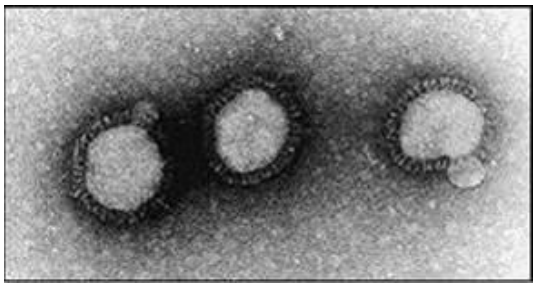
Imperial College London
26 October 2018

Practical: Estimating key epidemiological determinants of pandemic influenza – Part 1
Dr Ada Yan, Dr Ilaria Dorigatti, Dr Simon Cauchemez, Dr Anne Cori, Prof Steven Riley & Prof Christophe Fraser



The influenza virus

Background: Pandemic influenza

Pandemics of influenza arise when a new strain capable of human-to-human transmission emerges that is sufficiently distinct from circulating strains so that the level of population immunity is low or nil. New strains emerge by transfer from aquatic birds, which are the natural reservoir of influenza. Exactly how new strains emerge is not clear, and may differ for different pandemics. Possibilities include genetic re-assortment or recombination between human and avian viral strains, possibly via intermediary species (pigs, poultry, etc…), or by gradual accumulation of adaptive mutations.

The potential for pandemic influenza to cause significant mortality was illustrated by the 1918-20 H1N1 'Spanish flu' which is estimated to have killed at least 20,000,000 worldwide. Note however that the much lower mortality caused by the 1957 H2N2 'Asian flu', the 1966 H3N1 'Hong Kong flu' or the 2009 H1N1 pandemics shows that devastation is not an inevitable result of a pandemic, but depends on the biology of each new strain.

Preparedness is greatly helped by knowledge of the epidemiological determinants of influenza spread such as the basic reproduction number R_0, the duration of latency, and the duration of infectiousness. The aim of this practical is to estimate some of these quantities.

Objective: Estimating transmission parameters

From a methodological perspective, the practical will introduce you to the methods and issues surrounding parameter estimation in epidemic models. More specifically, we are going to use different methods to estimate the reproduction number from epidemic data. Today we will focus on the analysis of exponential growth rates.

We are going to consider a dataset, which gives the excess pneumonia and influenza mortality during the great 1918 H1N1 "Spanish flu" pandemic in New York City.

Estimating the epidemic growth rate of an outbreak.

The spreadsheet "flu_practical_part1_data_2018.xlsx" contains data on excess pneumonia and influenza mortality during the great 1918 H1N1 "Spanish flu" pandemic in New York City.

Plot the epidemic curve and its logarithm. Linear increase in cases on a logarithmic scale indicates exponential increase in the number of infections. For how long does the epidemic grow exponentially?

For the first 3 weeks.

Why does the epidemic cease to grow exponentially after this time?

Because of 1) the depletion of susceptible and 2) control measures (if any).

Estimate the growth rate (r) of the epidemic during this exponential growth phase, and the doubling time, using the formula relating the number infected at two time points t_1 and t_2 of your choice

$$r = \frac{1}{t_2 - t_1} \ln\left(\frac{I(t_2)}{I(t_1)}\right)$$

Here, the reports are *weekly* so you will need to divide your rates by 7 to obtain daily values. You can use Excel to calculate these quantities.

$$r_{weekly} = \frac{1}{week_2 - week_1} \ln\left(\frac{I(week_2)}{I(week_1)}\right) = \frac{1}{4-2} \ln\left(\frac{592}{6}\right) = 2.296 \quad \text{per week}$$

$r = \frac{1}{7} r_{weekly}$ *=2.296/7=0.328* per day

doubling time = ln(2)/r =ln(2)/0.328=2.11 days

A better solution to find r is to do linear regression on the natural-log (ln) transformed data. To do so, in Excel, the easiest is to plot the scatterplot of the data you want to use for your linear regression. Then right click on the scatterplot and click "add trendline". In "options", click "Display equation on chart", then "OK". What are the new estimated growth rates and doubling times? In general, when do you expect those results to differ from the previous ones and why is the latter method better?

r_{weekly} = 2.296 per week

r =2.296/7=0.328 per day

doubling time = ln(2)/r =ln(2)/0.328=2.11 days

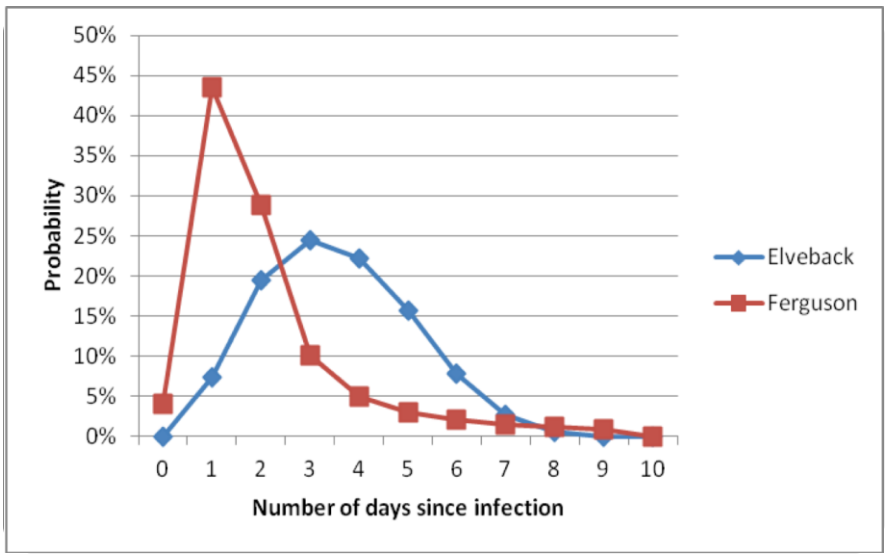
Results are similar to before. Generally speaking, results are going to differ more if the growth is further off the linear trend. Estimating r from the linear regression is better because it uses all data over the period identified as the period of exponential growth, whereas the previous method only uses 2 data points (note that results can vary depending on which pair of data points you choose).

The generation time distribution

The generation time is the time it takes between infection of one individual and subsequent infection of others by that individual.

Because an individual infects others randomly during their infectious period, generation times for infector-infectee pairs are drawn from a distribution, known as the generation time distribution, which we denote $g(t)$. ($g(t)$ is the probability that the generation time is equal to t.) An important summary statistic for the generation time distribution is the mean generation time T_G.

The graph below shows the distribution of the generation time for influenza from 2 highly cited papers: Elveback et al, AJE, 1976 (very commonly used until recently; but no data presented) and Ferguson et al, Nature, 2005 (estimates are derived from epidemiological data on household transmission, consistent with data on viral shedding). The mean generation time is 4.1 days in Elveback et al, compared to 2.6 in Ferguson et al.



Why do the differences between Elveback's & Ferguson's distributions matter for control? – You can for example think about the logistical issues associated with the delivery of antiviral drugs during a pandemic. What will be the likely impact on transmission if the drug is delivered 2 days after infection of the primary case of the household?

For Ferguson's distribution, most of transmission occurs very quickly after infection of the primary case. So, a 2-days delay in delivery considerably reduces the efficacy of the intervention. By contrast, under Elveback's assumption, the drug can still have an impact on spread even with a 2 day-delay.

The generation time distribution for simple models

In general, the generation time distribution can take any form, and each generation time distribution gives rise to a different model. However, if the generation time distribution is approximated using certain mathematical equations, the resulting model is easier to analyse, and as we will see later, R_0 is easier to calculate.

Three common approximations are:

1. The generation time is the same for all infector-infectee pairs (discrete generation model).
2. Infected individuals are instantaneously infectious, have constant infectivity throughout their infectious period, and recover at rate $\gamma = 1/D$ (SIR model). The mean generation time T_G is then equal to the mean infectious period D.
3. Infected individuals become infectious at rate $\alpha = 1/L$, then recover at rate $\gamma = 1/D$ (SEIR model). Then mean generation time T_G is then equal to $L + D$.

The relationship between the epidemic growth rate r, the basic reproduction number R_0 and the generation time distribution $g(t)$.

If the epidemic growth rate r and the generation time distribution $g(t)$ are known, it is possible to derive R_0 from the following equation:

$$R_0 = \frac{1}{\int_0^\infty \exp(-r.t)\, g(t)\, dt} \qquad (1)$$

[if you want to learn more about this, the relationship between the reproduction number R, the exponential growth rate r and the generation time T_g is discussed in the following paper: Wallinga and Lipsitch, Proceedings of the Royal Society B, 2007].

This definition assumes that all individuals are susceptible at the start of the epidemic.

In general, for the 3 following models, the equation for R_0 simplifies as follows:

Box: "Equation for R_0 in 3 simple models"

Discrete generation:	$R_0 = \exp(rT_G)$
SIR:	$R_0 = 1 + rD$
SEIR:	$R_0 = (1 + rL)(1 + rD)$

Estimating the basic reproduction number in New York.

You are going consider a total of 6 models, 3 with a mean generation time of 2.6 days (like Ferguson et al) and 3 with a mean generation time of 4.1 days (like Elveback et al). Assumptions in each model are given in the following table:

	Ferguson	**Elveback**
Discrete generation	T_G=2.6 days	T_G=4.1 days
SIR	D=2.6 days	D=4.1 days
SEIR	L=1 days D=1.6 days	L=1.5 days D=2.6 days

For each model, use the box "equation for R_0 in 3 simple models" to estimate R_0 in New-York, 1918.

	Ferguson	**Elveback**
Discrete generation	$R_0 = 2.35$	$R_0 = 3.84$
SIR	$R_0 = 1.85$	$R_0 = 2.34$
SEIR	$R_0 = 2.02$	$R_0 = 2.76$

Are estimates of R_0 affected by assumptions made on the distribution of the generation time $g(t)$? Are the differences epidemiologically relevant?

Estimates of R_0 are affected by assumptions made on $g(t)$.

1) They depend on the mean value of $g(t)$, T_G: A larger T_G leads to larger estimates of R_0 (you can check this by comparing the Elveback estimate with the Ferguson estimate)

2) Estimates of R_0 are also affected by the variance and shape of $g(t)$. Indeed, for the 4 models with the same T_G (=2.6 days) but $g(t)$, R_0 estimates are different.

In their sensitivity analysis, pandemic influenza models often consider values of R_0 in the range 1.6-2.4. Some mitigation strategies might provide good results for R_0=1.8 but not for R_0=2.3. So, the differences are epidemiologically relevant.

What are the other sensitivities? Do you expect estimates to be biased if there is under-reporting? What if the reporting rate is constant over time? What if it varies over time?

1) Estimates are also sensitive to the time period we choose to characterize exponential growth.

2) Estimates should not be affected by under-reporting as long as the reporting rate is constant over time. But they will be affected if reporting changes over time.