of clusters, or sub-trees, fulfilling criteria empirically determined so as to represent recent transmission. Although a variety of criteria are used to identify **transmission clusters**, they usually include (i) a minimum number of clustered sequences (e.g. 2 or more), (ii) minimal intra-cluster genetic differences (e.g. ≤ 0.045 substitutions per sites) and/or (iii) strong support for the branch leading to the most recent common ancestor of a cluster (e.g. $\geq 90\%$ confidence). This method has been empirically validated in 'proof of concept' studies comparing viral sequences from known partners, i.e. individuals who reported to have infected one another^{5,12}.

The phylogenetic reconstruction of transmission chains is increasingly used in epidemiological studies where pathogen genotypes are available¹³. Transmission clusters usually agree with patient-derived contact information, as was the case, for instance, for the Singapore 2003 SARS outbreak^{13,14}. Phylogenetic inference is often applied to the identification of transmission networks

when no other source of information is available. This is sometimes done for forensic purposes, despite the controversial nature of this practice. In this case, virological evidence obtained from the defendant and complainant is used to establish criminal liability that the defendant caused the complainant's infection. The first known case of criminalised transmission of HIV, the famous Florida dentist case, is a textbook example and fuelled a huge controversy at the time¹⁵.

3 - The Molecular Clock

An important dimension of pathogen transmission is time. A pathogen's transmission frequency, outbreak duration or time of introduction in a population are factors of great epidemiological relevance. The time frame within which transmission events occur can be determined from pathogen genetic data under the assumption of a **molecular clock**.

Since its proposal in the 1960s¹⁶, the molecular clock has become an essential tool of evolutionary biology. The molecular clock hypothesis states that nucleotide and amino-acid sequences evolve at a rate that is relatively constant over time, thus defining a relationship between genetic distance and time. Since observations suggest that evolutionary rates can vary across species, genes or even within lineages, several models have been developed to 'relax' this assumption¹⁷.

The '**strict**' (or constant) clock model assumes that all lineages in a tree evolve at the same rate. In order to estimate a constant rate of evolution, one can measure the genetic distance between species sampled at known time points (represented by the distance between a tip and the root of a phylogeny) and plot these measures against the corresponding sampling dates. This approach is called **root-to-tip linear regression**. The slope of the regression line corresponds to the strict rate of substitutions per unit of time. The estimated rate can in turn be used to 'calibrate' a phylogeny and convert genetic distance (i.e. branch lengths) into time units. This method runs into statistical problems if non-independent comparisons are used in the regression.

More complex models allow evolutionary rates to vary through time or among lineages, resulting in variation around an average rate. These are called '**relaxed**' molecular clock models. Several rate-variable molecular clock models have been developed and applied to pathogen sequences. Some methods use maximum likelihood to optimize the substitution rate over a phylogeny of sequences with known isolation dates¹⁸. A likelihood ratio test is then used to compare the fit of a single substitution rate with a multiple-rate model. Similar methods use Bayesian statistics to select the most likely parameters of molecular evolution for a set of viral sequences, including the substitution rates and timing of ancestral nodes¹⁹. Note that the reliability of molecular dating depends on the accuracy with which genetic distance is estimated (dependent on a model of nucleotide substitution), and on the appropriateness of the calibration rate (dependent on a molecular clock model).

Molecular dating was recently used to investigate the origin of the Guinea 2014 Ebolavirus outbreak²⁰. This study suggested that the viruses responsible for the Sierra Leone (SL) and Guinea (GN) epidemic were the result of a single introduction from the natural reservoir of Ebola virus. It further suggested that the founder virus diverged from Central African Variants in 2004 and crossed from GN to SL in May 2014. Molecular dating is also used to estimate the tempo of transmission of a pathogen in different contexts or risk groups, like the transmission of HIV between men who have sex with men or heterosexuals^{21,22}.

4 - Population dynamics

Another landmark of phylodynamics is the quantification of **demographic changes**. By observing sequence variation in pathogens sampled from a population, one can infer the past history of that population from the relationship that exists between genetic diversity and population size. As for molecular dating, various models have been developed to infer population dynamics. The most popular of these models are derived from Kingman's **coalescent theory**^{23,24}.

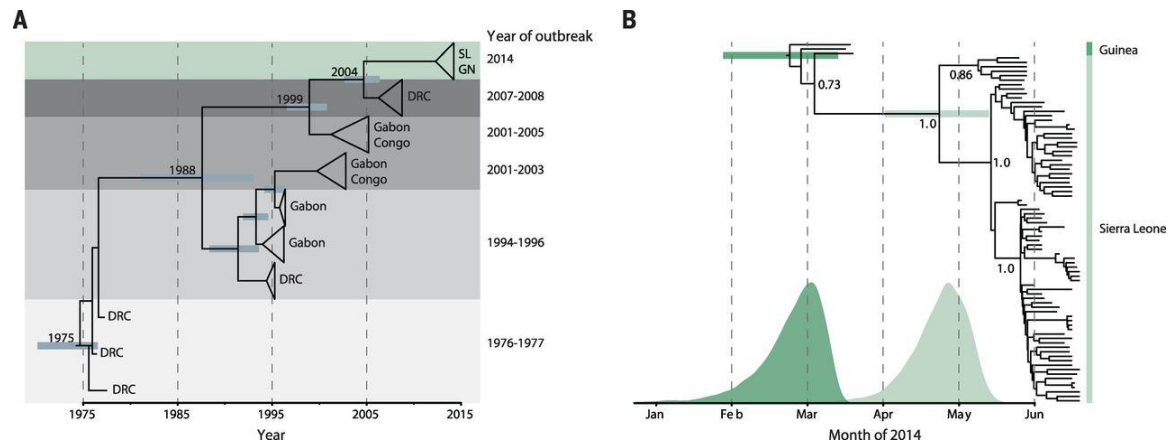


Figure 2. Molecular dating of the 2014 outbreak. (A) Separation of the 2014 lineage from central African lineages [SL, Sierra Leone; GN, Guinea; DRC, Democratic Republic of Congo; time of most recent common ancestor (tMRCA), September 2004; 95% highest posterior density (HPD), October 2002 to May 2006]. (B) Dating of the most recent common ancestor of the 2014 West African outbreak (23 February; 95% HPD, 27 January to 14 March) and of the Sierra Leone lineages (23 April; 95% HPD, 2 April to 13 May). Posterior support for major nodes is shown. Adapted from Gire *et al.* Science 2014

The coalescent theory describes a stochastic process where historical states of a population are inferred from the genealogy of individuals randomly sampled from it. The idea behind this theory is that, in the absence of selection, sampled lineages are assumed to randomly 'coalesce' (i.e. to merge into a common ancestor) as we go back in time (**Figure 3A**). In the case of a pathogen, the reconstructed phylogeny can be interpreted as a proxy for an incomplete transmission tree, where internal nodes represent coalescent points (**Figure 3B**). The rate at which lineages coalesce depends on how many lineages are coalescing (the more lineages, the faster the rate), and on the effective size of the population (the more parents to choose from, the slower the rate). The **effective population size** is the number of individuals in a (census) population who contribute offspring to the next generation.

If we consider a sample of n gene sequences from a total population of N , we can reconstruct the genealogy of these sequences (**Figure 3**). Looking backward in time, the number of ancestral sequences decreases as the lineages coalesce, until all lineages coalesce into the most recent common ancestor (the root) of the sample. Through this process, the probability of coalescence at the previous generation (i.e. the probability that two sequences in the current generation share a single ancestor in the previous generation) is $1/(2N)$, where N is the effective population size. The probability that coalescence occurred $t + 1$ generations ago is given by the distribution $1/2N (1 - 1/2N)^t$. If we assume that the number of mutations that occurred on a sequence in a given period of time is a Poisson variable, the mean time of $2N$ generations separating the two sequences implies that the mean number of mutations in the two sequences is $\theta = 4N\mu$, where μ

is the mutation rate per sequence per generation.

In the same way that phylogenetic inference is contingent to the modelling of molecular evolution, the coalescent is highly dependent on the assumption of a **demographic model**, i.e. a mathematical function describing the evolution of the size of a population over time (e.g. constant growth, exponential growth, piecewise expansion). The choice of a model will determine the set of parameters needed to estimate in order to accurately reconstruct the population history of the sampled lineages. These demographic models are hierarchically nested, allowing the performance of statistical tests to select the best fit for a given dataset.

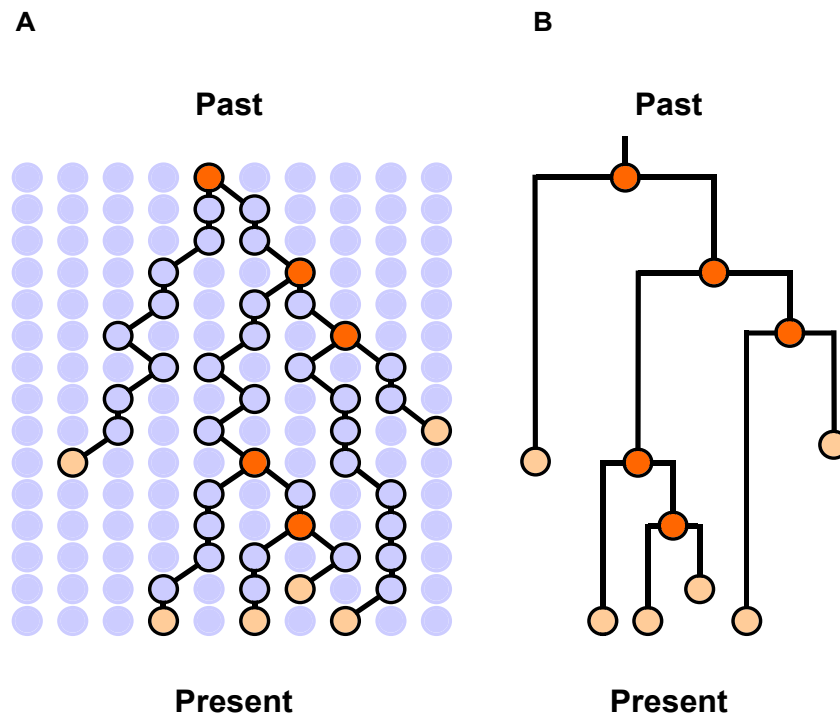


Figure 3. Principle of the coalescent. (A) Complete genealogy of a population of 6 individuals (yellow dots), over 15 generations (rows). (B) Corresponding phylogeny where yellow and red dots represent sampled individuals and hypothetical common ancestors, respectively.

The joint inference of the phylogeny and population dynamics of a population is frequently applied to the analysis of infectious disease outbreaks. Examples include the study of influenza A seasonality ²⁵, of the early epidemic growth of HIV-1 in the US or the UK ^{26,27}, or of the iatrogenic transmission of Hepatitis C virus in Egypt ²⁸.

5 - Phylogeography

The **geographical spread** a pathogen during an epidemic is also inferable from phylogenetic trees annotated with sampling locations. When individuals are infected in one location then move to another, or when they infect someone whilst travelling, this is apparent as a “change” in the location ascribed to one branch of the tree. These changes in location along a phylogenetic tree are determined from the location values at the tips and the shape of the tree (see **Figure 4**).

The simplest phylogeographic reconstruction methods use a **maximum parsimony** framework,

which infers the minimum set of migration events required to explain the observed phylogeny. Maximum parsimony is related to the principle formulated by William of Ockham in the 13th century, the so called Ockham's razor, according to which the simplest explanation for a phenomenon is to be preferred over a more complex (and often less likely) one.

Probabilistic, and in particular Bayesian, statistical frameworks have recently been developed for phylogeographic inference. These allow flexibility in hypothesis testing and the integration of epidemiological information. In this case, location exchange processes are modelled using continuous Markov chains. All possible transitions from a location state to another are inferred, fitted to the data, and the most likely location of the internal nodes of a tree, given the observed locations at the tips, is estimated together with its probability²⁹.

In these models, location states are either discretely or continuously distributed. The choice of the approach depends on whether the sampling scheme is amenable to discretization or not. For instance, if only the country or town of sampling is known, a **discrete diffusion** model may be preferred. If sequences are drawn from unique locations that are continuously distributed over a two-dimensional geographic area (i.e. like postcodes or GPS coordinates) **continuous diffusion** models can be used. Continuous diffusion approaches are based on Brownian diffusion models and account for variability on the branch dispersal rates. Diffusion models can also be **symmetric** (the rate of migration from A to B equals the rate of migration from B to A) or **asymmetric**.

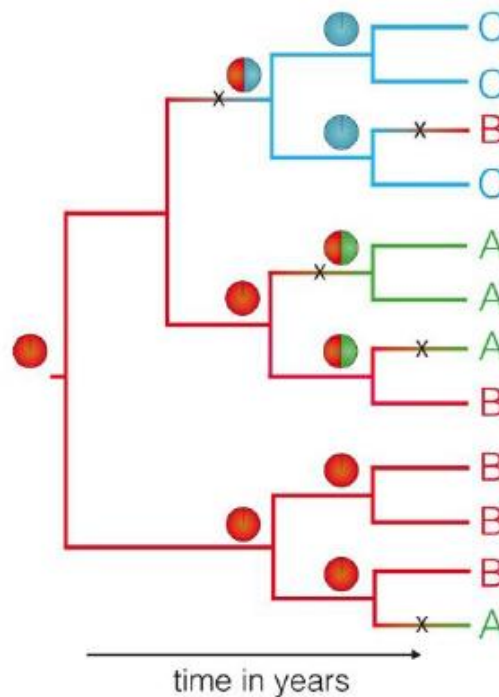


Figure 4. Principle of phylogeographic inference. Traits A, B and C at the tips of the phylogeny represent geographic locations from which genetic sequence data was collected. Crosses on the branches represent estimated changes in location. The color-coded pie charts represent posterior probability support for the location estimates. Adapted from Faria *et al.* 2014²⁹.

References

1. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441 (1999).
2. Ruan, Y. J. *et al.* Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* **361**, 1779–1785 (2003).
3. Liu, W. *et al.* Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**, 420–425 (2010).
4. Brenner, B. G. *et al.* Transmission networks of drug resistance acquired in primary/early stage HIV infection. *AIDS Lond. Engl.* **22**, 2509–2515 (2008).
5. Pao, D. *et al.* Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS Lond. Engl.* **19**, 85–90 (2005).
6. Hué, S. *et al.* Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS Lond. Engl.* **28**, 1967–1975 (2014).
7. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
8. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
9. Shoemaker, J. S., Painter, I. S. & Weir, B. S. Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet.* **15**, 354–358 (1999).
10. Whelan, S., Liò, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**, 262–272 (2001).
11. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013).
12. Edo-Matas, D. *et al.* The evolution of human immunodeficiency virus type-1 (HIV-1) envelope molecular properties and coreceptor use at all stages of infection in an HIV-1 donor–recipient pair. *Virology* **422**, 70–80 (2012).
13. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput Biol* **10**, e1003457 (2014).
14. Vega, V. B. *et al.* Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infect. Dis.* **4**, 32 (2004).
15. Ou, C. Y. *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**, 1165–1171 (1992).
16. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
17. Bromham, L. & Penny, D. The modern molecular clock. *Nat. Rev. Genet.* **4**, 216–224 (2003).
18. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399 (2000).
19. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* **4**, e88 (2006).
20. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
21. Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A. & Leigh Brown, A. J. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* **5**, e50 (2008).
22. Hughes, G. J. *et al.* Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog.* **5**, e1000590 (2009).
23. Kingman, J. F. C. The coalescent. *Stoch. Process. Their Appl.* **13**, 235–248 (1982).
24. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
25. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619 (2008).
26. Robbins, K. E. *et al.* U.S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* **77**, 6359–6366 (2003).
27. Hué, S., Pillay, D., Clewley, J. P. & Pybus, O. G. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4425–4429 (2005).

28. Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. & Rambaut, A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* **20**, 381–387 (2003).
29. Faria, N. R., Suchard, M. A., Rambaut, A. & Lemey, P. Toward a quantitative understanding of viral phylogeography. *Curr. Opin. Virol.* **1**, 423–429 (2011).

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 25: Real-time modelling: an introduction Lecture

Objectives

By the end of this lecture, you should:

- Understand what is a real-time model;
- Understand why real-time models are needed;
- Have an idea of some existing methods;
- Know the issues associated with real-time modelling;

Definition of a real-time model

A **real-time model** is a mathematical model that has been updated, or fitted, or calibrated, to an ongoing epidemic. Real-time models usually share several characteristics; they incorporate knowledge from previous experience (prior knowledge), they use statistical fitting and analysis and epidemic modelling to reflect what we know about the current epidemic. Bayesian inference methods are best suited for such a task as they use Bayes' rule to update the probability estimates of the model from the prior as additional evidence is obtained, thus synthesizing at each step the available information and its uncertainty.

If done properly, a real-time model should give more valid projections about the future course of the epidemic as it is happening than a model which has not been developed in real-time.

Why use models in real-time?

Many different institutions (e.g. country governments, local authorities, agencies or hospitals) plan for emergencies. These plans are based on models which describe scenarios presented to the planners. These scenarios are based on past experience (which may be quite limited) and assumptions.

The main dangers facing pre-emptive planning are that institutions may over-react to a mild epidemic or under-react to a severe one. Over-reacting may appear to be the safest option. However, besides wasting money, it can deplete public confidence in public health interventions and thus jeopardize future interventions, should more severe epidemics occur.

Plans are often associated with a number of trigger points with important public health and/or economical consequences. Hospitals can, for example, opt to delay elective surgery if an increasing need of intensive care beds is planned, a region can ban cattle movements if it is predicted to stop the progression of an outbreak or, even more radically, impose measures such as contiguous culling. Real-time analysis and prediction can help refine the plans.

When designing a real-time model, the current situation first needs to be addressed. During the last pandemic of influenza A/H1N1 in 2009, people started to use a new term, namely "nowcasting", as opposed to forecasting. Nowcasting tries to synthesise information from each of the available sources of data to estimate the progress of the current epidemic. Sources of data are numerous and none of them gives a complete, up-to-date unbiased picture of epidemic. Example of common sources used in epidemiological surveillance are

lab-confirmed cases, suspected cases, GP consultations and deaths. All these are partial, biased and subject to delay.

Modellers will face two types of questions, a first set of critical questions related to the revision of plans and a second set of what-if questions used for advising policies. Possible "critical" questions would be about the number of cases, deaths, hospitalisations, or ICU admissions, their time period, the timing of their peak (not necessarily happening at the same time due to delays) and the situation at this peak (peak demand). Possible "what-if" questions relate to possible interventions such as school closure, restriction of treatments (supply and capacity issues), vaccination or culling of animals within an exclusion zone.

A brief description of some common methods used in real-time

The set of methods used in real-time modelling is closely linked to the phases of the surveillance of the epidemic. These phases are the result of the constraints from the surveillance systems in place. Phase 1, the early phase, is defined by case-based reporting (usually laboratory confirmed). This phase allows estimation of key epidemiological quantities such as R_0 , the serial interval, etc. During this phase, key indicators of severity, such as the case fatality rates/ratio (CFR) or hospitalisation rate/ratios will first be assessed to get an idea of the likely burden associated with the pathogen. Due to the exponential growth in the incidence during an epidemic, the epidemic eventually reaches a second phase (the generalised epidemic) where only aggregate reporting (not confirmed) is available. Methods are necessary even at this stage to refine estimates (e.g. tracking the reproduction number or correct/adjust CFRs and hospitalisation rates). Understanding the current state of the epidemic during this phase is critical to projecting the future course of the epidemic.

The main method that is used to estimate the reproduction number during the early phase has been designed by Wallinga & Teunis 2004 [1] for the epidemic of SARS. It allows tracking the reproduction number in real time to see if it is changing and analyse the factors which affect it. The method, which will be described in more detail in the next lecture, is based on the reconstruction of possible infection trees (tree linking cases with their infector and their infectees). A likelihood is attached to each of the possible trees, based on the probability of the infection events derived from the serial interval distribution.

During the generalised epidemic, we are interested in keeping track of the reproduction number and to assess when the epidemic will peak, how high the peak will be and ultimately, the number of cases in the entire epidemic. Very simple methods can be used to track the growth rates of the epidemic as the early phase of the epidemic grows exponentially. Simple calculations shows that if serial interval is not changing, the growth rate is related to the reproduction number (see [2] or next lecture for more details). Any change in growth rate might indicate that the epidemic is peaking.

Modellers may also fit transmission models (e.g. SIR) to the data. Usually these data sources (e.g. GP consultations, deaths, number of farms infected) include non-specific (unconfirmed) diagnoses and will include a background level of diseases. Uncertainty reduces as the epidemic progresses [3].

Key issues related to real-time analysis

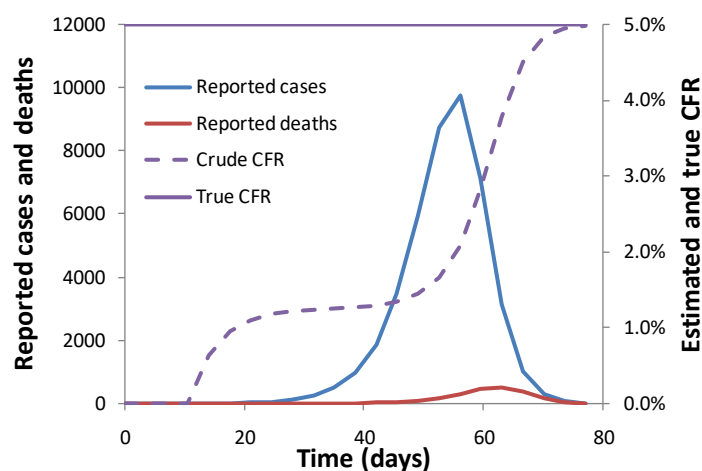
As compared with the traditional situation, modelling an epidemic in real-time presents several issues which need to be considered.

Problem with rates

Most problems concern the definition of the numerators. If severe outcomes (the numerator) are typically easier to detect than are mild outcomes, the total number of cases (the denominator) can be difficult to track. This problem can be solved by observing a well-defined population, although it can be difficult to study a population which is sufficiently large and representative, especially if the investigated outcomes are rare.

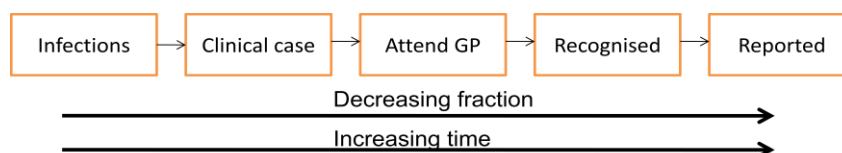
Delays

In real-time, the epidemic pattern that will eventually be seen in the data is only just beginning to emerge. It is usually unclear as to whether all the cases that are going to arise have occurred, or if they have all been reported? Due to right censorship (i.e. the most recent cases are not included as they have not yet been reported or confirmed), R_n appears to be decreasing. There can be numerous delays between events occurring and them being reflected in the data; which is especially the case for mortality data, with delays occurring between onset and death (survivorship data are usually censored). We need to account for this to avoid biased estimates of CFR. Parametric and non-parametric methods exist which can account for these delays, such as Kaplan-Meier analyses, or it is possible to jointly estimate CDF for time to discharge & death and CFR. Each of these methods (including crude estimates) should give same CFR estimate at the end of the epidemic, but significant differences exist in early stages. Additional delay to registration of death can be particularly long for some age groups, which can change during the epidemic and delays are often long in relation to the doubling time of epidemic. As a consequence, the unadjusted CFR would be expected to change over the course of epidemic, and will be especially exaggerated if weekly data are used.



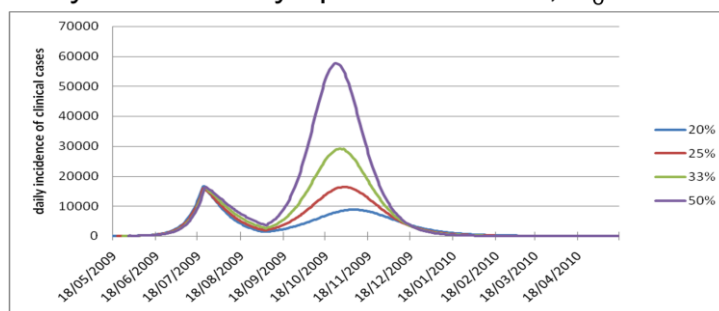
The problem of scale

The key questions that real-time modelling aims to address are: When will the epidemic peak? How high will the peak be? How many cases will there be in total? The way to answer these questions is to use an epidemic (SIR) model. In such a model, cases peak when the proportion of the population that is susceptible reach the threshold density, or the epidemic threshold (see the session on the natural dynamics in block 2), which depends on R_0 . The size of the epidemic is related to R_0 , so in principle, it is possible to predict the size of the peak and epidemic size if we know R_0 . To parameterise such a model, we need to track the depletion of susceptibles (i.e. infections) at the same time as the data (i.e. # cases, or GP consultations, or deaths) give us just a fraction of the infections. We can then identify the size of the peak and of the epidemic if we know what fraction of infections are reported by the surveillance system used.



As we saw in the session on the natural dynamics in block 1, the epidemic peaks because the size of susceptible population has been depleted sufficiently when $R_n = 1$. Thus, in order to predict when an epidemic will peak, we will need to know how much the susceptible population has been depleted. As mentioned earlier in these notes, due to under- and delays in reporting (or “scaling”) the number of susceptibles that have been removed from the population does not equal to the number of reported cases.

Real-time projections H1N1v Sept 2009,
by fraction of symptomatic cases, $R_0 = 1.45$



M Baguelin (unpublished)

Changing behaviour

We do not have a direct measure of the extent of morbidity in the community, but instead have a measure of attendance at health care facilities. This might change over time especially in case of heightened awareness and diagnosis or if people are discouraged from attending. This will give a distorted picture of the epidemic. Changes in behaviour can be detected only by monitoring individuals directly. Recently (launched in the UK in 2009), an internet surveillance tool called Flusurvey (internet based cohort www.flusurvey.org.uk) demonstrated changes in the use of health services during the 2009 pandemic. This type of direct surveillance tool can adjust for biases in surveillance and can record the impact of important clinical and public health measures in its own right (e.g. antiviral use).

Also, contact rates are not necessarily constant over the course of the epidemic. It may be incidental (e.g. School holidays), or a deliberate consequence of policy (e.g. school closure or travel restrictions). Individuals may choose to change their behaviour over the course of the epidemic by attempting to reduce contact or attempting to circumvent control measures (particularly veterinary).

Conclusions

Real-time analyses should mirror what may happen during the course of an epidemic reasonably closely, as the model is fitted to data while describing the progress of the epidemic to date with predictions. It can be used to update plans and monitor trigger points. It can also be used to assess control policies (e.g. vaccination [4]). It makes use of multiple sources of data but is fitted to a 'shadow' of the real epidemic. It may require novel sources of data (e.g. cohort-based data, serology) to be able to detect and correct for changing biases in observed data and needs to take account due to delays and incomplete and biased picture from surveillance.

References

1. Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American journal of epidemiology* 160: 509–516. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15353409>.
2. Wallinga J, Lipsitch M (2007) How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings Biological sciences / The Royal Society* 274: 599–604. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1766383&tool=pmcentrez&rendertype=abstract>. Accessed 1 August 2011.
3. Hall IM, Gani R, Hughes HE, Leach S (2007) Real-time epidemic forecasting for pandemic influenza. *Epidemiology and infection* 135: 372–385. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2870596&tool=pmcentrez&rendertype=abstract>. Accessed 4 June 2013.
4. Baguelin M, Hoek AJV, Jit M, Flasche S, White PJ, et al. (2010) Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine* 28: 2370–2384. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20096762>. Accessed 5 July 2011.