Block 2:

Analysis of seroprevalence data and their application...

Lecture notes

Introduction to Infectious Disease Modelling and its Applications – 2018

Overview of block 2

In block 1, we learnt about the basic methods for setting up models and you were introduced to some of the insights into the natural dynamics of infections which models can provide.

In this block, we discuss how some of the key input parameters into models, such as the basic reproduction number and the rate at which two specific individuals come into effective contact per unit time can be obtained. In addition, we discuss how the models we have considered so far can be extended to incorporate more realistic assumptions about contact i.e. non-random mixing, and the implications of non-random mixing for the impact of control strategies.

The sessions in this block follow the following sequence (see also Figure 1):

Sessions 8-9: Applying modelling techniques to analyse (seroprevalence) data
Lecture 8 provides an overview of the methods for analysing data on the past history of infection, which enable us to estimate the force of infection and basic reproduction number of an infection, which we can use in models. This is followed by lecture 9, which explains the basics of fitting models to data.

In the accompanying practical, we will fit a model to seroprevalence data for rubella to estimate the force of infection and basic reproduction number in different settings.

Session 10: Contrasting the effects of rubella vaccination between high and low transmission settings
This session consists of a practical, which illustrates how vaccination may affect the age distribution of individuals who are susceptible to infection and the implications for adverse impacts. This will use the estimates of R_0 calculated in the previous session in a model of the transmission dynamics of rubella (which still assumes that individuals mix randomly), and discuss how vaccination in childhood may affect the infection incidence among adults.

Session 11: Incorporating heterogeneous mixing to describe the transmission dynamics and control of infectious diseases
In this session, we learn about the methods for incorporating non-random mixing into models.
In the accompanying practical, we use the estimates of the force of infection for rubella obtained in practical 8/9 to estimate age-dependent contact parameters. These contact parameters are then used in a model to explore the effect of non-random mixing on the impact of a vaccination control programme.

Session 14: Estimating the basic reproduction number for heterogeneously mixing populations
In this session, we learn how the basic reproduction number can be calculated, taking account of non-random mixing between individuals.

In the accompanying practical, we use the contact parameters obtained using the force of infection calculated in session 8 to calculate the basic reproduction number and discuss how non-random mixing influences the herd immunity threshold.

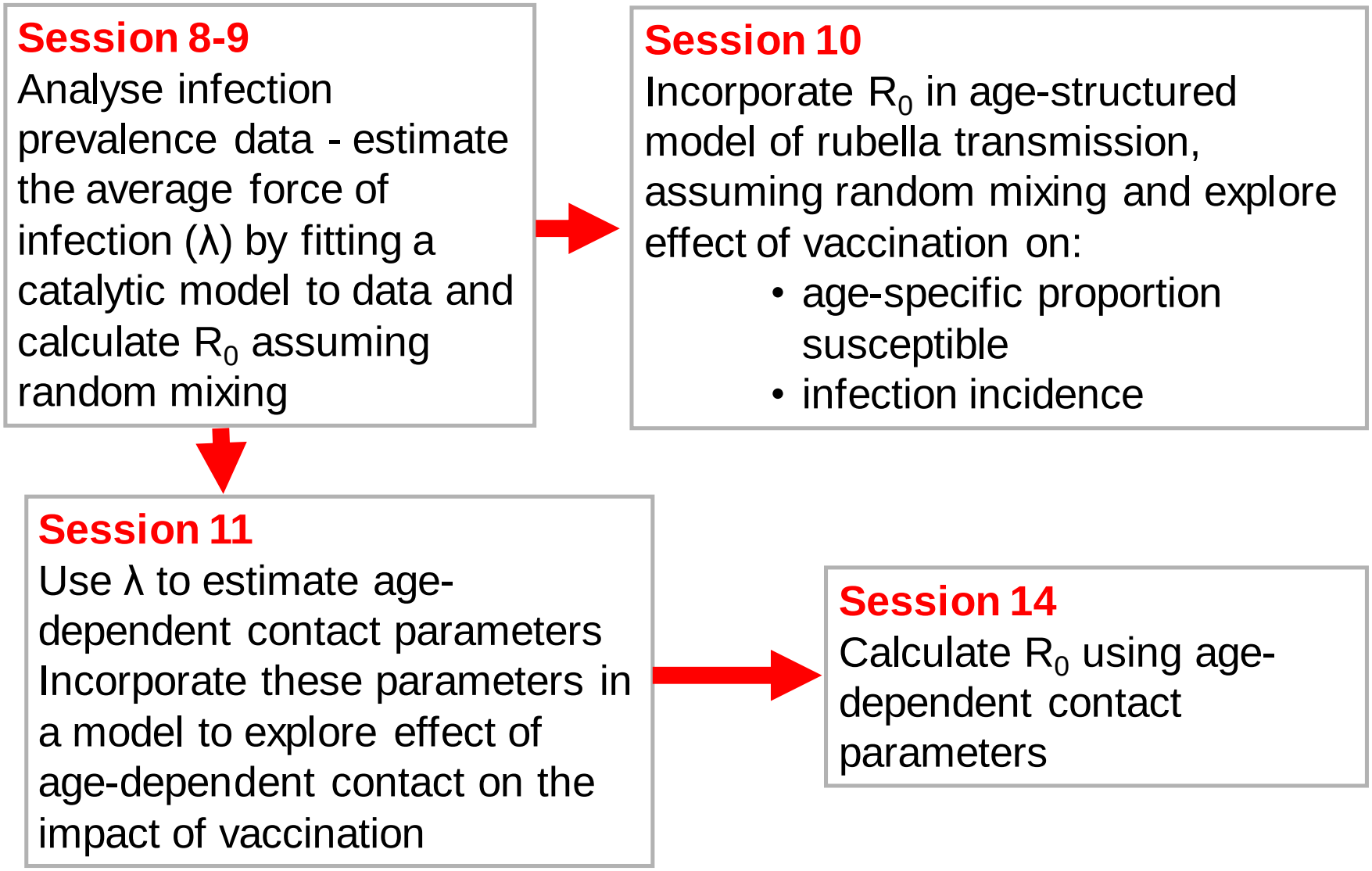


Figure 1: Summary of block 2

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 8: Applying modelling techniques to analyse (seroprevalence) data

Lecture

Overview and objectives

This session discusses how we might use data on the prevalence of previous infection (e.g. serological data) to estimate important epidemiological statistics (some of which are used in models), such as the average force of infection, the basic reproduction number, the average age at infection and the number of new infections which might occur in different age groups per unit time.

By the end of this session, you should:

- Understand the relationship between long-term data on the infection incidence and cross-sectional data on the age-specific prevalence of previous infection;
- Know how to use catalytic models to estimate the average force of infection from cross sectional data;
- Be able to use estimates of the average force of infection to calculate the proportion of the population that is susceptible, the basic reproduction number, the herd immunity threshold;
- Be able to use graphical and model-free methods to estimate the average force of infection and the extent to which it is age-dependent.

Introduction

Insights from serological data

So far on this course, we have developed a simple model of the transmission dynamics of an immunizing infection, which, after including births into and deaths out of the population, had the following structure:

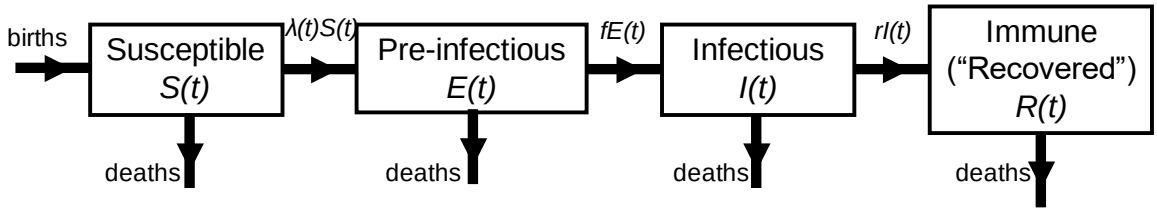


Figure 1: General structure of the model describing the transmission of an immunizing infection in a population, accounting for births and deaths, developed in block 1.

This model assumed that individuals mix randomly and required input on the following:

- the pre-infectious period;
- the infectious period;
- the birth and death rates;

- the rate at which two specific individuals come into effective contact per unit time (β).

The first four parameters in this list are usually known for most infections and populations. However, the parameter β is poorly understood and difficult to measure directly, and we have usually calculated it from the basic reproduction number, R_0, using the equation:

$$\beta = \frac{R_0}{ND}$$

Here, N is the total population size and D is the duration of infectiousness. During the exercises, we have usually provided the value for R_0 to use in models. How do we know the value for R_0?

We can often estimate R_0 for a given infection using data on the age-specific proportion of individuals who have previously experienced infection, such as those shown in Figure 2.

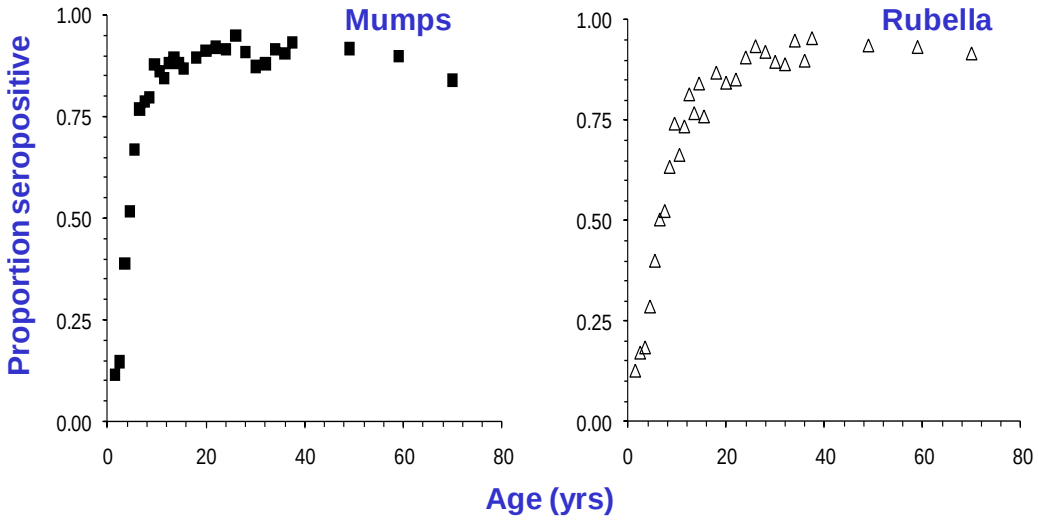


Figure 2: Observed proportion of individuals who were positive for mumps and rubella antibodies in the UK during the 1980s[1]. The data were collected from unvaccinated individuals.

These show that the proportion of individuals who had antibodies to mumps and rubella and who, in the absence of vaccination, had therefore probably been infected, increased with increasing age. In addition, the proportion of individuals who had antibodies to mumps increased more rapidly with age than that for rubella, suggesting that mumps was more infectious than was rubella. In fact, using data such as these, we can estimate several important epidemiological indices, in addition to the basic reproduction number, namely:

- the average force of infection;
- the average age at infection;
- the herd immunity threshold;
- the number of new infections which might occur in different age grops per unit time.

Before discussing the methods for analysing these data, we first review the relationship between these data and the predictions obtained from the dynamic model that we worked with in previous sessions.

The relationship between cross-sectional data on the age-specific proportion seropositive and long-term predictions from a dynamic model

Figure 3A shows predictions of the number of infectious individuals over time, obtained using the measles model developed in block 1. If we were to track individuals born in year 80 in the model, we would see that, as the cohort ages, the proportion of individuals who have ever been infected increases (Figure 3B). In fact, if the infection is endemic, the patterns shown in Figure 3B should be similar to those seen in cross-sectional data, such as those shown in Figure 2.

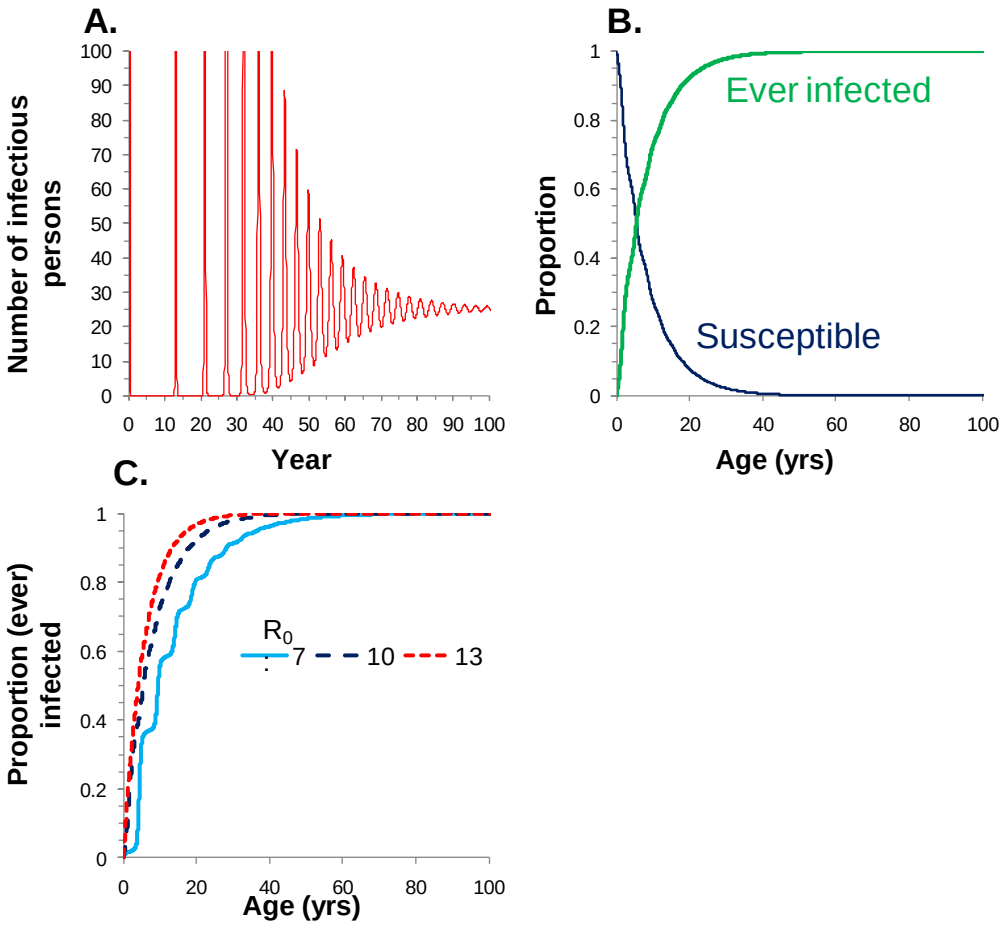


Figure 3: A. Predictions of the number of infectious persons obtained using the measles model developed in block 1, assuming that R_0=10. B and C. Predictions of the proportion of a cohort born in year 80 in the model discussed in Figure A who should have B) ever been infected or who are still susceptible by the time they reach different ages or C) ever been infected, assuming values for R_0 of 7-13. See file Cohort_transmission.mmd.

As might be expected, the proportion of individuals who have ever been infected by a given age (e.g. 10 years) decreases as R_0 (or the infectiousness of the agent) decreases (Figure 3C). The fact that the age-specific proportion ever infected is sensitive to the size of R_0 indicates that, given appropriate methods, we should be able to estimate R_0 from data such as those in Figure 2.

Data such as those in Figure 2 are typically analysed using so-called "catalytic models" to estimate the ***average force of infection***, which is then used to calculate the average age at infection, the proportion of the population that is susceptible, the basic reproduction number

and the herd immunity threshold and the number of new infections which might occur in different age groups per unit time.

Estimating and applying the average force of infection

Applying catalytic modelling techniques to estimate the average annual force of infection

As discussed in previous sessions, the force of infection at a given time t ($\lambda(t)$) is defined as the rate at which susceptible individuals are infected and can be calculated using the equation:

$$\lambda(t) = \beta I(t) \qquad \text{Equation 1}$$

Here, β is the rate at which two specific individuals come into effective contact per unit time and $I(t)$ is the number of infectious individuals at time t. In general for endemic immunizing infections, the force of infection changes over time, although its average value remains unchanged over time (Figure 4).

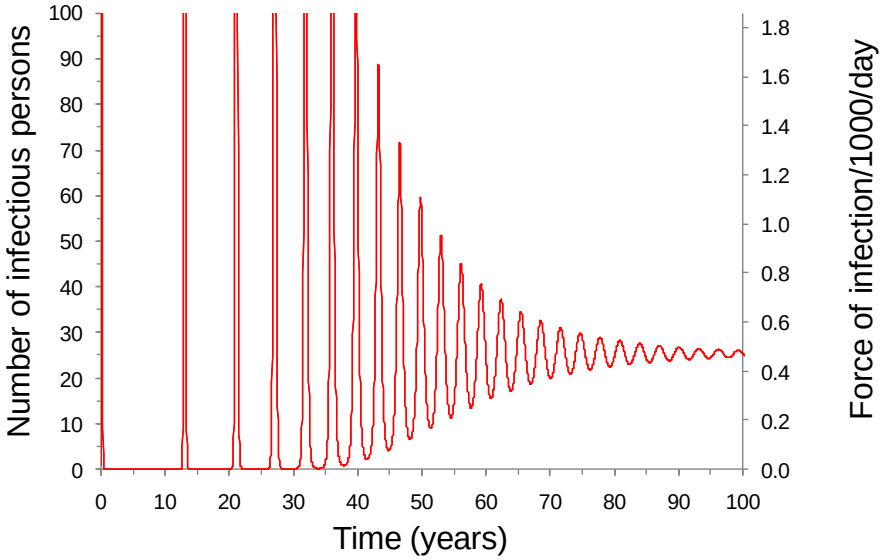
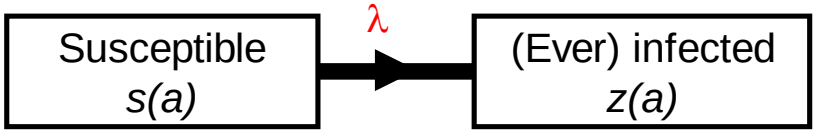


Figure 4: Comparison between predictions of the number of infectious persons and the daily force of infection, obtained using the model discussed in Figure 3. Here, the average force of infection is about 0.47 per 1000 per day, which is equivalent to an average annual force of infection of 17% per year.

The proportion of individuals who have ever been infected at different ages can often be described using the model with the following general structure:



This model is known as a simple catalytic model, and tracks individuals from birth, assuming that they are infected at some annual rate λ (i.e. the force of infection) which is independent of age and calendar year, and that those susceptible and infected have similar mortality

rates. The origin of the term "catalytic model" stems from Münch (1959)[2] (see Figure 5 and Hens et al (see recommended reading)).

The rates of change in the proportion of individuals who are susceptible (*s(a)*) and those ever infected (*z(a)*) with age are given by the equations:

$$\frac{ds(a)}{da} = -\lambda s(a)$$

$$\frac{dz(a)}{da} = \lambda s(a)$$

Assuming that all individuals are susceptible to infection at birth (forgetting about the contribution of maternal antibodies for now), then, as discussed in the session on differential equations, these equations can be solved to give the equation:

$$s(a) = e^{-\lambda a}$$

Assuming that the *proportion ever infected* by a given age *a* is given by 1-*proportion susceptible* at that age, we obtain the following equation:

$$z(a) = 1 - e^{-\lambda a}$$

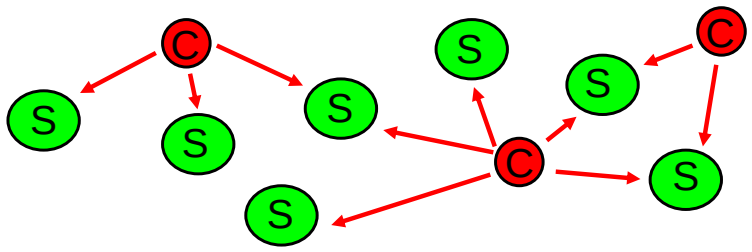


Figure 5: Illustration of the catalytic (or infection) process. In a "typical" catalytic process, the molecules of the original substance (denoted by the letter S) are exposed to the molecules of some catalyst (denoted by the letter C) which, in Munch's time, were thought to remain unchanged over time. Thus the molecules of the original substance are exposed to a constant force converting them into a new substance. By analogy, in an infection process, the circles labelled "S" are the susceptible individuals, and the circles labelled "C" are infectious persons ("cases"), who are converting susceptible individuals into those who have ever been infected. The analogy is imperfect, because, in contrast with the catalyst, the number of infectious persons (and thus the force of infection) changes over time.

Figure 6 shows the effect of different values for the force of infection on predictions obtained using the catalytic model of the age-specific proportions of individuals who are susceptible or who have ever been infected. Specifically if the force of infection is high (similar to that for rubella in high transmission settings [3], most individuals are infected by age 10 years. These values are similar to those predicted by the transmission model discussed above (see Figure 3). If it is 1% (similar to that for *M tuberculosis* in parts of Africa during the 1980s [4] less than 50% of adults are predicted to have been ever infected.

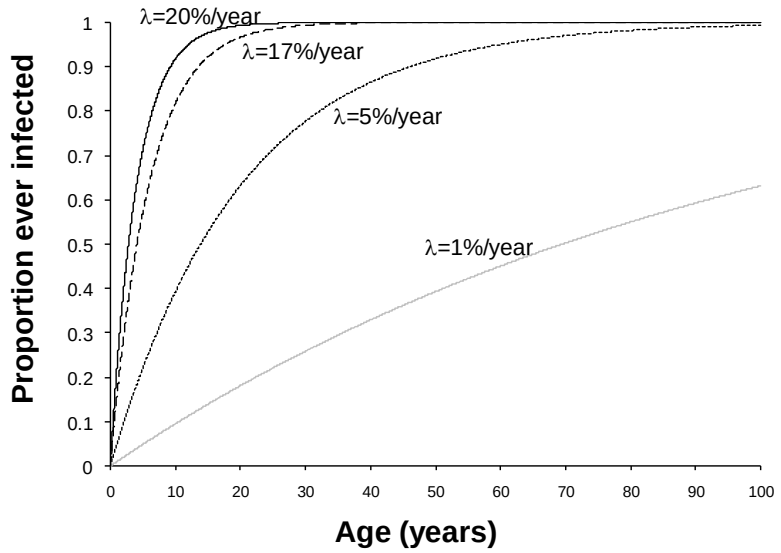


Figure 6: Comparison between estimates of the proportion of individuals who have ever been infected, assuming that the force of infection is the same for all age groups and is between 1% and 20% per year, as calculated using the equation $z(a) = 1 - e^{-\lambda a}$.

Question: Based on Figure 6 *and the data shown in* Figure 2*, what was likely to have been the average force of infection for mumps and rubella during the 1980s in the UK?*

Comparing predictions from the catalytic model against observed data by eye gives a reasonably good idea of what the average annual force of infection may have been. However, we can improve our estimates by formally "fitting" our model to the data, whereby the value for the force of infection is varied until the smallest "distance", as described by some "goodness of fit statistic" between the model prediction and the observed data is attained (Figure 7).

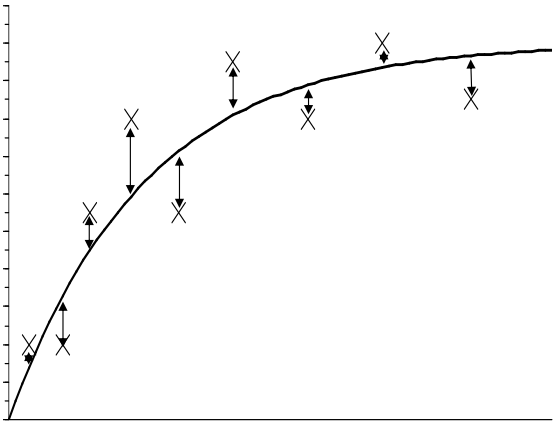


Figure 7: Illustration of the fitting process. Typically in this process, the value of a parameter is varied until the distance between the model prediction and the observed data (reflected by the vertical lines) is as small as possible.

There are many methods for fitting models to data, such as that of least squares or of maximum likelihood, although different methods will usually give similar parameter estimates. Mathematical model-building packages (e.g. Berkeley Madonna, ModelMaker)

usually provide tools for finding best-fitting parameters, although these should be applied cautiously as they can provide no information on either the goodness of fit or on the confidence intervals of the parameters. For a detailed discussion of the methods for fitting models, you should consult a statistics handbook[5]. These methods for fitting models are discussed later in the course and the method of maximum likelihood will be used during the practical.

Figure 8 compares the predictions based on the best-fitting estimates of the force of infection for mumps against the observed data. In this instance, the best-fitting force of infection is 19.76% (95% CI: 19.1-21.5) per year.

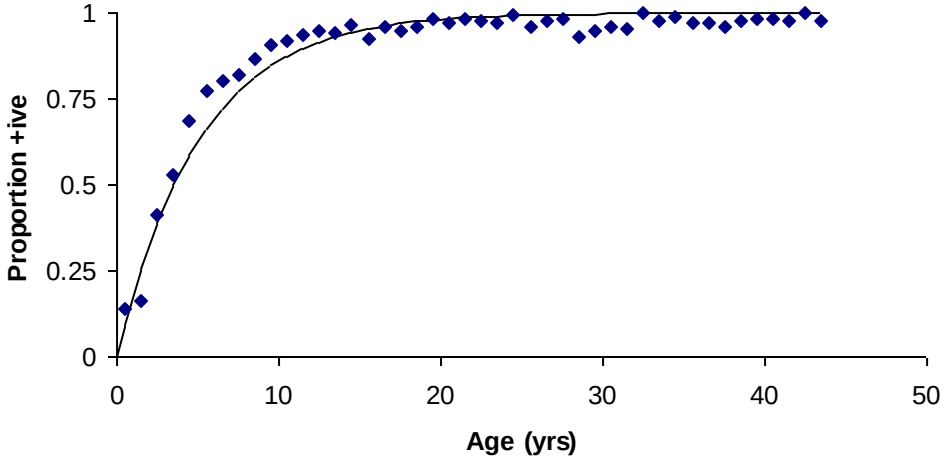


Figure 8: Comparison between the best-fitting estimates of the proportion of individuals that have ever been infected with mumps against the observed proportion of individuals who were positive to mumps antibodies during the 1980s in England and Wales. Data from Farrington (1990)[1].

The simple catalytic model makes several simplifying assumptions, e.g. that all individuals are susceptible at birth and that the force of infection is neither age nor time dependent (which may be inappropriate). Methods for adapting these simplifications are discussed later in this session. The analyses also need to make some assumptions about what the "infection marker" represents, e.g. past infection, current infection, recent infection etc and this (obviously) will influence the interpretation of the force of infection and the age-specific marker prevalence profiles.

Applying estimates of the average force of infection

The average age at infection

The average age at infection, *A*, is a helpful summary measure which can be used to compare the 'infectiousness' of different infections, or of the same infection across different populations (see ref [6]). Low average ages at infection are suggestive of increased transmission. For example, the average age of measles infection in the USA (1955-8) was about 5-6 years as compared with 2-3 years in Ghana (1960-8), suggesting that transmission was more intense in Ghana than in the USA[6].

Changes in the average age at infection also provide insight into whether an intervention (e.g. treatment of infectious persons) has had any impact. The average age at infection is also very important in guiding vaccination policy. For example, if the average age at infection is 4 years, vaccinating children aged over 4 years will have little impact on transmission.

The simplest method for calculating the average age at infection is to calculate the median age at infection, defined as the age by which 50% of individuals have been infected. This can be read off a scatter plot of the data. For example, 50% of individuals had antibodies to mumps by about age 5 years in England and Wales during the 1980s, suggesting that the average age at mumps infection was about 5 years.

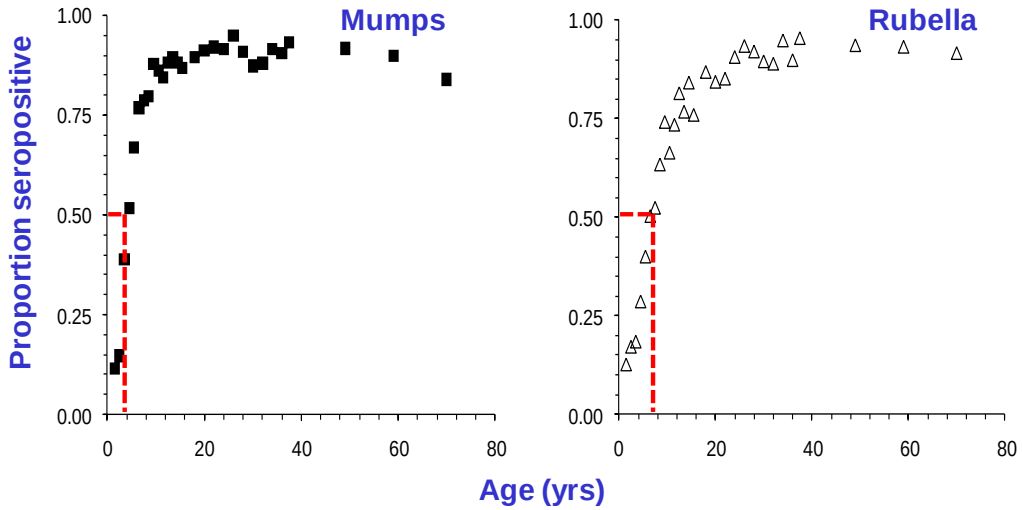


Figure 9: Illustration of how the median age at infection can be estimated from a plot of the age-specific proportion seropositive. The median age is given by the age at which the dashed vertical line crosses the x-axis.

We can also obtain an expression for the average age at infection (A) by using the following relationship between the average time to an event and the average rate at which it occurs (see block 1):

average time to event = 1/(average rate at which it occurs)

Applying this relationship, we obtain the following equation:

$$A \approx 1/\lambda$$

This relationship is approximate, since the average age at infection depends on how many individuals die before being infected and hence on the mortality rate. It also assumes that the force of infection is independent of age and that individuals mix randomly. However, for most practical purposes, this equation provides a reasonably good estimate of the average age at infection. Equations for improving the estimates are discussed in Anderson and May[6]; in general, these can be derived from the general mathematical formula for the average:

$$A = \frac{\int_0^\infty a\lambda(a)S(a)da}{\int_0^\infty \lambda(a)S(a)da} \quad \text{or} \quad A = \frac{\sum_a a\lambda(a)S(a)}{\sum_a \lambda(a)S(a)}$$

where $\lambda(a)$ is the force of infection among individuals of age *a,* and therefore $\lambda(a)S(a)$ is the number of new infections among individuals of age *a*.

Example

The average annual force of infection for mumps in England and Wales during the 1980s was about 19.8% per year (see above). Using the relationship $A \approx 1/\lambda$ implies that the average age at infection was 1/0.198≈5 years.

The proportion susceptible

The proportion of the population that is susceptible depends both on the rate at which individuals become infected, and the death rate in the population. If we assume that individuals mix randomly, then in populations with "rectangular" age distributions (similar to those in industrialized countries today – see below), in which it can be assumed that everybody lives only until age *L*, the proportion susceptible is related to the average age at infection *A* through the following expression:

$$s \approx A/L$$

Using the expression $A \approx 1/\lambda$ in this expression leads to the equation $s \approx 1/(\lambda L)$.

The formula $s \approx A/L$ follows from the "hand-waving" argument that in a population with a rectangular age distribution, "on average" everyone is susceptible until the average age A and everyone above that age is immune. The proportion susceptible is then given by the area of the shaded box shown in Figure 10A (i.e. *A*) divided by the sum of the shaded and white areas (i.e. *L*).

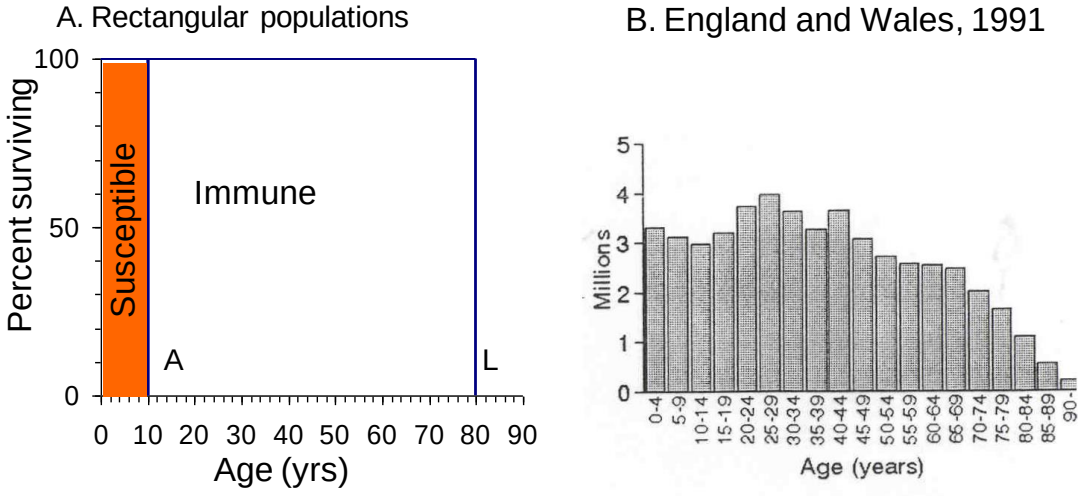


Figure 10: Illustration of the relationship between the proportion of the population that is susceptible and the life expectancy in a population with a rectangular age distribution. The shaded area reflects the proportion of the population that is susceptible. b) Population in England and Wales, 1991. Data source: Office for National Statistics.

When the age distribution follows a "type II" pattern (i.e. it is exponential, similar to that in "developing" countries – see Figure 11), the proportion susceptible is given by the expression:

$$s = \frac{1}{1+\frac{L}{A}}$$

The derivation of this expression is discussed in the Appendix. Note that if a population has an exponential age distribution, then the mortality rate is assumed to be constant (i.e. it is identical for all individuals).

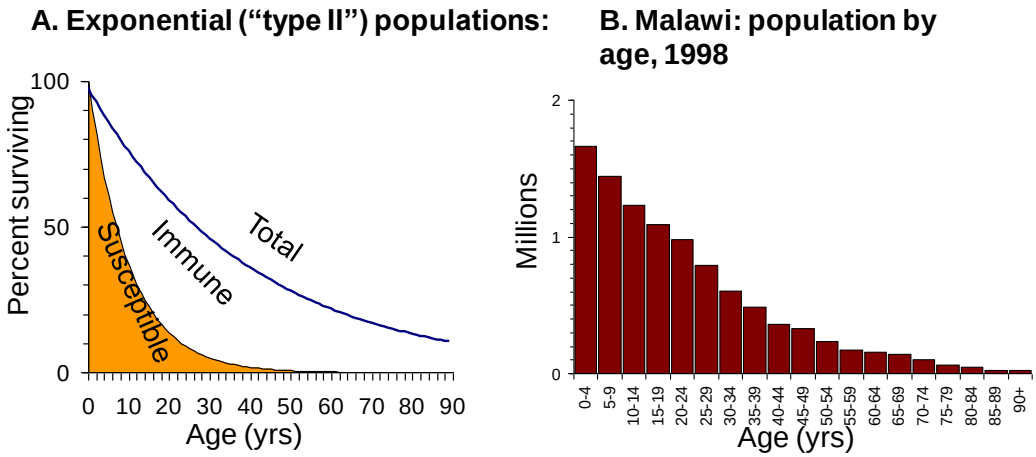


Figure 11: Illustration of the relationship between the proportion of the population that is susceptible and the life expectancy in a population with an exponential age distribution. The shaded area reflects the proportion of the population that is susceptible. b) Population in Malawi, 1998. Data source: National Statistics Office, Malawi.

Example

The average age at mumps infection in England and Wales during the 1980s was calculated to be about 5 years (see above). Assuming a life expectancy L of 70 years and that the age distribution is rectangular implies that the average proportion of the population that was susceptible was given by

$$s \approx A/L \approx 5/70 \approx 0.07.$$

The basic reproduction number (R_0)

If we assume that individuals mix randomly, so that the force of infection is identical for all age groups, we can calculate the basic reproduction number R_0 of an endemic infection as the reciprocal of the proportion susceptible:

$$R_0 = 1/s$$

This follows from the reasoning that the net reproduction number is equal to $R_0 s$, and for an endemic infection R_n=1; after rearranging the expression $R_0 s$=1, we obtain the result R_0=1/s.

Assuming further that the age distribution of the population is rectangular ("Type I"), we can use the relationship between the proportion of the population that is susceptible, the average age at infection and the force of infection discussed above to obtain the expression:

$$R_0 \approx L/A \approx \lambda L$$

In contrast, if the population has an approximately exponential age distribution then R_0 can be calculated using the following equation:

$$R_0 = 1 + L/A = 1 + \lambda L$$

These expressions have been used to calculate the basic reproduction number for many different infections, as summarized in Table 1. Once the basic reproduction number is known, we can then calculate the herd immunity threshold using the expression $1-1/R_0$.

Example

The average proportion of the population in England and Wales that was susceptible to mumps during the 1980s was 0.07 (see above). Using the relationship $R_0=1/s$ implies that $R_0=1/0.07\approx14$.

Table 1: Summary of estimated values for the basic reproduction number R_0, for different infections. Reproduced from Anderson and May (1991)

Infection	**Location**	Time period	**R_0**
Measles	Cirencester, England	1947-50	13-14
	England and Wales	1950-68	16-18
	Kansas, USA	1918-21	5-6
	Ghana	1960-8	14-15
	Eastern Nigeria	1960-8	16-17
Pertussis	England and Wales	1944-78	16-18
	Ontario, Canada	1912-13	10-11
Chicken pox	Maryland, USA	1913-17	7-8
	Baltimore, USA	1943	10-11
Diphtheria	New York, USA	1918-19	4-5
	Maryland, USA	1908-17	4-5
Scarlet fever	Maryland, USA	1908-17	7-8
	New York, USA	1918-17	5-6
	Pennsylvania, USA	1910-16	6-7
Mumps	Baltimore, USA	1943	7-8
	England and Wales	1960-80	11-14
Rubella	England and Wales	1960-70	6-7
	Poland	1970-7	11-12
	Gambia	1976	15-16
Poliomyelitis	USA	1955	5-6
	Netherlands	1960	6-7

The (age-specific) number of new infections per unit time

The force of infection can be used to predict the age-specific proportion of individuals susceptible to infection (see below) and then the number of new infections in different age groups per unit time. This is particularly important if infection at a particular age is associated with adverse outcomes. For example, rubella infection during pregnancy is associated with the child being born with Congenital Rubella Syndrome; infection with the polio virus during adulthood is associated with an increased risk of paralytic polio; the risk of developing measles encephalitis depends on the age at measles infection (see ref [6]).

In general, as discussed in previous sessions, the number of new infections per unit time in a population is given by the expression

$$\lambda(t)S(t)$$

where $S(t)$ is the number of susceptible individuals at time t and $\lambda(t)$ is the force of infection at time t. Adapting this expression, we obtain the following expression for the number of new infections among individuals of age a:

$$\lambda(a)S(a)$$

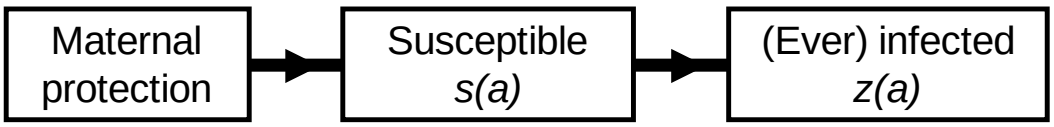
Here, $\lambda(a)$ is the force of infection among individuals of age a and $S(a)$ is the number of susceptible individuals of age a. These equations will be applied extensively during the next session.

Fine-tuning simple catalytic models

a) Maternally-derived immunity

The assumption that individuals are susceptible from birth may lead to a poor fit of the simple catalytic model to seroprevalence data for youngest age groups, who are protected by maternal antibodies. It also means that the force of infection estimated by fitting a simple catalytic model to the data may *underestimate* the true force of infection since the model *overestimates* the total person years of time that individuals are actually at risk of being infected. Therefore, in order to match the seroprevalence data, the best-fitting force of infection does not need to be as high as that required if it is assumed that individuals are susceptible, e.g. only after the first year of life.

To deal with maternally-derived immunity, the simple catalytic model would need to be amended to track individuals from birth as they lose their maternal protection to become susceptible and then infected, as follows:



In these models, maternal protection is typically assumed to be lost at a constant rate (specified by the average duration of protection by maternal antibodies), or individuals are assumed to be immune to infection, e.g. during the first few months of life and are susceptible thereafter. The expression for the proportion of individuals who are susceptible at a given age a if it is assumed that individuals are immune for the first 6 months of life is

$$s(a)= e^{-\lambda(a-0.5)}.$$

The term “a-0.5” reflects the person years during which individuals of age a were susceptible to infection.

The corresponding expression for the assumption that individuals lose maternal immunity at a constant rate e.g. μ is slightly more complicated than this expression, though it can be shown that it equals: $\frac{\mu(e^{-\mu a} - e^{-\lambda a})}{\lambda - \mu}$.

The two assumptions lead to similar predictions of the age-specific proportion infected, but slightly different estimates in the force of infection (though the confidence limits overlap) – see Figure 12 and Table 2.

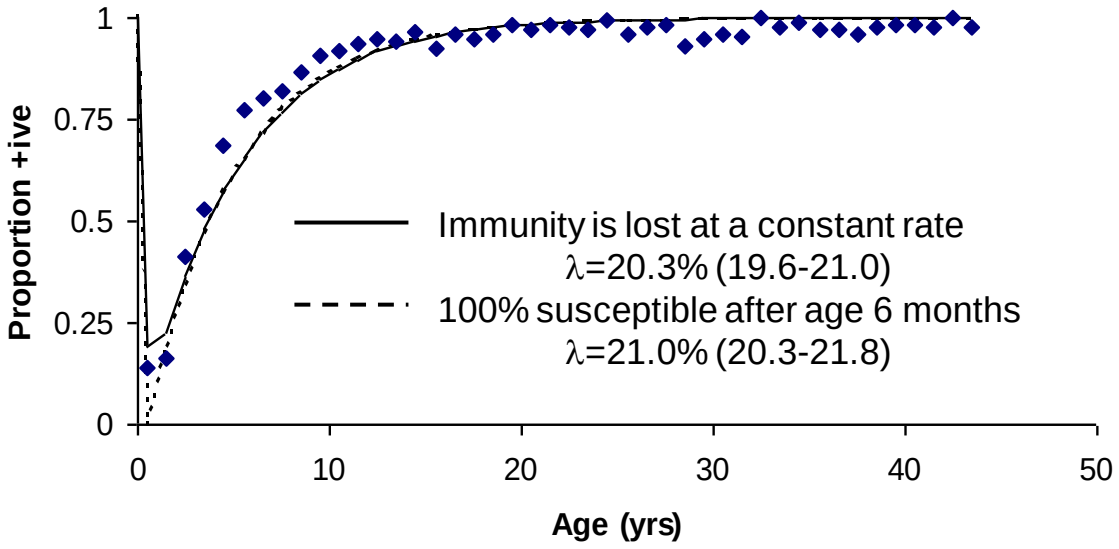


Figure 12: Comparison between predictions from the best-fitting catalytic model of the age-specific proportion of individuals who have ever been infected, against data on the proportion of individuals who had antibodies to mumps during the 1980s in the UK. Maternally-derived protection is assumed to decline at a constant rate or lasts for 6 months before individuals are fully-susceptible again.

Table 2: Comparison between best-fitting forces of infection, for different assumptions about protection from maternal antibodies:

Type of model	**Force of infection (% per year)** *(95% CI)*
Simple catalytic	19.8 (*19.1-20.5*)
Constant decline in maternal immunity	20.3 (*19.6-21.0*)
100% susceptible after age 6 months	21.0 (*20.3-21.8*)

The duration of maternally-derived immunity relative to the force of infection has implications for designing vaccination strategies, e.g. vaccinating at too young an age, when infants are still protected by maternal immunity means that many doses of vaccine are "wasted"; however, vaccinating older individuals in populations in which the force of infection is high may have little effect on reducing morbidity, as many children will have already been infected and will therefore be immune (see Anderson and May (1992)[6]).

b) Age-dependency in the force of infection

The assumption that the force of infection does not depend on age may be inappropriate for some infections. For the mumps example described above, model predictions underestimate the proportion of 5-14 year olds that are seropositive, suggesting that the true force of infection may be higher in this age group than that estimated. For older individuals,

model predictions overestimate the observed proportion seropositive. This overestimate may be attributable to an overestimate in the force of infection and/or to the sensitivity of the test (which, for example, may be unable to detect low antibody titres).

Graphical methods can be used to determine whether or not the force of infection is age-dependent[7]. This involves plotting –ln(observed proportion seronegative) of each datapoint against the age midpoint. If the resulting plot is a straight line, then:

1. the force of infection is the same for all age groups
2. the gradient of the line equals the force of infection, λ

Otherwise, the force of infection cannot be assumed to be the same for all age groups...

The rationale for this graphical test is as follows:

If the force of infection is not age-dependent, then, as discussed above, the proportion of individuals of a given age *a* who are susceptible (*s(a)*) is given by the expression:

$$s(a) = e^{-\lambda a}$$

Taking the natural logarithms of both sides (see the maths refresher to revise logs if necessary), we see that:

$$\ln(s(a)) = -\lambda a$$

Multiplying both sides of this equation by -1, we obtain the result:

$$-\ln(s(a)) = \lambda a$$

This equation is analogous to that of a straight line, y=mx+c, which indicates that the log of the proportion susceptible at age *a* is linearly related to the force of infection.

This principle is illustrated in Figure 13 using an “ideal” dataset for which the force of infection is 10% per year, identical for all ages and that the proportion susceptible equals the proportion seronegative. If we take the –ln(proportion seronegative), we obtain a straight line with a gradient of 10%, i.e. equal to the assumed force of infection.

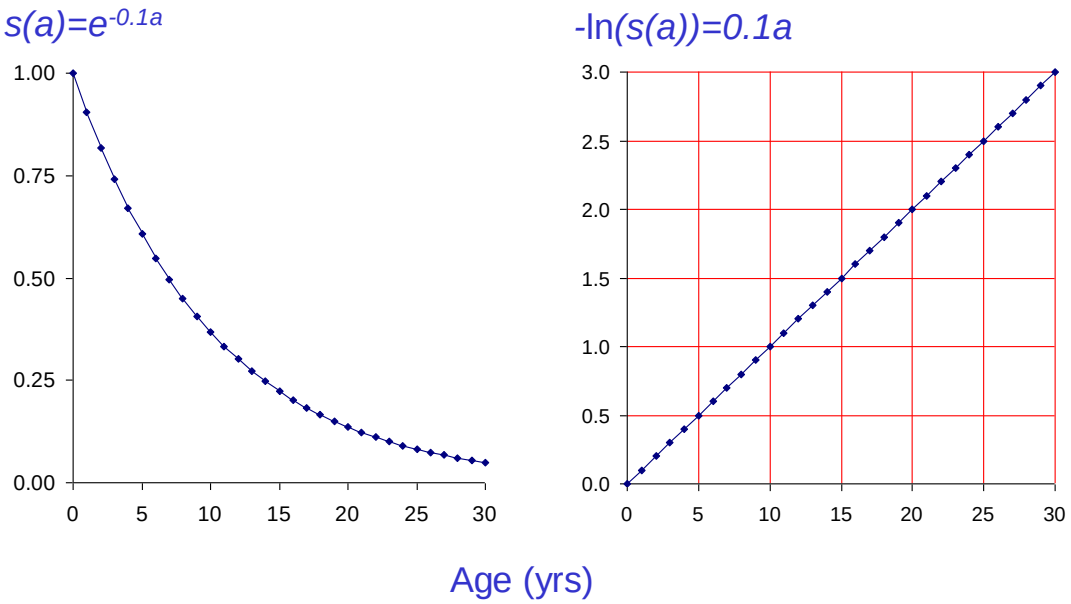


Figure 13: A. Predictions of the proportion of individuals that are susceptible at different ages, as obtained using the simple catalytic model, assuming that the force of infection is 10% per year year. B. Predictions of –ln(proportion susceptible) using values for the proportion susceptible from figure A.

The steps for determining whether the force of infection is age-dependent can be summarized as follows:

1. Estimate *s(a)*, the proportion susceptible in age group *a*, by S_a/N_a, where S_a is the number of susceptible individuals in age group *a* and N_a is the number of individuals in age group a who were tested.
2. Calculate the values $-\ln(S_a/N_a)$ (if $S_a = 0$, replace by 0.5).
3. Plot these values against the midpoints of the age groups.

If the force of infection is identical for all age group, then the plot should approximate to a straight line through the origin. The slope of the straight line is the force of infection λ.

If the graph deviates substantially from a straight line, then the assumption of a constant force of infection is inappropriate. Note again that the prevalence of prior infection in the first few months of life, due to the presence of maternal antibodies, should be set to zero in these analyses.

Example

The following figure shows the plot of –ln(observed proportion seronegative) for the mumps dataset described in Figure 2. Based on this plot, it would be reasonable to assume that the force of infection is constant in the age ranges <15 and ≥15 years, but that it differs between these two age groups.

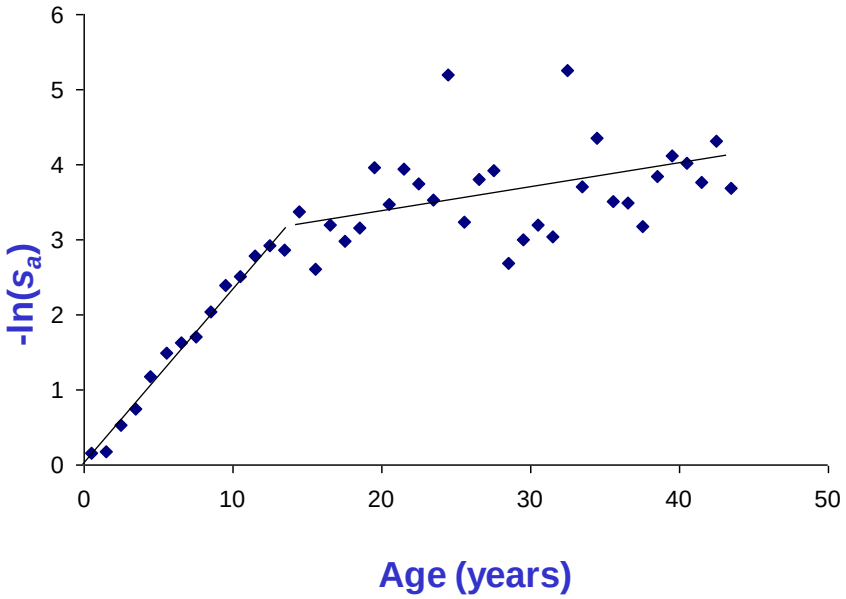
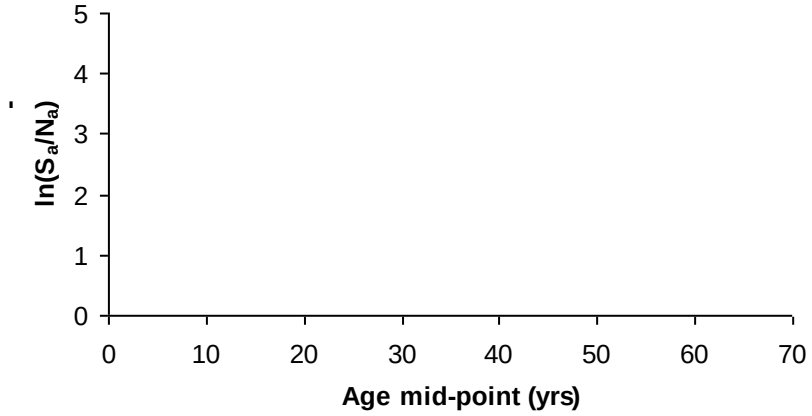


Figure 14: Plot of $-\ln(s_a)$ for the mumps data described in Figure 2.

Exercise

Complete the following table and use a graphical check to see if the force of infection is age-dependent.

Age (yrs)	N_a	S_a	S_a/N_a	$-\ln(S_a/N_a)$
1-5	52	32		
6-10	63	25		
11-20	46	4		
21-50	58	1		
51-80	42	0		



Once the way in which the force of infection depends on age has been determined, expressions for the age-specific proportion susceptible then need to be fitted to the data to estimate the actual force of infection. The expression depends on assumptions of how the force of infection changes with age. For example, the following is the expression that we

would use if we assume that the force of infection differs between those aged <15 and ≥15 years (denoted by λ_1 and λ_2 respectively):

$$s(a) = \begin{cases} e^{-\lambda_1 a} & a < 15 \text{ years} \\ e^{-15\lambda_1} e^{-\lambda_2 (a-15)} & a \geq 15 \text{ years} \end{cases}$$

Note that the expression for individuals aged over 15 years equals the product of:

1. the proportion of individuals who are susceptible at age 15 years (i.e. $e^{-15\lambda_1}$)
2. the proportion of individuals who escape infection between the ages 15 and a years i.e. $e^{-\lambda_2 (a-15)}$

As shown in the following figure, the assumption that the force of infection is age dependent leads to an improved fit to the mumps data described above, with the best-fitting force of infection being much lower for children than for adults:

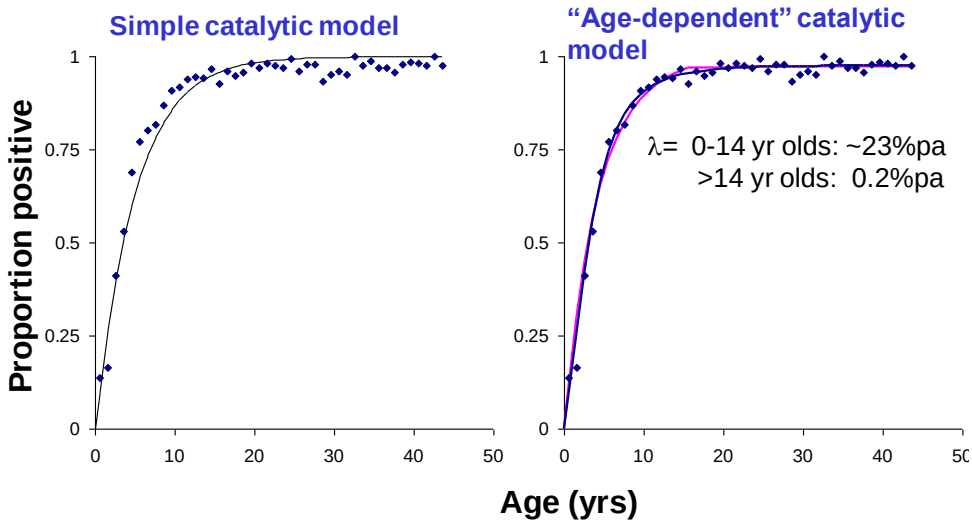


Figure 15: Comparison between predictions from the best-fitting simple catalytic model and an age-dependent catalytic model and the mumps data shown in Figure 2.

Similar age-specific patterns of a higher force of infection among children, as compared with that among children, have been found for several infections[1]. The age differences are probably due to differences in susceptibility or in exposure (e.g. young individuals may be more likely to contact other young individuals, than older individuals). Elucidation of the importance of such factors is important for predicting the likely effect of vaccination strategies targetted at children (see the later sessions on heterogeneous mixing).

The age-specific pattern will also depend on the study population eg whether it is urban/rural, developed/developing etc. Data on measles serology from New Haven, 1957[8] also suggested that there are differences in age-specific patterns between large and small families

Methods for fitting the above model to the seroprevalence data are discussed in the supplementary questions of the practical. Methods for dealing with alternative assumptions about the force of infection (e.g. that it changes continuously with age) are discussed in ref [1].

c) Time dependency in the force of infection

For infections for which the incidence cycles periodically, the fit of a simple catalytic model may be especially poor for the younger age classes if the data have been collected immediately after an epidemic. In such instances, estimates of the force of infection are best obtained by restricting the fitting to data from older individuals, who would have survived both epidemic and non-epidemic years, with high and low forces of infection respectively. For these individuals, the estimated force of infection should be interpreted as an "average" (annual) force of infection. These issues are discussed in detail by Whitaker and Farrington[9].

For some infections, apparent age-differences in the force of infection may be due to secular changes in the force of infection. Insight into whether this is the case can be obtained by analysing data collected over several years (see e.g. Sutherland et al [10] for Dutch tuberculin data analyses, Nokes et al and Ljungström et al [11, 12] for toxoplasmosis).

d) Non-immunizing infections

Other types of catalytic models (see Munch, 1959) may be more appropriate than simple catalytic models for some infections :

For example, the reversible or SIS (susceptible – infected/infectious - susceptible) model assumes that individuals become infected at a rate λ, and once infected, lose their infected status at a rate μ. This leads to the age-specific pattern in the prevalence of infection shown in Figure 16 :

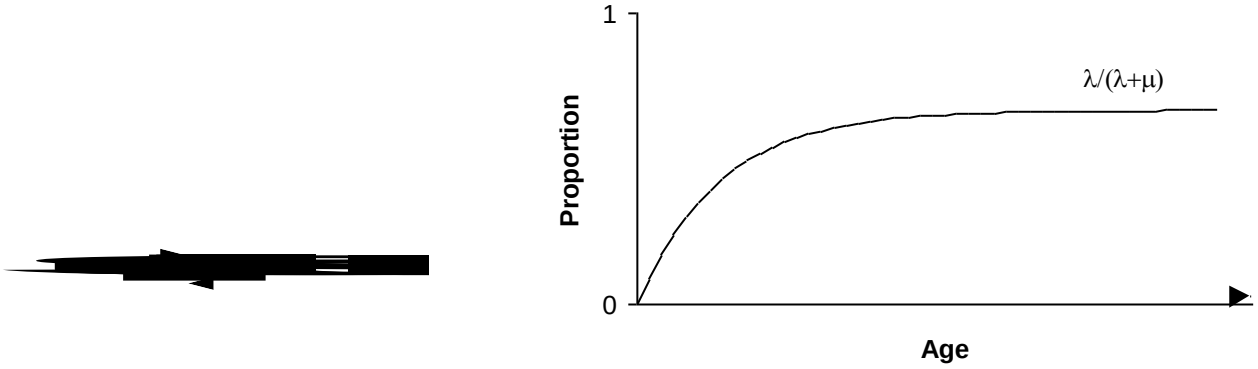


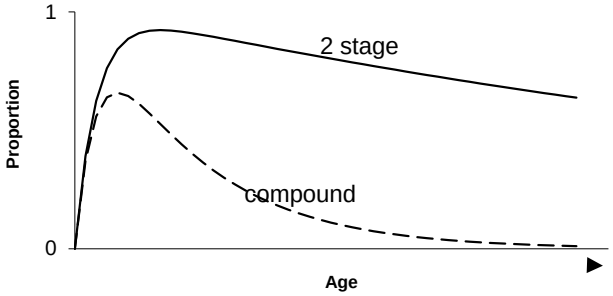
Figure 16: General structure of a reversible model and its predictions of the proportion infected.

The reversible model has been applied to cross-sectional tuberculin data [2, 4], diphtheria, malaria [13], filariasis [14].

Other variants include the two-stage or SIR (susceptible – infected/infectious - recovered) model:

and the compound model

which have been applied to yaws and histoplasmosis data[2] and lead to the following patterns in the age-specific prevalence of current infection:



Exercise:

Prove the result that the level of the plateau in the prevalence of infection for the reversible model is given by the equation $\lambda/(\lambda+\mu)$. *Hint: write down the differential equations for the rate of change by age in the prevalence of susceptible and infected individuals and use the facts that i) the proportion infected at any given age equals 1-proportion susceptible at that age and ii) at any point on the plateau,* $\frac{ds}{da} = \frac{dz}{da} = 0$. *i.e. the rate of change in the proportion of susceptible and infected individuals with age is zero.*

Model-free methods

Several useful parameters can be calculated directly without using a model. In particular, if it can be assumed that the proportion of individuals who are negative to the biological test for infection is the same as the proportion susceptible, the proportion susceptible, s, may be estimated by

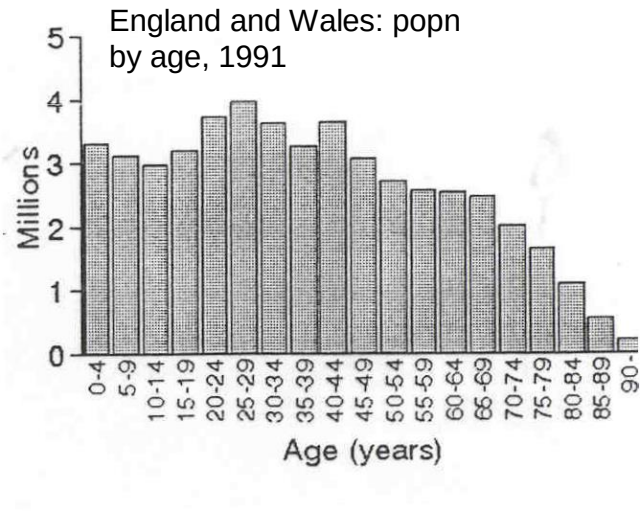
$$s = \Sigma_a \pi_a \times S_a / N_a$$

where π_a is the proportion of the population in age group a (obtained from life tables), S_a is the number of susceptibles among the N_a individuals of age a who were tested.

If it is appropriate to assume that the population has a rectangular age distribution with life expectancy *L*, then π_a is given by $\pi_a = w_a/L$, where w_a is the width of age group a. The proportion susceptible is then given by

$$s = (\Sigma_a w_a \times S_a / N_a)/L$$

which can be calculated directly from the data.



For example, considering the population of England and Wales shown on the left, if the life expectancy is 75 years, then the width of the age group 0-4, 5-9 and 10-19 years is 5, 5 and 10 years respectively, which gives the result that

$\pi_{0\text{-}4} = 5/75$

$\pi_{5\text{-}9} = 5/75$

$\pi_{10\text{-}19} = 10/75$

etc

Finally, if there are maternal antibodies in infancy, it is best to assume that the proportion susceptible is zero to some age M at which the maternal antibodies wear off (e.g. M = 0.5 years).

Whatever the mortality rate, if it is reasonable to assume that the force of infection is constant with age, then you can also directly calculate R_0 and the herd immunity threshold using the expressions $R_0 = 1/s$ and HIT=1-1/R_0.

Exercise

The following are some data on the numbers of individuals found to be seronegative to measles antibodies in a given population.

Age (yrs)	N_a	S_a	S_a/N_a	π_a	$\pi_a \times S_a/N_a$
1-5	52	32			
6-10	63	25			
11-20	46	4			
21-50	58	1			
51-80	42	0			

Using the above expressions and assuming that the age distribution of the population is rectangular and the life expectancy is 80 years, estimate

a) the proportion of individuals who are susceptible

b) the average age at infection

Assuming that the force of infection is not age-dependent, estimate

c) the force of infection

d) the basic reproduction number

Summary

Cross-sectional data on the prevalence of previous infection (for an endemic infection) can be used to estimate the average force of infection, which can then be used to estimate the average age at infection, the basic reproduction number, the proportion of the population that is susceptible. Such data are typically analysed using catalytic modelling methods, which can be adapted to deal with maternally-derived immunity, and an age-dependent force

of infection. Graphical methods can be used to infer whether the force of infection is age-dependent. Model-free methods can also be used to obtain approximate estimates of the proportion susceptible, the average age at infection and the force of infection.

Further reading

Vynnycky E and White RG (2010) An introduction to infectious disease modelling. Oxford University Press. Chapter 5 (sections 5.1 and 5.2).

The following paper by Hens et al, published in 2009, is available as supplementary reading: Hens N, Aerts M, Faes C, Shkedy Z, Lejeune O, Van Damme P, Beutels P. Seventy-five years of estimating the force of infection from current status data. Epidemiol Infect. 2010 Jun;138(6):802-12. Epub 2009 Sep 21.

Reference List

1. Farrington CP. Modelling forces of infection for measles, mumps and rubella. Stat Med 1990; 9(8):953-967.

2. Munch H. Catalytic models in epidemiology. Cambridge, Massachusetts: Harvard University Press; 1959.

3. Cutts FT, Abebe A, Messele T et al. Sero-epidemiology of rubella in the urban population of Addis Ababa, Ethiopia. Epidemiol Infect 2000; 124(3):467-479.

4. Fine PE, Bruce J, Ponnighaus JM, Nkhosa P, Harawa A, Vynnycky E. Tuberculin sensitivity: conversions and reversions in a rural African population. Int J Tuberc Lung Dis 1999; 3(11):962-975.

5. Clayton D, Hills M. Statistical methods in epidemiology. Oxford: Oxford University Press; 1993.

6. Anderson RM, May RM. Infectious Diseases of Humans. Dynamics and control. Oxford: Oxford University Press; 1992.

7. Griffiths DA. A catalytic model for infection for measles. Appl Statist 1974; 23(3):330-339.

8. Grenfell BT, Anderson RM. The estimation of age-related rates of infection from case notifications and serological data. J Hyg (Lond) 1985; 95(2):419-436.

9. Whitaker HJ, Farrington CP. Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. Stat Med 2004; 23(15):2429-2443.

10. Sutherland I, Bleiker MA, Meijer J, Styblo K. The risk of tuberculous infection in The Netherlands from 1967 to 1979. Tubercle 1983; 64(4):241-253.

11. Nokes DJ, Forsgren M, Gille E, Ljungstrom I. Modelling toxoplasma incidence from longitudinal seroprevalence in Stockholm, Sweden. Parasitology 1993; 107(Pt 1):33-40.

12. Ljungstrom I, Gille E, Nokes J, Linder E, Forsgren M. Seroepidemiology of Toxoplasma gondii among pregnant women in different parts of Sweden. Eur J Epidemiol 1995; 11(2):149-156.

13. Kitua AY, Smith T, Alonso PL et al. Plasmodium falciparum malaria in the first year of life in an area of intense and perennial transmission. Trop Med Int Health 1996; 1(4):475-484.

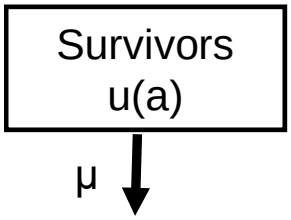
14. Vanamail P, Subramanian S, Das PK et al. Estimation of age-specific rates of acquisition and loss of Wuchereria bancrofti infection. Trans R Soc Trop Med Hyg 1989; 83(5):689-693.

Appendix: Illustration of the result that in a population with an exponential age distribution, the proportion susceptible is 1/(1+L/A)

This result can be proved by using the following steps:

1. Considering a cohort of individuals from birth and working out the proportion of the original cohort which is i) still alive at age a and ii) still alive and susceptible by age a.
2. Working out the total numbers of individuals in the population who are susceptible and alive, using the results from step 1.
3. Dividing the expression for the total number of individuals in the population who are susceptible by the total population size, calculated in step 2.

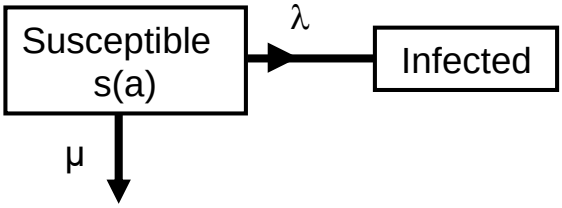
1i) Calculating the number of individuals who are still alive by age a.



If a population has an exponential age distribution, then it means that individuals die at a constant rate e.g. *m*. Considering a cohort of individuals who are born at the same time, the change in the proportion of the original cohort which is still alive by age a (denoted by *u(a)*) can be represented by the diagram on the left.

Applying the reasoning from the lecture notes, the expression for the rate of change in u(a) is given by the expression $\frac{du(a)}{da} = -mu(a)$. As you saw in the lecture, this gives the result that $u(a)=e^{-ma}$.

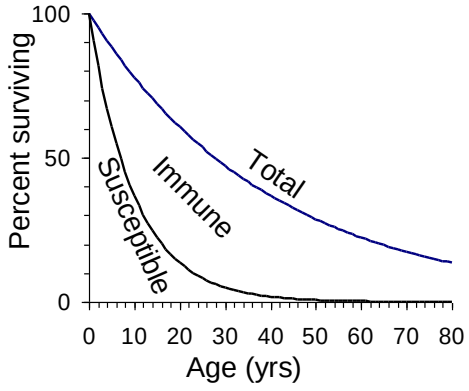
1ii) Calculating the number of individuals who are still susceptible by age a.



Assuming that individuals are infected at a constant rate λ, the change in the proportion of the original cohort which is still alive and susceptible (denoted by s(a)) can be represented by the diagram on the left.

Applying the reasoning in section II of the lecture, the expression for the rate of change in s(a) is given by the expression $\frac{ds(a)}{da} = -(m+\lambda)s(a)$. As you saw in the lecture, this gives the result that $s(a)=e^{-(m+\lambda)a}$.

2. Calculating the total number of individuals who are alive and susceptible



As you saw in step 1i), the proportion of the original cohort which survives until age a is given by the expression $u(a)=e^{-ma}$. This equation can be represented diagrammatically by the "Total" line on the figure on the left.

Assuming that a total of N_0 individuals are born each year and that the number of births into the population balances out the number of deaths, the total population size is given by the area under the "total" curve shown here, multiplied by N_0.

More formally, this area is given by the integral between zero and infinity of the expression $N_0e^{-\mu a}$, or, using mathematical notation, as $N_0\int_0^\infty e^{-ma}da$.

From your previous mathematical knowledge (see also the maths refresher), you may recall that the integral of e^{-ma} equals 1/m, and hence this integral expression equals N_0/m. Since L (the average life expectancy) is equal to 1/m, we see that the total population size equals N_0L.

By the same argument, the total number of individuals who are susceptible is given by the area under the susceptible curve shown above, multiplied by N_0, or equivalently, the integral between zero and infinity of the expression $N_0e^{-(m+\lambda)a}$. This is expressed formally as $N_0\int_0^\infty e^{-(m+\lambda)a}da$.

Using your previous mathematical knowledge and applying the same logic as you used to work out the total population size, we see that this expression equals $N_0/(m+\lambda)$. As discussed in the lecture, $\lambda=1/A$, where A is the average age at infection.

This implies that the expression for the total number of susceptible individuals, $N_0/(m+\lambda)$ can be re-expressed as $N_0/(1/A +1/L)$.

3. The overall proportion of the population which is susceptible is given by the expression:

Total number of individuals who are susceptible/ Total population size

Substituting the expressions $N_0/(1/A +1/L)$ and N_0L for the number of individuals who are susceptible and alive respectively into the above expression gives the result that

$s = \left(\frac{N_0}{\frac{1}{A}+\frac{1}{L}}\right) \div N_0L$, which simplifies to $\frac{1}{1+\frac{L}{A}}$.

Introduction to Infectious Disease Modelling and its Applications - 2018

Session 9: Fitting models to data: I. Finding a goodness of fit metric

Lecture

Overview and Objectives

This is the first lecture in a two-part series on fitting models to data. In this first lecture, we cover the methods used to measure the goodness of fit of a model to data. In the second lecture, we will continue with algorithms to achieve the best fits, and will also discuss methods for undertaking sensitivity analysis.

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

1. Explain the need to fit models to data.
2. Define key terms used in model fitting/calibration and sensitivity analysis.
3. Explain the principles behind the least squares, weighted least squares and maximum likelihood methods of fitting.
4. Implement these methods in MS Excel and Berkeley Madonna.

Introduction

As discussed in earlier sessions, once the general stricture of a model describing the transmission of an infection has been specified, we typically need to decide on the input parameters and initial conditions of the model. In addition, if we need the model to describe the epidemiology of the infection in some population, we also need to fit the model to some data. We illustrate these requirements using two examples.

Example 1: Describing age-specific seroprevalence data using catalytic models (block 2)

At the beginning of Block 2 of this course, we looked at a catalytic model of which described the acquisition of serum antibodies in a population. This model has a simple **structure** with only two compartments: a susceptible compartment S (for people who have yet to acquire an infection) and an ever infected compartment Z (for people who have previously had an infection and have seroconverted). It was assumed that everyone who acquires the infection seroconverts, and there is no waning of serum antibody titres.

Hence the model only requires one **parameter**, the force of infection λ, which represents the rate at which susceptible individuals acquire the infection, and hence move from compartment S to compartment Z. It also has **initial conditions** representing the state of the population at a particular age. For instance, we could assume that everyone in the population is susceptible at age 0, so $S(0) = Z(0) = 0$.

To determine the value of the force of infection λ, the model requires **data to fit to**. This may be, for example, the results of a cross-sectional serological survey giving the proportion of individuals in a sample who are seropositive at different ages. The force of infection λ is then

given different values, and for each value of λ, what the model predicts about the proportion of the population at different age groups who are in compartment Z (ever infected, and hence seropositive) is compared to data. The value of λ which results in the best fit of model predictions to data is then considered the "best fitting" parameter value.

Example 2: Describing the transmission of pandemic influenza in the UK (2009)

Another example, which is more complex than the catalytic model that we have just discussed, is provided by a model describing the transmission of pandemic influenza in the UK built by modellers at the HPA and LSHTM during the 2009 H1N1v pandemic[1]. This has an SEIR (susceptible-pre-infectious-infectious-recovered) compartment structure, except that there are two pre-infectious compartments and two infectious compartments. Having two pre-infectious compartments rather than just one means that the distribution of the time interval between infection and onset of infection no longer follows the exponential distribution; instead, it follows the gamma distribution (see e.g. Wearing et al[2]).

Hence, this model has six compartments, and five parameters representing the flows between the compartments (see Figure 1 below). The interesting thing about this model is that it was built in real-time, as the pandemic itself was unfolding. It was therefore initially fitted to estimates of the number of H1N1v cases in the UK up to November 15 2009. That is, the model parameters were chosen in order to ensure that what the model says about the number of cases from June 1 (the start of the pandemic) to November 15 match data as much as possible. The model is then used with those parameters to predict what the pandemic would look like after November 15, and turned out to be fairly accurate when compared to new data a few weeks later.

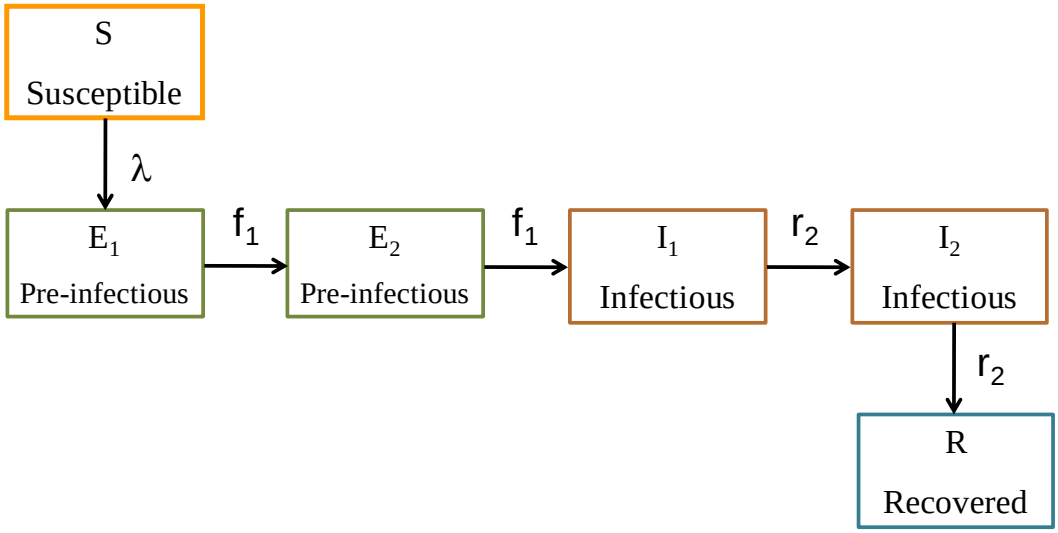


Figure 1. Simplified compartmental flow diagram for the HPA-LSHTM model of pandemic influenza.

[1]For further details, see Baguelin M, van Hoek AJ, Jit M, Flasche S, White PJ, Edmunds WJ. Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. Vaccine 2010; 28(12):2370-84.

[2]. Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005 Jul;2(7):e174. Epub 2005 Jul 26.

Terminology

The terminology used to describe model fitting and related processes can be confusing as related terms are often used (sometimes incorrectly) to mean different things. Some key concepts are the following:

Parameterisation is the process of finding numerical values for the parameters in a model, regardless of what method is used to do this. Some common methods include fitting the model to outcome data, inserting the parameters directly from available evidence (eg. in the literature), and simply guessing!

Fitting (or calibration) is a method of parameterisation of a model by finding a parameter set that produces model results with a "good fit" to outcome data.

Validation is the process of comparing the outcomes of a model to data *after* it is parameterised to see if the results are "valid" (either have face validity or satisfy some statistical test). An example is the real-time pandemic influenza model mentioned earlier, which was parameterised to case numbers up to November 15, and then validated a few weeks later by comparing it to what actually happened. Often, part of a data set used to parameterise a model is "held back" (not used for parameterisation) so that it can be used for validation instead.

Sensitivity analysis is the process of altering parts of a model (such as the parameters, structure or even type of model) to see what effect this has on results. The most common form is parametric sensitivity analysis, where parameters are adjusted to examine the effect on results.

Some formalisms

In order to understand the process of fitting a model to data, there are some important terms we need to understand. To start with, the process of fitting a model to data is made up of two stages:

1) The first stage is that of answering a statistical question. How do we decide whether a model provides a "good" fit to data? In order to answer that question, we need to construct a **goodness of fit** metric (that is, a measurement that represents how well a model fits data).

2) The second stage requires answering a computational quesion. 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).

What is a goodness of fit metric? We can describe it formally in this way. Suppose we have a model with m input parameters x_1, ..., x_m. We can write this using mathematical notation as $\boldsymbol{x} = (x_1, ..., x_m)$. The technical term for **x** is an "m-dimensional input vector". **x** needs to obey

certain constraints (conditions on values that each input parameter can take), which, using mathematical notation, is written as $\boldsymbol{x} \in X$, where X refers to the set of possible constraints.

We also have *n* data points (observations) which we call O_1, ..., O_n. These may, for instance, correspond to some outcome measure at various time points t_1, ..., t_n.

These data points are compared to the corresponding model outputs at the same time points t_1, ..., t_n. The value of these outputs depend on the value of the input vector $\boldsymbol{x}$, so we call them $E_1(\boldsymbol{x})$, ..., $E_n(\boldsymbol{x})$, where $E_1(\boldsymbol{x})$ is the model output at time point t_1 when the input parameters take the value $\boldsymbol{x}$ and so forth. Hence we want to compare O_1 to $E_1(\boldsymbol{x})$, O_2 to $E_2(\boldsymbol{x})$ and so on, such that O_i is as close as possible to $E_i(\boldsymbol{x})$.

Example

To illustrate these ideas, we return to our catalytic model example. This has a single input parameter, λ, which needs to lie between 0% and 100% inclusive, so we write $\lambda \in [0\%,100\%]$. Instead of data at different time points, we have data on seroprevalence at a single time point but for different age groups a_1, ..., a_n, which we call O_1, ..., O_n. These are compared to model predictions of seroprevalence for the same age group for a given value of λ, which we label $E_1(\lambda)$, ..., $E_n(\lambda)$.

Returning to our general description, in order to compare model predictions $E_1(\boldsymbol{x})$, ..., $E_n(\boldsymbol{x})$ to actual observations O_1, ..., O_n, we need a **goodness of fit function** which takes as its inputs both the model predictions as well as the observations, and returns a single number that signifies how close $E_1(\boldsymbol{x})$, ..., $E_n(\boldsymbol{x})$ is to O_1, ..., O_n.

Hence we may write this function as $g(E_1(\boldsymbol{x}), ..., E_n(\boldsymbol{x}), O_1, ..., O_n)$. The smaller (or larger, depending on the function) it is, the better the model fit to data. Hence it is called the **objective function** or **minimand** (maximand). For instance, we might choose a function that takes the value 0 when $E_1(\boldsymbol{x}) = O_1$, $E_2(\boldsymbol{x}) = O_2$ and so forth, and get bigger the further away we get from this perfect fit. Figure 2 shows how the input parameters, model outcomes, observations and goodness of fit function are related.

Formally, we write this **optimisation problem** in the following way: Minimise $f(\boldsymbol{x})$ subject to $\boldsymbol{x} \in X$, where $f(\boldsymbol{x}) = g(E_1(\boldsymbol{x}), \ldots, E_n(\boldsymbol{x}), O_1, \ldots, O_n)$.

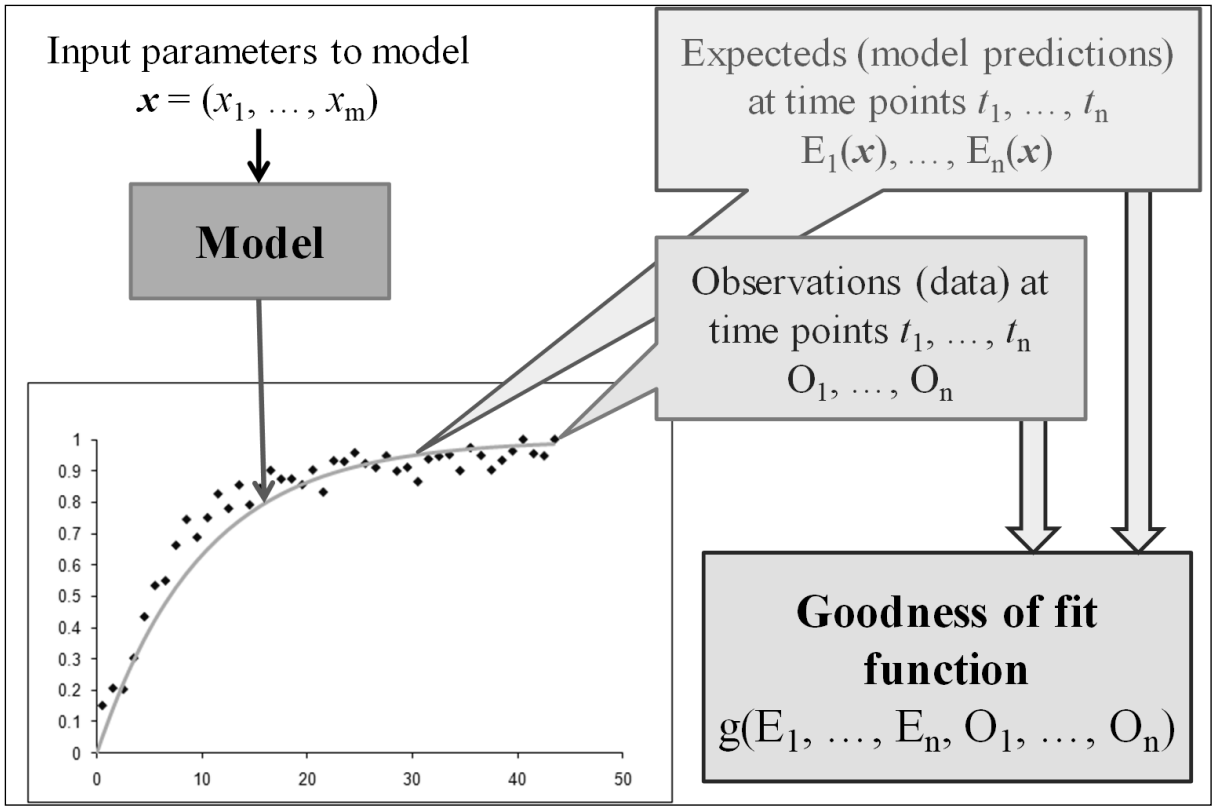


Figure 2. Relationship between input parameters, model outcomes, observations and goodness of fit function.

The method of least squares

How do we choose $g(E_1(\boldsymbol{x}), \ldots, E_n(\boldsymbol{x}), O_1, \ldots, O_n)$ in such a way that it gets smaller (or larger) the better the model fits the data, that is, the smaller the distance between $E_1(\boldsymbol{x})$ and O_1, $E_2(\boldsymbol{x})$ and O_2 and so on?

An obvious way is to define it as the sum of the **residuals** of each model prediction point, that is, the difference between the value of that point and the corresponding actual observation. Hence we have $g(\boldsymbol{x}) = \sum_i \left(E_i(\boldsymbol{x}) - O_i\right)$. This is nice and simple, but the problem is that positive and negative differences cancel out when we sum them up, so we may end up with a small value of $g(x)$ even though individually none of the model predictions are particularly close to their data points (see Figure 3).

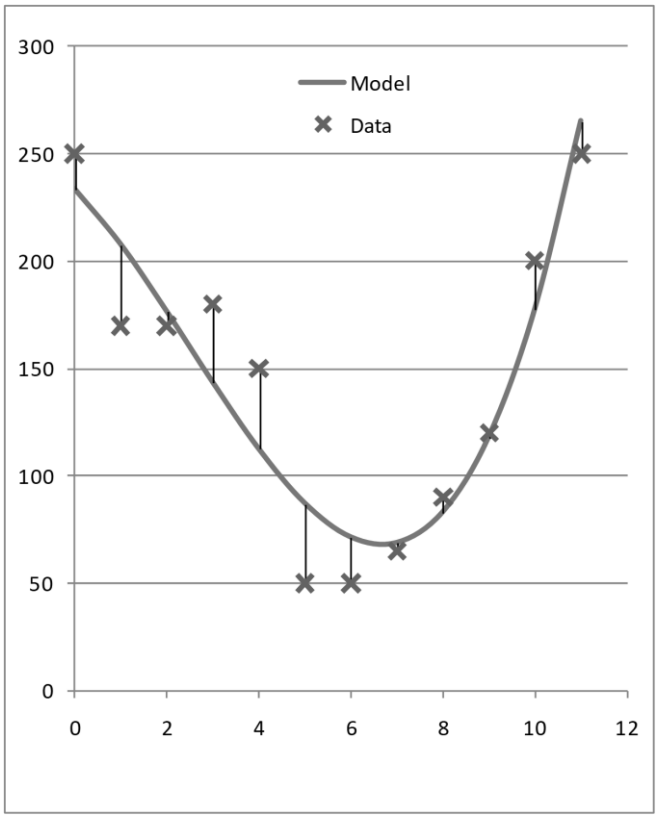


Figure 3. Residuals of each model prediction point compared to data. Notice that some of the residuals have positive sign while others are negative.

The obvious solution is to ignore the sign (positive or negative), so that the goodness of fit function is defined as the sum of the absolute differences between each data point and corresponding model prediction. This is written formally as $g(\boldsymbol{x}) = \sum_i \left| E_i\ (\boldsymbol{x}) - O_i \right|$.

The problem here is the modulus function $|a|$. This has a sharp point at the origin a=0 (we say that it is non-smooth at a=0) that makes the function difficult to deal with, both for mathematical analysis and also for computer calculations.

A clever way around is to instead take the sum of squared distances between the data point and model prediction. Hence we have $g(\boldsymbol{x}) = \sum_i \left(E_i\ (\boldsymbol{x}) - O_i \right)^2$, which will also always be positive but doesn't have the same sharp point at the origin. This is called the **residual sum of squares (SSR or SSQ)**, and is a very popular goodness of fit metric. The method of minimising this metric in order to obtain the best fit is called the **method of least squares**.

Least squares fitting can be conducted using most modelling packages, including MS Excel and Berkeley Madonna. In Excel, you can make use of Solver in order to work out the optimal value of a parameter. You use Excel formulas to calculate the residuals (differences between your model outcomes and actual data points at the same time points), use "=SUM" to sum them up, and then make the sum the cell to optimise in Solver, while varying the model parameters (see Figure 4).

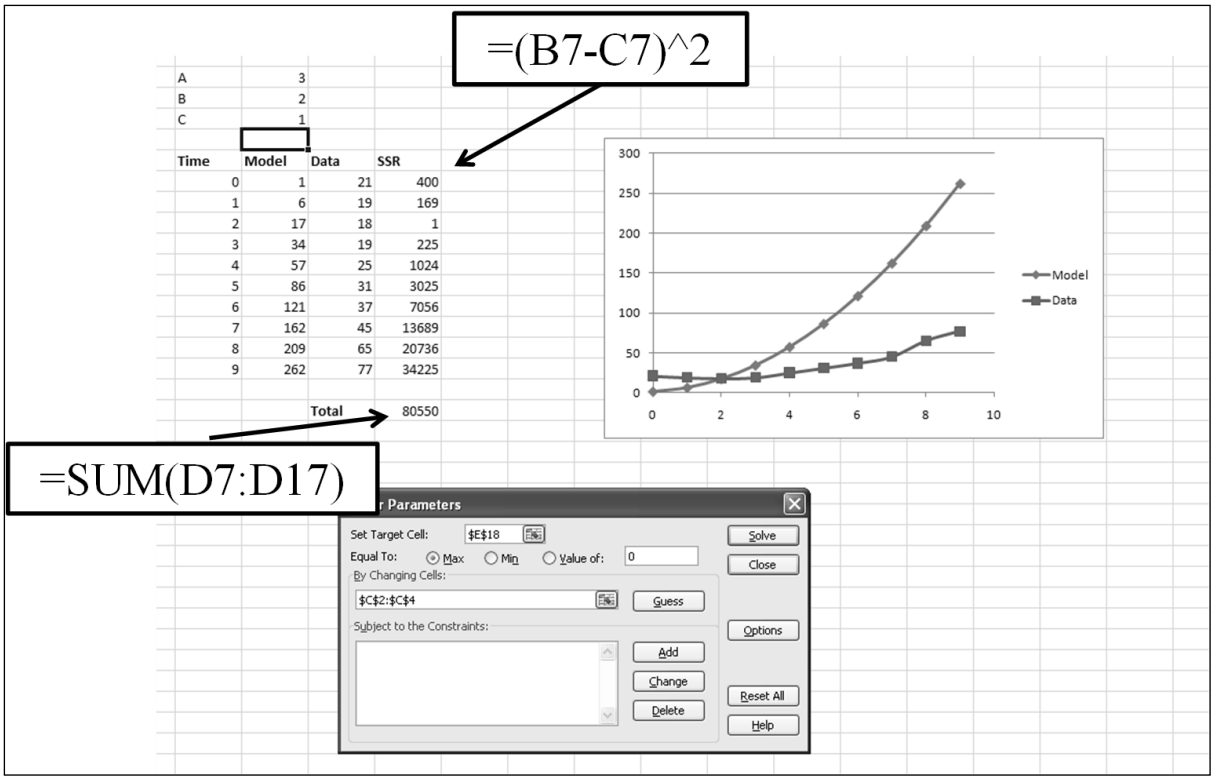


Figure 4. Using Excel to carry out least squares fitting.

Least squares fitting is not straightforward in Berkeley Madonna. Recall that elsewhere in the course, you used Berkeley Madonna to fit an epidemic curve to disease incidence data (e.g. pandemic flu, Ebola) by changing parameters. Berkeley Madonna can fit one or more parameters, although it is not always reliable when fitting multiple parameters. You need to choose the parameter value(s) which give the smallest value of "fit" (if doing several fits).

However, to confuse things further, the "fit" in Berkeley Madonna is actually not the residual sum of squares, but the root mean squared difference between the data and model prediction, $\sqrt{\frac{\sum_i \left(E_i(x) - O_i\right)^2}{n}}$. If we compare this to the formula for the residual sum of squares, we notice that this is actually the square root of SSR divided by n (the number of model parameters), so you can work backwards and calculate the SSR from the "fit".

Berkeley Madonna is set up in such a way that it is much easier to fit model predictions of the cumulative numbers of cases, than it is to fit to the incidence of cases. Fitting to the cumulative numbers of cases is not ideal since it encourages the fitting routine to fit the "later" data points best (because in doing so, it can make the SSR as small as possible).

The advantage of the least squares method is that it is simple, intuitive and easy to set up in MS Excel or a programming language. However, there are also disadvantages. Firstly, the SSR doesn't actually tell us how "good" a fit the model is to data (aside from the fact that we want it to be as small as possible!). Secondly, we aren't taking into account how uncertain we are about our data (observations). Every data point is given equal weight, even if we are (for example) extremely certain about one point and less certain about the others. It can be shown (but we won't do it here as it is a bit technical) that the second issue is "not a problem" if the

errors around the observations are uncorrelated, have roughly equal variances and are (preferably) normally distributed.

If they obviously don't have equal variances, one way around is to weigh the observations according to how much we "believe" each observation. Hence, the SSR becomes $g(\boldsymbol{x}) = \sum_i w_i \left(E_i\ (\boldsymbol{x}) - O_i\right)^2$, where w_i is the weight that we give to observation *i*. For instance, we may choose to weight each observation by the inverse of its variance (if we know it), i.e. $w_i = 1/\sigma_i$. The points with more uncertainty around them will have larger variances and therefore be downweighted (less important). The main problem with this is that we often do not know the variance of each data point. One way around this is to instead use the inverse of the model value, i.e. $w_i = 1/E_i(x)$. This is called the **Pearson chi-squared statistic**.

Maximum likelihood estimation

An entirely different approach from least squares fitting (but one that turns out actually not to be so different after all, as we shall see later), is to use a method called **maximum likelihood estimation**. To understand this method, we first need to review the statistical ideas of **probability** and **likelihood**. We talk about the probability of an event occurring given particular parameters, but the likelihood of a parameter taking a value given that a particular event has occurred.

Example

To illustrate, suppose we had a panel of serological samples. If each serum from the panel has a probability of 0.5 of being seropositive, then we can talk about (for example), the probability of obtaining 5 seropositive sera in a random sample of 10. In fact, the number of seropositive sera obtained is binomially distributed[2] so this is easy to work out.

Suppose X is a random variable representing the number of seropositive sera obtained, p is the probability of each serum being positive and n is the total number of sera in the sample. Then X is binomially distributed, and we write X ~ Bin(n,p).

The probability that X takes a particular value given some values of p and n, P(X|p,n), is given by

$$P(X|p,n) = {}^nC_r\, p^r(1-p)^{n-r}.$$

Here ${}^nC_r$ is the binomial coefficient, defined by ${}^nC_r = \frac{n!}{r!(n-r)!}$. It represents the number of ways of choosing r positive sera from a sample of n sera. p^r represents the probability that

[2] A binominal distribution is a probability distribution that represents the number of successes we get when we carry out a number of independent Bernoulli trials, that is experiments where the only possible outcomes are success or failure. For instance, the number of heads in 10 coin tosses is binomially distributed. See also the stochastic modelling lecture.

exactly r sera are positive and $(1-p)^{n-r}$ represents the probability that the remaining n-r sera are negative.

In this example, n=10 and p=0.5, so:

$$P(X=5 \mid p=0.5, n=10) = {}^{10}C_5\,(0.5)^5(0.5)^5 = 0.25.$$

Hence the probability of obtaining 5 seropositive sera is 0.25. Figure 6 shows the probability mass function of X (the probability that the number of positive sera obtained takes different values).

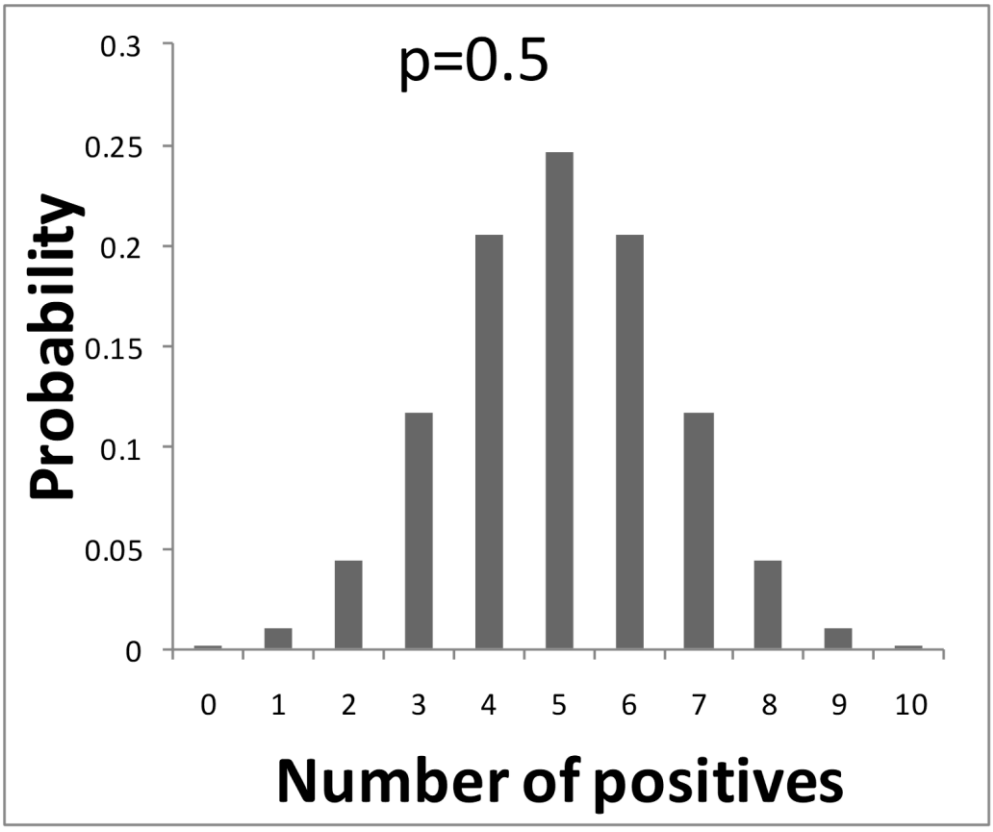


Figure 5. Probability mass function of X (the probability that the number of positive sera obtained takes different values).

Now let's consider the problem in reverse.

If we find 5 seropositive sera in a sample of 10, what is the probability of p having the value of (say) 0.5?

Now we are talking about the probability of a parameter of a distribution taking a value, so we refer to this as the likelihood of p (the probability of each serum being positive) taking the value of 0.5.

To work that out, think about the probability of obtaining 5 seropositive sera in a sample of 10 for different values of p. We can then plot this probability as a function of p (see Figure 6) –

this is the likelihood function, L(p|X=5). Notice that the likelihood function is maximised at p=0.5. We say that p=0.5 is the maximum likelihood estimate for X=5 (having 5 positive sera).

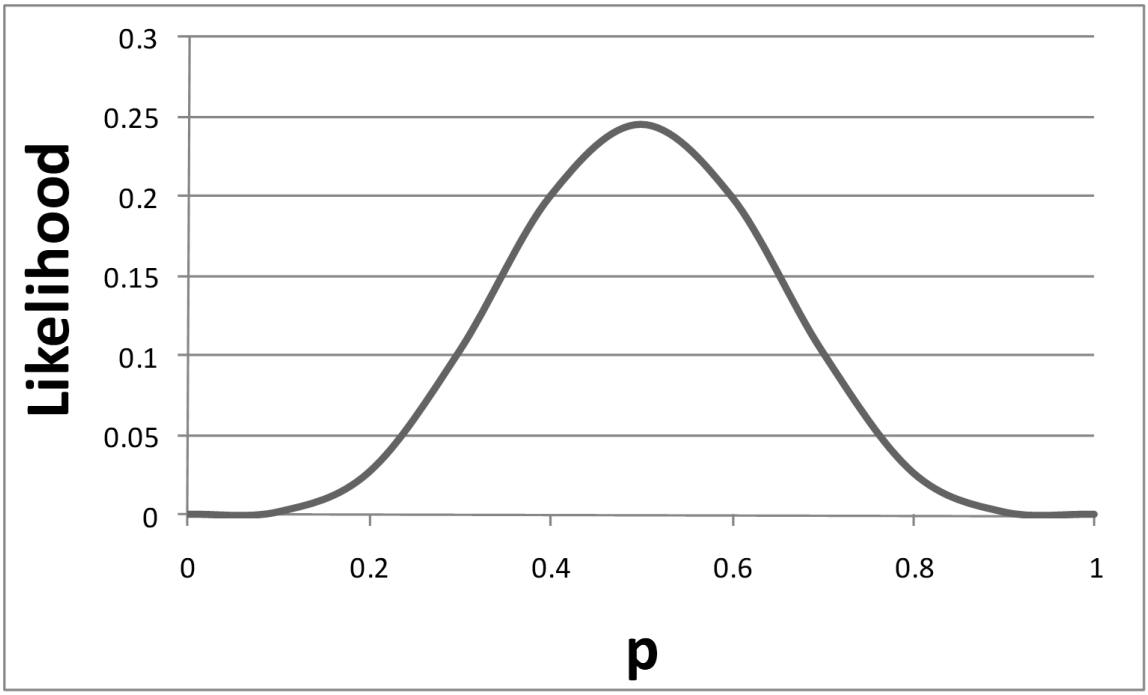


Figure 6. Likelihood function for p (probability of a serum sample being positive), given that we observe 5 positive sera in a panel of 10 sera.

How do we calculate the likelihood function?

We know that X ~ Bin(10,p). i.e. number of positive sera (X) follows the binomial distribution, where the sample size is 10, and with probability p.

Hence the likelihood function of p, given that 5 positive sera have been observed, is given by the probability of observing 5 positive sera, in a sample of 10, for given values of p. This probability is given by the following expression:

{ The number of ways of choosing 5 sera out of a sample of 10 to be positive ($=^{10}C_5$) }
×
{The probability that 5 are positive ($=p^5$)}
×
{The probability that the other 5 are negative ($=(1-p)^5$)}

Using mathematical notation, this would be written as follows:

$$L(p|X=5) = P(X=5|n=10,p) = {}^{n}C_r\, p^r(1-p)^{n-r} = {}^{10}C_5\, p^5(1-p)^5.$$

Notice that L(P=p,N=n|X=r) is proportional to P(X=r|P=p,N=n) .

Maximum likelihood estimation is a method of finding best fitting parameter values by maximising the likelihood function of the model given the observed data. We mentioned earlier

that it is actually related to the least squares method. This is due to the fact that if the observations to which we are fitting a model obey certain conditions, then the residual sum of squares (SSR) is actually the maximum likelihood estimate for the model. These conditions are as follows:

(i) the errors around each observation are uncorrelated,
(ii) the errors have equal variances and
(iii) the errors are normally distributed.

Example: The catalytic model

Let us return to to our catalytic model example. In this instance, we are trying to find the force of infection, λ, so that model predictions of the number of positive samples at each age, $E_a(\lambda)$, is close to observed number of positive samples, O_a.

The number of positive samples in age group *a* is binomially distributed, i.e. $E_a(\lambda) \sim \text{Bin}(N_a, p_a)$. This means that $E_a(\lambda)$ takes the distribution when there is a fixed number of trials (samples) N_a, each with fixed probability of being a success (positive) p_a. The likelihood that λ takes a particular value is the product of the likelihood that λ takes that value for each age group:

$$\begin{aligned} L(\lambda) &= \prod_a L(\lambda, a) \\ &= \prod_a P(\lambda \mid N_a, O_a) \\ &= \prod_a {}^{N_a}C_{O_a} p_a^{O_a} (1-p_a)^{N_a - O_a} \text{ where } p_a = \frac{E_a(\lambda)}{N_a} \end{aligned}$$

So now it is simply a matter of choosing the value of λ that will give us the largest value for the resulting expression.

The likelihood function tells us that we have found the “best fitting” parameters when we have maximised the likelihood. However, a different model may allow us to get an even better fit to data. Of course the ultimate best fit would be a model with a separate variable for every data point. This would reproduce the data perfectly - but at the cost of not giving us any useful insight at all! Such a model is called a **saturated model**.

The **model deviance** compares the goodness of fit of a model with that of the saturated model, and is defined as deviance $D = -2 \log \frac{L(\mathbf{x})}{L(\hat{\mathbf{x}})}$, where L(***x***) is the likelihood of a particular model given input parameters ***x***, L($\hat{x}$) is the likelihood of the saturated model (called the **saturated likelihood**).

This looks like a peculiar expression, but the deviance has useful properties. If we have two models M1 and M2 then we can find the difference (deviance of M1) – (deviance of M2). This is χ^2-distributed with degrees of freedom equal to (number of parameters in M1) – (number of parameters in M2). The deviance can also be used to do several useful things, such as calculating confidence intervals around best fitting parameters, testing hypotheses (for instance, to see if additional parameters in a more complex model significantly increase its

goodness of fit) and choose the most parsimonious model using criteria such as the Akaike Information Criterion (AIC). These issues are beyond the scope of the current lecture.

For the catalytic model discussed at the start of Block 2, you may recall that we had a rather complicated expression for the model deviance. You now have enough information to actually work out the deviance, although the algebra is a bit messy. First, write down the model loglikelihood log L(λ) and saturated loglikelihood L($\hat{\lambda}$) as follows:

$$\begin{aligned}\log L(\lambda) &= \log \prod_a {}^{N_a}C_{O_a} p_a{}^{O_a} (1-p_a)^{N_a - O_a} \\ &= \sum_a O_a \log p_a + (N_a - O_a)\log(1-p_a) + \log {}^{N_a}C_{O_a}\end{aligned}$$

where $p_a = E_a(\lambda)/N_a$

$$\begin{aligned}\log L(\hat{\lambda}) &= \log \prod_a {}^{N_a}C_{O_a} \hat{p}_a{}^{O_a} (1-\hat{p}_a)^{N_a - O_a} \\ &= \sum_a O_a \log \hat{p}_a + (N_a - O_a)\log(1-\hat{p}_a) + \log {}^{N_a}C_{O_a}\end{aligned}$$

where $\hat{p}_a = O_a/N_a$

Since the deviance is $D = 2(\log L(\hat{\lambda}) - \log L(\lambda))$, the constant $\log {}^{N_a}C_{O_a}$ (which appears in both model and saturated loglikelihood) cancels out, so it can be ignored. You'll then end up with the expressions used in the practical at the start of Block 2.

At this stage, it is not essential to follow the algebra. What is important is to understand the concepts behind fitting models to data. Opportunities to fit models are provided later on in the course.

Further reading

If you wish to explore the concepts that were introduced here in more depth, one very readable book is "The Ecological Detective: Confronting Models with Data" by Hilborn and Mangel. This book it is written for the ecologist (or epidemiologist) rather than the statistician. Chapter 5 deals with the least squares method while chapter 7 talks about maximum likelihood estimation.

A recent paper (Vanni et al. Calibrating models in economic evaluation. A seven-step approach. Pharmacoeconomics 2011; 29(1):35-49) discusses various goodness of fit metrics and methods of numerical optimisation, including very modern Bayesian techniques. This goes further than we cover in this course.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 11: Modelling the effects of non-random mixing on the transmission dynamics and control of infections

Lecture

Objectives

By the end of this session, you should:

- be able to define and set up matrices of "Who Acquires Infection from Whom" (WAIFW) to describe non-random ("heterogeneous") mixing between individuals
- be able to use force of infection estimates to calculate WAIFW matrices
- understand the possible effect of non-random (heterogeneous) mixing patterns between individuals on the transmission dynamics and control of infectious diseases

Introduction

During the last few sessions, you saw that the force of infection for several infections appeared to be higher for children than for adults. There are several possible reasons for these differences e.g.:

- age-dependent mixing patterns e.g. children are most likely to mix with other children (rather than with adults), who are also most likely to be infectious, at least for measles, rubella etc.
- age-dependent differences in susceptibility, e.g. are children more likely to be susceptible to infection than adults?
- genetic or other differences in susceptibility or exposure e.g. those most susceptible to infection are infected at a young age.
- etc?

Age-dependent mixing is probably the most likely explanation and (as a result) has received the most attention in the literature. It also has important implications for determining control strategies, especially those targeting certain subgroups.

Example:

Figure 1 shows contact patterns between children and adults for two different hypothetical populations. In population A, each child contacts four other children and two adults, and each adult contacts one child and two adults. In population B, each child contacts one other child and three adults, and each adult contacts three children and one adult.

Suppose the same proportion of children in both populations have been vaccinated against a new pandemic strain of influenza, but no adults are vaccinated. *In which population is the incidence likely to be smallest once the new strain is introduced?*

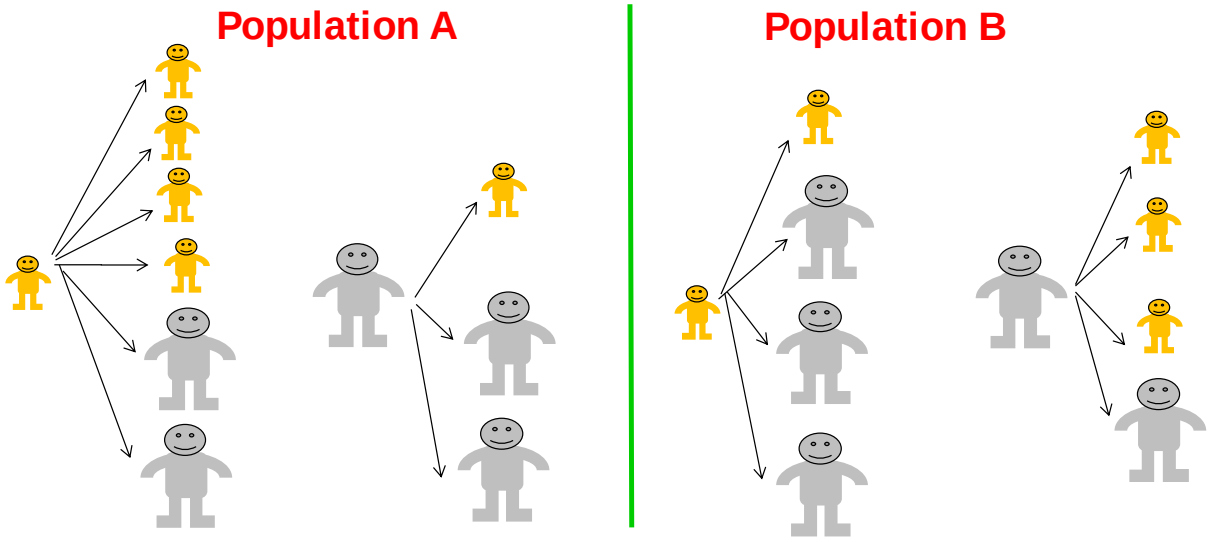


Figure 1: Example of two hypothetical contact patterns between children and adults. Children and adults are indicated by the little (yellow) and large (grey) characters respectively.

The previous question is intended to illustrate the fact that predicting the impact of an intervention that is introduced into a non-randomly mixing population is not straightforward.

For example, on the one hand, you might argue that the impact of vaccinating children will be greater in population A than in population B, since children in population A contact more individuals than do children in population B (i.e. 6 vs. 4).

However, we need to weigh this up against the fact that children in population A contact fewer adults than do children in population B (i.e. 2 vs. 3). Consequently, reducing the number of infectious children through vaccination will have a smaller impact on the infection incidence among adults in population A than in population B.

In fact, using a simple argument (see below), we can show that the impact of vaccinating children should be greatest in population B. However, we would need to use a transmission model if we wish to answer detailed questions such as "What would the incidence be in population A and B following the introduction of the new strain if, for example, 30% of children had been vaccinated?

Answer

We can answer this question by considering what the value for the net reproduction number would be if we were to vaccinate all children with a vaccine that has 100% efficacy. As discussed in previous sessions, the net reproduction number correlates with the trend in incidence. Considering population A in this situation, each adult would continue to infect two other adults, and so the net reproduction number would be 2. In population B, the net reproduction number would be 1 after vaccinating all children. Since the net reproduction number will be smaller in population B than in population A after vaccinating children, the incidence in population B should therefore be correspondingly smaller.

In order to make reasonable predictions of the impact of interventions against infections in real populations (in which individuals do not mix randomly) models need to include

assumptions about the amount of contact between individuals in different subgroups of the population.

Before describing the methods for incorporating non-random mixing into models, we first review the evidence for age-dependent mixing.

Evidence for age-dependent contact patterns

There is much evidence to suggest that contact patterns are age-dependent.

One study in The Netherlands examined the ages of pairs of tuberculosis cases who had onset during the period 1993-1996 and whose isolates shared identical *M tuberculosis* DNA fingerprint patterns (Figure 2a). It is reasonable to assume that these pairs represented a primary and a secondary case, although both cases could have also been infected by another person outside of the study (Borgdorff, et al 1999).

The study found that the ages of the two cases in the pairs were highly correlated, indicating that individuals are most likely to transmit infection to others of a similar age (Figure 1a). For example, the average age difference between individuals in a pair was 13.9 years (SD 12.2 years), which is statistically significantly smaller than the difference between that of randomly paired cases (25.5 years, 95% CI 21.5-29.5 years). Analogous age patterns were observed between presumed primary and secondary cases of measles and meningococcal meningitis in the UK from the period 1995-1998 (Figure 1b) (Edmunds, et al 2006).

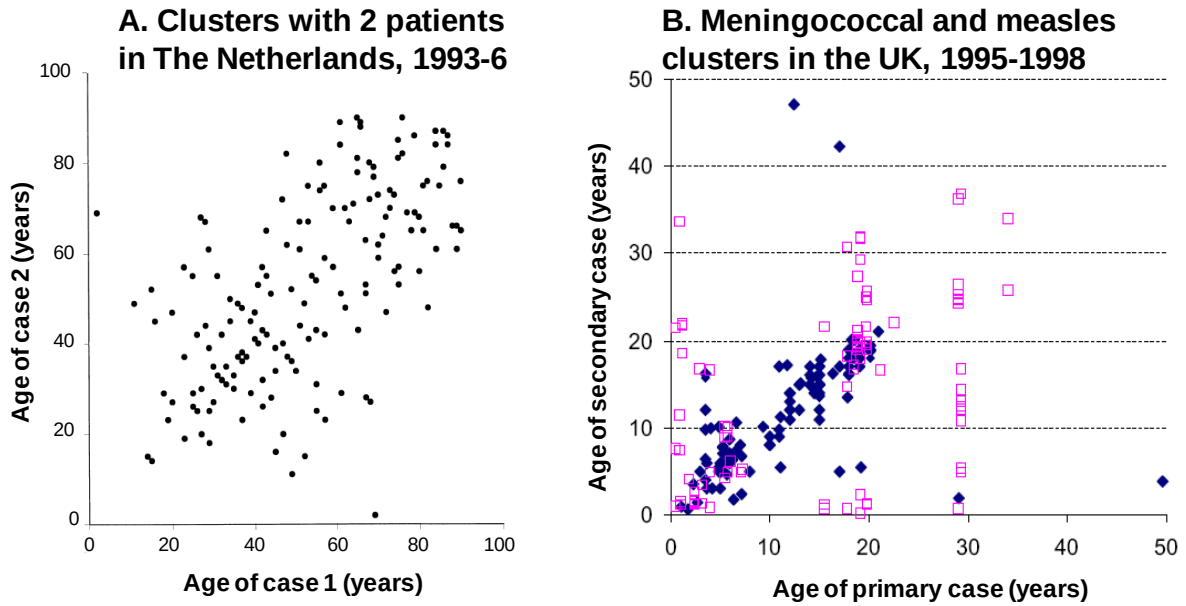


Figure 2: Age of first and second cases in clusters of A. TB cases in The Netherlands, defined using DNA fingerprinting (Borgdorff, et al 1999) B. Meningococcal and measles cases in the UK (Edmunds, et al 2006).

The largest study of contact patterns (the POLYMOD study) was published in 2008. Individuals from 8 European countries completed a diary detailing their physical and non-physical contacts on a single day between May 2005 and September 2006. A physical contact was defined as skin-to-skin contact (e.g. a kiss or a handshake) with another person;

a non-physical contact was defined as a two-way conversation with two or more words in the physical presence of another individual but with no physical contact.

In total, 7,290 diaries were collected, and the number of diaries collected per country ranged from 267 in The Netherlands to 1,328 in Germany (Mossong, et al. 2008). Similar contact patterns were observed for different countries, with individuals being most likely to contact others of a similar age (Figure 3). Substantial contact between 30-39 year olds and young children, as well as between middle-aged adults and older children was also found, presumably reflecting the contact between parents and their children.

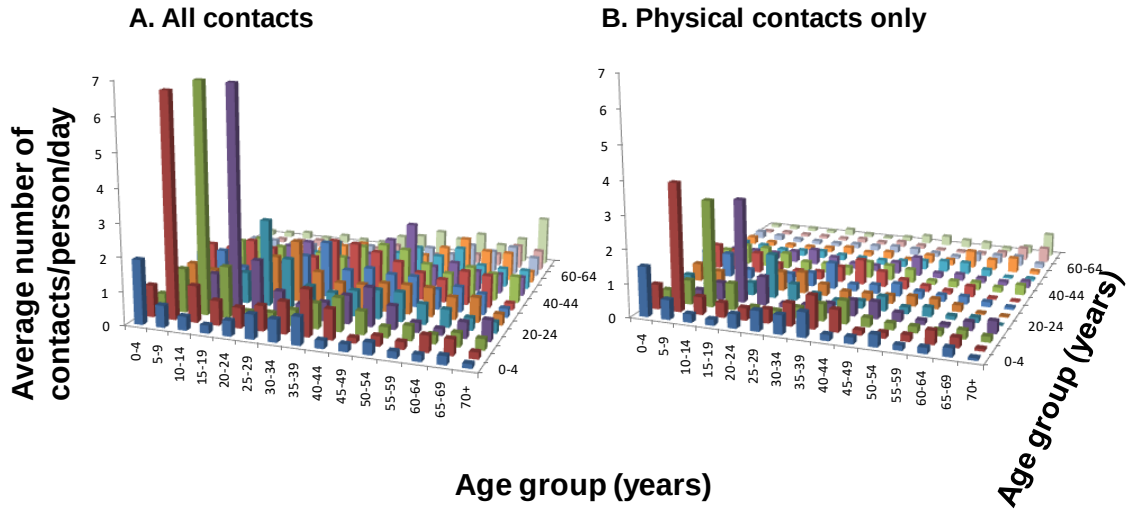


Figure 3: Data on average numbers of individuals contacted per day, May 2005-Sept 2006 in the UK from the POLYMOD study (Mossong et al (2008)

Revision of the relationship between the force of infection, the contact parameter (β) and the number of infectious individuals for a randomly mixing population

As discussed in previous sessions, when we assume that individuals mix randomly, the force of infection at a given time t (the rate at which susceptible individuals are infected per unit time) is given by the expression

$$\lambda(t) = \beta I(t)$$

where β is the rate at which 2 specific individuals come into effective contact per unit time, and $I(t)$ is the number of infectious individuals at time t. An effective contact is defined, as by Abbey (1948), as one that is sufficient to lead to transmission if it occurs between a susceptible and an infectious person.

β, in turn, is related to the number of contacts that each person makes with others and the proportion of contacts which are sufficient to lead to transmission of infection (which depends on the infection). However, as empirical data on the numbers of contacts that individuals typically make have been rare (at least, until recently) and the proportion of contacts that are sufficient to lead to transmission is poorly understood, β has usually been inferred from epidemiological data.

If we assume that individuals mix randomly, β can be calculated from the basic reproduction number using the following expression (see block 1):

$$\beta = R_0/(\text{average duration of infectiousness} \times \text{Total population size})$$

As we saw in the last session, R_0 can be calculated using the average force of infection or the average age at infection, as estimated from serological data.

Methods for incorporating heterogeneous mixing into models

In populations in which individuals (e.g. children and adults) do not mix randomly, the relationship between the force of infection and the contact parameter β is analogous to that for populations in which individuals are assumed to mix randomly, except that β needs to be stratified according to the subgroups considered in the model.

Example

Considering a population in which mixing patterns for children differ from those of adults, the overall force of infection experienced by children (the "young") at a given time t ($\overline{\lambda_y(t)}$) is given by the sum of the force of infection attributable to *contact with other children* ($\lambda_{yy}(t)$) and that attributable *to contact with adults* (the "old") ($\lambda_{yo}(t)$) as follows:

$$\overline{\lambda_y(t)} = \lambda_{yy}(t) + \lambda_{yo}(t) \qquad (1)$$

Similarly, the force of infection experienced by adults (the "old") at a given time t ($\overline{\lambda_o(t)}$) is given by the sum of the force of infection attributable to *contact with children* ($\lambda_{oy}(t)$) and that attributable *to contact with adults* (the "old") ($\lambda_{oo}(t)$) as follows:

$$\overline{\lambda_o(t)} = \lambda_{oy}(t) + \lambda_{oo}(t) \qquad (2)$$

Each of the components $\lambda_{yy}(t)$, $\lambda_{yo}(t)$, $\lambda_{oy}(t)$ and $\lambda_{oo}(t)$ can be expressed in terms of the product of the rate at which, e.g. a specific child comes into effective contact with another specific child per unit time and the number of infectious children as follows.

Derivation of the expression for the force of infection experienced by children which is attributable to contact with other children ($\lambda_{yy}(t)$)

As we saw in previous sessions, if we assume that individuals mix randomly, the number of new infections per unit time is given by the expression:

$$\beta S(t) I(t)$$

Extending this logic, the number of new infections among children that is attributable to contact with other children is given by the expression:

$$\beta_{yy} S_y(t) I_y(t) \qquad (A)$$

where β_{yy} is the rate at which a specific susceptible child and a specific infectious child come into effective contact, and $S_y(t)$ and $I_y(t)$ are the numbers of susceptible and infectious children at time t.

We can also express the number of new infections among children that is attributable to contact with other children using the following expression:

$$\lambda_{yy}(t)S_y(t) \qquad \text{(B)}$$

Equating expressions (A) and (B) we see that:

$$\lambda_{yy}(t)S_y(t) = \beta_{yy}\, S_y(t)I_y(t)$$

Cancelling out the term for the number of susceptible children $S_y(t)$ from both sides of the equation, we see that the force of infection experienced by children which is attributable to contact with other children is given by the expression:

$$\lambda_{yy}(t) = \beta_{yy}\, I_y(t) \qquad (C)$$

Derivation of the expression for the force of infection experienced by children which is attributable to contact with adults ($\lambda_{yo}(t)$)

By a similar argument to that applied to derive $\lambda_{yy}(t)$, the number of new infections among children which are attributable to contact with adults is given by the expressions

$$\beta_{yo}\, S_y(t)I_o(t) \qquad \text{(A')}$$

and

$$\lambda_{yo}(t)S_y(t) \qquad \text{(B')}$$

where β_{yo} is the rate at which a specific (susceptible) child comes into effective contact with a specific (infectious) adult per unit time and $S_y(t)$ and $I_o(t)$ are the numbers of susceptible children and infectious adults respectively at time t.

Equating expressions (A') and (B') we see that

$$\lambda_{yo}(t)S_y(t) = \beta_{yo}\, S_y(t)I_o(t)$$

Cancelling out the term for the number of susceptible children from both sides of the equation, we obtain the result that:

$$\lambda_{yo}(t) = \beta_{yo}I_o(t) \qquad (C')$$

Substituting expressions (C) and (C') for the force of infection among children attributable to contact with other children and with adults into expression (1), we obtain the result that:

$$\begin{aligned}\overline{\lambda_y(t)} &= \lambda_{yy}(t) + \lambda_{yo}(t) \\ &= \beta_{yy}\, I_y(t) + \beta_{yo}I_o(t)\end{aligned}$$

The same logic can be applied in expression (2) to derive the following expression for the force of infection among adults:

$$\overline{\lambda_o(t)} = \beta_{oy}I_y(t) + \beta_{oo}I_o(t)$$

where β_{oy} *is* the rate at which a specific (susceptible) adult and a specific (infectious) child come into effective contact per unit time and β_{oo} is the rate at which a specific (infectious) adult comes into effective contact with a specific (susceptible) adult per unit time.

Sidenote on notation for the subscripts
Note that throughout the text, the first subscript for β reflects the category of the susceptible person or the recipient of the infection, and the second subscript reflects the category of the infectious person. Thus, β_{yo} is the rate at which a a specific (susceptible) child and a specific (infectious) adult come into effective contact per unit time.

Exercise:
Derive the result that $\overline{\lambda_o(t)} = \beta_{oy}I_y(t) + \beta_{oo}I_o(t)$

Revision of matrices

Matrices provide a convenient means of summarizing sets of equations which need to be satisfied simultaneously. For example, if the following equations have to be satisfied simultaneously:

$$5x + 3y = 6$$
$$3x + 4y = 3$$

they could be summarized using the following notation:

$$\begin{pmatrix} 5 & 3 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$$

Notice that the matrix $\begin{pmatrix} 5 & 3 \\ 3 & 4 \end{pmatrix}$ just consists of the numbers in front of the variables x and y (ie their "coefficients") in the simultaneous equations and the "vector" $\begin{pmatrix} 6 \\ 3 \end{pmatrix}$ contains the constant terms of these equations.

Similarly, the equations

$$4x + 8y + 3z = 18$$
$$2x + y + 5z = 12$$
$$x + 3y + 8z = 4$$

could be written using matrix notation as follows:

$$\begin{pmatrix} 4 & 8 & 3 \\ 2 & 1 & 5 \\ 1 & 3 & 8 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 18 \\ 12 \\ 4 \end{pmatrix}$$

Extending this logic, the equations for the simultaneous equations for the force of infection among children and adults

$$\overline{\lambda_y(t)} = \beta_{yy} I_y(t) + \beta_{yo} I_o(t)$$
$$\overline{\lambda_o(t)} = \beta_{oy} I_y(t) + \beta_{oo} I_o(t)$$

would normally be summarized using the following matrix notation:

$$\begin{pmatrix} \overline{\lambda_y(t)} \\ \overline{\lambda_o(t)} \end{pmatrix} = \begin{pmatrix} \beta_{yy} & \beta_{yo} \\ \beta_{oy} & \beta_{oo} \end{pmatrix} \begin{pmatrix} I_y(t) \\ I_o(t) \end{pmatrix}$$

and the "beta" matrix is known as the "WAIFW" matrix or the matrix of "Who Acquires Infection From Whom" (Anderson and May, 1991) (pronounced "WAYFU", "WHYFU" or "WAYFWER"). You may also come across other notations in the literature

eg $\lambda(a,t) = \sum_j \beta_j I_j(t)$

$\lambda(a,t) = \int \beta(a,a')I(a',t)da'$.

Exercise:

a) Write down the following simultaneous equations using matrix notation:

i) $3x + 2y = 6$, $5x + 5y = 3$ *ii)* $2x + 6y = 5$, $3x + 4y = 3$

b) Write down the simultaneous equations corresponding to the following matrix equations:

i) $\begin{pmatrix} 1 & 5 \\ 3 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ *ii)* $\begin{pmatrix} 8 & 1 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$

Possible structures for WAIFW matrices

Values for β_{yy}, β_{yo}, β_{oy}, *and* β_{oo} can be calculated using the above matrix equation, given estimates of age-specific force of infections ($\overline{\lambda_y(t)}$ and $\overline{\lambda_o(t)}$) (eg estimated from age-specific seroprevalence data (see the session on analysing seroprevalence data) and of the number of infectious individuals in the different age-bands (i.e. $I_y(t)$ *and* $I_o(t)$*).* However, since the matrix equation summarizes *2 equations in 4 unknowns,* it is not possible to calculate unique values for each of the 4 beta parameters β_{yy}, β_{yo}, β_{oy}, *and* β_{oo}.

The *structure* of the WAIFW matrix therefore has to be constrained so that the equations

$$\overline{\lambda_y(t)} = \beta_{yy} I_y(t) + \beta_{yo} I_o(t)$$
$$\overline{\lambda_o(t)} = \beta_{oy} I_y(t) + \beta_{oo} I_o(t)$$

reduce to *2 equations in 2 unknowns*. The constraints applied depend on what appears to be most realistic in that population as follows.

The commonest (and most realistic) constraint applied is one which assumes that *the rate at which that a child contacts and transmits infection to an adult is the same as the rate at which an adult contacts and transmits infection to a child*, i.e. $\beta_{oy} = \beta_{yo}$. This assumption reduces the matrix equation to *2 equations in 3 unknowns*; thus a further constraint is required to reduce the equation further.

One possible constraint is that the probability that an *adult contacts and infects a child* is the same as the probability than an *adult contacts and infects an adult* (i.e. $\beta_{oy} = \beta_{oo}$). This leads to the following structure for the WAIFW matrix: $\begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_2 \end{pmatrix}$.

Another possible structure for the WAIFW matrix is $\begin{pmatrix} \beta_1 & \beta_1 \\ \beta_1 & \beta_2 \end{pmatrix}$, which reflects the assumption that a child is equally likely to contact and infect another child or an adult, and that adults have "special" kinds of mixing patterns with other adults.

Other possible structures are: $\begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_1 \end{pmatrix}$, and $\begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix}$. The second structure is unrealistic when considering age-dependent mixing patterns -- it is unlikely that children disassociate themselves completely from adults.

Each of the above matrices is *symmetrical.*

The following are examples of *asymmetric* matrices: $\begin{pmatrix} \beta_1 & \beta_2 \\ \beta_1 & \beta_2 \end{pmatrix}$, and $\begin{pmatrix} \beta_1 & \beta_1 \\ \beta_2 & \beta_2 \end{pmatrix}$

Asymmetric matrices are generally considered to be unrealistic for many infections, since they assume, e.g. that the rate at which *a child contacts and infects an adult* differs from the rate at which *an adult contacts and infects a child*. On the other hand, they may be realistic for infections transmitted via the faecal-oral route.

Exercise:
Write down the WAIFW matrix for the assumption that adults only mix with children and children only mix with adults.

Example of a WAIFW calculation

Since the parameters β_{yy}, β_{yo}, β_{oy} *and* β_{oo} are constants they can be estimated for a given WAIFW structure if the values for the force of infection and the number of infectious individuals at any given time, usually taken to be the equilibrium, are known.

Suppose that in your population, the average annual force of infection for rubella is 0.12 and 0.05 per year for individuals aged under and over 15 years respectively, and that in your

model population, these age-specific forces of infections result in 29 and 6 infectious persons among children and adults respectively. (see methods below). The equations for the average force of infection in relation to the equilibrium number of infectious individuals can therefore be written as: $\begin{pmatrix} 0.12 \\ 0.05 \end{pmatrix} = \begin{pmatrix} \beta_{yy} & \beta_{yo} \\ \beta_{oy} & \beta_{oo} \end{pmatrix} \begin{pmatrix} 29 \\ 6 \end{pmatrix}$

If you decide that the WAIFW matrix should have the following (unrealistic) structure: $\begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix}$ then you would need to solve the matrix equations: $\begin{pmatrix} 0.12 \\ 0.05 \end{pmatrix} = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix} \begin{pmatrix} 29 \\ 6 \end{pmatrix}$ which can be written simply as:

$$0.12 = 29\ \beta_1$$
$$0.05 = 6\ \beta_2$$

Solving these equations gives: β_1 = 0.00413 per year and β_2 = *0.008* per year.

Alternatively, if the WAIFW matrix has the following structure: $\begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_2 \end{pmatrix}$ then you would need to solve the matrix equations: $\begin{pmatrix} 0.12 \\ 0.05 \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_2 \end{pmatrix} \begin{pmatrix} 29 \\ 6 \end{pmatrix}$ which can be written as:

$$0.12 = 29\beta_1 + 6\beta_2$$
$$0.05 = (29 + 6)\ \beta_2$$

You can calculate β_2 directly from the second equation (which gives β_2= 0.0014 per year). By substituting for β_2 into the first equation, and rearranging you should get β_1 = 0.0038 year.

Exercise

Calculate the appropriate values for β_1 and β_2 assuming that mixing is described using WAIFW structures $\begin{pmatrix} \beta_1 & \beta_1 \\ \beta_1 & \beta_2 \end{pmatrix}$ and $\begin{pmatrix} \beta_1 & \beta_2 \\ \beta_1 & \beta_2 \end{pmatrix}$.

Note on alternatives to WAIFW matrices

WAIFW matrices have become the standard method of incorporating non-random mixing in transmission models, though some workers have also explored the possibility of using continuous functions to describe how the contact parameter changes with age (Massad (1994)).

Methods for calculating the number of susceptible and infectious individuals in a population

In general, if we assume that most individuals become infectious shortly after infection, the average number of infectious individuals in a population can be calculated using the approximation:

Prevalence ≈ Incidence × duration of infectiousness

Therefore, the number of infectious individuals can be calculated using the expression:

Average number of infectious persons ≈
Average number of new infections per unit time × duration of infectiousness

As discussed previously, if we assume that individuals mix randomly, the number of new infections at a given time *t* is given by the expression

$$\lambda(t)S(t)$$

where $S(t)$ is the number of susceptible individuals at time *t* and $\lambda(t)$ is the force of infection at time *t*.

Extending this logic, if we assume that individuals mix randomly, the average number of new infections is given by the expression

$$\lambda S$$

where λ is the average force of infection and S is the average number of susceptible individuals.

Likewise, the average numbers of new infections among children and adults are given by the expressions

$$\lambda_y S_y \quad \textit{(children)}$$
$$\lambda_o S_o \quad \textit{(adults)}$$

where S_y and S_o are the average numbers of children and adults that are susceptible.

These can be calculated by integrating expressions for the age-specific numbers of individuals that are susceptible over the age ranges of interest. This is equivalent to summing the area under the curve of the age-specific numbers of individuals that are susceptible over the age range of interest.

For example, as discussed in the last session, the proportion of individuals of age *a* that are susceptible ($s(a)$) follows a pattern similar to that shown in Figure 4a, depending on how the force of infection changes with age. If the force of infection is the same for all age groups, the equation for this line is

$$s(a) = e^{-\lambda a}$$

Expressions for $s(a)$ when the force of infection is age-dependent are discussed in the previous sesson.

The expression for the number of individuals of age *a* that are susceptible is *N(a)s(a)* where *N(a)* is the number of individuals that are of age *a* in the population. The curve of the number of individuals that are susceptible follows a similar pattern to that for *s(a)*, depending further on the age distribution of the population.

For example, if it is rectangular, i.e there are equal numbers of individuals in each age group, the number of susceptible individuals of age *a* also follows this pattern (Figure 4b). The total numbers of individuals that are susceptible in the population can be calculated as the area under the curve of Figure 4b between the ages 0 and life expectancy. Likewise, the total numbers of susceptible children in the population can be estimated by calculating the area under the curve of Figure 4b between the ages 0 and the maximum age of a child. Methods for integrating expressions are provided in the maths refresher.

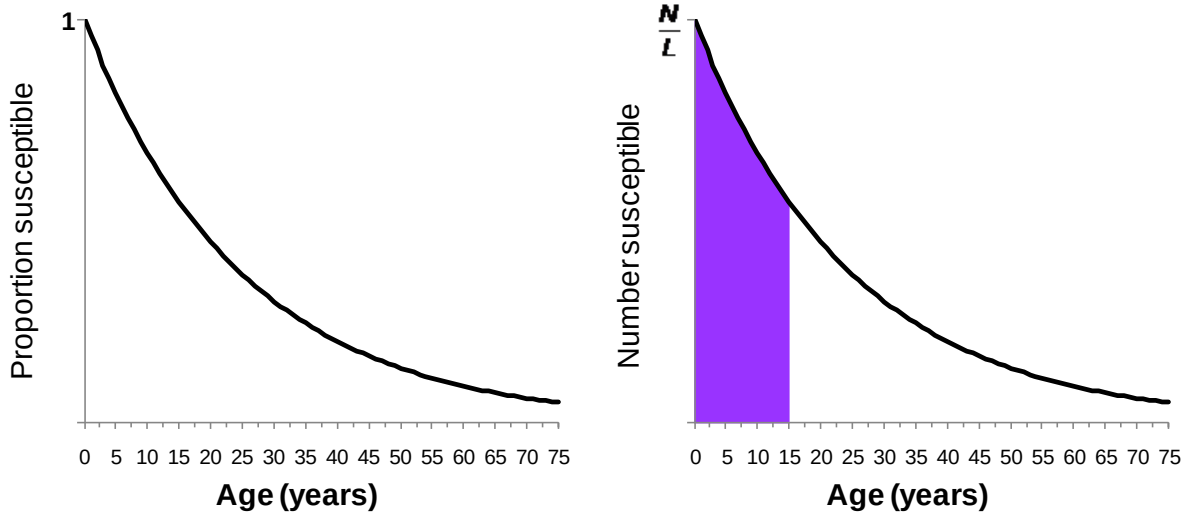
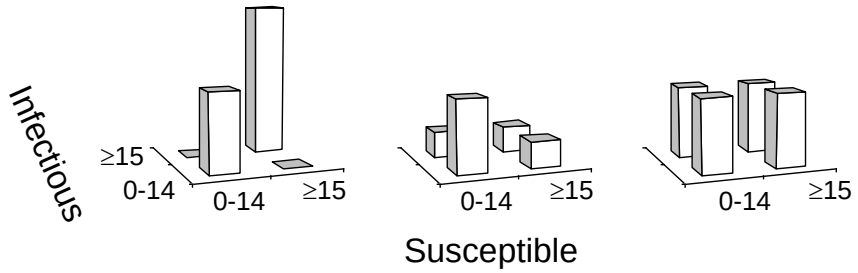


Figure 4: A. General pattern in the proportion of individuals of a given age *a* that are susceptible. B. Number of individuals of a given age *a* that are susceptible assuming that the proportion that is susceptible at each age *a* follows the pattern in Figure A and that the age distribution is rectangular, of size *N*, and a life expectancy of *L*. If children are aged 0-15 years, the shaded area denotes the total number of susceptible children.

The effect of different WAIFW structures on the transmission dynamics and control of infections

WAIFW matrices are often represented graphically and this helps to identify the implications of different contact patterns between individuals on control of transmission. The following diagram compares the magnitude of different β values describing age-dependent transmission in three different populations.



Exercise:
In which of these populations should it be easiest to control transmission through infant vaccination? Why?

This theme will be continued in the next lecture, which focuses on methods for estimating the basic reproduction number for non-randomly mixing populations.

Summary

Several studies suggest that contact patterns between individuals are strongly age-dependent, although the nature of this age-dependency is still poorly understood. Contact patterns greatly influence the impact of interventions against infections, and therefore models need to take account of this if they are to be used to predict the effect of control. Assumptions of non-random mixing between individuals are incorporated into models using matrices of "Who Acquired Infection from Whom". These matrices can be calculated using estimates of the age-specific force of infection and the average numbers of infectious persons after imposing some constraints on the structure of the matrices.

Further reading

Vynnycky E and White RG (2010) An introduction to infectious disease modelling. Oxford University Press. Chapter 7, pp 177-203.

References

Anderson RM and May RM Infectious Diseases in Humans: Dynamics and Control. Oxford University Press. 1992, [ch 8,9]

Anderson RM and May (1985) Age-related changes in the rate of disease transmission: implications for the design of vaccination programmes. *J of Hygiene (Camb)*, 1985, 94, 365-436

Borgdorff MW, Nagelkerke NJ, van Soolingen D, Broekmans JF. Transmission of tuberculosis between people of different ages in The Netherlands: an analysis using DNA fingerprinting. Int J Tuberc Lung Dis. 1999 Mar;3(3):202-6.

Diekmann O, Heesterbeek JAP, Metz JAJ (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, *J Math Biol*, 28, 365-82.

Dietz K (1993) The estimation of the basic reproduction number for infectious diseases, *Statistical Methods in Medical Research*, 2, 23-41

Edmunds WJ, Kafatos G, Wallinga J, Mossong JR. Mixing patterns and the spread of close-contact infectious diseases. Emerg Themes Epidemiol. 2006 Aug 14;3:10.

Massad E, Burattini MN, de Azevedo Neto RS, Yang HM, Coutinho FA, Zanetta DM A model-based design of a vaccination strategy against rubella in a non-immunized community of Sao Paulo State, Brazil *Epidemiol Infect* 1994; 112:579-94

Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLoS Med, 2008, 5(3): e74 doi:10.1371/journal.pmed.0050074

Introduction to Infectious Disease Modelling and its Applications – 2018

Sessions 12-13: Groupwork

In these sessions, you are asked to work in your allocated groups to address the problem that you have been assigned.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 14: Estimating the basic reproduction number for non-randomly mixing populations

Lecture

Overview and Objectives

During the last session you saw how the effect of vaccination on reducing transmission in a population depended greatly on assumptions about contact between individuals. This suggests that the herd immunity threshold and therefore R_0 must depend on the amount of contact between different subgroups in the population. In this session we illustrate how R0 can be calculated by taking account of contact patterns between different population subgroups.

By the end of this session you should understand:

- the importance of accounting for heterogeneity in contact between individuals when calculating the basic reproduction number and critical vaccination coverage for controlling transmission,
- how the basic reproduction number and herd immunity threshold are calculated for both randomly and heterogeneously mixing populations.

Revision of the relationship between the net and basic reproduction numbers, and herd immunity threshold for randomly (homogeneously) mixing populations

As covered in earlier sessions of the course, the basic reproduction number for a randomly mixing population can be calculated using the expression $R_0=\beta ND$, where β is the rate at which two specific individuals come into effective contact per unit time, N is the population size, and D is the average duration of infectiousness.

The net and basic reproduction numbers are related through the following equation:

$$R_n = R_0 \times s$$

where s is the proportion of individuals who are susceptible in the population. This expression provides a useful method for calculating the basic reproduction number. For example, when the infection is at equilibrium, each infectious person must be leading to one secondary infectious person, and thus the net reproduction number must be equal to 1. Substituting for R_n=1 into the above expression and then rearranging it gives:

$$R_0 = 1/ s^*$$

where s^* is the proportion of the population which is susceptible to infection at equilibrium.

Other expressions for the basic reproduction number for randomly mixing populations, such as the following, derive from this expression.

a) $R_0 = L/A$

This equation is applicable for populations with a rectangular age distribution similar to that in industrialized settings, in which the average age at infection is A and L is the life expectancy.

b) *$R_0 = N/(B(A-m))$*

This equation applies for populations with any general demography in which the average age at infection (*A*) is small, *N* is the total population size, *B* is the number of infants who survive and *m* is the duration of maternal immunity. By definition, when the average at infection, *A*, is small, $B \times (A-m)$ is the number of individuals who are susceptible in the population, and thus $B \times (A-m)/N$ is the proportion of individuals in the population who are susceptible.

c) *$R_0 = 1/(1-I^*)$*

This equation applies for infections which do not confer immunity ("SIS" infections). Here I^* is the equilibrium prevalence of infection in the population and thus $1-I^*$ is the equilibrium proportion of individuals who are susceptible.

Rearranging the expression $R_0 = 1/s^*$ leads to the following equation for the (critical) proportion of the population which needs to be susceptible to control transmission:

$$s^* = 1/R_0$$

To control transmission, the net reproduction number R_n must be maintained below one and thus the proportion of the population which is susceptible, *s* must be maintained below the critical value s^*. i.e. $R_n<1$ and $s<s^*$. This can be achieved by immunizing the proportion of the population given by the herd immunity threshold (H):

$$H = 1-s^* = 1-1/R_0$$

Estimating the basic reproduction reproduction number for heterogeneously mixing populations

For a randomly mixing population, the R_0 would typically be estimated by carrying out using the following steps:

1. Measure the prevalence of previous infection in the population, using a serosurvey.
2. Assuming that the infection is at equilibrium, calculate s^*, the proportion of the population which is susceptible.
3. Estimate R_0 using the expression $R_0 = 1/s^*$.

Note that it is also possible to estimate the R_0 using the growth rate in the cumulative numbers of infections (see session on the basic dynamics of infectious diseases in block 1) or using data on the secondary attack rate (see Fine et al, 1988).

For a heterogeneously mixing population, it would be estimated by carrying out the following steps:

1. Measure the prevalence of previous infection in the population, using a serosurvey.
2. Estimate the forces of infection.
3. Choose the structure of the matrix of "Who Acquires Infection From Whom".
4. Calculate the transmission coefficients β corresponding to the chose WAIFW matrix.
5. Formulate the "Next Generation Matrix" (NGM).
6. Calculate R_0 from the Next Generation Matrix".

Steps 1-4 have been covered in earlier sessions of the course. We elaborate on what is meant by the next generation matrix and how it is calculated.

Calculating the Next Generation Matrix

For populations in which individuals mix heterogeneously e.g. in which there are 2 subgroups, with individuals in one group mixing intensively with their own group, and less with individuals in the other group, the number of secondary infectious persons resulting from one infectious person when it is introduced into a totally susceptible population must depend on the subgroup to which that individual belongs.

For example, returning to the population considered during the lecture on defining WAIFW matrices, in which individuals were stratified into the young and the old, the number of secondary infectious persons resulting from each young infectious person will be different from the number of secondary infectious persons resulting from each old infectious person.

Also, the number of *old* infectious persons generated by each young infectious person will be different from the number of *young* infectious persons generated by each young infectious person. Likewise, the number of *old* infectious persons generated by each old infectious person will be different from the number of *old* infectious persons generated by each young infectious person. Each of these numbers can be expressed in terms of the β coefficients of the WAIFW matrix describing mixing between individuals, the number of individuals in each age group and the duration of infectiousness, as follows.

For example, as you saw earlier, the basic reproduction number for a population in which individuals are assumed to mix randomly can be written as follows:

$$R_0=\beta N D$$

where β is the rate at which two specific individuals come into effective contact per unit time, N is the total population size, and D is the duration of infectiousness.

Extending this logic, the number of infectious persons among *young* individuals resulting from the introduction of one infectious young person into a totally susceptible population is given by:

$$R_{yy} = \beta_{yy} N_y D$$

where β_{yy} is the rate at which 2 specific young individuals come into effective contact per unit time, and N_y is the total number of young individuals in the population.

Similarly, the number of infectious persons among *old* individuals resulting from the introduction of one infectious young person into a totally susceptible population is given by:

$$R_{oy} = \beta_{oy} N_o D$$

where β_{oy} is the rate at which a specific young infectious person comes into effective contact with a specific old susceptible individual, and N_o is the total number of old individuals in the population.

The expressions for the number of secondary infectious persons among young individuals resulting from each old infectious person and the number of secondary infectious persons among old individuals resulting from each old infectious person (R_{yo} and R_{oo} respectively) are analogous:

$$R_{yo} = \beta_{yo} N_y D$$

and

$$R_{oo} = \beta_{oo} N_o D$$

Notice the direction of the subscripts used in the notation for the reproduction numbers above. In each equation (e.g. R_{oy}) the first component of the subscript reflects the category of individuals among whom the secondary infectious persons occur (i.e. old individuals when considering R_{oy}), and the second component of the subscript reflects the category of the infectious person who is transmitting the infection (i.e. young individuals when considering R_{oy}) .

The next generation matrix is defined as the matrix which summarizes the number of secondary infectious persons in a given category resulting from individuals in each of the categories. In the above example, the next generation matrix would be given by the following matrix:

$$\begin{pmatrix} R_{yy} & R_{yo} \\ R_{oy} & R_{oo} \end{pmatrix} = \begin{pmatrix} \beta_{yy} N_y D & \beta_{yo} N_y D \\ \beta_{oy} N_o D & \beta_{oo} N_o D \end{pmatrix}$$

In populations such as the above, the number of secondary infectious persons resulting from the introduction of an infectious infectious person into a totally susceptible population will be some average of each of the numbers in this matrix. The theory was developed by Heesterbeek et al during the 1990s (see Diekmann et al (1990)).

Examples

1. In a population with the following next generation matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, an infectious person in either of the subgroups (i.e. a typical infectious person) causes one secondary infectious infectious person in its own subgroup and one secondary infectious person in the other

subgroup and therefore leads to two secondary infectious persons. The basic reproduction in this population is therefore 2.

2. In a population with the following next generation matrix $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ one infectious person in either of the subgroups (i.e. a typical infectious person) leads to 2 infectious persons in its own subgroup and 1 in the other subgroup and therefore causes 3 infectious persons in total. The R_0 in this population is therefore 3.

3. In a population with the following next generation matrix $\begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix}$, an infectious person in the first subgroup causes 1 secondary infectious person in the second subgroup, and a infectious person in the second subgroup causes 4 secondary infectious persons in the 1st subgroup. In this instance, the value for the basic reproduction number is not straightforward, and a "typical" infectious person is best defined as one which partially belongs to both groups 1 and 2. If a fraction x of this typical infectious person belongs to group 1, by definition, a fraction (1-x) must belong to group 2. We can represent this infectious individual using "vector" notation: $\begin{pmatrix} x \\ 1-x \end{pmatrix}$ and we consider how we can identify the value for x below.

Defining the "typical" infectious person

There is a mathematical proof (derived by Heesterbeek et al, which is beyond the scope of this course – see reference list) which shows that, by repeatedly applying the Next Generation Matrix to some vector representing an initial infectious person introduced into a totally susceptible population, in which there is an "infinite supply of susceptible individuals", then

- **a)** The number of secondary infectious persons resulting from each infectious person in each generation converges to the basic reproduction number and
- **b)** The distribution of the infectious persons in each generation converges to some distribution. which reflects that of the "typical" infectious person.

This result is illustrated in table 1 considering a population in which the Next Generation Matrix is $\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$.

This particular finding leads to the result that it is possible to find the basic reproduction number and the distribution of a "typical" infectious person either by:

i) Repeatedly multiplying the Next Generation Matrix to the vector representing an infectious individual (i.e. $\begin{pmatrix} x \\ 1-x \end{pmatrix}$) and calculating the ratio between the number of infectious persons in each generation and that in the preceding generation, or

ii) Finding the values for R_0 and x for which the following matrix equation holds:

$$\begin{pmatrix} R_{yy} & R_{yo} \\ R_{oy} & R_{oo} \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix}$$

Example 3 (continued)

Returning to example 3 above and considering the second method for calculating the basic reproduction number, we see the number of secondary infectious persons resulting from the introduction of this typical infectious person into a totally susceptible population has to satisfy the following equation:

$$\begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix}$$

i.e. a typical infectious person generates R_0 replicas of itself.

This equation can be written equivalently as $\begin{pmatrix} 4-4x \\ x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix}$

or, using simultaneous equations, as:

$4-4x=R_0x$ Equation 1

$x=R_0(1-x)$ Equation 2

These equations can be solved (see Appendix) to give x=2/3 (i.e. 2/3 of each typical infectious person belongs to group 1 and the other third belongs to group 2) and an R_0 of 2, which implies that 50% (1-1/R_0) of the overall population would need to be immune to control transmission.

If you substitute x=2/3 and R_0=2 into the equation $\begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix}$, you should see that they are solutions of this equation.

The fact that 50% of the population would need to be immune to control transmission is illustrated as follows. If 50% of the population is immune in the population, then the next generation matrix can be written as $\begin{pmatrix} 0 & 0.5\times 4 \\ 0.5\times 1 & 0 \end{pmatrix}$, which is equivalent to $\begin{pmatrix} 0 & 2 \\ 0.5 & 0 \end{pmatrix}$.
Each typical infectious person generates R_n (the net reproduction number) replicas of itself and the net reproduction number has to satisfy the following equation:

$$\begin{pmatrix} 0 & 2 \\ 0.5 & 0 \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_n\begin{pmatrix} x \\ 1-x \end{pmatrix}$$

Based on the logic described above, 2/3 of the typical infectious person belongs to group 1 and the other third belongs to group 2 (i.e. x=2/3). Substituting for x=2/3 into the above equation leads to the result that the net reproduction number is one.

Table 1: Summary of the effect of introducing one young infectious person into a population in which individuals are either young or old and the next generation matrix is $\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$

	Generation number:									
	0	1[a]	2[b]	3[c]	4	5	6	7	8	9
Number of infectious persons who are:										
young	1	1	2	7	29	124	533	2293	9866	42451
old	0	1	5	22	95	409	1760	7573	32585	140206
total	**1** (=1+0)	**2** (=1+1)	**7** (=2+7)	**29** (=7+22)	**124** (=29+95)	**533** (=124+409)	**2293** (=533+1760)	**9866** (=2293 +7573)	**42451** (=9866 +32585)	**182657** (=42451 +140206)
Ratio between the number of infectious persons in the current generation and that in the preceding generation:	-	**2** (=2/1)	**3.5** (=7/2)	**4.1429** (=29/7)	**4.2759** (=124/29)	**4.2984** (=533/124)	**4.3021** (=2293/533)	**4.3027** (=9866/ 2293)	**4.3028** (=42451/ 9866)	**4.3028** (=182657/ 42451)
Proportion of infectious persons who are:										
young	1	0.5	0.2857	0.2414	0.2339	0.2326	0.2324	0.2324	0.2324	0.2324
old	0	0.5	0.7143	0.7586	0.7661	0.7674	0.7676	0.7676	0.7676	0.7676

[a] **Calculating the values for generation 1:** The young infectious person introduced into the totally susceptible population should generate 1 young infectious person and 1 old infectious person in the next (first) generation. This number can also be obtained by multiplying the matrix $\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$ by the vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. In total there will be 2 infectious persons in the first generation and the ratio between the number of infectious persons in the first generation and that in the initial ("0^{th}") generation is 2/1 = 2.

[b] **Calculating the values for generation 2:** The young infectious person from the second generation should generate 1 young secondary infectious person and 1 old secondary infectious person. The old infectious person from the second generation should generate 1 young secondary infectious person and 4 old secondary infectious persons. In total, there should be 2 young infectious persons and 5 old secondary infectious persons in the second generation. These numbers can also be obtained by multiplying the matrix $\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$ by the vector representing the number of infectious persons seen in the first generation i.e. $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. In total, there will be 7 infectious persons in the second generation and the ratio between the number of infectious persons in the second generation and that in the first generation is 7/2 =3.5.

[b] **Calculating the values for the 3rd and subsequent generations:** Repeating the same process to obtain the numbers of infectious persons seen in the third and subsequent generations, you should see that the ratio between the number of infectious persons seen in each generation and that in the preceding generation converges to 4.30 (which turns out to be the basic reproduction number for this next generation matrix) and that the age distribution of the infectious persons in each generation converges to some constant distribution (which is defined to be that of the "typical" infectious person). This distribution is independent of the distribution of the infectious persons introduced into the population at the start.

Calculating R_0 for "disassortative" Next Generation Matrices

In example 3, we saw that for the Next Generation Matrix $\begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix}$, R_0 equalled the square root of 4, (written $\sqrt{4}$). In fact, we can show that a Next Generation Matrix of the form $\begin{pmatrix} 0 & R_{12} \\ R_{21} & 0 \end{pmatrix}$, representing a population in which people (or groups) just mix with people (or groups) of the opposite type, R_0 is given by the square root of the two numbers in the matrix, i.e.

$$R_0 = \sqrt{R_{12}R_{21}} \qquad \text{Equation 3}$$

In this matrix, we have chosen to denote the types of people in the population using numbers 1 and 2, rather than "*y*" and "*o*" to avoid associating this matrix with contact patterns between "children" and "adults". For example, it is unlikely that children never contact old people. A matrix of the form $\begin{pmatrix} 0 & R_{12} \\ R_{21} & 0 \end{pmatrix}$ can be used to describe contact between a human and an animal vector population, or transmission of a sexually-transmitted infection in a heterosexually mixing population.

There are different ways of deriving Equation 3. For practice, you should try to work through the derivation below to convince yourself that the equation $R_0 = \sqrt{R_{12}R_{21}}$ holds. In reality, once you see that you have a Next Generation Matrix of the form $\begin{pmatrix} 0 & R_{12} \\ R_{21} & 0 \end{pmatrix}$, then, if you just remember that $R_0 = \sqrt{R_{12}R_{21}}$, you just need to apply it to calculate R_0. Also note that the square root of a number can be positive or negative and since it is not possible for R_0 to be negative, R_0 has to be the positive value of $\sqrt{R_{12}R_{21}}$.

Derivation of the result $R_0 = \sqrt{R_{12}R_{21}}$ for the Next Generation Matrix $\begin{pmatrix} 0 & R_{12} \\ R_{21} & 0 \end{pmatrix}$

The following derivation is one that perhaps requires the fewest steps:

1. Begin by noting that R_0 has to satisfy the following equation:

$$\begin{pmatrix} 0 & R_{12} \\ R_{21} & 0 \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix}$$

2. If we write out this equation, we obtain the following two equations:

$$R_{12}(1-x) = R_0 x \qquad \text{Equation 4}$$

$$R_{12}x = R_0(1-x) \qquad \text{Equation 5}$$

3. At this stage, we could rearrange Equation 4 to get an expression for R_0 in terms of everything else, substitute that expression for R_0 into Equation 5 and then rearrange that expression. An alternative approach, which requires fewer steps than this is to rearrange both equations to obtain an expression for x/(1-x), set those two equations equal to each other and rearrange that expression to get an expression for R_0, as follows:.

a) Dividing both sides of Equation 4 by (1-x) and then by R_0, we obtain the following equation for *x/(*1-*x):*

$$\frac{x}{1-x} = \frac{R_{12}}{R_0}$$ Equation 6

b) Dividing both sides of Equation 5 by (1-x) and then by R_{21}, we obtain the following equation for x/(1-x):

$$\frac{x}{1-x} = \frac{R_0}{R_{21}}$$ Equation 7

c) Setting Equation 6 equal to Equation 7 we see that

$$\frac{R_{12}}{R_0} = \frac{R_0}{R_{21}}$$ Equation 8

d) If we rearrange this equation by multiplying both sides of the equation by R_0 and R_{21}, we obtain the following result for R_0:

$$R_0^{\,2} = R_{12}R_{21}$$ Equation 9

or, equivalently:

$$R_0 = \sqrt{R_{12}R_{21}}$$ Equation 10

Exercise

1. Use the alternative approach mentioned in point c) to obtain the result $R_0 = \sqrt{R_{12}R_{21}}$

2. Calculate R_0 for the following Next Generation Matrices:

a) $\begin{pmatrix} 0 & 48 \\ 3 & 0 \end{pmatrix}$ b) $\begin{pmatrix} 0 & 2 \\ 32 & 0 \end{pmatrix}$ c) $\begin{pmatrix} 0 & 49 \\ 4 & 0 \end{pmatrix}$ d) $\begin{pmatrix} 0 & 13 \\ 4 & 0 \end{pmatrix}$

Calculating the net or effective reproduction number

The above methods can be applied to calculate the net or effective reproduction for a given population. Considering a population in which individuals are stratified into the young and the old, for example, the methods would be applied to the following next generation matrix:

$$\begin{pmatrix} \beta_{yy} S_y D & \beta_{yo} S_y D \\ \beta_{oy} S_o D & \beta_{oo} S_o D \end{pmatrix}$$

where S_y and S_o are the numbers of young and old susceptible individuals in the population. Calculations of the net reproduction number for the UK population based on matrices such as these, taking into account past levels of measles vaccination coverage in the UK, are currently used to work out the potential for a measles epidemic to occur (Gay (1995))

An example of the application of reproduction number estimates – measles in England

Measles vaccine was introduced in England in 1968, but its uptake was variable. The introduction of vaccination resulted in a decline in the notification rates for measles and by the early 1990s they had reached a very low level. During the first half of 1994, slight increases were seen in the notification rates; a large outbreak had occurred in Scotland during the period 1993-4 and there were concerns that a large epidemic, with more than 100,000 cases, was imminent.

Work was carried out by Gay et al (Gay et al (1995)) to estimate the net (or effective) reproduction number (R_n) in England and Wales and whether an epidemic might occur.

Figure 1 shows examples of the WAIFW matrices calculated using values for the force of measles infection estimated from data from England and Wales from before the introduction of measles vaccination. Due to uncertainty about the force of infection among those aged 10-14 years, several different assumptions about the amount of contact between individuals in this age group and other age groups were explored.

These estimates were then combined with estimates of the proportion of individuals in different age groups who were susceptible to measles infection in 1994. Figure 2 shows the estimates for the net reproduction number (R_n) obtained using different model assumptions for contact between 10-14 year olds. These results highlighted that the net reproduction number in 1994 was very close to 1 (the horizontal dashed line) and that there was potential for an epidemic to occur.

Further calculations carried out by Gay et al suggested if an outbreak occurred, it could involve more than 100,000 cases. These conclusions were supported by other studies using dynamic models (Babad et al (1995)). As a result of these analyses, a measles-rubella vaccination campaign was carried out in November 1994, targeting 95% of the 7 million 5-16 yr olds. No measles epidemic was recorded in the period afterwards.

This was perhaps the first time that modelling was used to guide policy in the UK. Since then, the potential for a measles epidemic to occur in England and Wales continues to be evaluated in the same way (Choi et al (2008)).

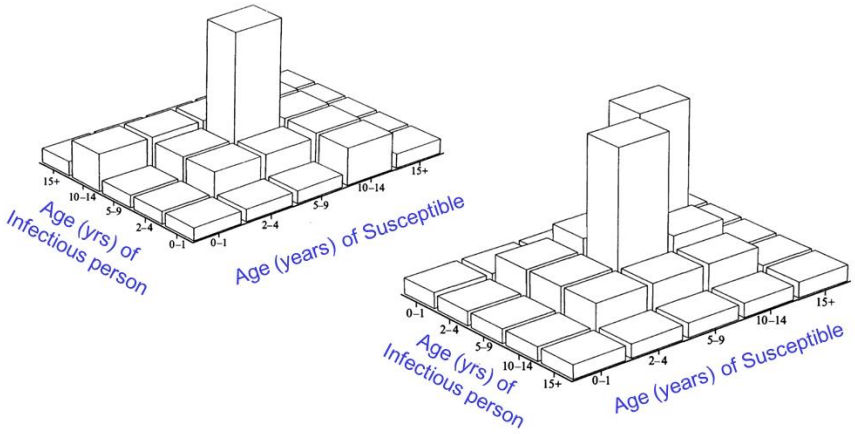


Figure 1: Examples of WAIFW matrices describing contact between different age groups in England and Wales obtained by Gay et al using estimates of the force of infection for measles calculated using data collected before the introduction of vaccination.

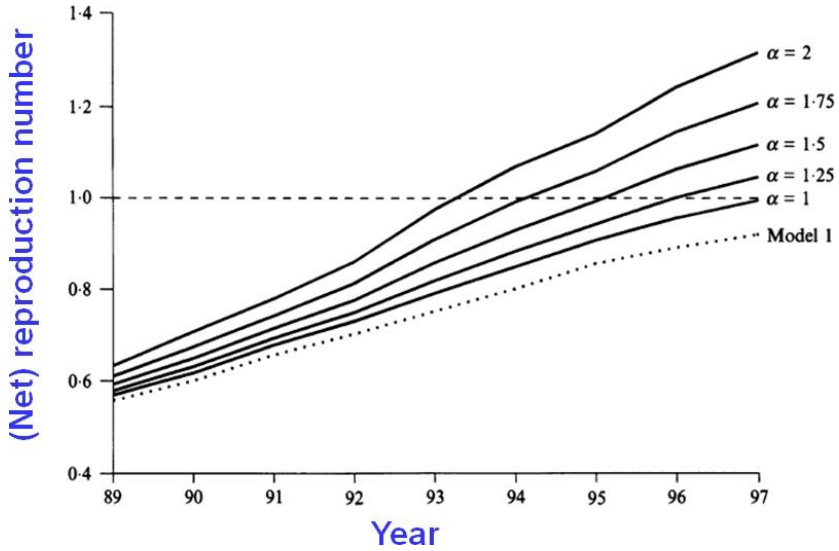


Figure 2: Estimates for R_n during the 1990s in England and Wales obtained by Gay et al (Gay et al (1995)) using different assumptions about contact between individuals aged 10-14 years. α reflects the factor by which the rate at which 10-14 year olds come into contact with each other differs from the rate at which 5-9 year olds come into contact with each other.

Exercise (challenge?)

Calculate R_0 and herd immunity threshold for the following Next Generation Matrix:

$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$. (Reminder: the solution to the equation of the form ax^2+bx+c is given by $\frac{-b \pm \sqrt{b^2-4ac}}{2}$)

Alternative methods for calculating the R_0 (Optional)

The above method for calculating R_0 can be laborious. Another method of deriving the basic reproduction number would be to use the fact that it is the maximum value of ρ which satisfies the equation:

$$(R_{0_{11}} - \rho)(R_{0_{22}} - \rho) - R_{0_{12}}R_{0_{21}} = 0 \qquad \text{(A)}$$

This equation follows from the fact that, when introduced into a totally susceptible population, a typical infectious person generates R_0 replicas of itself, and R_0 has to satisfy the equation:

$$\begin{pmatrix} R_{011} & R_{021} \\ R_{012} & R_{022} \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = R_0\begin{pmatrix} x \\ 1-x \end{pmatrix} \qquad \text{(B)}$$

This equation can be rearranged to give:

$$\begin{pmatrix} R_{011} - R_0 & R_{021} \\ R_{012} & R_{022} - R_0 \end{pmatrix}\begin{pmatrix} x \\ 1-x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

By matrix theory, it can be shown that when the above equation holds, the determinant of the matrix $\begin{pmatrix} R_{011} - R_0 & R_{021} \\ R_{012} & R_{022} - R_0 \end{pmatrix}$, which is given by $(R_{0_{11}} - R_0)(R_{0_{22}} - R_0) - R_{0_{12}}R_{0_{21}}$ has to equal 0. An equation of the form $(R_{0_{11}} - \rho)(R_{0_{22}} - \rho) - R_{0_{12}}R_{0_{21}} = 0$ has two solutions. The R_0 is the larger value of these two solutions (and referred to formally as the "dominant eigenvalue of the Next Generation Matrix", since it leads to the greater value for the proportion of the population which needs to be immune in order to control transmission (through the equation $H=1-1/R_0$) and which must therefore be sufficiently high to control transmission even in the highest risk group.

Extending the logic to populations containing more than 2 subgroups

The above logic can be extended relatively easily to deal with populations consisting of more than two subgroups. For example, if the population comprised 3 subgroups, a fraction x, y and 1-x-y of the typical infectious person could be considered to belong to the first, second and third subgroups. With a Next Generation Matrix such as $\begin{pmatrix} 1 & 5 & 4 \\ 2 & 3 & 6 \\ 3 & 7 & 8 \end{pmatrix}$, the basic reproduction number would be calculated by solving the following matrix equation:

$$\begin{pmatrix} 1 & 5 & 4 \\ 2 & 3 & 6 \\ 3 & 7 & 8 \end{pmatrix}\begin{pmatrix} x \\ y \\ 1-x-y \end{pmatrix} = R_0\begin{pmatrix} x \\ y \\ 1-x-y \end{pmatrix}$$

which would be written equivalently as:

$$4x+5y+4(1-x-y) = R_0x$$

$2x+3y+6(1-x-y) = R_0y$
$3x+7y+8(1-x-y) = R_0(1-x-y)$

Further reading

Vynnycky E and White RG (2010) An introduction to infectious disease modelling. Oxford University Press. Chapter 7, pp 203-222.
Also, exercise 7.4 at the end of chapter 7 and the exercises associated with model 7.5 in the online exercises provide the opportunity for you to follow the calculations of Gay et al (1995) in determining the potential for a measles epidemic to occur in England during the 1990s.

References

Babad, H.R., et al., Predicting the impact of measles vaccination in England and Wales: model validation and analysis of policy options. Epidemiol Infect, 1995. 114(2): p. 319-44.

Choi, Y.H., et al., The potential for measles transmission in England. BMC Public Health, 2008. 8: p. 338.

Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations J Math Biol. 1990;28(4):365-82.

Diekmann O, Heesterbeek JA. Mathematical epidemiology of infectious diseases. Model building, analysis and interpretation. Wiley series in Mathematical and Computational Biology, 2000

Farrington CP, Whitaker HJ. Estimation of effective reproduction numbers for infectious diseases using serological survey data. Biostatistics. 2003 Oct;4(4):621-32.

Farrington CP, Kanaan MN, Gay NJ Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. Appl Statist. 2001, 50(3):251-292

Fine PEM, Jezek Z, Grab B, Dixon H. The transmission potential of monkeypox virus in human populations. Int J Epidemiol 1988, 17(3): 643-50

Gay NJ, Hesketh LM, Morgan-Capner P, Miller E .Interpretation of serological surveillance data for measles using mathematical models: implications for vaccine strategy. Epidemiol Infect. 1995 Aug;115(1):139-56.

Vynnycky E and White RG (2010) An introduction to infectious disease modelling. Oxford University Press. Chapter 7, pp 203-222.

Appendix A: Proof of the result that for Example 3, x=2/3 and R_0=2

The steps in the derivation of this result are as follows:

1. Rearrange Equation 1 to obtain an expression for $\frac{x}{1-x}$ in terms of R_0.

In our case after dividing the right-hand side of Equation 1 by 1-x , and dividing the left-hand side of equation 1 by R_0, we obtain the equation:

$$\frac{4}{R_0} = \frac{x}{1-x} \qquad \text{Equation 11}$$

2. Similarly, we can rearrange Equation 2 to obtain the following expression for $\frac{x}{1-x}$ in terms of R_0:

$$\frac{x}{1-x} = R_0 \qquad \text{Equation 12}$$

Equating Equation 11 to Equation 12, we obtain the following:

$$\frac{4}{R_0} = R_0$$

This equation can be rearranged to give the result:

$$R_0^{\,2} = 4$$

The solution to this equation is the square root of 4, (written $\sqrt{4}$), so R_0=2 or -2.

Since R_0 cannot be negative, this implies that R_0=2.

Substituting for R_0=2 into equation 4, we obtain the following equation in terms of x:

$$\frac{x}{1-x} = 2$$

After rearranging this equation, we obtain the following equation:

$$x = 2(1-x)$$

which, after some rearrranging, leads to the result x=2/3

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 15: Paper discussion

In this session you are asked to critically discuss one of the papers provided. Please be ready to indicate your order of preference for these at the start of session 5. You will be allocated to one of the papers during the start of block 2. On the day, please go to the room indicated in the timetable.

Tuberculosis modelling

> Lin HH, Dowdy D, Dye C, Murray M, Cohen T. ***The impact of new tuberculosis diagnostics on transmission: why context matters.*** Bull World Health Organ. 2012 Oct 1;90(10):739-747A

Zika modelling

> **Ferguson** NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez MG, Nouvellet P, Lessler J. ***EPIDEMIOLOGY. Countering the Zika epidemic in Latin America.*** Science. 2016 Jul 22;353(6297):353-4..

You are expected to think about the strengths and weaknesses of your chosen paper, including:

1. What is the research question and is it clearly explained?
2. What are the primary findings?
3. Are the findings original?
4. What model techniques are used?
5. What is the model structure?
6. What are the major model assumptions, and are they clearly explained?
7. Do you think modelling is a useful method to explore the research question? Or would other more conventional epidemiological methods be more appropriate?
8. Do you think the model is 'valid'?
9. Was a sensitivity analysis performed?
10. What was explored, what was not?

Electronic versions are provided on Moodle.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 16: Review

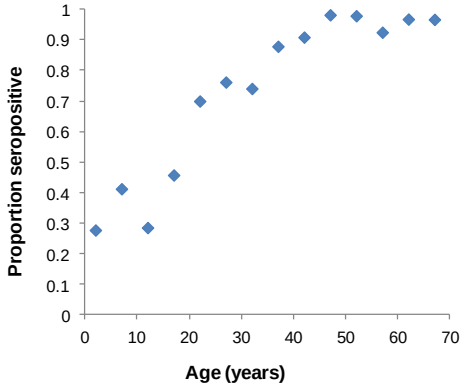
Lecture

This is an optional session going over the key points arising in the sessions in block 2. Please bring any questions that you may have about the course material in this block to this session.

The next few pages include some multiple choice questions that you may like to try. We will go over the questions during the review session.

Q1. The following Table and figure show data on the age-specific proportion of individuals who were seropositive for hepatitis A in country Z. The life expectancy in the population was about 70 years and the age distribution was rectangular. Which of the following statements is incorrect? Note that there may be more than one incorrect answer.

Age mid point	Proportion positive	Age mid point	Proportion positive
2	0.278	37	0.879
7	0.413	42	0.909
12	0.286	47	0.982
17	0.458	52	0.979
22	0.7	57	0.925
27	0.762	62	0.968
32	0.741	67	0.967



a) The force of infection was probably age-dependent.
b) The force of infection was probably not age-dependent.
c) The average force of infection was about 20%/year.
d) Assuming that individuals mix randomly, 20-30% of the population was probably susceptible.
e) Assuming that individuals mix randomly, then we would need to immunize at least 70-80% of the population to control transmission

Q2. Country Z has recently introduced rubella vaccination among very young children. If we assume that individuals mix randomly and that the vaccination coverage is below the herd immunity threshold, which of the following statements is likely to be true. Note that there may be more than one correct answer:

a) The average age at rubella infection is likely to increase.
b) The average age at rubella infection is likely to remain unchanged.
c) The overall proportion of the population that is susceptible may remain unchanged
d) The overall proportion of the population that is susceptible will decrease
e) The overall proportion of the population that is susceptible will increase

Q3. The following WAIFW matrix describes contact between individuals in urban and rural areas. (Note that the letters u and r next to the rows and above the columns reflect urban and rural areas respectively.) Assuming that β_1 is not equal to β_2, which of the statements below is incorrect? (Note that there may be more than one incorrect answer).

$$\begin{array}{c} \\ u \\ r \end{array} \begin{array}{c} \begin{array}{cc} u & r \end{array} \\ \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_2 \end{pmatrix} \end{array}$$

a) Individuals in urban areas effectively contact each other at a different rate from the rate at which they effectively contact individuals in rural areas.
b) The rate at which individuals from rural areas contact each other is different from the rate at which individuals from urban areas contact each other.
c) Individuals from rural areas contact individuals from urban areas at the same rate at which they contact other individuals from rural areas.
d) The rate at which individuals from rural areas contact each other is equal to the rate at which individuals from urban areas contact each other.
e) Individuals from rural areas contact individuals from urban areas at a rate which is different from the rate at which they contact other individuals from rural areas.

Q4. The following is the Next Generation Matrix relating to an infection that is transmitted between children and adults, in population Y. Children and adults are denoted by the letters c and a respectively. Which of the following statements is correct?

$$\begin{array}{c} \\ c \\ a \end{array}\begin{array}{c} \begin{array}{cc} c & a \end{array} \\ \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \end{array}$$

a) Each adult leads to fewer infections in adults than they do in children.
b) Each adult leads to more infections in adults than they do in children.
c) The basic reproduction number is 2
d) The basic reproduction number is 1
e) The basic reproduction number 6

Q5. The following is the Next Generation Matrix for an infection which is transmitted from vectors to humans and from humans to vectors, but which cannot be transmitted either from humans to humans or from a vector to another vector. (Note that the letters v and h next to the rows and above the columns reflect vectors and humans respectively.) Which of the following statements is incorrect?

$$\begin{array}{c} \\ v \\ h \end{array}\begin{array}{c} \begin{array}{cc} v & h \end{array} \\ \begin{pmatrix} 0 & 3 \\ 1.5 & 0 \end{pmatrix} \end{array}$$

a) The basic reproduction number is approximately 2.12.

b) One of the following statements is correct:
i) The fraction of the typical infectious “person” that is a vector is approximately 0.59.
ii) The basic reproduction number is 4.5.

c) If vaccination is introduced just among humans, with two thirds of humans becoming completely protected against infection, the infection will eventually disappear among humans, assuming that no other interventions are introduced.

d) If vaccination is introduced just among humans, with one third of humans becoming completely protected against infection, the net reproduction number of the infection will be approximately equal to 1.7.

e) If no humans are vaccinated but the vector population is sprayed with a chemical agent, so that all vectors are half as infectious as they were previously, the net reproduction will be approximately equal to 1.5.